

Common brain networks between major depressive disorder and symptoms of depression that are validated for independent cohorts

Authors: Ayumu Yamashita^{1*}, Yuki Sakai¹, Takashi Yamada^{1,2}, Noriaki Yahata^{1,3,4,5}, Akira Kunitatsu^{6,7}, Naohiro Okada^{3,8}, Takashi Itahashi², Ryuichiro Hashimoto^{1,2,9}, Hiroto Mizuta¹⁰, Naho Ichikawa¹¹, Masahiro Takamura¹¹, Go Okada¹¹, Hirotaka Yamagata¹², Kenichiro Harada¹², Koji Matsuo^{12,13}, Saori C Tanaka¹, Mitsuo Kawato^{1,14*}, Kiyoto Kasai^{1,3,8}, Nobumasa Kato^{1,2}, Hidehiko Takahashi^{10,15}, Yasumasa Okamoto¹¹, Okito Yamashita^{1,14} and Hiroshi Imamizu^{1,16}

Affiliations:

¹ Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institutes International, Kyoto, Japan

² Medical Institute of Developmental Disabilities Research, Showa University, Tokyo, Japan

³ Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

⁴ Institute for Quantum Life Science, National Institutes for Quantum and Radiological Science and Technology, Chiba, Japan

⁵ Department of Molecular Imaging and Theranostics, National Institute of Radiological Sciences, National Institutes for Quantum and Radiological Science and Technology, Chiba, Japan

⁶ Department of Radiology, IMSUT Hospital, Institute of Medical Science, The University of Tokyo, Tokyo, Japan

⁷ Department of Radiology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

⁸ The International Research Center for Neurointelligence (WPI-IRCN) at the University of Tokyo Institutes for Advanced Study (UTIAS), Tokyo, Japan

⁹ Department of Language Sciences, Tokyo Metropolitan University, Tokyo, Japan

¹⁰ Department of Psychiatry, Kyoto University Graduate School of Medicine, Kyoto, Japan

¹¹ Department of Psychiatry and Neurosciences, Hiroshima University, Hiroshima, Japan

¹² Division of Neuropsychiatry, Department of Neuroscience, Yamaguchi University Graduate School of Medicine, Yamaguchi, Japan

¹³ Department of Psychiatry, Faculty of Medicine, Saitama Medical University, Saitama, Japan

¹⁴ Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

¹⁵ Department of Psychiatry and Behavioral Sciences, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan

¹⁶ Department of Psychology, Graduate School of Humanities and Sociology, The University of Tokyo, Tokyo, Japan

*Correspondence to: ayumu@atr.jp (A.Y.), kawato@atr.jp (M.K.)

1 **Abstract**

2 Many studies have highlighted the difficulty inherent to the clinical application of
3 fundamental neuroscience knowledge based on machine learning techniques. It is
4 difficult to generalize machine learning brain markers to the data acquired from
5 independent imaging sites, mainly due to large site differences in functional magnetic
6 resonance imaging. We address the difficulty of finding a generalizable major
7 depressive disorder (MDD) brain network markers which would distinguish patients
8 from healthy controls (a classifier) or would predict symptom severity (a prediction
9 model) based on resting state functional connectivity patterns. For the discovery dataset
10 with 713 participants from 4 imaging sites, we removed site differences using our
11 recently developed harmonization method and developed a machine learning MDD
12 brain network markers. The classifier achieved 70% generalization accuracy, and the
13 prediction model moderately well predicted symptom severity for an independent
14 validation dataset with 449 participants from 4 different imaging sites. Finally, we
15 found common 2 functional connections between those related to MDD diagnosis and
16 those related to depression symptoms. The successful generalization to the perfectly
17 independent dataset acquired from multiple imaging sites is novel and ensures scientific
18 reproducibility and clinical applicability.

19

20 **Keywords: resting-state functional magnetic resonance imaging, resting-state**
21 **functional connectivity, machine learning, major depressive disorder**

22

1 A recent initiative, the Research Domain Criteria (RDoC), has sought to redefine and
 2 subtype psychiatric disorders in terms of biological systems, without relying on a
 3 diagnosis based solely on symptoms and signs. This initiative is expected to inform our
 4 understanding of overlapping and heterogeneous clinical presentations of psychotic
 5 disorders ¹⁻⁴. In particular, resting state functional magnetic resonance imaging
 6 (rs-fMRI) is a useful modality to this end because it enables us to non-invasively
 7 investigate whole brain functional connectivity (FC) in diverse patient populations ^{5,6}.
 8 Rs-fMRI allows for the quantification of the FC of correlated, spontaneous
 9 blood-oxygen-level dependent (BOLD) signal fluctuations ⁷. According to the original
 10 idea of the RDoC initiative, redefinition and subtyping of psychiatric disorders should
 11 be achieved by applying the so-called unsupervised learning technique to the FCs
 12 without relying on a diagnosis as a ground truth ⁸⁻¹⁰. However, the number of
 13 explanatory variables, FCs, is usually between 10,000 and 100,000, while the sample
 14 size, i.e. the number of participants, is usually between 100 and 1,000. Thus, overfitting
 15 to noise in the data by machine learning algorithms and the resultant inflation of
 16 prediction performance can easily occur unless special precautions are taken ¹¹. This
 17 situation makes it difficult to directly apply unsupervised learning algorithm to FC data.

18 To address this problem, we proposed the following hierarchical supervised /
 19 unsupervised approach, having partially succeeded in several studies ¹²⁻¹⁵. First, we
 20 identified a small number of FCs that reliably distinguish healthy controls (HCs) and
 21 psychiatric disorder patients using a supervised learning algorithm. We can use the
 22 identified FCs not only for a brain network biomarker of the psychiatric disorder but
 23 also for biologically meaningful dimensions of the disorder. Second, we applied
 24 unsupervised learning to these low biological dimensions to further understand

1 psychiatric disorders. For instance, we were able to achieve subtyping of major
2 depressive disorder (MDD) by locating MDD patients in these dimensions (subtyping).
3 It may be possible to evaluate the drug effect by locating patients before and after the
4 pharmacological treatment in these biological dimensions ¹². Furthermore, locating
5 different psychiatric disorder patients (e.g. MDD, schizophrenia [SCZ] and autism
6 spectrum disorder [ASD]) in these dimensions may reveal the relationships among the
7 disorders (multi-disorder spectrum) ¹²⁻¹⁵. In this way, although our approach starts with
8 supervised learning based on diagnosis, our final goal is to understand psychiatric
9 disorders in the biological dimensions while avoiding overfitting to noise in the
10 discovery dataset and ensuring generalization performance for the independent data in
11 completely different multiple imaging sites.

12 Furthermore, an increasing number of studies have highlighted the difficulty in
13 finding a clear association between existing clinical diagnostic categories and
14 neurobiological abnormalities ^{8,16,17}. Therefore, the necessity of a symptom-based
15 approach, which directly describes the association with neurobiological abnormalities,
16 is increasingly recognized rather than a diagnosis-based approach ¹⁸. Here, we also
17 construct a brain network marker which would predict symptom severity.

18 Whether a brain network marker constructed in the first stage generalizes to the
19 data acquired from multiple completely different imaging sites is a very important issue
20 for the above hierarchical supervised/unsupervised approach ¹⁹⁻²¹. However, an
21 increasing number of studies have highlighted the difficulty in generalization of the
22 brain network marker to the data acquired from multiple completely independent
23 imaging sites, even using the supervised learning method ^{22,23}. For example, in a recent
24 paper by Drysdale, which is one of the most successful brain network markers of MDD,

1 the classification accuracy for MDD in completely independent imaging sites was
2 68.8% for 16 patients from 1 site, which represents only 3% of the validation cohort
3 (Drysdale et al., 2017, Supplementary Tables 3 and 6).

4 Here, we targeted MDD, the world's most serious psychiatric disorder in terms
5 of its social repercussions ^{24,25}, and investigated whether we could construct a brain
6 network marker which generalizes to the data acquired from multiple completely
7 different imaging sites. We considered and satisfied 3 issues and conditions to ensure
8 generalization of our network marker of MDD to the independent validation dataset,
9 which does not include imaging sites of the discovery dataset. First, we used our
10 recently developed harmonization method, which could reduce site differences in FC ²⁶.
11 According to our recent study, the differences in resting state FCs for different imaging
12 sites consist of measurement bias due to differences in fMRI protocols and MR scanners,
13 and sampling bias due to recruitment of different participant populations. The
14 magnitude of the measurement bias was larger than the effects of disorders including
15 MDD, and the magnitude of the sampling bias was comparable to the effects of
16 disorders ²⁶. Therefore, a reduction in the site difference in FC is essential for the
17 generalization of network models in the validation dataset. Second, we validated our
18 network marker using a perfectly independent and large cohort collected from multiple
19 completely different imaging sites from a Japanese nation-wide database project called
20 the Strategic Research Program for Brain Science (<https://bicr.atr.jp/decnefpro/>). We
21 used a rs-fMRI discovery dataset with 713 participants (149 MDD patients) from 4
22 imaging sites and an independent validation dataset with 449 participants (185 MDD
23 patients) from 4 imaging sites that were not included in the discovery dataset. We
24 further used another dataset of 75 HCs, 154 SCZ patients and 121 ASD patients to

1 investigate the multi-disorder spectrum. In total, we used 1,512 participants' data in this
 2 study. Furthermore, unlike previous studies that restricted the subtype of MDD^{9,12}, we
 3 targeted all MDD patients without restricting according to depression subtype in order
 4 to enable future subtyping in the biological dimensions, which is beyond the purpose of
 5 the current paper. Third, we carefully avoided overfitting noise in the discovery dataset.
 6 As explained above, the number of explanatory variables is typically larger than the
 7 sample size in the rs-fMRI study, thus overfitting to noise in the discovery dataset by
 8 machine learning algorithms and resultant inflation of prediction performance can
 9 happen easily unless special precautions are taken. We used a sparse machine learning
 10 algorithm with the least absolute shrinkage and selection operator (LASSO) to avoid
 11 overfitting to noise and selected only essential FCs²⁷. As a result, for the first time, to
 12 our knowledge, we developed a generalizable brain network marker for MDD diagnosis
 13 without restricting to certain subtypes such as treatment-resistant or melancholy types
 14 and a generalizable brain network marker for depression symptoms.

15

16 **Results**

17 **Datasets.** We used two rs-fMRI datasets for the analyses. The “discovery dataset”
 18 contained data from 713 participants (564 HCs from 4 sites, 149 MDD patients from 3
 19 sites; Table 1), and the “independent validation dataset” contained data from 449
 20 participants (264 HCs from independent 4 sites, 185 MDD patients from independent 4
 21 sites; Table 1). Most data utilized in this study can be downloaded publicly from the
 22 DecNef Project Brain Data Repository (<https://bicr-resource.atr.jp/srpbsopen/> and
 23 <https://bicr.atr.jp/dcn/en/download/harmonization/>). The imaging protocols and data
 24 availability statement of each site is described in Supplementary Table 1. Depression

1 symptoms were evaluated using the Beck Depression Inventory-II (BDI-II) score
2 obtained from most participants in each dataset. Clinical details such as medication
3 information and the presence of comorbidities in patients with MDD are described in
4 Supplementary Table 2.

5

6 **Site difference control in FC.** Classical preprocessing was performed, and FC was
7 defined based on a functional brain atlas consisting of 379 nodes (regions) covering the
8 whole brain ²⁸. The Fisher's z-transformed Pearson correlation coefficients between the
9 preprocessed BOLD signal time courses of each possible pair of nodes were calculated
10 and used to construct 379 x 379 symmetrical connectivity matrices in which each
11 element represents a connection strength, or edge, between two nodes. We used 71,631
12 connectivity values ($379 \times 378 / 2$) of the lower triangular matrix of the connectivity
13 matrix. To control for site differences in the FC, we applied a traveling subject
14 harmonization method to the discovery dataset ²⁶. In this method, the measurement bias
15 (the influence of the difference in the properties of MRI scanners, such as the imaging
16 parameters, field strength, MRI manufacturer, and scanner model) was estimated by
17 fitting the regression model to the FC values of all participants from the discovery
18 dataset and the traveling subject dataset, wherein multiple participants travel to multiple
19 sites to assess measurement bias (see *Control of site differences* in Methods section).
20 This method enabled us to subtract only the measurement bias while leaving important
21 information due to differences in subjects among imaging sites. We applied the ComBat
22 harmonization method ²⁹⁻³² to control for site differences in the FC of the independent
23 validation dataset because we did not have a traveling subject dataset for those sites.

24

1 **Reproducible FCs related to MDD diagnosis and depression symptoms.** Utilizing a
2 simple mass univariate analysis, we investigated the reproducibility of the effect sizes
3 by diagnosis and depression symptoms on individual FC across the discovery and
4 validation datasets. For the effect of the diagnosis on each FC, we calculated the
5 difference in the FC value across participants between the HCs and the MDDs (*t*-value).
6 The Pearson's correlation coefficient between FC strength and BDI scores (*r*-value) was
7 calculated for the effect of the depressed symptom on each FC. Fig. 1a left scatter plot
8 shows the diagnosis effect size for the discovery dataset in the abscissa and that for the
9 validation dataset in the ordinate for each FC. Fig. 1 a right is a scatter plot for the
10 symptom effect size. Effect sizes for the two datasets were positively correlated,
11 implying reproducibility of these effects. We compared the distributions of diagnosis
12 and symptom statistics of the discovery dataset to the distributions in the shuffled data
13 in which diagnosis and symptom severity were permuted across subjects. We found the
14 larger effects of the diagnosis in the original data in comparison to the shuffled data (Fig.
15 1a left histograms). We confirmed that the results were similar for the symptom (Fig. 1a
16 right histograms). These results indicate that resting-state FCs contain consistent
17 information across the two datasets regarding MDD diagnosis and depression
18 symptoms.

