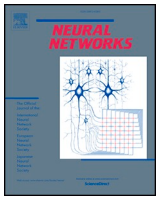
Neural Networks 187 (2025) 107335

Contents lists available at ScienceDirect

Neural Networks

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



Full Length Article

Comprehensive evaluation of pipelines for classification of psychiatric disorders using multi-site resting-state fMRI datasets

Yuji Takahara [a,b,*], Yuto Kashiwagi [a,c], Tomoki Tokuda [a], Junichiro Yoshimoto [a,d,e], Yuki Sakai [a,f], Ayumu Yamashita [a,g], Toshinori Yoshioka [a,f], Hidehiko Takahashi [h,i], Hiroto Mizuta [j], Kiyoto Kasai [k,l,m], Akira Kunimitsu [n], Naohiro Okada [k,l], Eri Itai [o], Hotaka Shinzato [o], Satoshi Yokoyama [o], Yoshikazu Masuda [o], Yuki Mitsuyama [o], Go Okada [o], Yasumasa Okamoto [o], Takashi Itahashi [p], Haruhisa Ohta [p], Ryu-ichiro Hashimoto [p,q], Kenichiro Harada [r], Hirotaka Yamagata [r], Toshio Matsubara [r], Koji Matsuo [s], Saori C. Tanaka [a,t], Hiroshi Imamizu [a,u], Koichi Ogawa [b], Sotaro Momosaki [b], Mitsuo Kawato [a,f], Okito Yamashita [a,v,**]

[a] Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institute International, Kyoto, Japan
[b] Laboratory for Drug Discovery and Disease Research, Shionogi & Co., Ltd., Osaka, Japan
[c] Laboratory for Drug Discovery and Development, Shionogi & Co., Ltd., Osaka, Japan
[d] Department of Biomedical Data Science, School of Medicine, Fujita Health University, Aichi, Japan
[e] International Center for Brain Science, Fujita Health University, Aichi, Japan
[f] XNef, Inc., Kyoto, Japan
[g] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[h] Department of Psychiatry and Behavioral Sciences, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Japan
[i] Center for Brain Integration Research, Tokyo Medical and Dental University, Japan
[j] Department of Psychiatry, Graduate School of Medicine, Kyoto University, Japan
[k] Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[l] The International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo, Tokyo, Japan
[m] UTokyo Institute for Diversity and Adaptation of Human Mind (UTIDAHM), The University of Tokyo, Tokyo, Japan
[n] Department of Radiology, International University of Health and Welfare, Mita Hospital, Tokyo, Japan
[o] Department of Psychiatry and Neurosciences, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan
[p] Medical Institute of Developmental Disabilities Research, Showa University, Tokyo, Japan
[q] Department of Language Sciences, Tokyo Metropolitan University, Tokyo, Japan
[r] Division of Neuropsychiatry, Department of Neuroscience, Yamaguchi University Graduate School of Medicine, Yamaguchi, Japan
[s] Department of Psychiatry, Faculty of Medicine, Saitama Medical University, Saitama, Japan
[t] Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan
[u] Department of Psychology, Graduate School of Humanities and Sociology, The University of Tokyo, Tokyo, Japan
[v] RIKEN, Center for Advanced Intelligence Project, Tokyo, Japan

ARTICLE INFO

Keywords:
Classification biomarker
Pipeline
Generalization
fMRI
Multi-site dataset
Major depressive disorder

ABSTRACT

Objective classification biomarkers that are developed using resting-state functional magnetic resonance imaging (rs-fMRI) data are expected to contribute to more effective treatment for psychiatric disorders. Unfortunately, no widely accepted biomarkers are available at present, partially because of the large variety of analysis pipelines for their development. In this study, we comprehensively evaluated analysis pipelines using a large-scale, multi-site fMRI dataset for major depressive disorder (MDD). We explored combinations of options in four sub-processes of the analysis pipelines: six types of brain parcellation, four types of functional connectivity (FC) estimations, three types of site-difference harmonization, and five types of machine-learning methods. A total of