19 Furthermore, to statistically evaluate the reproducibility of these effects on FCs,
20 we calculated Pearson's correlation between the discovery and validation datasets
21 regarding the above two statistics (*t*-values for diagnosis and *r*-values for symptom). We
22 found significant correlations between the two datasets for diagnosis (*t*-value: $r_{(71631)} =$
23 0.51, 95 % confidence interval (CI) = [0.508 0.519], $R^2=0.26$, [permutation test, $P <$
24 0.001, one-sided]), as well as for symptom (*r*-value: $r_{(71631)} = 0.39$, 95 % CI = [0.380

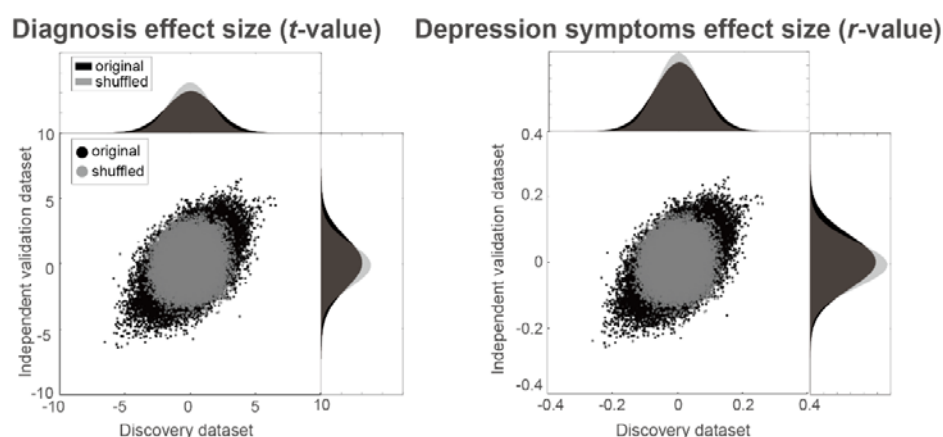
1 0.393], $R^2=0.15$, [permutation test, $P < 0.001$, one-sided], Fig. 1a). This result indicates
 2 that the effects of MDD diagnosis on FCs and the effects of symptom severity were
 3 reproducible even in the independent dataset acquired from completely different sites.

4

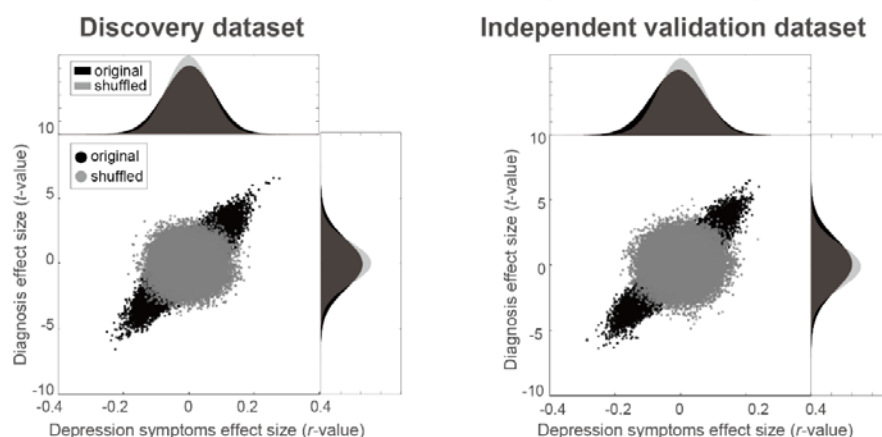
5 **Shared information on FCs between MDD diagnosis and depression symptoms.** We
 6 investigated whether the FCs related to MDD diagnosis and the FCs related to
 7 depression symptoms share identical information or partially overlapping information.
 8 To this end, we calculated Pearson's correlation between the t -values and r -values on
 9 FCs in the same dataset. We found high correlations but not completely identical
 10 (Discovery dataset: $r = 0.86$, Independent validation dataset: $r = 0.91$, Fig. 1b). This
 11 result indicates that shared information exists on FCs between MDD diagnosis and
 12 depression symptoms.

13

a Reproducibility of diagnosis and symptoms effects



b Shared information between diagnosis and symptoms effects



1
2 **Figure 1: Results of mass univariate analysis.** (a) Reproducibility across the two
3 datasets regarding diagnosis (left) and symptom (right) effects. (left) Scatter plot and its
4 histograms of the diagnosis effect size (the difference in mean functional connectivity
5 strengths between depressed patients and healthy groups: t -value). Each point in the
6 scatter plot represents the diagnosis effect in the discovery dataset in the abscissa and
7 that for the independent validation dataset in the ordinate for each functional
8 connectivity. (right) Same format for the depression symptoms effect size (Pearson's
9 correlation between BDI-II and functional connectivity strength: r -value). The original
10 data is in black, while the shuffled data in which subject information was permuted is in
11 gray. (b) Shared information between diagnosis and symptom effects. Scatter plots and
12 its histogram of the diagnosis effect size (t -value) in the ordinate and the depression
13 symptoms effect size (r -value) in the abscissa for all functional connectivity within the
14 discovery dataset (left) and the validation dataset (right). Each point represents

symptom effect size in the abscissa and that for diagnosis in the ordinate for each functional connectivity. The original data is in black, while the shuffled data in which subject information was permuted is in gray.

Brain network marker of MDD diagnosis generalized to MDD data obtained from completely different multisites. We constructed a brain network marker for MDD, which distinguished between HCs and MDD patients, using the discovery dataset based on 71,631 FC values. Based on our previous studies^{12-15,33}, we assumed that psychiatric disorder factors were not associated with whole brain connectivity, but rather with a specific subset of connections. Therefore, we used logistic regression with LASSO, a sparse machine learning algorithm, to select the optimal subset of FCs³⁴. We have already succeeded in constructing generalizable brain network markers of ASD, melancholic MDD, SCZ and obsessive compulsive disorder^{12-15,33} by using a similar sparse estimation method that automatically selects the most important connections. We also tried a support vector machine (SVM) for classification instead of LASSO. However, the performance was not improved compared to that with LASSO (Supplementary Note 1).

To estimate the weights of logistic regression and a hyperparameter that determines how many FCs were used, we conducted a nested cross validation procedure (Fig. 2) (see *Constructing MDD classifier using the discovery dataset* in the Methods section). We first divided the whole discovery dataset into the training set (9 folds out of 10 folds), which was used for training a model, and the test set (1 fold out of 10 folds), which was used for testing the model. To avoid bias due to the difference in the numbers of MDD patients and HCs, we used an undersampling method for equalizing the numbers between the MDD and HC groups³⁵. Since only a subset of training data is

1 used after undersampling, we repeated the random sampling procedure 10 times (i.e.,
2 subsampling). We then fitted a model to each subsample while tuning a regularization
3 parameter in an inner loop of nested cross validation, resulting in 10 classifiers. The
4 mean classifier-output value (diagnostic probability) was considered indicative of the
5 classifier output. Diagnostic probability values > 0.5 were considered to be indicative of
6 a MDD diagnosis. We calculated the area under the curve (AUC), accuracy, sensitivity,
7 specificity, positive predictive value (PPV), and negative predictive value (NPV).
8 Furthermore, we evaluated classifier performance for the unbalanced dataset using the
9 Matthews correlation coefficient (MCC)^{36,37}, which takes into account the ratio of the
10 confusion matrix size.
11
12

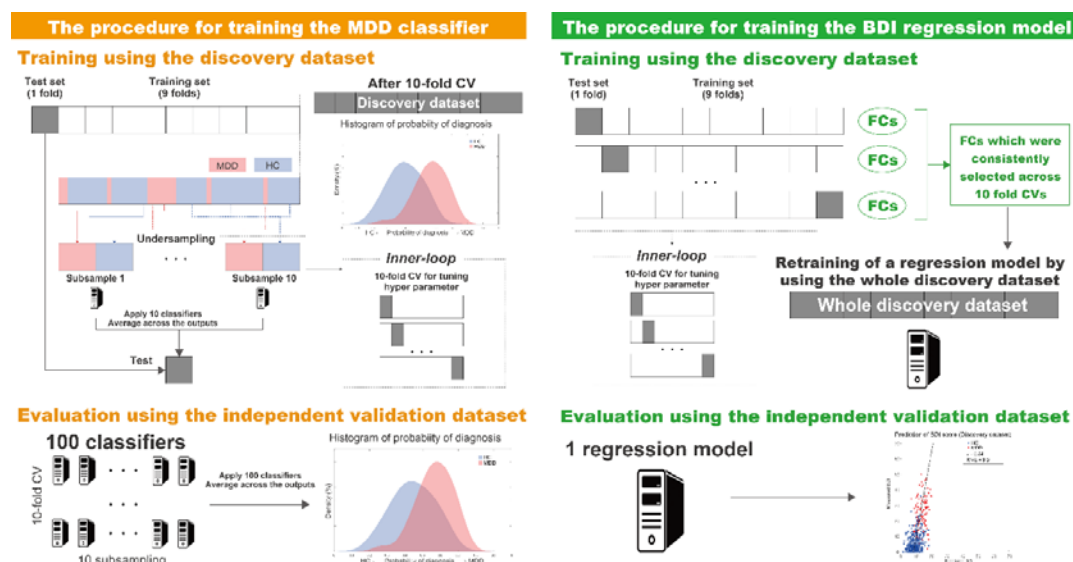


Figure 2: Schematic representation of the procedure for training brain network markers and evaluation of their predictive power. The MDD classifier was constructed using a nested cross validation procedure, undersampling, and subsampling technique in the discovery dataset. The BDI regression model was constructed using the union of FC values selected by the embedded method in the discovery dataset. Generalization performances were evaluated by applying the constructed classifiers and to the independent validation dataset. The machine learning classifiers are represented by PC cartoons. BDI: Beck Depression Inventory-II, CV: cross validation, MDD: major depressive disorder, HC: healthy control, FC: functional connectivity.

The classifier distinguished MDD and HC populations with an accuracy of 67% in the discovery dataset. The corresponding AUC was 0.77, indicating acceptable discriminatory ability. Fig. 3a shows that the two diagnostic probability distributions of the MDD and HC populations were clearly separated by the 0.5 threshold (right, MDD; left, HC) for the discovery dataset. The sensitivity, specificity, PPV, and NPV were 75%, 65%, 0.35, and 0.91, respectively. This classifier led to an acceptable MCC of 0.33. We found that acceptable classification accuracy was achieved for the full dataset as well as for the individual datasets from 3 of the imaging sites (Fig. 3b) to similar degrees. Only HC individuals were identified in the SWA dataset; however, notably, its

1 probability distribution was comparable to the HC populations at other sites.

2 We tested the generalizability of the classifier using an independent validation

3 dataset. We created 100 classifiers of MDD (10-fold $\times$ 10 subsamples); therefore, we

4 applied all trained classifiers to the independent validation dataset. Next, we averaged

5 the 100 outputs (diagnostic probability) for each participant and considered the

6 participant as a patient with MDD if the averaged diagnostic probability value was > 0.5 .

7 The classifier distinguished the MDD and HC populations with an accuracy of 69% in

8 the independent validation dataset. If the accuracy for the validation dataset is much

9 smaller than that of the discovery dataset, overfitting is strongly suggested and the

10 reproducibility of the results is put into doubt. In our case, 69% accuracy for the

11 validation dataset was actually higher than 67% accuracy for the discovery dataset, and

12 this concern does not apply. The corresponding AUC was 0.77 (permutation test, $P <$

13 0.01, one-sided), indicating an acceptable discriminatory ability. Fig. 4a shows that the

14 two diagnostic probability distributions of the MDD and HC populations were clearly

15 separated by the 0.5 threshold (right, MDD; left, HC). The sensitivity, specificity, PPV,

16 and NPV were 74%, 65%, 0.60, and 0.78, respectively. This approach led to an

17 acceptable MCC of 0.38 (permutation test, $P < 0.01$, one-sided). In addition, acceptable

18 classification accuracy was achieved for the individual datasets of the 4 imaging sites

19 (Fig. 4b). To investigate whether our classifier can be generalized to milder depression,

20 we applied our classifier to MDD patients with low BDI scores (score ≤ 20 , $n = 30$) in

21 the independent validation dataset. As a result, 22 of the 30 patients were correctly

22 classified as having MDD (accuracy of 73%), a similar performance level to when the

23 classifier was applied to all patients with MDD. Furthermore, all patients with MDD at

24 the KUT imaging site were treatment-resistant patients (treatment-resistant depression:

adequate use of two or more antidepressants for 4-6 weeks is not efficacious, or intolerance to two or more antidepressants exists). We calculated the classification accuracy only at KUT and obtained the same performance level (accuracy = 75%). These results suggest that the current MDD classifier can be generalized to milder depression, as well as to treatment-resistant patients with MDD.

Regarding the effectiveness of the developed network marker, although discriminability was acceptable (AUC = 0.77) in the independent validation dataset, the performance of the PPV was low in the discovery dataset (0.35). This occurred because the number of patients with MDD was much smaller than that of HCs (about 4 times as many HCs as MDDs) in the discovery dataset. By contrast, in the independent validation dataset, in which the number of HCs is about 1.5 times as high as the number of MDDs, the PPV, at 0.60, was acceptable. When applying a developed network marker in clinical practice, we assume this marker to be applied to those who actually visit the hospital. Therefore, the actual PPV will be acceptable in clinical practice because the prevalence of MDD may be relatively high compared to the general prevalence of MDD. Furthermore, in the independent validation dataset, when we divided the dataset into low- and high-risk groups based on the cutoff value (probability of MDD being 0.51) determined in the discovery dataset³⁸, the odds (sensitivity / 1-sensitivity) were 1.58 in the high-risk group. Moreover, the odds ratio was 5.8 when the odds in the low group were set to 1. That is, the output of the classifier (probability of MDD) will be useful information for psychiatrists as a physical measure supplementing the patients' symptoms and signs in order to make a diagnosis.

We further investigated whether the discrimination performances were different across imaging sites in the independent validation dataset. We calculated the

1 95% confidence intervals (CIs) of the discrimination performances (AUC, accuracy,
2 sensitivity, and specificity) using a bootstrap method for every imaging site. We
3 repeated the bootstrap procedure 1,000 times and calculated the 95% CI for each site.
4 We then checked whether there was a site whose CI did not overlap with the CIs of
5 other imaging sites. We were unable to find such an imaging site, suggesting no
6 significant systematic difference. However, we noted that the sensitivity at the HUH site
7 was inferior to that at the two other imaging sites (see Supplementary Note 2 and
8 Supplementary Fig. 1: CI of sensitivity in the HUH does not overlap with CI in the
9 HRC or UYA). We discuss the differences in performances among imaging sites in the
10 Discussion section.

11 Finally, we checked the stability of our developed network marker to see if the
12 same subject was consistently classified in the same class when the subject was scanned
13 multiple times at various imaging sites. We applied our marker to a traveling subject
14 dataset (Supplementary Table 6) in which 9 healthy participants (all male participants;
15 age range, 24–32 years; mean age, 27 ± 2.6 years) were scanned about 50 times at 12
16 different sites, producing a total of 411 scan sessions. We achieved a high accuracy in
17 this dataset (mean accuracy = 84.5, 1SD = 12.8, across participants). This result
18 indicates that our developed network marker has high stability even if the same subject
19 is scanned multiple times at various imaging sites.

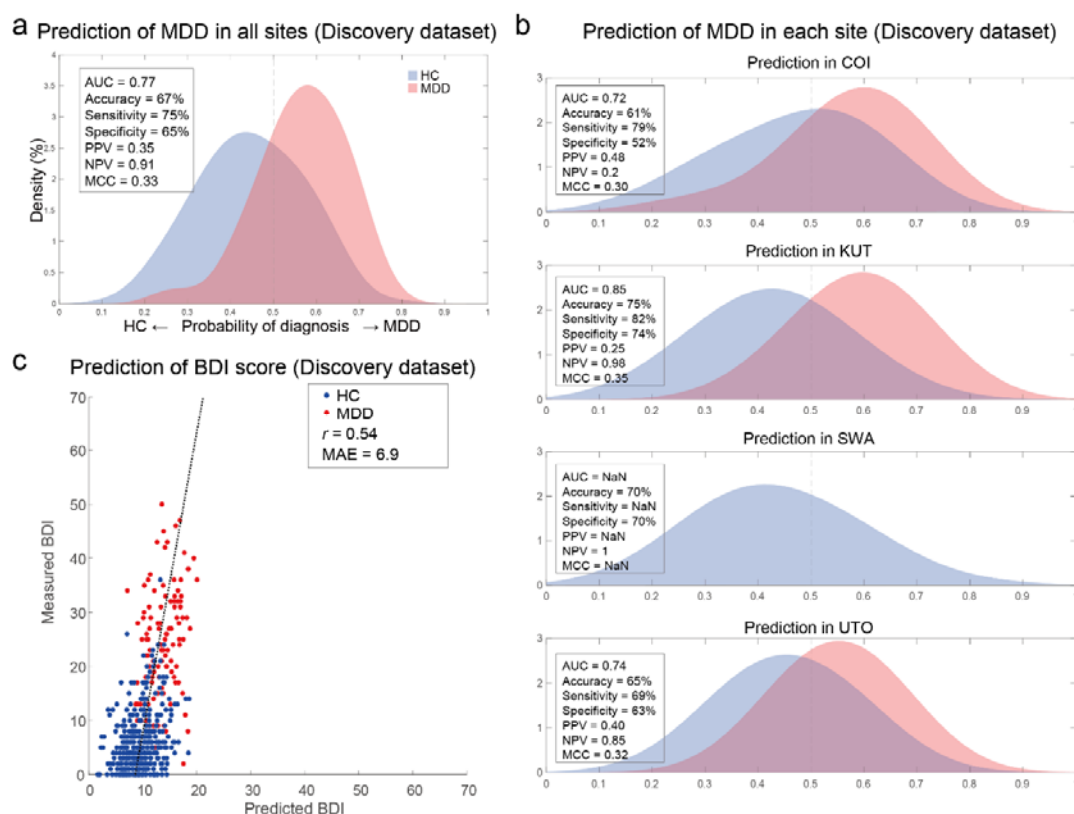


Figure 3: MDD classifier and BDI regression model performances in the discovery dataset. (a) The probability distribution for the diagnosis of MDD in the discovery dataset and (b) probability distributions for each imaging site. MDD and HC distributions are depicted in red and blue, respectively. (c) Scatter plots of measured and predicted BDI. The solid line indicates the linear regression of the measured BDI from the predicted BDI. The correlation coefficient (r) and mean absolute error (MAE) are shown. Each data point represents one participant. BDI: Beck Depression Inventory-II; HC: healthy control; MDD: major depressive disorder; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value; MCC: Matthews correlation coefficient; COI: Center of Innovation in Hiroshima University; KUT: Kyoto University; SWA: Showa University; UTO: University of Tokyo.

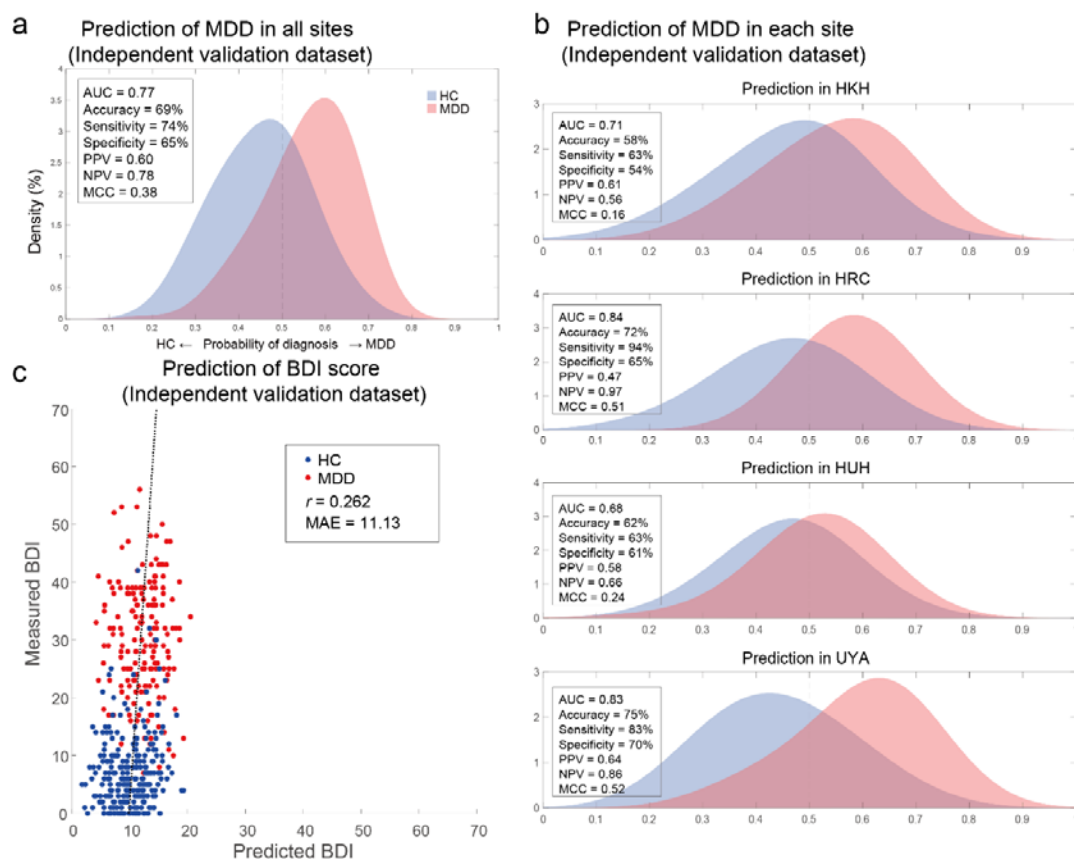


Figure 4: MDD classifier and BDI regression model performances in the independent validation dataset. (a) The probability distribution for MDD diagnosis in the independent validation dataset and (b) probability distributions for each imaging site. MDD and HC distributions are depicted in red and blue, respectively. (c) Scatter plots of measured and predicted BDI. The solid line indicates the linear regression of the measured BDI from the predicted BDI. The correlation coefficient (r) and mean absolute error (MAE) are shown. Each data point represents one participant. BDI: Beck Depression Inventory-II; HC: healthy control; MDD: major depressive disorder; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value; MCC: Matthews correlation coefficient; HKH: Hiroshima Kajikawa Hospital; HRC: Hiroshima Rehabilitation Center; HUH: Hiroshima University Hospital; UYA: Yamaguchi University.

Brain network prediction model of depression symptoms generalized to completely different multisite data. We constructed a brain network prediction model of the BDI

1 score using the discovery dataset based on 71,631 FC values. We employed linear
2 regression using the LASSO method. At first, we tried to evaluate the prediction
3 accuracy on the discovery dataset using a 10-fold CV procedure following our method
4 for the MDD classifier. However, no FC was selected by the LASSO in 7 out of 10 folds
5 during the hyperparameter determination. This result indicates that the regularization in
6 the LASSO worked too strongly. Therefore, we constructed a regression model using
7 the FC values selected by the embedded method in the discovery dataset (Fig. 2)³⁹. This
8 approach caused information leakage because we evaluated the model using the
9 discovery dataset, 30% of which was used for selecting the FCs; therefore, the results in
10 the discovery dataset may be overfitted. This reservation meant that it was important to
11 confirm generalization performance by applying this regression model to an
12 independent validation dataset, as described below. Finally, we calculated the mean
13 absolute error (MAE) and Pearson's correlation coefficients between the predicted and
14 measured BDI scores. The BDI score was well predicted with a significant correlation
15 ($r_{(477)} = 0.54$, 95 % CI = [0.473 0.601], $R^2=0.29$, $P = 1.6 \times 10^{-37}$, one-sided; MAE = 6.9;
16 Fig. 3c). Furthermore, a significant correlation was achieved for HC and MDD
17 populations separately (HC, $r_{(367)} = 0.28$, 95 % CI = [0.185 0.374], $R^2=0.08$, $P = 3.7 \times$
18 10^{-8} , one-sided; MDD, $r_{(110)} = 0.30$, 95 % CI = [0.116 0.459], $R^2=0.09$, $P = 0.0016$).
19 Once again, cautiously, these results may be overfitted because the evaluation data are
20 not independent data. The correct assessment should be based on results from the
21 following independent validation dataset.

22 We tested the generalizability of the regression model using the independent
23 validation dataset. We created one BDI regression model using all the discovery dataset
24 samples; therefore, we applied the trained regression model to the independent

1 validation dataset and considered its output as the predicted BDI score. The BDI score
2 was moderately well predicted, with a significant correlation in the independent
3 validation dataset ($r = 0.26$; $MAE = 11.1$; Fig. 4c; permutation test, $P < 0.01$, one-sided).
4 We could not construct any regression model for the whole permutation procedure
5 because no FC were selected at the nested CV in the LASSO procedure. This result
6 indicated that the performance of the BDI regression model in the independent
7 validation data without permutation was statistically significant.

8
9 **Important FCs for the brain network markers.** We examined important resting state
10 FCs for MDD diagnosis and depression symptoms by extracting the important FCs
11 related to the MDD classifier and BDI regression model, respectively. Briefly, we
12 counted the number of times an FC was selected by LASSO during the 10-fold CV. We
13 considered this FC to be important if this number was significantly higher than the
14 threshold for randomness, according to a permutation test. For the MDD classifier, we
15 permuted the diagnostic labels of the discovery dataset and conducted a 10-fold CV and
16 10-subsampling procedure, and we repeated this permutation procedure 100 times. We
17 then used the number of counts for each connection selected by the sparse algorithm
18 during 10 subsamplings x 10-fold CV (max 100 times) as a statistic in every
19 permutation dataset. To control for the multiple comparison problem, we set a null
20 distribution as the max distribution of the number of counts over all FCs and set our
21 statistical significance to a certain threshold (permutation test, $P < 0.05$, one-sided). We
22 also performed a permutation test for the BDI regression model. We permuted the BDI
23 scores of the discovery dataset, conducted a 10-fold CV, and repeated this permutation
24 procedure 100 times.

1 Figures 5a and 5b shows the spatial distribution of the 31 FCs and 13 FCs
2 related to the MDD diagnosis and depression symptoms, respectively, that were
3 automatically and unbiasedly identified from the data by the machine learning
4 algorithms. Two FCs were common between the diagnosis and symptom models. These
5 connections were the connection ① between right insula and right frontal medial
6 orbital cortex, and ② between the right insula and right cingulum anterior cortex. A
7 detailed list of the FCs is provided in Supplementary Tables 3 and 4. We discussed
8 details of these FCs in the Discussion section.