* Corresponding author at: 1-1, Futaba-cho 3-chome, Toyonaka, Osaka 561-0825, Japan.
** Corresponding author at: 2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan.
E-mail addresses: yuji.takahara@shionogi.co.jp (Y. Takahara), oyamashi@atr.jp (O. Yamashita).

https://doi.org/10.1016/j.neunet.2025.107335
Received 11 July 2024; Received in revised form 17 February 2025; Accepted 27 February 2025
Available online 28 February 2025

360 different MDD classification biomarkers were constructed using the SRPBS dataset acquired with unified protocols (713 participants from four sites) as the discovery dataset, and datasets from other projects acquired with heterogeneous protocols (449 participants from four sites) were used for independent validation. We repeated the procedure after swapping the roles of the two datasets to identify superior pipelines, regardless of the discovery dataset. The classification results of the top 10 biomarkers showed high similarity, and weight similarity was observed between eight of the biomarkers, except for two that used both data-driven parcellation and FC computation. We applied the top 10 pipelines to the datasets of other psychiatric disorders (autism spectrum disorder and schizophrenia), and eight of the biomarkers exhibited sufficient classification performance for both disorders. Our results will be useful for establishing a standardized pipeline for classification biomarkers.

1. Introduction

Most psychiatric disorders, including major depressive disorder (MDD), are diagnosed based on the Diagnostic and Statistical Manual of Mental Disorders, such as DSM-5, or the International Classification of Diseases. Their diagnosis is based on visible symptoms and medical interviews. Although the overlapping phenotypes among multiple psychiatric disorders complicate the selection of appropriate treatment at an early stage, no objective and practical markers have yet led to more refined diagnoses and targeted treatment (Arnow et al., 2015; Kapur et al., 2012).

Non-invasive brain-imaging techniques are expected to elucidate the patterns of brain structure and function; that is, the neurophenotypes, that characterize these disorders (Craddock et al., 2013; van Essen & Ugurbil, 2012). In particular, acquiring resting-state functional magnetic resonance imaging (rs-fMRI) is easy and can be used to develop classification markers between healthy individuals and those with psychiatric or developmental disorders (Clare Kelly et al., 2008; van Essen & Ugurbil, 2012), such as Alzheimer's disease (Chen et al., 2011; Greicius et al., 2004), schizophrenia (Calhoun et al., 2012; Garrity et al., 2007; Jafri et al., 2008; Yoshihara et al., 2020; Zhou et al., 2007), autism spectrum disorder (Mellema et al., 2022; Plitt et al., 2015; Yahata et al., 2016), depression (Craddock et al., 2009; Ichikawa et al., 2020; Yamashita et al., 2020), and attention-deficit hyperactivity disorder (Milham et al., 2012). In recent years, several global projects have acquired large-scale rs-fMRI data (Koike et al., 2021; Martino et al., 2013; Tanaka et al., 2021; van Essen et al., 2013), and the understanding of psychiatric disorders and development of markers have been progressing. However, unfortunately, no practical diagnostic markers have yet been identified (Castellanos et al., 2013; Kapur et al., 2012), perhaps because of the absence of superior pipelines and the lack of generalization capability.

The development of classification biomarkers using rs-fMRI comprises several processes, and multiple methods as well as innumerable pipelines are included in each process (Arbabshirani et al., 2017; Brown & Hamarneh, 2016; Wolfers et al., 2015). The diversity of pipelines has a significant effect on the diagnostic and generalization performance, and few studies have searched for ideal pipelines (Dadi et al., 2019; Graña & Silva, 2021; Mellema et al., 2022; Pervaiz et al., 2020). The absence of standard pipelines reduces the reliability of rs-fMRI biomarkers. As no widely accepted standard pipeline has been established for optimal biomarker development, identifying a suitable pipeline in the field of psychiatric disorders is critical for developing diagnostic markers for rs-fMRI.