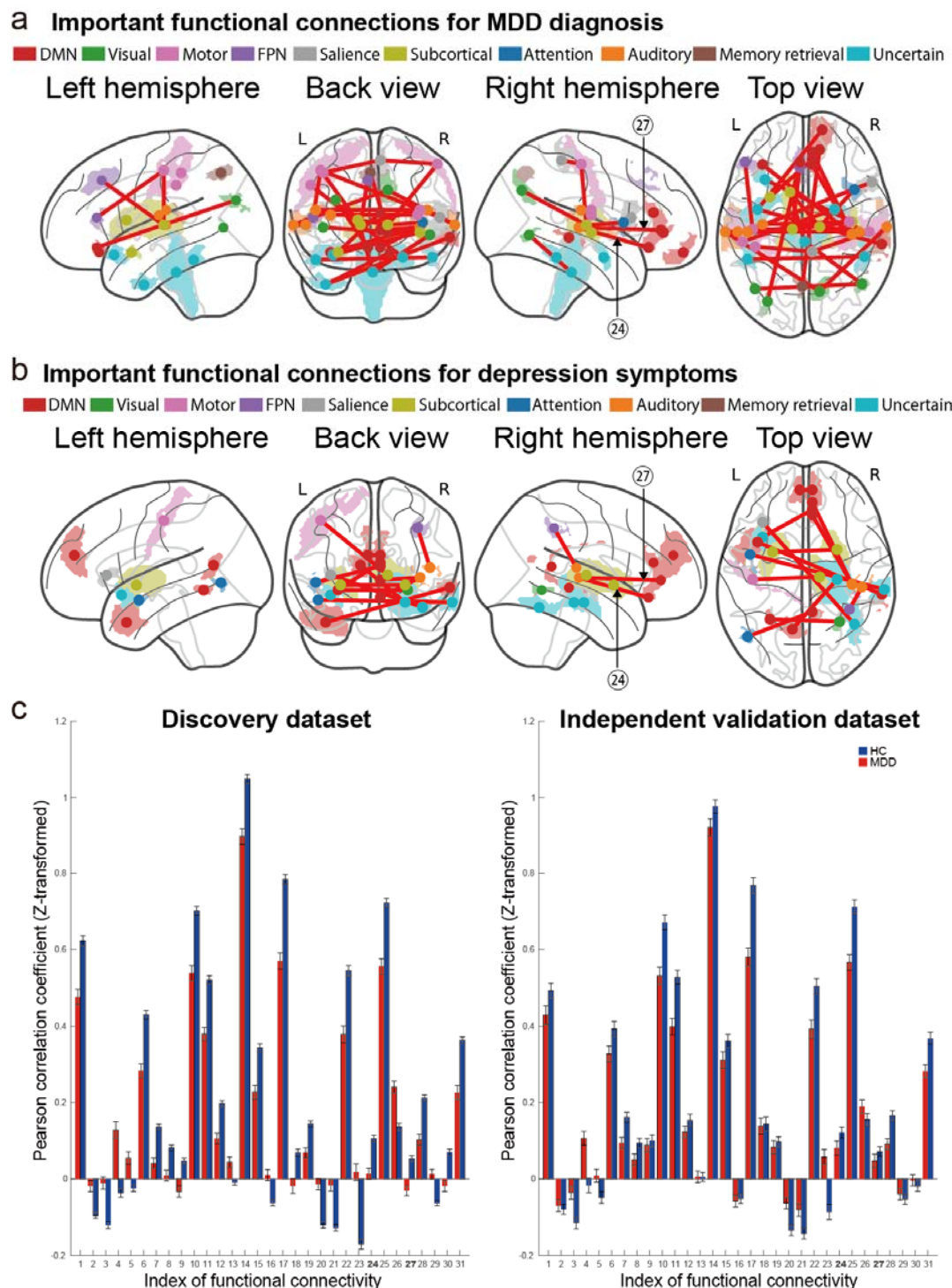


Figure 5: Important FCs for MDD diagnosis and depression symptoms. (a) The 31 functional connections (FCs) which are important for MDD diagnosis viewed from left, back, right, and top. Interhemispheric connections are shown in the back and top views

only. Regions are color-coded according to the intrinsic network. **(b)** The 13 FCs which are important for depression symptoms viewed from left, back, right, and top. Interhemispheric connections are shown in the back and top views only. Regions are color-coded according to the intrinsic network. Two connections were common (▪ between the right insula and the right frontal medial orbital cortex, and ▪ between the right insula and the right cingulum anterior cortex). **(c)** The FC values of 31 FCs for both HCs and MDD patients in the discovery dataset and the independent validation dataset. MDD: major depressive disorder; DMN: default mode network; FPN: fronto-parietal network.

Discussion

In the present study, we thoroughly considered conditions and resolved difficulties in order to ensure the generalization of our brain network marker in the independent validation dataset, which does not include any imaging sites of the discovery dataset. We succeeded in generalizing our network marker to the big independent validation dataset. This generalization ensures scientific reproducibility and the clinical applicability of rs-fMRI. Without this fundamental evidence, we cannot proceed to the development of rs-fMRI-based subtyping, evaluation of drug effects, or exploration of multi-spectrum disorder in the biological dimensions, as mentioned in the Introduction section. Therefore, our study found generalizable psychiatric biomarkers which the fields of psychiatry, neuroscience and computational theory have long sought out, to no avail, since the RDoC initiative.

We developed generalizable brain network markers without restriction to treatment-resistant or melancholy MDD types. Most previous studies have reported the performance of a prediction model using data from the same imaging sites using a CV technique. However, because of large imaging-site differences in rs-fMRI data^{26,40}, CV methods generally induce inflations in performance. To ensure reproducibility, it is

critical to demonstrate the generalizability of the models with an independent validation dataset acquired from completely different imaging sites^{11,19-21}. To overcome the above-mentioned site differences, we reduced site differences in a multisite large-scale rs-fMRI dataset using our novel harmonization method. Next, we constructed an MDD classifier that was acceptably generalized to the independent validation dataset. Acceptable generalized prediction performance was also achieved for the 4 individual imaging site datasets (Fig. 4b). This generalization was achieved even though the imaging protocols in the independent validation datasets were different from the discovery dataset. There are only two studies in which generalization of FC-based MDD classifiers to independent validation data was demonstrated^{9,12}. To the best of our knowledge, our work is the first to construct a generalized classifier of MDD without restriction to certain MDD subtypes: Drysdale concentrated MDD patients who were treatment resistant and Ichikawa restricted patients with the melancholic subtype of MDD. Constructing the whole MDD marker is important for subsequent MDD subtyping analyses. This was achieved for the first time by collecting data on a large variety of MDD patients from multiple imaging sites and objectively harmonizing them with a traveling subject dataset.

With respect to site differences in prediction performance, we found that the sensitivity at the HUH site was inferior to that of the two other imaging sites (see Supplementary Fig. 1). The reason for which the sensitivity was low at the HUH site is that the threshold for separating HCs and patients with MDD was shifted to the MDD side (the probability of diagnosis = 0.5: see the vertical line in Fig. 4b). In contrast, the thresholds were shifted to the HC side at the HKH and UYA sites. These shifts may be due to the fact that the removal of site differences was insufficient. If FC includes a

1 measurement bias, which represents the site difference, the threshold will shift. Our
2 previous study showed that the measurement bias, which includes the site difference,
3 was the largest at the HUH site ²⁶. These results indicate that it is important to remove
4 site differences using precise harmonization methods, such as the traveling subject
5 harmonization method if possible, when we apply a classifier to new subjects collected
6 from a new imaging site. Because of the absence of the traveling subject dataset,
7 traveling subject harmonization was not possible for the independent validation dataset,
8 and we were forced to use ComBat in this case.

9 The machine learning algorithms reliably identified the 31 FCs which are
10 important for MDD diagnosis (Fig. 5a, and Supplementary Table 3). We hereafter
11 summarize the characteristics of the 31 FCs. First, 25 of 31 FCs exhibited
12 hypo-connectivity in the MDD population in the independent validation dataset (the
13 absolute value of the FC was closer to 0 in MDD than in HC individuals; Fig. 5c). The
14 connectivity between the left and right insula had the largest difference among 31 FCs
15 between MDD patients and HCs (FC 17 in Fig. 5c). Abnormalities in the insula were
16 found not only in MDD patients ^{41,42} but also reported as common abnormalities
17 (reduced gray-matter volume) among psychiatric disorders ¹. Therefore, the
18 connectivity associated with the insula is a potential candidate for the neurobiological
19 dimension to understand a multi-spectrum disorder. Second, only 3 FCs (FCs 13, 16,
20 and 26) exhibited hyper-connectivity in the MDD population (the absolute value of the
21 FC was greater in MDD than HC individuals) in the independent validation dataset.
22 However, the differences in those 3 FC values between HC and MDD were not
23 significant (Supplementary Table 3). Finally, only 3 FCs (FCs 4, 5, and 23 in Fig. 5c)
24 had reversed FC values between MDD patients and HCs (positive values in MDD

1 patients and negative values in HCs). Two of 3 FCs were the FCs between the right
2 postcentral cortex and the right thalamus, and the left postcentral cortex and the left
3 thalamus. In the study of Drysdale et al.⁹, the FCs related to the thalamus showed
4 stronger and more positive connectivity in MDD patients as a common feature across
5 the 4 biotypes of MDD, consistent with our current results. A previous study also
6 reported increased FC values between the thalamus and sensory motor cortex
7 (postcentral cortex) in MDD patients⁴³. Participants who have an increased FC value
8 between the thalamus and sensory motor cortex have a greater decline in cognitive
9 function and affective experience⁴³. Finally, two FCs were common between the
10 diagnosis and symptom models. These connections were the connection between right
11 insula and right frontal medial orbital cortex, and between the right insula and right
12 cingulum anterior cortex. We need further analyses to clarify how abnormalities in each
13 FC are associated with cognitive and affective functions in a future study.

14 Ultimately, it would be very important to understand the relationships across
15 disorders (multi-disorder spectrum). For example, investigating the heterogeneous
16 clinical presentations of psychiatric disorders, the MDD-ness, which is the output of the
17 MDD classifier, may provide a useful biological dimension across the multiple-disorder
18 spectrum. To explore this possibility, we applied our MDD classifier to SCZ patients
19 and ASD patients included in the DecNef Project Brain Data Repository
20 (<https://bicr-resource.atr.jp/srpbsopen/>). We found that SCZ had a high tendency
21 (similarity) toward MDD while ASD had no such a tendency toward MDD
22 (Supplementary Figure 2a). This result suggests that the MDD classifier generalizes to
23 SCZ but not to ASD. We note that our discovery dataset for the construction of the
24 MDD classifier included no patients with MDD who were comorbid with SCZ and only

1 1 patient with MDD who was comorbid with ASD. Therefore, our classifier was not
2 affected by either SCZ or ASD diagnosis. Thus, the above generalization of the MDD
3 classifier may point to a certain neurobiological relevance among diseases. Our SCZ
4 patients were in the chronic phase and had negative symptoms. Considering that the
5 negative symptoms of schizophrenia are similar to depression symptoms, the
6 generalization hypothesizes the existence of neurobiological dimensions underlying the
7 common symptoms between SCZ and MDD. For example, anhedonia exists as a
8 transdiagnostic symptom between SCZ and MDD ^{44,45}, and some studies have been
9 conducted to understand the neurological basis of anhedonia across psychiatric
10 disorders including SCZ and MDD ^{45,46}. We need further analyses to quantitatively
11 examine this hypothesis and investigate the neurobiological relationship between SCZ
12 and MDD by gathering more precise information on SCZ (symptoms and medication
13 history). To further understand the multi-disorder spectrum, we developed markers of
14 SCZ and ASD using the same method as in this study in addition to a brain network
15 marker of MDD (Supplementary Note 3). As a result, we found an interesting
16 asymmetric relationship among these disorders: the classifier of SCZ did not generalize
17 to patients with MDD (Supplementary Figures 2b). This kind of asymmetry in the
18 classifiers had also been found between the SCZ classifier and the ASD classifier (the
19 ASD classifier generalized to SCZ, but the SCZ classifier did not generalize to ASD)
20 ¹³⁻¹⁵. These results provide us with important information for understanding the
21 biological relationships between diseases. For example, the above asymmetry between
22 the SCZ and ASD or MDD classifiers suggests that the brain network related to SCZ is
23 characterized by a larger diversity than that of ASD or MDD, and that it partially shares
24 information with the smaller brain network related to ASD or MDD than that of SCZ

1 ^{14,15}.

2 Although biomarkers have been developed with the aim of diagnosing patients,
3 the focus has shifted to the identification of biomarkers that can determine therapeutic
4 targets, such as theranostic biomarkers ^{47,48}, which would allow for more personalized
5 treatment approaches. The 31 FCs discovered in this study are promising candidates for
6 theranostic biomarkers for MDD because they are related to the MDD diagnosis. Future
7 work should investigate whether modulation of FC could be an effective treatment of
8 MDD by using an intervention method with regard to FC, such as functional
9 connectivity neurofeedback training ⁴⁷⁻⁵¹.

1 **Materials and Methods**

2 **Participants.** We used 2 rs-fMRI datasets for the analyses: (1) The “discovery dataset”
 3 contained data from 713 participants (564 HCs from 4 sites, 149 MDD patients from 3
 4 sites; Table 1). Each participant underwent a single rs-fMRI session which lasted 10
 5 min. Within the Japanese SRPBS DecNef project, we planned to acquire the rs-fMRI
 6 data using a unified imaging protocol (Supplementary Table 1;
 7 <http://bicr.atr.jp/rs-fmri-protocol-2/>). However, there were 2 erroneous phase-encoding
 8 directions ($P \rightarrow A$ and $A \rightarrow P$). In addition, different sites had different MRI hardware
 9 (Supplementary Table 1). During the rs-fMRI scans, participants were instructed to
 10 “Relax. Stay Awake. Fixate on the central crosshair mark, and do not concentrate on
 11 specific things”. (2) The “independent validation dataset” contained data from 449
 12 participants (264 HCs from independent 4 sites, 185 MDD patients from independent 4
 13 sites; Table 1). Data were acquired following protocols reported in Supplementary
 14 Table 1. The sites used were different from the discovery dataset. Each participant
 15 underwent a single rs-fMRI session lasting 5 or 8 min. In both datasets, depression
 16 symptoms were evaluated using the BDI-II score obtained from most participants in
 17 each dataset. This study was carried out in accordance with the recommendations of the
 18 institutional review boards of the principal investigators’ respective institutions
 19 (Hiroshima University, Kyoto University, Showa University, University of Tokyo, and
 20 Yamaguchi University) with written informed consent from all subjects in accordance
 21 with the Declaration of Helsinki. The protocol was approved by the institutional review
 22 boards of the principal investigators’ respective institutions (Hiroshima University,
 23 Kyoto University, Showa University, University of Tokyo, and Yamaguchi University).
 24 Most data utilized in this study can be downloaded publicly from the DecNef Project
 25 Brain Data Repository at <https://bicr-resource.atr.jp/srpbsopen/> and
 26 <https://bicr.atr.jp/dcn/en/download/harmonization/>. The data availability statements of
 27 each site are described in Supplementary Table 1.