As most older biomarkers are based on single-site data to avoid unknown effects of the heterogeneity of discovery datasets, they often exhibit a lack of generalization capabilities. Differences in scanners, imaging procedures, and instructions to participants may also affect rs-fMRI data. Discovery datasets must comprise large-size, multi-site data to develop markers with a high generalization capability. Once a marker has been developed using a multi-site dataset, it must be validated using another independent multi-site dataset. This validation test enhances the reliability of the marker. Several methods have minimized the inter-site differences in rs-fMRI, including ComBat (Fortin et al., 2017, 2018; Johnson et al., 2007; Yu et al., 2018) and traveling-subject harmonization (Yamashita et al., 2019). These methodological approaches are also expected to enhance the generalization capability of markers.

This study aims to explore the pipeline for identifying MDD classification biomarkers with strong performance for independent validation datasets using a large, multi-site rs-fMRI dataset. This study presents three key points of novelty. First, to the best of our knowledge, this is the first study to systematically evaluate advanced methods—such as Glasser's surface-based parcellation, distance correlation, and a site-difference harmonization approach—within a unified framework for biomarker discovery. Second, we assess the generalization performance of the biomarkers using an independent validation dataset, providing evidence for their applicability across datasets. Third, we swap the roles of the discovery and validation datasets, confirming that the pipeline's performance was not dependent on a specific discovery dataset and highlighting its robustness and flexibility under varying conditions.

2. Materials and methods

2.1. Ethics statement

All participants in this study provided written informed consent. All recruitment procedures and experimental protocols were conducted in accordance with the Declaration of Helsinki and approved by the institutional review boards of the respective institutions of the principal investigators (Advanced Telecommunications Research Institute International [approval numbers: 13–133, 14–133, 15–133, 16–133, 17–133, and 18–133], Hiroshima University [E-38], Kyoto Prefectural University of Medicine [RBMR-C-1098], Showa University [B-2014–019 and UMIN000016134], The University of Tokyo Faculty of Medicine [3150], Kyoto University [C809 and R0027], and Yamaguchi University [H23–153 and H25–85]).

2.2. Patients and subjects

We analyzed the datasets that were previously presented in Yamashita et al. (2020): (1) Dataset I contained data from 713 participants (564 healthy controls (HCs) from four sites and 149 MDDs from three sites); (2) Dataset II contained data from 449 participants (264 HCs and 185 MDDs from four independent sites); and (3) Dataset III contained data from 231 participants (125 with autism spectrum disorder (ASD) from two sites and 106 with schizophrenia (SCZ) from three sites). In the BDI-II scale, MDD patients included in Dataset II had a higher severity of depression compared to patients in Dataset I. Details of the datasets are presented in Tables 1, 2, and S1. The HC group was treated as the typical development (TD) group when combined with the ASD group from Dataset III.

2.3. Preprocessing

We preprocessed the rs-fMRI data using fMRIPrep version 1.0.8 (Esteban et al., 2019). The first 10 s of the data were discarded to allow for T1 equilibration. The preprocessing steps were as follows: slice-timing correction, realignment, co-registration, distortion

correction using a field map, segmentation of the T1-weighted structural images, and normalization to the Montreal Neurological Institute space. The surface projection step was performed only for surface-based parcellation. The "fieldmap-less" distortion correction was performed for Dataset II owing to a lack of field map data. For more details on these pipelines, refer to http://fmriprep.readthedocs.io/en/1.0.8/workflows.html. Co-registration was unsuccessful for the data of six participants in Dataset II; therefore, we excluded them from further analysis.

2.4. Parcellation

We considered six methods for extracting the brain regions of interest (ROIs): five pre-defined parcellations, which were previously used for the development of depression diagnosis and stratification biomarkers (Ichikawa et al., 2020; Tokuda et al., 2018; Yamashita et al., 2020; Yoo et al., 2018), and one data-driven parcellation that was reported as superior (Dadi et al., 2019, Fig. 1).

Pre-defined parcellation:

1) Glasser's surface-based method: 379 ROIs, including 360 cortical parcels and 19 subcortical parcels (Glasser et al., 2016), which were utilized in ciftify toolbox version 2.0.2–2.0.3 (Dickie et al., 2019).
2) Glasser's volume-based method [https://figshare.com/articles/dataset/HCP-MMP1_0_projected_on_fsaverage/3498446], which has only 360 cerebral cortical ROIs.
3) Shen's atlas, derived using a group-wise spectral clustering algorithm, which contains 268 ROIs (Shen et al., 2013).
4) BrainVISA, anatomically defined in the BrainVISA Sulci Atlas (BSA), which has 145 anatomically defined ROIs covering the entire cerebral cortex (http://brainvisa. Info, Perrot et al., 2011).
5) FIND laboratory's parcellation (Shirer et al., 2012), based on functional (rather than structural) ROIs, from which we selected 78 ROIs, excluding cerebellum-related ROIs.

Table 2
Demographic characteristics of participants for other disorders.
The rs-fMRI data acquired from 125 ASDs (two sites) and 106 SCZs (three sites) were used to develop ASD and SCZ biomarkers. Dataset 1, used as the control group, features the same group composition for TD and HC, with the only difference being that the naming changes depending on whether the disease group is classified as ASD. All data were acquired using a unified imaging protocol (SRPBS DecNef project).

	Dataset III					
	ASD			SCZ		
Site	Number	Male/ Female	Age (y)	Number	Male/ Female	Age (y)
Center of Innovation, Hiroshima University (COI)	0	-	-	0	-	-
Kyoto University (KUT)	0	-	-	51	26/25	41.0 ± 10.7
Showa University (SWA)	115	100/15	32.1 ± 7.8	19	15/4	42.9 ± 8.4
University of Tokyo (UTO)	10	9/1	37.0 ± 9.6	36	24/12	31.4 ± 10.3
Summary	125	109/16	32.5 ± 8.0	106	65/41	38.1 ± 11.2

Data-driven parcellation: We defined the ROIs using a linear decomposition method, namely dictionary learning (Mensch et al., 2016). A data-driven atlas was constructed using the discovery dataset. We set the number of components to dim = 80, which is the optimal number (Dadi et al., 2019).

Table 1
Demographic information of participants. The data in Dataset I were acquired using a unified imaging protocol (the Japanese Strategic Research Program for the Promotion of Brain Science (SRPBS) Decoded Neurofeedback (DecNef) Project). Dataset II was acquired from four completely different sites from those of Dataset I. Please refer to Table S1 for more detailed information on the protocols.

	HC or TD				MDD				All			
Site	Number	Male/ Female	Age (y)	BDI	Number	Male/ Female	Age (y)	BDI	Number	Male/ Female	Age (y)	BDI
Dataset I												
Center of Innovation, 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
Dataset II												
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

TD: typical development; BDI: Beck Depression Inventory

Raw data	Parcellation (number of ROIs)			FC	Harmonization for discovery dataset	Machine learning		Output
	Surfaced-based	Functional	Glasser + 19 ROI (379)	Pearson correlation	Traveling-subject harmonization	Sparse	LASSO	
	Volume-based	Functional	Glasser (360)	Tangent-space covariance	ComBat		SLR	
			Shen (268)	Partial correlation	No correction	Non-sparse	RIDGE	
			FIND lab. (77)	Distance correlation			SVM	
		Functional & data-driven	Dictionary learning (80)			Tree-based	Random forest	
		Structural	BSA (145)					

Fig. 1. Design for comprehensive exploration of analysis pipelines of resting-state FC MDD classification biomarker.
Following preprocessing with fmriprep-ciftify, 360 biomarkers were constructed using 360 pipelines (6 parcellations × 4 FC × 3 site-difference harmonization × 5 machine learning).

2.5. Functional connectivity (FC) matrix

Four FC calculation methods were considered (Fig. 1): Pearson's full correlation coefficient, tangent-space covariance (Varoquaux et al., 2010), partial correlation, and distance correlation (Yoo et al., 2019). The FC for each participant was calculated from the rs-fMRI BOLD signals across the ROIs for each parcellation. All FCs were calculated using the Nilearn Python library (version 0.6.1) [https://nilearn.github.io/stable/index.html]. We used the FC values of the lower triangular matrix of the connectivity matrix and applied Fisher's z-transformation to each FC except for the tangent-space covariance. We calculated the tangent-space covariances of both the discovery and validation datasets based on the group mean values of the former.