28
 29 **Preprocessing and calculation of the resting state FC matrix.** We preprocessed the
 30 rs-fMRI data using FMRIPREP version 1.0.8⁵². The first 10 s of the data were
 31 discarded to allow for T1 equilibration. Preprocessing steps included slice-timing
 32 correction, realignment, coregistration, distortion correction using a field map,
 33 segmentation of T1-weighted structural images, normalization to Montreal Neurological
 34 Institute (MNI) space, and spatial smoothing with an isotropic Gaussian kernel of 6 mm
 35 full-width at half-maximum. “Fieldmap-less” distortion correction was performed for
 36 the independent validation dataset due to the lack of field map data. For more details on

1 the pipeline, see <http://fmripiprep.readthedocs.io/en/latest/workflows.html>. For 6
2 participants' data in the independent validation dataset, the coregistration was
3 unsuccessful, and we therefore excluded these data from further analysis.

4 *Parcellation of brain regions:* To analyze the data using Human Connectome Project
5 (HCP) style surface-based methods, we used ciftify toolbox version 2.0.2⁵³. This
6 allowed us to analyze our data, which lacked the T2-weighted image required for HCP
7 pipelines, using an HCP-like surface-based pipeline. Next, we used Glasser's 379
8 surface-based parcellations (cortical 360 parcellations + subcortical 19 parcellations) as
9 regions of interest (ROIs), considered reliable brain parcellations²⁸. The BOLD signal
10 time courses were extracted from these 379 ROIs. To facilitate the comparison of our
11 results with previous studies, we identified the anatomical names of important ROIs and
12 the names of intrinsic brain networks that included the ROIs using anatomical
13 automatic labeling (AAL)⁵⁴ and Neurosynth (<http://neurosynth.org/locations/>).

14 *Physiological noise regression:* Physiological noise regressors were extracted by
15 applying CompCor⁵⁵. Principal components were estimated for the anatomical
16 CompCor (aCompCor). A mask to exclude signals with a cortical origin was obtained
17 by eroding the brain mask and ensuring that it contained subcortical structures only.
18 Five aCompCor components were calculated within the intersection of the subcortical
19 mask and union of the CSF and WM masks calculated in the T1-weighted image space
20 after their projection to the native space of functional images in each session. To
21 remove several sources of spurious variance, we used a linear regression with 12
22 regression parameters, such as 6 motion parameters, average signals over the whole
23 brain, and 5 aCompCor components.

24 *Temporal filtering:* A temporal bandpass filter was applied to the time series using a
25 first-order Butterworth filter with a pass band between 0.01 Hz and 0.08 Hz to restrict
26 the analysis to low-frequency fluctuations, which are characteristic of rs-fMRI BOLD
27 activity⁵⁶.

28 *Head motion:* Framewise displacement (FD)⁵⁷ was calculated for each functional
29 session using Nipype (<https://nipype.readthedocs.io/en/latest/>). FD was used in the
30 subsequent scrubbing procedure. To reduce spurious changes in FC from head motion,
31 we removed volumes with $FD > 0.5$ mm, as proposed in a previous study⁵⁷. The FD
32 represents head motion between 2 consecutive volumes as a scalar quantity (i.e., the
33 summation of absolute displacements in translation and rotation). Using the
34 aforementioned threshold, $6.3\% \pm 13.5$ volumes (mean $\pm$ SD) were removed per
35 rs-fMRI session in all datasets. If the ratio of the excluded volumes after scrubbing
36 exceeded the mean + 3 SD, participants were excluded from the analysis. As a result, 32

1 participants were removed from all datasets. Thus, we included 683 participants (545
2 HCs, 138 MDD patients) in the discovery dataset and 440 participants (259 HCs, 181
3 MDD patients) in the independent validation dataset for further analysis.

4 *Calculation of FC matrix:* FC was calculated as the temporal correlation of rs-fMRI
5 BOLD signals across 379 ROIs for each participant. There are a number of different
6 candidates to measure FC, such as the tangent method and partial correlation; however,
7 we used a Pearson's correlation coefficient because they are the most commonly used
8 values in previous studies. Fisher's z-transformed Pearson's correlation coefficients
9 were calculated between the preprocessed BOLD signal time courses of each possible
10 pair of ROIs and used to construct 379×379 symmetrical connectivity matrices in
11 which each element represents a connection strength between 2 ROIs. We used 71,631
12 FC values $[(379 \times 378)/2]$ of the lower triangular matrix of the connectivity matrix for
13 further analysis.

14 *Control of site differences:* Next, we used a traveling subject harmonization method to
15 control for site differences in FC in the discovery dataset. This method enabled us to
16 subtract pure site differences (measurement bias) which are estimated from the traveling
17 subject dataset wherein multiple participants travel to multiple sites to assess
18 measurement bias. The participant factor ($\mathbf{p}$), measurement bias ($\mathbf{m}$), sampling biases
19 ($\mathbf{s}_{hc}$, $\mathbf{s}_{mdd}$), and psychiatric disorder factor ($\mathbf{d}$) were estimated by fitting the regression
20 model to the FC values of all participants from the discovery dataset and the traveling
21 subject dataset. For each connectivity, the regression model can be written as follows:

$$Connectivity = \mathbf{x}_m^T \mathbf{m} + \mathbf{x}_{s_{hc}}^T \mathbf{s}_{hc} + \mathbf{x}_{s_{mdd}}^T \mathbf{s}_{mdd} + \mathbf{x}_d^T \mathbf{d} + \mathbf{x}_p^T \mathbf{p} + const + e,$$

$$\text{such that } \sum_j^9 p_j = 0, \sum_k^4 m_k = 0, \sum_k^4 s_{hc_k} = 0, \sum_k^3 s_{mdd_k} = 0, d_1(HC) = 0,$$

22 in which $\mathbf{m}$ represents the measurement bias (4 sites $\times$ 1), $\mathbf{s}_{hc}$ represents the sampling
23 bias of HCs (4 sites $\times$ 1), $\mathbf{s}_{mdd}$ represents the sampling bias of patients with MDD (3
24 sites $\times$ 1), $\mathbf{d}$ represents the disorder factor (2 $\times$ 1), $\mathbf{p}$ represents the participant factor
25 (9 traveling subjects $\times$ 1), $const$ represents the average functional connectivity value
26 across all participants from all sites, and $e \sim \mathcal{N}(0, \gamma^{-1})$ represents noise. Measurement
27 biases were removed by subtracting the estimated measurement biases. Thus, the
28 harmonized functional connectivity values were set as follows:

$$Connectivity^{Harmonized} = Connectivity - \mathbf{x}_m^T \hat{\mathbf{m}},$$

29
30 in which $\hat{\mathbf{m}}$ represents the estimated measurement bias. More detailed information has
31 been previously described²⁶.

32 We used the ComBat harmonization method²⁹⁻³² to control for site differences in FC in

the independent validation dataset because we did not have a traveling subject dataset for those sites. We performed harmonization to correct only for the site difference using information on MDD diagnosis, BDI score, age, sex, and dominant hand as auxiliary variables in ComBat. Notably, compared with the conventional regression method, the ComBat method is a more advanced method to control for site effects²⁹⁻³².

Constructing the MDD classifier using the discovery dataset. We constructed a brain network marker for MDD that distinguished between HCs and MDD patients using the discovery dataset based on 71,631 FC values. To construct the network marker, we applied a machine learning technique. Although SVM is often used as a classifier, SVM is not suitable for investigating the contribution of explanatory variables because it is difficult to calculate the contribution of each explanatory variable. Based on our previous study¹³, we assumed that psychiatric disorder factors were not associated with whole brain connectivity, but rather with a specific subset of connections. Therefore, we conducted logistic regression analyses using the LASSO method to select the optimal subset of FCs³⁴. A logistic function was used to define the probability of a participant belonging to the MDD class as follows:

$$P_{sub}(y_{sub} = 1 | \mathbf{c}_{sub}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{c}_{sub})},$$

in which y_{sub} represents the class label (MDD, $y = 1$; HC, $y = 0$) of a participant, $\mathbf{c}_{sub}$ represents an FC vector for a given participant, and $\mathbf{w}$ represents the weight vector. The weight vector $\mathbf{w}$ was determined to minimize

$$J(\mathbf{w}) = -\frac{1}{n_{sub}} \sum_{j=1}^{n_{sub}} \log P_j(y_j = 1 | \mathbf{c}_j; \mathbf{w}) + \lambda \|\mathbf{w}\|_1,$$

in which $\|\mathbf{w}\|_1 = \sum_i^N |w_i|$ and λ represent hyperparameters that control the amount of shrinkage applied to the estimates. To estimate weights of the logistic regression and a hyperparameter λ , we conducted a nested cross validation procedure (Fig. 2). In this procedure, we first divided the whole discovery dataset into a training set (9 folds of 10 folds) which used for training a model and a test set (a fold of 10 folds) for testing the model. To minimize bias due to the differences in the numbers of MDD patients and HCs, we used an undersampling method³⁵. Almost 125 MDD patients and 125 HCs were randomly sampled from the training set, and classifier performance was tested using the test set. Since only a subset of training data is used after undersampling, we repeated the random sampling procedure 10 times (i.e., subsampling). We then fitted a model to each subsample while tuning a regularization parameter in the inner loop of

1 the nested cross validation, resulting in 10 classifiers. For the inner loop, we used the
 2 “*lassoglm*” function in MATLAB (R2016b, Mathworks, USA) and set “NumLambda”
 3 to 25 and “CV” to 10. In this inner loop, we first calculated a value of λ just large
 4 enough such that the only optimal solution is the all-zeroes vector. A total of 25 values
 5 of λ were prepared at equal intervals from 0 to $\lambda_{\max}$ and the λ was determined
 6 according to the one-standard-error-rule in which we selected the largest λ within the
 7 standard deviation of the minimum prediction error (among all λ)²⁷. The mean classifier
 8 output value (diagnostic probability) was considered indicative of the classifier output.
 9 Diagnostic probability values > 0.5 were considered indicative of MDD patients. We
 10 calculated the AUC using the “*perfcurve*” function in MATLAB. In addition, we
 11 calculated the accuracy, sensitivity, specificity, PPV, and NPV. Furthermore, we
 12 evaluated classifier performance for the unbalanced dataset using the MCC^{36,37}, which
 13 takes into account the ratio of the confusion matrix size.

14
 15 **BDI score regression model in the discovery dataset.** We constructed a linear
 16 regression model to predict the BDI score using the discovery dataset based on 71,631
 17 FC values. To construct the linear regression model, we applied a machine-learning
 18 technique to participants with BDI scores in the discovery dataset. Although SVR is
 19 often used as a regression model, SVR is not suitable for investigating the contribution
 20 of explanatory variables because it is difficult to calculate the contribution of each
 21 explanatory variable. Therefore, we employed linear regression using the LASSO
 22 method as follows:

$$\text{Predicted } BDI_{sub} = \mathbf{w}^T \mathbf{c}_{sub},$$

23 in which *Predicted BDI_{sub}* represents the BDI score of a participant; *c_{sub}* represents
 24 an FC vector for a given participant, and *w* represents the weight vector of the linear
 25 regression. The prediction model was constructed while feature selection using the
 26 embedded method with LASSO was performed (Fig. 2)³⁹. We conducted a 10-fold CV
 27 procedure for this regression model. We constructed a regression model using the
 28 combination of FC values selected in all 10 folds in the training dataset (Fig. 2). This
 29 caused information leakage across the folds; therefore, the training dataset may be
 30 overfitting. This issue meant that it was important to confirm generalization
 31 performance by applying this regression model to an independent validation dataset, as
 32 described below. Finally, we calculated the mean absolute error (MAE) and Pearson’s
 33 correlation coefficients between the predicted and measured BDI scores.

34

35

Generalization performance of the classifier and regression model. We tested the generalizability of the classifier and regression model using an independent validation dataset. We created 100 classifiers of MDD (10-fold $\times$ 10 subsamples); therefore, we applied all trained classifiers to the independent validation dataset. Next, we averaged the 100 outputs (diagnostic probability) for each participant and considered the participant as a patient with MDD if the averaged diagnostic probability value was >0.5 . In contrast, we created the BDI regression model using all the discovery dataset samples; therefore, we applied the trained regression model to the independent validation dataset and considered its output as the predicted BDI score.

To test the statistical significance of the MDD classifier performance, we performed a permutation test. We permuted the diagnostic labels of the discovery dataset and conducted a 10-fold CV and 10-subsampling procedure. Next, we took an average of the 100 outputs (diagnostic probability); a mean diagnostic probability value > 0.5 was considered indicative of a diagnosis of MDD. We repeated this permutation procedure 100 times and calculated the AUC and MCC as the performance of each permutation. We also performed a permutation test for the BDI regression model. We permuted the BDI scores of the discovery dataset, conducted a 10-fold CV, and repeated this permutation procedure 100 times.

Identification of the FCs linked to diagnosis and symptoms. We examined resting-state functional connectivity for MDD diagnosis and depression symptoms by extracting the important FCs related to the MDD classifier and BDI regression model, respectively. Briefly, we counted the number of times an FC was selected by LASSO during the 10-fold CV. We considered that this FC was important if this number was significantly higher than chance, according to a permutation test. We permuted the diagnostic labels of the discovery dataset and conducted a 10-fold CV and 10-subsampling procedure. We then used the number of counts for each connection selected by the sparse algorithm during 10 subsampling $\times$ 10 CV(max 100 times) as a statistic in every permutation dataset. To control the multiple comparison problem, we set a null distribution as the max distribution of the number of counts over all functional connections and set our statistical significance to a threshold ($p < 0.05$, one-sided). FCs selected ≥ 17 times of 100 times were regarded as diagnostically important. We also performed a permutation test for the BDI regression model. We permuted the BDI scores of the discovery dataset, conducted a 10-fold CV, and repeated this permutation procedure 100 times. FCs selected ≥ 1 times of 10 times were regarded as relevant to depression symptoms.