2.6. Preprocessing for ROI time series

Physiological noise regressors were extracted using CompCor (Behzadi et al., 2007). We used linear regression with 12 regression parameters to remove several sources of spurious variance: six motion parameters, the average signals over the entire brain (global signal), and five anatomical CompCor components.

We applied a temporal bandpass filter to the time series using a second-order Butterworth filter with a pass band between 0.01 and 0.08 Hz to restrict the analysis to low-frequency fluctuations, which are characteristic of rs-fMRI BOLD activity (Ciric et al., 2017).

A scrubbing process based on head motion was conducted using frame displacement (FD; Power et al., 2014), which was calculated using Nipype (https://nipype.readthedocs.io/en/latest/). We removed the volumes with FD $>$ 0.5 mm, as proposed in a previous study (Power et al., 2014). Using this threshold, 6.3% $\pm$ 13.5 vol (mean $\pm$ SD) per rs-fMRI session were removed from all datasets. Subjects whose ratio of excluded volumes by scrubbing exceeded the mean + 3 SD were removed, resulting in 48 participants being removed from all datasets. We confirmed that there was no bias in the FD exclusion rate between MDD patients and healthy subjects in both Datasets I and II ($p = 0.72$ for Dataset I and $p = 0.39$ for Dataset II, Kruskal–Wallis test, data not shown). This resulted in 683 participants (545 HCs and 138 MDDs) in Dataset I, 444 (263 HCs and 181 MDDs) in Dataset II, and 218 (116 ASDs and 102 SCZs) in Dataset III.

2.7. Harmonization

We used three site-difference harmonization methods in the discovery dataset (Fig. 1): traveling-subject harmonization, ComBat, and no correction. Traveling-subject harmonization enabled us to estimate the measurement (**m**) by controlling the participant bias (**p**) using the traveling-subject dataset (Table S2), in which multiple participants traveled to multiple recording sites that recorded their rs-fMRIs using identical recording protocols (Yamashita et al., 2019). For each connectivity corresponding to each subject in the traveling-subject dataset, the regression model can be expressed as follows:

$$Connectivity = \boldsymbol{x}_m^T \boldsymbol{m} + \boldsymbol{x}_p^T \boldsymbol{p} + const + e,$$

$$such\ that \sum_j^9 p_j = 0,\ \sum_k^4 m_k = 0,$$

where **m** represents the measurement bias (four sites × 1), **p** represents the participant factor (nine traveling subjects × 1), *const* represents the average FC values across all participants from all sites, and $e \sim \mathcal{N}(0,\ \gamma^{-1})$ represents noise. In addition, $\mathbf{x}_m$ and $\mathbf{x}_p$ are vectors represented by 1-of-K binary coding. $\mathbf{x}_m$ for the measurement bias $\boldsymbol{m}$ belonging to site k is a binary vector, with all elements equal to 0, except for element k, which is equal to 1. The measurement biases were removed by subtracting the estimated measurement biases. Thus, the harmonized FC values were set as follows:

$$Connectivity^{Harmonized} = Connectivity - \boldsymbol{x}_m^T \widehat{\boldsymbol{m}},$$

which denotes the estimated measurement bias. The applicable data are not limited to the traveling-subject dataset but also include data acquired at the facilities where the traveling-subject dataset was collected. Detailed information has been described previously (Yamashita et al., 2019). The ComBat harmonization method (Dansereau et al., 2017; Fortin et al., 2017, 2018; Johnson et al., 2007) is a well-known control method for site differences in FC. Although the traveling-subject method requires the traveling-subject dataset to be acquired in advance, the ComBat approach allows site differences to be corrected using only discovery data. We performed ComBat harmonization to correct only for site differences, while maintaining the FC correlated with two biological

covariates: age and sex. To prevent potential information leakage, diagnostic information was intentionally excluded from the covariates. The deleterious effects of site differences on prediction accuracy decrease as the total sample size increases (Dansereau et al., 2017). Additionally, the ComBat method performs scale adjustment by harmonizing the variance of data across sites. Harmonization was not applied to the validation dataset to determine the effects of harmonization on biomarker construction, which constituted no correction.

2.8. Machine learning

We considered five supervised machine-learning methods (Fig. 1): *1)* the least absolute shrinkage and selection operator (LASSO) was implemented using the *"lassoglm"* function, and we set *"NumLambda"* to 25 and *"CV"* to 10. λ was determined according to the one standard error rule in which the largest λ within the standard deviation of the minimum prediction error was selected. *2)* Sparse logistic regression (SLR) was performed using the *"biclsfy_slrvar"* function, and we set *"nlearn"* to 300 and *"usebias"* to 1 (https://bicr.atr.jp/~oyamashi/SLR_WEB.html) (Yamashita et al., 2008). *3)* RIDGE was implemented using the *"fitclinear"* function, and we set *"Solver"* to {*"sgd", "lbfgs"*}, *"OtimizeHyperparameters"* to *"Lambda,"* and *"Kfold"* to 10. *4)* The support vector machine (SVM) was employed using the *"fitcsvm"* function, and we set *"KernelScale"* and *"Boxconstraint"* to the optimization parameters derived from the *"bayesopt"* function. *5)* The random forest was implemented using the *"TreeBagger"* function, and we set *"NumTrees"* to 1000 and *"PredictorSelection"* to *"interaction-curvature"*. All functions described here were executed in MATLAB (R2018b, MathWorks, USA).

2.9. Construction and validation of MDD classifier

We constructed a brain network biomarker for MDD that was classified between HCs and MDDs using the discovery dataset (Dataset I) based on each FC value (Fig. S1). We used the 10-fold nested cross-validation (CV) method, which is based on a previously proposed method (Yamashita et al., 2020) with a slight modification to standardize the validation datasets to prevent overfitting. We first divided Dataset I into a training set (nine of 10 folds), which was used to train the model, and a test set (one of 10 folds), which was used for testing. We used an undersampling method to equalize the numbers between the MDD and HC groups (Wallace et al., 2011) to prevent bias owing to differences in the numbers and ages between the two groups. As only a subset of the training set was used after undersampling, we repeated the random sampling procedure 10 times (i.e., subsampling). Undersampling was performed by matching the mean ages of the MDD and HC groups in each undersampling and standardizing both the undersampled training subset and test set with the mean and standard deviation values of each FC in each undersampling. Subsequently, we fitted a model to each subsample and created 10 classifiers. The mean classifier output value (classification probability) indicated the classifier output. Subjects with a classification probability greater than 0.5 were considered as MDD patients. To construct classification markers for ASD and SCZ, subsampling and undersampling were performed on the same healthy or typical development subject group as for MDD and patients with each disease, and the classification performances were evaluated using the 10-fold nested CV method.

We tested the generalization capability of the classification biomarkers using the validation dataset (Dataset II). One biomarker consisted of 100 classifiers, derived from 10-fold cross-validation with 10 iterations of undersampling per fold, all of which were used. The classification probability of each subject was calculated after the FC was standardized using the mean value and standard deviation of the training subset.

2.10. Evaluation criteria

We calculated the area under the curve (AUC), accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC; Chicco, 2017). Because it is difficult to express the marker's superiority with a single metric, we defined and used the two custom metrics. Recognizing that the AUC alone may not provide an accurate comparison of the classification performance for imbalanced data, we calculated the "composite score," which combined the AUC and scale-adjusted MCC:

$$Composite\ score = 0.5 * AUC + 0.5 * (MCC + 1)/2.$$

We also quantified the instability of the classification results between the 10-fold CV result for the discovery dataset and the application result for the validation dataset and defined the "instability":

$$Instability = \sqrt{\left(Sensitivity_{discovery} - Sensitivity_{validation}\right)^2 + \left(Specificity_{discovery} - Specificity_{validation}\right)^2}.$$

The above two-custom metrics helped narrow down the number of metrics and clarified the selection criteria.

2.11. Confirmation of pipeline stability

We confirmed the pipeline stability by swapping the roles of the datasets. We constructed MDD classification biomarkers using identical pipelines, except for the traveling-subject harmonization method, in which Dataset II was used as the discovery dataset and Dataset I was used as the validation dataset. We calculated the composite scores, instability, and classification probabilities for each marker. The correlation coefficient of the classification outputs for the marker, calculated before and after swapping the roles of the datasets, is referred to as the 'consistency of classifier output.'