References

- 1 Goodkind, M. *et al.* Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry* **72**, 305-315, doi:10.1001/jamapsychiatry.2014.2206 (2015).
- 2 Jacobi, F. *et al.* Prevalence, co-morbidity and correlates of mental disorders in the general population: results from the German Health Interview and Examination Survey (GHS). *Psychol Med* **34**, 597-611, doi:10.1017/S0033291703001399 (2004).
- 3 Lee, S. H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-994, doi:10.1038/ng.2711 (2013).
- 4 McTeague, L. M. *et al.* Identification of Common Neural Circuit Disruptions in Cognitive Control Across Psychiatric Disorders. *Am J Psychiatry* **174**, 676-685, doi:10.1176/appi.ajp.2017.16040400 (2017).
- 5 Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proc Natl Acad Sci U S A* **107**, 4734-4739, doi:10.1073/pnas.0911855107 (2010).
- 6 Xia, M. & He, Y. Functional connectomics from a "big data" perspective. *Neuroimage* **160**, 152-167, doi:10.1016/j.neuroimage.2017.02.031 (2017).
- 7 Smith, S. M. *et al.* Functional connectomics from resting-state fMRI. *Trends Cogn Sci* **17**, 666-682, doi:10.1016/j.tics.2013.09.016 (2013).
- 8 Clementz, B. A. *et al.* Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *Am J Psychiatry* **173**, 373-384, doi:10.1176/appi.ajp.2015.14091200 (2016).
- 9 Drysdale, A. T. *et al.* Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med* **23**, 28-38, doi:10.1038/nm.4246 (2017).
- 10 Tokuda, T. *et al.* Identification of depression subtypes and relevant brain regions using a data-driven approach. *Scientific reports* **8**, 14082, doi:10.1038/s41598-018-32521-z (2018).
- 11 Whelan, R. & Garavan, H. When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol Psychiatry* **75**, 746-748, doi:10.1016/j.biopsych.2013.05.014 (2014).
- 12 Ichikawa, N. *et al.* Primary functional brain connections associated with melancholic major depressive disorder and modulation by antidepressants.

- 1 *Scientific reports* **10**, 3542, doi:10.1038/s41598-020-60527-z (2020).
- 2 13 Yahata, N. *et al.* A small number of abnormal brain connections
- 3 predicts adult autism spectrum disorder. *Nat Commun* **7**, 11254,
- 4 doi:10.1038/ncomms11254 (2016).
- 5 14 Yoshihara, Y. *et al.* Overlapping but asymmetrical relationships
- 6 between schizophrenia and autism revealed by brain connectivity. *bioRxiv*,
- 7 403212, doi:10.1101/403212 (2018).
- 8 15 Yoshihara, Y. *et al.* Overlapping but asymmetrical relationships
- 9 between schizophrenia and autism revealed by brain connectivity. *Schizophrenia*
- 10 *Bulletin* (2020).
- 11 16 Insel, T. R. & Cuthbert, B. N. Brain disorders? Precisely. *Science* **348**,
- 12 499-500, doi:10.1126/science.aab2358 (2015).
- 13 17 Singh, I. & Rose, N. Biomarkers in psychiatry. *Nature* **460**, 202-207,
- 14 doi:10.1038/460202a (2009).
- 15 18 Ayuso-Mateos, J. L., Nuevo, R., Verdes, E., Naidoo, N. & Chatterji, S.
- 16 From depressive symptoms to depressive disorders: the relevance of thresholds.
- 17 *Br J Psychiatry* **196**, 365-371, doi:10.1192/bjp.bp.109.071191 (2010).
- 18 19 Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature*
- 19 *Human Behaviour* **1**, 0021, doi:10.1038/s41562-016-0021 (2017).
- 20 20 Nosek, B. A. & Errington, T. M. Making sense of replications. *Elife* **6**,
- 21 doi:10.7554/eLife.23383 (2017).
- 22 21 Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and
- 23 reproducible neuroimaging research. *Nat Rev Neurosci* **18**, 115-126,
- 24 doi:10.1038/nrn.2016.167 (2017).
- 25 22 Dinga, R. *et al.* Evaluating the evidence for biotypes of depression:
- 26 Methodological replication and extension of. *Neuroimage Clin* **22**, 101796,
- 27 doi:10.1016/j.nicl.2019.101796 (2019).
- 28 23 He, Y., Byrge, L. & Kennedy, D. P. Nonreplication of functional
- 29 connectivity differences in autism spectrum disorder across multiple sites and
- 30 denoising strategies. *Hum Brain Mapp*, doi:10.1002/hbm.24879 (2020).
- 31 24 Ferrari, A. J. *et al.* Burden of depressive disorders by country, sex, age,
- 32 and year: findings from the global burden of disease study 2010. *PLoS Med* **10**,
- 33 e1001547, doi:10.1371/journal.pmed.1001547 (2013).
- 34 25 Hay, S. I. *et al.* Global, regional, and national disability-adjusted
- 35 life-years (DALYs) for 333 diseases and injuries and healthy life expectancy
- 36 (HALE) for 195 countries and territories, 1990–2016: a systematic

- 1 analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**,
2 1260-1344, doi:10.1016/S0140-6736(17)32130-X (2017).
- 3 26 Yamashita, A. *et al.* Harmonization of resting-state functional MRI
4 data across multiple imaging sites via the separation of site differences into
5 sampling bias and measurement bias. *PLoS Biol* **17**, e3000042,
6 doi:10.1371/journal.pbio.3000042 (2019).
- 7 27 Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical learning with*
8 *sparsity: the lasso and generalizations*. (Chapman and Hall/CRC, 2015).
- 9 28 Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral
10 cortex. *Nature* **536**, 171-178, doi:10.1038/nature18933 (2016).
- 11 29 Fortin, J. P. *et al.* Harmonization of cortical thickness measurements
12 across scanners and sites. *Neuroimage* **167**, 104-120,
13 doi:10.1016/j.neuroimage.2017.11.024 (2018).
- 14 30 Fortin, J. P. *et al.* Harmonization of multi-site diffusion tensor imaging
15 data. *Neuroimage* **161**, 149-170, doi:10.1016/j.neuroimage.2017.08.047 (2017).
- 16 31 Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in
17 microarray expression data using empirical Bayes methods. *Biostatistics* **8**,
18 118-127, doi:10.1093/biostatistics/kxj037 (2007).
- 19 32 Yu, M. *et al.* Statistical harmonization corrects site effects in
20 functional connectivity measurements from multi-site fMRI data. *Hum Brain*
21 *Mapp*, doi:10.1002/hbm.24241 (2018).
- 22 33 Takagi, Y. *et al.* A Neural Marker of Obsessive-Compulsive Disorder
23 from Whole-Brain Functional Connectivity. *Scientific reports* **7**, 7538,
24 doi:10.1038/s41598-017-07792-7 (2017).
- 25 34 Tibshirani, R. Regression Shrinkage and Selection via the Lasso.
26 *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288
27 (1996).
- 28 35 Wallace, B. C., Small, K., Brodley, C. E. & Trikalinos, T. A. Class
29 Imbalance, Redux. 754-763, doi:10.1109/icdm.2011.33 (2011).
- 30 36 Chicco, D. Ten quick tips for machine learning in computational
31 biology. *BioData Min* **10**, 35, doi:10.1186/s13040-017-0155-3 (2017).
- 32 37 Matthews, B. W. Comparison of the predicted and observed secondary
33 structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein*
34 *Structure* **405**, 442-451, doi:[https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
35 (1975).
- 36 38 Akobeng, A. K. Understanding diagnostic tests 3: Receiver operating

1 characteristic curves. *Acta Paediatr* **96**, 644-647,
2 doi:10.1111/j.1651-2227.2006.00178.x (2007).

3 39 Guyon, I. & Elisseeff, A. An introduction to variable and feature
4 selection. *Journal of Machine Learning Research* **3**, 1157-1182,
5 doi:10.1162/153244303322753616 (2003).

6 40 He, Y., Byrge, L. & Kennedy, D. P. Non-replication of functional
7 connectivity differences in ASD: a multi-site study. *bioRxiv*, 640797,
8 doi:10.1101/640797 (2019).

9 41 Yin, Z. *et al.* Decreased Functional Connectivity in Insular Subregions
10 in Depressive Episodes of Bipolar Disorder and Major Depressive Disorder.
11 *Front Neurosci* **12**, 842, doi:10.3389/fnins.2018.00842 (2018).

12 42 Sliz, D. & Hayley, S. Major depressive disorder and alterations in
13 insular cortical activity: a review of current functional magnetic imaging
14 research. *Front Hum Neurosci* **6**, 323, doi:10.3389/fnhum.2012.00323 (2012).

15 43 Kang, L. *et al.* Functional connectivity between the thalamus and the
16 primary somatosensory cortex in major depressive disorder: a resting-state fMRI
17 study. *BMC Psychiatry* **18**, 339, doi:10.1186/s12888-018-1913-6 (2018).

18 44 Hallford, D. J. & Sharma, M. K. Anticipatory pleasure for future
19 experiences in schizophrenia spectrum disorders and major depression: A
20 systematic review and meta-analysis. *Br J Clin Psychol* **58**, 357-383,
21 doi:10.1111/bjc.12218 (2019).

22 45 Spano, M. C. *et al.* Anhedonia across borders: Transdiagnostic
23 relevance of reward dysfunction for noninvasive brain stimulation
24 endophenotypes. *CNS Neurosci Ther*, doi:10.1111/cns.13230 (2019).

25 46 Ferenczi, E. A. *et al.* Prefrontal cortical regulation of brainwide circuit
26 dynamics and reward-related behavior. *Science* **351**, aac9698,
27 doi:10.1126/science.aac9698 (2016).

28 47 Yahata, N., Kasai, K. & Kawato, M. Computational neuroscience
29 approach to biomarkers and treatments for mental disorders. *Psychiatry Clin*
30 *Neurosci* **71**, 215-237, doi:10.1111/pcn.12502 (2017).

31 48 Yamada, T. *et al.* Resting-State Functional Connectivity-Based
32 Biomarkers and Functional MRI-Based Neurofeedback for Psychiatric
33 Disorders: A Challenge for Developing Theranostic Biomarkers. *Int J*
34 *Neuropsychopharmacol* **20**, 769-781, doi:10.1093/ijnp/pyx059 (2017).

35 49 Koush, Y. *et al.* Learning Control Over Emotion Networks Through
36 Connectivity-Based Neurofeedback. *Cereb Cortex* **27**, 1193-1202,

doi:10.1093/cercor/bhv311 (2017).

50 Megumi, F., Yamashita, A., Kawato, M. & Imamizu, H. Functional MRI neurofeedback training on connectivity between two regions induces long-lasting changes in intrinsic functional network. *Front Hum Neurosci* **9**, 160, doi:10.3389/fnhum.2015.00160 (2015).

51 Yamashita, A., Hayasaka, S., Kawato, M. & Imamizu, H. Connectivity Neurofeedback Training Can Differentially Change Functional Connectivity and Cognitive Performance. *Cereb Cortex* **27**, 4960-4970, doi:10.1093/cercor/bhx177 (2017).

52 Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods* **16**, 111-116, doi:10.1038/s41592-018-0235-4 (2019).

53 Dickie, E. W. *et al.* ciftify: A framework for surface-based analysis of legacy MR acquisitions. *bioRxiv*, 484428, doi:10.1101/484428 (2019).

54 Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273-289, doi:10.1006/nimg.2001.0978 (2002).

55 Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90-101, doi:10.1016/j.neuroimage.2007.04.042 (2007).

56 Ciric, R. *et al.* Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* **154**, 174-187, doi:10.1016/j.neuroimage.2017.03.020 (2017).

57 Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320-341, doi:10.1016/j.neuroimage.2013.08.048 (2014).

Acknowledgements. We would like to thank Editage (www.editage.com) for English language editing.

Funding. This study was conducted under the “Development of BMI Technologies for Clinical Application” of the Strategic Research Program for Brain Sciences, and the

1 contract research Grant Number JP18dm0307008, supported by the Japan Agency for
2 Medical Research and Development (AMED). This study was also partially supported
3 by the ImPACT Program of the Council for Science, Technology and Innovation
4 (Cabinet Office, Government of Japan). This study was also supported by AMED under
5 Grant Number JP18dm0307001 & JP18dm0307004, JSPS KAKENHI 16H06280
6 (Advanced Bioimaging Support), and the International Research Center for
7 Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced
8 Study (UTIAS). H.I. was partially supported by JSPS KAKENHI 26120002, 18H01098,
9 and 18H05302.

10

11 **Author contributions.** A.Y. and M.K. designed the study. T.Y., N.Y., A.K., N.O., T.I.,
12 R.H., H.M., N.I., M.T., G.O., H.Y., K.H., K.M., S.T., K.K., N.K., H.T., and Y.O.
13 recruited participants for the study, collected their clinical and imaging data, and
14 constructed the database. A.Y. performed data preprocessing. A.Y. and O.Y. performed
15 analysis under the supervision of M.K. and H.I.; A.Y., M.K., O.Y., and H.I. primarily
16 wrote the manuscript.

17

18 **Competing financial interests.** Competing financial interests: M.K., N.Y., R.H., H.I.,
19 N.K. and K.K. are inventors of a patent owned by Advanced Telecommunications
20 Research (ATR) Institute International related to the present work [PCT/JP2014/061543
21 (WO2014178322)]. M.K., N.Y., R.H., N.K. and K.K. are inventors of a patent owned by
22 ATR Institute International related to the present work [PCT/JP2014/061544
23 (WO2014178323)]. M.K. and N.Y. are inventors of a patent application submitted by
24 ATR Institute International related to the present work [JP2015-228970]. A.Y. and M.K.

- 1 are inventors of a patent application submitted by ATR Institute International related to
- 2 the present work [JP2018-192842].
- 3

Table 1. Demographic characteristics of participants in both datasets

Site	HC				MDD				ALL			
	Number	Male/ Female	Age (y)	BDI	Number	Male/ Female	Age (y)	BDI	Number	Male/ Female	Age (y)	BDI
Discover dataset												
Center of Innovation in Hiroshima University (COI)	124	46/78	51.9 ± 13.4	8.2 ± 6.3	70	31/39	45.0 ± 12.5	26.2 ± 9.9	194	77/117	49.4 ± 13.5	14.7 ± 11.7
Kyoto University (KUT)	169	100/69	35.9 ± 13.6	6.0 ± 5.4	17	11/6	43.9 ± 13.3	27.7 ± 10.1	186	111/75	36.7 ± 13.7	8.3 ± 9.1
Showa University (SWA)	101	86/15	28.4 ± 7.9	4.4 ± 3.8	0	-	-	-	101	86/15	28.4 ± 7.9	4.4 ± 3.8
University of Tokyo (UTO)	170	78/92	35.6 ± 17.5	6.7 ± 6.5	62	36/26	38.7 ± 11.6	20.4 ± 11.4	232	114/118	36.4 ± 16.2	14.5 ± 11.8
Summary	564	310/254	38.0 ± 16.1	6.3 ± 5.6	149	78/71	42.3 ± 12.5	24.9 ± 10.7	713	388/325	38.9 ± 15.5	10.7 ± 10.6
Independent validation dataset												
Hiroshima Kajikawa Hospital (HKH)	29	12/17	45.4 ± 9.5	5.1 ± 4.6	33	20/13	44.8 ± 11.5	28.5 ± 8.7	62	32/30	45.1 ± 10.5	17.6 ± 13.7
Hiroshima Rehabilitation Center (HRC)	49	13/36	41.7 ± 11.7	9.1 ± 8.5	16	6/10	40.5 ± 11.5	35.3 ± 9.5	65	19/46	41.4 ± 11.5	15.6 ± 14.3
Hiroshima University Hospital (HUH)	66	29/37	34.6 ± 13.0	6.9 ± 5.9	57	32/25	43.3 ± 12.2	30.9 ± 9.0	123	61/62	38.6 ± 13.3	18.0 ± 14.1
Yamaguchi University (UYA)	120	50/70	45.9 ± 19.5	7.1 ± 5.6	79	36/43	50.3 ± 13.6	29.7 ± 10.7	199	86/113	47.6 ± 17.5	16.0 ± 13.6
Summary	264	104/160	42.2 ± 16.5	7.2 ± 6.3	185	94/91	46.3 ± 13.0	30.3 ± 9.9	449	198/251	43.9 ± 15.3	16.7 ± 13.9

- 1 All demographic distributions are matched between the MDD and HC populations in the discovery dataset ($P > 0.05$) except for BDI.
- 2 The demographic distribution of age is matched between MDD and HC populations in the independent validation dataset ($P > 0.05$).
- 3 The demographic distribution of the sex ratio and BDI were not matched between the MDD and HC populations in the independent validation dataset ($P < 0.05$). BDI: Beck Depression Inventory-II; HC: Healthy Control; MDD: Major Depressive Disorder.
- 4

1 Supplementary Materials

2

3 **Supplementary Note 1**

4 **Prediction performance using SVM**

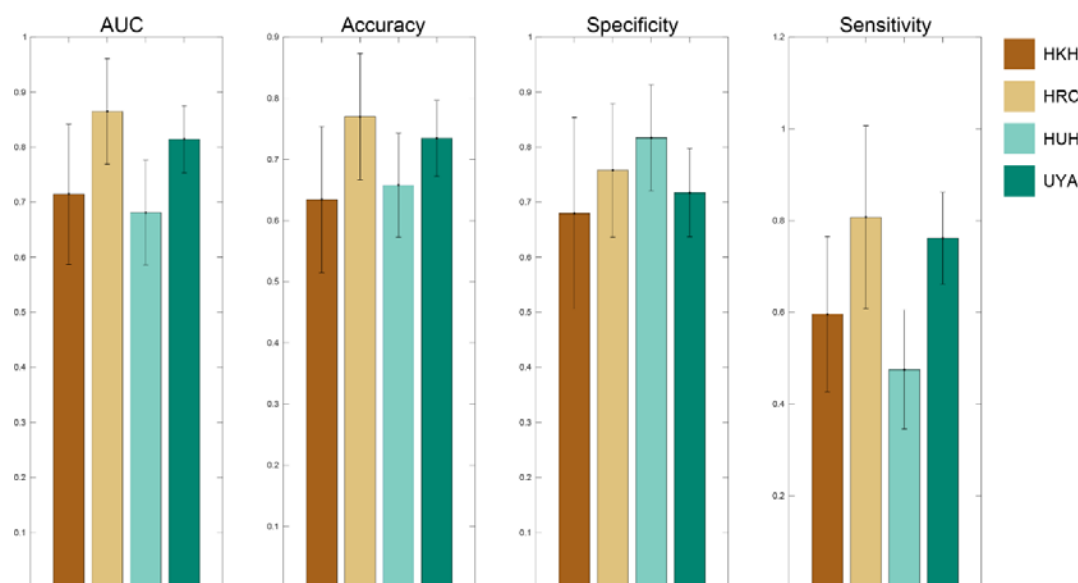
5 Since the performance levels of the prediction models were acceptable, we further tried
6 a support vector machine (SVM) for classification. However, the performance was not
7 improved compared to that in LASSO. In the discovery dataset, the classifier of SVM
8 separated major depressive disorder (MDD) and healthy control (HC) populations with
9 an accuracy of 71%. The corresponding AUC sensitivity and specificity were 0.78, 72%
10 and 70%, respectively. This approach led to a Matthews correlation coefficient (MCC)
11 of 0.35. In the independent validation dataset, the classifier of SVM separated the MDD
12 and HC populations with an accuracy of 70%. The corresponding AUC, sensitivity, and
13 specificity were 0.75, 62%, and 76%, respectively. This approach led to an MCC of
14 0.38.

15

1 **Supplementary Note 2**

2 **Differences in prediction performance among imaging sites**

3 To investigate whether the prediction performances were different among imaging sites
4 in the independent validation dataset, we calculated the 95% confidence interval (CI) of
5 discrimination performances (area under the curve [AUC], accuracy, sensitivity, and
6 specificity) in every imaging site using a bootstrap method. We repeated the bootstrap
7 procedure 1,000 times and calculated the 95% CI for every site. We then checked
8 whether there is a site whose CI does not overlap with the CIs of other imaging sites.
9 We could not find such an imaging site, suggesting no significant systematic difference.
10 However, we noted that the sensitivity at the HUH site was inferior to that at the two
11 other imaging sites (Supplementary Fig. 1: CI of sensitivity in the HUH does not
12 overlap with CI in the HRC or UYA). We discussed the differences in performance
13 among imaging sites in the Discussion section in the main texts.



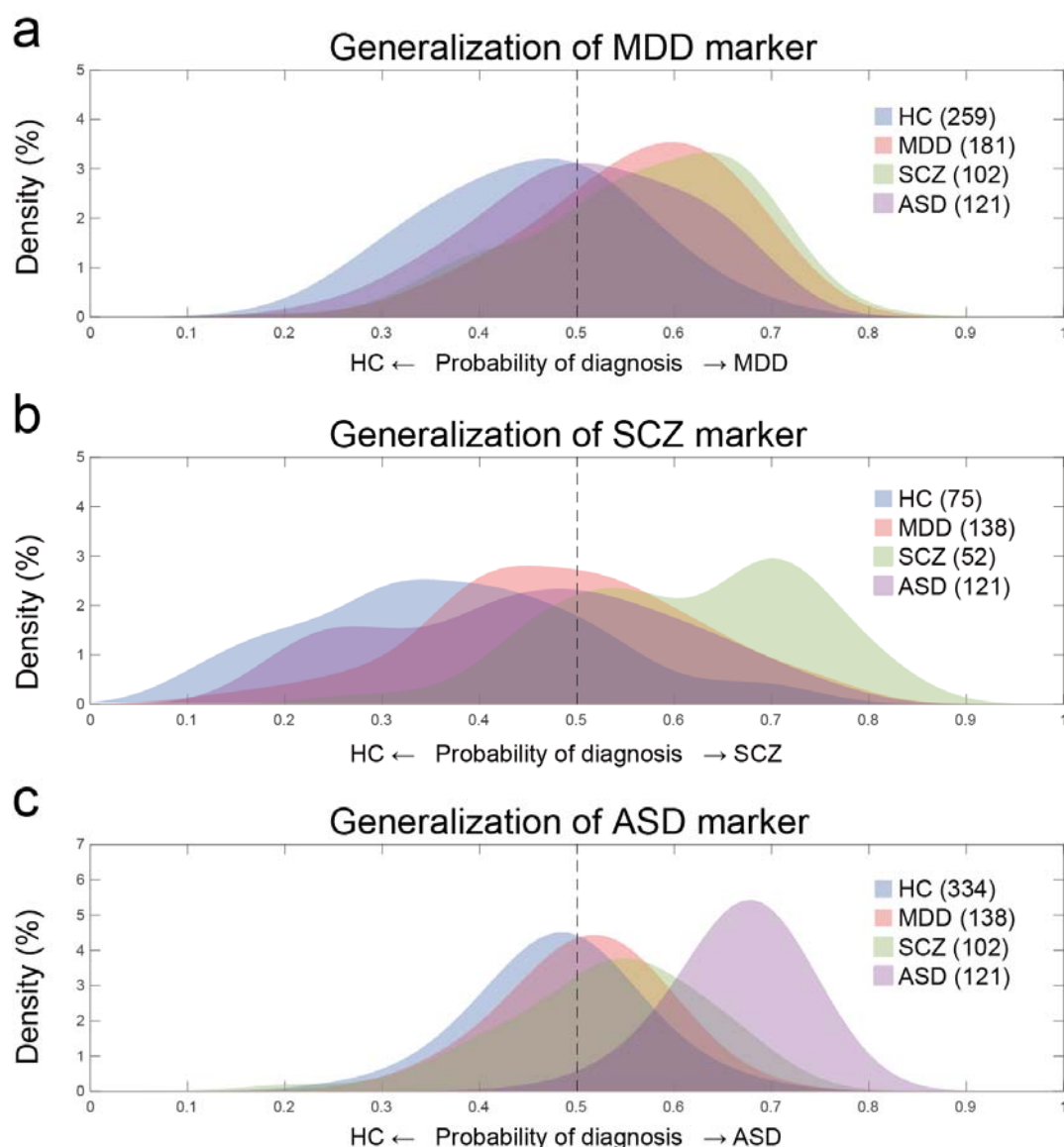
14 **Supplementary Figure 1: Bootstrap prediction performances in the independent**
15 **validation dataset.** Prediction performances of the major depressive disorder (MDD)
16 classifier in the independent validation dataset in each site. Each color bar indicates
17 each site. Error bar shows the 95 % confidence interval from the bootstrap. AUC: Area
18 under the curve, HKH: Hiroshima Kajikawa Hospital; HRC: Hiroshima Rehabilitation
19 Center; HUH: Hiroshima University Hospital; UYA: Yamaguchi University.
20

1 **Supplementary Note 3**

2 **Generalization of the classifiers to other disorders.** In addition to a brain network
3 marker of MDD, we developed brain network markers of schizophrenia (SCZ) and
4 autism spectrum disorder (ASD) using the same method as in the main text. We sought
5 to investigate and confirm the spectral structure among the disorders as revealed by
6 previous studies (12-16 in the main script). For example, if the MDD classifier predicts
7 patients with a different disorder as MDD patients, then the probability of diagnosis for
8 patients with that disorder should be over 0.5. In this case, we may say that the patients
9 possess some degree of MDD-ness and that this disorder is related to MDD according to
10 the imaging biological dimension.

11 Specifically, we first developed SCZ and ASD markers that distinguished
12 between HCs and patients. We used 564 HCs from the discovery dataset in the main
13 text, 102 SCZ patients from 3 sites, and 121 ASD patients from 2 sites (Supplementary
14 Table 5). Data were acquired using the same protocols as for the discovery dataset. We
15 tested the generalizability of the SCZ marker using an independent validation dataset for
16 SCZ patients (52 SCZ patients and 75 HCs from one site, Supplementary Table 5).
17 Since we did not have an independent validation dataset for ASD patients, we tested the
18 performance of the ASD marker using the 10-fold CV. We achieved acceptable
19 performance for both the SCZ marker (Discovery dataset: AUC = 0.85, accuracy = 78%,
20 sensitivity = 75%, specificity = 79%, Independent validation dataset: AUC = 0.89,
21 accuracy = 80%, sensitivity = 81%, specificity = 80%) and ASD marker (Discovery
22 dataset: AUC = 0.76, accuracy = 65%, sensitivity = 70%, specificity = 64%). We then
23 applied these brain network markers to other disorder patients. We computed the
24 probability of diagnosis in the MDD classifier, that is, the MDD-ness of individual
25 patient within the SCZ and ASD data, and vice versa (Supplementary Figure 2).

26 As a result, we found that SCZ patients have high MDD-ness (accuracy = 76%,
27 $p = 2.0 \times 10^{-12}$, two-way binomial test) and ASD-ness (accuracy = 68%, $p = 2.1 \times 10^{-4}$,
28 two-way binomial test). On the other hand, MDD patients did not have high SCZ-ness
29 (accuracy = 46%, $p = 0.35$, two-way binomial test) or ASD-ness (accuracy = 57%, $p =$
30 0.11, two-way binomial test), and ASD patients did not have high SCZ-ness (accuracy =
31 42%, $p = 0.10$, two-way binomial test) or MDD-ness (accuracy = 55%, $p = 0.20$,
32 two-way binomial test).



Supplementary Figure 2: Generalization of the classifiers to other psychiatric disorders. Density distributions of the probability of diagnosis obtained by applying (a) the MDD marker, (b) SCZ marker, and (c) ASD marker to the HC, MDD, SCZ, and ASD patients. In each panel, the patient distribution and the healthy control distribution are plotted separately, with the colored areas representing one or the other. The numbers in parentheses next to HC, MDD, ASD, and SCZ in each panel indicate the number of subjects in the distributions. The independent validation dataset was used in (a) and (b). Healthy controls in (a), (b), and (c) were scanned at the same sites as their corresponding patient data. HC: healthy control; MDD: major depressive disorder; ASD: autism spectrum disorder; SCZ: schizophrenia.

Supplementary Table 1 | Imaging protocols for resting-state fMRI in both datasets

Site	Center of Innovation in Hiroshima University	Kyoto University TimTrio	Showa University	University of Tokyo	Hiroshima Kajikawa Hospital	Hiroshima Rehabilitation Center	Hiroshima University Hospital	Yamaguchi University	Kyoto University Trio
Abbreviation	COI	KUT	SWA	UTO	HKH	HRC	HUH	UYA	KTT
MRI scanner	Siemens Verio	Siemens TimTrio	Siemens Verio	GE MR750w	Siemens Spectra	GE Signa HDxt	GE Signa HDxt	Siemens Skyra	Siemens Trio
Magnetic field strength	3.0 T								
Channels per coil	12	32	12	24	12	8	8	20	8
Field-of-view (mm)	212 × 212				192 × 192	256 × 256	256 × 256	220 × 220	256 × 192
Matrix					64 × 64				64 × 48
Number of slices	40				38	32	32	34	30
Number of volumes	240				107	143	143	200	180
In-plane resolution (mm)	3.3125 × 3.3125				3.0 × 3.0	4.0 × 4.0	4.0 × 4.0	3.4 × 3.4	4.0 × 4.0
Slice thickness (mm)	3.2				3.0	4	4.0	4.0	4.0
Slice gap (mm)	0.8				0	0	0	1.0	0
TR (ms)	2500				2,700	2,000	2,000	2,500	2,000
TE (ms)	30				31	27	27	30	30
Total scan time (min:s)	10:00				5:00	4:46	5:00	8:28	6:00
Flip angle (degree)	80				90	90	90	80	90
Slice acquisition order	Ascending				Ascending	Ascending (Interleaved)	Ascending (Interleaved)	Ascending	Ascending (Interleaved)
Phase encoding	AP	PA	PA	PA	AP	AP	PA	PA	AP
Eyes closed/ open/ fixate	Fixate				Fixate	Fixate	Fixate	Closed	Fixate
*Type of data availability	1	2	2	2	1	1	1	4	2

*Type of data availability, 1) freely available without restriction, allowing commercial reuse, 2) freely available, but not allowing commercial reuse, 3) available after registration to our record, but not allowing commercial reuse, 4) available only to our research group

Supplementary Table 2 Clinical characteristics of major depressive disorder patients in the discovery dataset			
Site	Center of Innovation in Hiroshima University (COI)	Kyoto University (KUT)	University of Tokyo (UTO)
HAMD17 total (mean±1SD)	15.7 ± 5.1	13.1 ± 5.1	10.8 ± 6.3
Diagnostic criteria	MINI	SCID	SCID
Duration of disease (since the first onset)	NA	11.0 ± 5.4 (yr)	9.0 ± 7.7 (yr)
Presence of suicide attempt	55 %	6 %	18 %
Psychiatric comorbidities			
GAD	3 %	6 %	0 %
OCD	7 %	6 %	0 %
ASD	0 %	6 %	0 %
Panic	0 %	13 %	0 %
Psychiatric medications			
Anxiolytic	51 %	63 %	77 %
Antipsychotic	24 %	31 %	32 %
Mood stabilizer	6 %	6 %	45 %
Antidepressant	90 %	94 %	69 %
Subtype of major depressive disorder			
Melancholic	64 %	NA	39 %
Treatment resistance	NA*	100%	NA

*Not applicable because of early treatment data. MINI: Mini International Neuropsychiatric Interview, SCID: Structured Clinical Interview for DSM-IV, NA: Not applicable, HAMD: Hamilton Depression Rating Scale.

1
2

Supplementary Table 3 Description of important FCs										
ID	ROI1			ROI2			r_{HC}	r_{MDD}	t -value	p -value
	Glasser	AAL label	Network	Glasser	AAL label	Network				
1	L.MST	Occipital_Mid	Visual	R.FST	Temporal_Mid	Visual	0.492	0.429	-2.01	0.04
2	L.3b	Postcentral	Motor	L.IFSa	Frontal_Inf_Tri	FPN	-0.079	-0.07	0.48	0.63
3	L.3b	Postcentral	Motor	R.44	Frontal_Inf_Oper	Salience	-0.115	-0.036	3.47	0.0006
4	L.3b	Postcentral	Motor	L.Thalamus	Thalamus	Subcortical	-0.017	0.107	4.52	>0.0001
5	L.7Pm	Precuneus	MR	R.p32	Frontal_Sup_Medial	DMN	-0.049	0.009	2.53	0.01
6	L.2	Postcentral	Motor	R.5m	Paracentral_Lobule	Auditory	0.393	0.328	-2.28	0.02
7	L.8BM	Frontal_Sup_Medial	FPN	L.Thalamus	Thalamus	Subcortical	0.162	0.094	-3.43	0.0007
8	L.47m	Frontal_Inf_Orb	DMN	R.52	Insula	Salience	0.095	0.05	-2.31	0.02
9	L.47s	Frontal_Inf_Orb	Uncertain	R.PoI1	Insula	Subcortical	0.1	0.089	-0.47	0.64
10	L.OP1	Rolandic_Oper	Auditory	R.OP1	Rolandic_Oper	Auditory	0.672	0.531	-4.69	>0.0001
11	L.OP1	Rolandic_Oper	Auditory	R.OP2-3	Rolandic_Oper	Motor	0.528	0.399	-4.57	>0.0001
12	L.Pir	Insula	Subcortical	R.p24	Cingulum_Ant	DMN	0.153	0.123	-1.29	0.20
13	L.PFt	Parietal_Inf	Motor	R.Amy	Hippocampus	Uncertain	0.006	0.006	0.03	0.98
14	L.PBelt	Temporal_Sup	Auditory	L.A4	Temporal_Sup	Auditory	0.977	0.923	-1.94	0.05
15	L.TE2p	Temporal_Inf	Uncertain	R.TE2p	Temporal_Inf	Uncertain	0.361	0.311	-1.81	0.07
16	L.IP0	Occipital_Mid	Visual	L.s32	Frontal_Med_Orb	DMN	-0.051	-0.058	-0.32	0.75
17	L.Ig	Insula	Salience	R.Ig	Insula	Salience	0.768	0.581	-5.98	>0.0001
18	L.TGv	Temporal_Inf	Uncertain	R.10pp	Frontal_Sup_Orb	DMN	0.144	0.139	-0.18	0.86
19	L.TGv	Temporal_Inf	Uncertain	R.STSvp	Temporal_Mid	DMN	0.097	0.083	-0.67	0.50

20	L.A4	Temporal_Sup	Auditory	R.POS2	Cuneus	Visual	-0.134	-0.063	3.37	0.0008
21	R.POS2	Cuneus	Visual	R.A4	Temporal_Sup	Auditory	-0.143	-0.08	2.83	0.0049
22	R.5m	Paracentral_Lobule	Saliency	R.1	Postcentral	Motor	0.504	0.392	-3.55	0.0004
23	R.1	Postcentral	Motor	R.Thalamus	Thalamus	Subcortical	-0.086	0.058	5.01	>0.0001
24	R.a24	Cingulum_Ant	DMN	R.52	Insula	Auditory	0.122	0.081	-1.78	0.076
25	R.OP1	Rolandic_Oper	Auditory	R.OP2-3	Rolandic_Oper	Motor	0.712	0.566	-5.29	>0.0001
26	R.OP2-3	Rolandic_Oper	Motor	L.Thalamus	Thalamus	Subcortical	0.155	0.19	1.39	0.17
27	R.52	Insula	Auditory	R.s32	Frontal_Med_Orb	DMN	0.072	0.047	-1.17	0.24
28	R.FOP4	Insula	Attention	R.Thalamus	Thalamus	Subcortical	0.165	0.092	-3.78	0.0002
29	R.FST	Temporal_Mid	Visual	B.Stem	<out of bound>	Uncertain	-0.053	-0.039	0.67	0.50
30	L.Caudate	Caudate	Subcortical	B.Stem	<out of bound>	Uncertain	-0.019	-0.004	0.74	0.46
31	L.Caudate	Caudate	Subcortical	R.Thalamus	Thalamus	Subcortical	0.368	0.282	-3.72	0.0002

- 1 ROI labels were determined by referring to AAL and Neurosynth (<http://neurosynth.org/locations/>)
- 2 DMN: Default mode network; FPN: Fronto-parietal task control; MR: Memory retrieval. *t*-value and *p*-value are the results of *t*-test of
- 3 functional connectivity values between MDD patients and HCs in the independent validation dataset.

Supplementary Table 4 All functional connections related to only BDI score regression model								
ID	ROI1			ROI2			r-value with BDI (Discovery)	r-value with BDI (Validation)
	Glasser	AAL label	Network	Glasser	AAL label	Network		
1	L.3b	Postcentral_L	Motor	R.Thalamus	Thalamus_R	Subcortical	0.28	0.04
2	L.POS1	Precuneus_L	DMN	R.HC	Hippocampus_R	Uncertain	0.22	-0.04
3	L.9m	Frontal_Sup_Medial_L	DMN	R.9m	Frontal_Sup_Medial_R	DMN	-0.21	-0.06
4	L.AAIC	Insula_L	Uncertain	R.p24	Cingulum_Ant_R	DMN	-0.21	-0.02
5	L.TGd	Temporal_Pole_Mid_L	DMN	R.STSvp	Temporal_Mid_R	DMN	-0.21	-0.08
6	L.FST	Temporal_Mid_L	Attention	R.RSC	Cingulum_Post_R	DMN	0.21	0.04
7	L.VMV2	Lingual_L	DMN	R.VMV2	Lingual_R	Visual	-0.23	0.02
8	L.FOP5	Insula_L	Salience	R.FFC	Fusiform_R	Uncertain	0.18	-0.01
9	L.PI	Temporal_Sup_L	Attention	R.TE1m	Temporal_Mid_R	Uncertain	-0.20	-0.01
10	R.a24	Cingulum_Ant	DMN	R.52	Insula	Auditory	-0.24	0.06
11	R.52	Insula	Auditory	R.s32	Frontal_Med_Orb	DMN	-0.24	0.03
12	R.AIP	Parietal_Inf_R	FPN	R.LBelt	Temporal_Sup_R	Auditory	0.18	-0.05
13	L.Putamen	Putamen_L	Subcortical	R.Putamen	Putamen_R	Subcortical	-0.25	0.03

ROI labels were determined by referring to AAL and Neurosynth (<http://neurosynth.org/locations/>)

DMN: Default mode network; FPN: Fronto-parietal task control; MR: Memory retrieval.

1
2
3

Supplementary Table 5 | Demographic characteristics of participants in both datasets

Site	Number	SCZ		Number	ASD		Number	HC	
		Male/ Female	Age (y)		Male/ Female	Age (y)		Male/ Female	Age (y)
Kyoto University (KUT)	48	24/24	41.5± 10.4	0	-	-	see Table 1 in the main text		
Showa University (SWA)	18	14/4	42.8± 8.6	111	96/15	32.0 ± 7.5	see Table 1 in the main text		
University of Tokyo (UTO)	36	24/12	31.4± 10.3	10	9/1	37.0± 9.6	see Table 1 in the main text		
Kyoto University Trio (KTT)	52	27/25	37.2± 9.4	0	-	-	75	48/27	28.9± 9.1

1 SCZ: Schizophrenia, ASD: Autism spectrum disorder, HC: Healthy control.

2

Supplementary Table 6 | Imaging protocols for resting state fMRI in the traveling subject dataset.

Site	ATR TimTrio	ATR Verio	Center of Innovation in Hiroshima University	Hiroshima University Hospital	Hiroshima Kajikawa Hospital	Kyoto Prefectural University of Medicine	Showa University	Kyoto University TimTrio	Kyoto University Skyra	University of Tokyo	Yaesu-clinic scanner 1	Yaesu-clinic scanner 2
Abbreviation	ATT	ATV	COI	HUH	HKH	KPM	SWA	KUT	KUS	UTO	YC1	YC2
MRI scanner	Siemens TimTrio	Siemens Verio	Siemens Verio	GE Signa HDxt	Siemens Spectra	Philips Achieva	Siemens Verio	Siemens TimTrio	Siemens Skyra	GE MR750W	Philips Achieva	Philips Achieva
The number of scans	132	27	27	18	18	27	27	27	27	27	27	27
Magnetic field strength						3T						
Number of channels per coil	12	12	12	8	12	8	12	32	32	24	8	8
Field-of-view (mm)						212 x 212						
Matrix						64 × 64						
Number of slices	40	39	40	35	35	40	40	40	40	40	40	40
Number of volumes						240						
In-plane resolution (mm)						3.3125 × 3.3125						
Slice thickness (mm)						3.2						
Slice gap (mm)						0.8						
TR (ms)						2,500						
TE (ms)						30						
Total scan time (min:s)						10:00						
Flip angle (deg)						80						
Slice acquisition order						Ascending						
Phase encoding	PA	PA	AP	PA	PA	AP	PA	PA	AP	PA	AP	AP
Eye closed / fixate						Fixate						
Field map	✓	✓	✓	-	-	✓	✓	✓	✓	✓	✓	-

1 Nine healthy participants (all male participants; age range, 24–32 years; mean age, 27 ± 2.6 years) were scanned at each of 12 sites, producing a total of 411 scan
2 sessions. Each participant underwent three rs-fMRI sessions of 10 min each at nine sites, two sessions of 10 min each at two sites (HKH & HUH), and five cycles
3 (morning, afternoon, next day, next week, next month) consisting of three 10-minute sessions each at a single site (ATT). In the latter situation, one participant
4 underwent four rather than five sessions at the ATT site due to poor physical condition. Thus, a total of 411 sessions were conducted. UTO: University of Tokyo;
5 HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications
6 Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; HKH:
7 Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KUS: Siemens Skyra scanner at Kyoto University; KPM: Kyoto Prefectural
8 University of Medicine; YC1: Yaesu Clinic 1; YC2: Yaesu Clinic 2; TR: repetition time; TE: echo time.
9