2.12. Marker and pipeline ranking

For the pipeline ranking, we used standardized values of five indicators: composite scores and instability values when Dataset I was used as the discovery dataset, composite scores and instability values when Dataset II was used as the discovery dataset, and consistency of classifier outputs. Unlike other metrics, "Instability" indicates better generalization performance with smaller values. Therefore, for the ranking calculation, only its sign was inverted. Because it is preferable for all five indicators to exhibit large values as markers, ranking based on the sum of these five scores was conducted for 240 pipelines, where both datasets could serve as the discovery dataset (Table 4).

2.13. Evaluation of similarity of important FCs

We examined the network-level similarity of the important FCs among the top 10 MDD biomarkers with the highest classification performance. As each classification biomarker was composed of 100 classifiers, the contribution of each FC to the classification biomarker could

be quantified by the sum of the absolute values of the weights or out-of-bag predictive importance. We applied the FCs that were extracted as the top 5 % of the high-contribution FCs in each marker to the same atlas of seven brain networks (Buckner et al., 2011; Choi et al., 2012; Thomas Yeo et al., 2011) to compare the similarity of FC patterns between biomarkers constructed using different parcellation options. The ROIs that constituted each FC were counted, classified on each brain network label, and divided by twice the total number of the top 5 % high-contribution FCs. As biases existed in the number of ROIs classified into each network depending on the parcellation, we calculated the existence probability of the ROIs of all FCs in the brain network for each parcellation, and divided the existence probability of the top 5 % FCs by the existence probability of the total FCs. The correlation coefficient of the brain network utilization between the high-contribution FCs of any two biomarkers was used as a quantitative index of marker similarity. Similarity is important if the correlation coefficient between two different utilization rates is significantly higher than a randomness threshold. We randomly selected 5 % of the FCs from each parcellation and calculated the corrected utilization rate of Yeo's brain network and the correlation coefficient between the two utilization rates. We performed this operation 10,000 times and set the statistical significance to a certain threshold (permutation test, $p < 0.05$, one-sided).

2.14. Evaluation of pipelines for other disorders

We confirmed the applicability of the top 10 pipelines that were suitable for constructing MDD classification markers to two other disorders: SCZ and ASD. We developed classification biomarkers for SCZ and ASD using the same process as that for MDD. The same data as the HCs described in Section 2.2 were used for the HC and TD subjects, and all data were collected from the DecNef Project Brain Data Repository (https://bicr.atr.jp/decnefpro/data; Tables 1 and 2; 564 HC or TD, 106 SCZ, and 125 ASD patients).

3. Experiments and results

3.1. Comparison of classification performance with all markers

Focusing on the four steps of the overall marker construction process (Fig. 1), we searched for the superior pipeline for the construction of MDD classification biomarkers with high performance on both the discovery and validation datasets (Fig. S1). We utilized 360 different pipelines, constructed 360 different brain network markers for MDD, and distinguished between HCs and MDDs using Dataset I as the discovery dataset. Subsequently, we applied all markers to the validation dataset (Dataset II). We obtained the AUC, MCC, sensitivity and specificity for each dataset. In addition, we calculated the composite scores and instability index to evaluate the stability of the discrimination results between the discovery and validation datasets.

Based on the composite scores, we first studied the importance of the selection of each development process in the marker pipelines: parcellations, FC, harmonization, and machine learning (Fig. 1). In the parcellation process (Fig. 2A), Glasser + 19 ROIs (379) and dictionary learning (80) showed higher composite scores for both Dataset I (discovery) and Dataset II (validation). In the FC process (Fig. 2B), Pearson's full correlation and tangent-space covariance outperformed the others. In the harmonization process (Fig. 2C), some markers using pipelines with the traveling-subject harmonization approach scored very high, although those using traveling-subject pipelines generally scored slightly lower than markers using pipelines that did not apply any inter-site correction. However, the classification performance of the markers in the discovery dataset using a pipeline containing the ComBat option was significantly lower than that of the other markers. This result may be owing to the sample imbalance between HCs and MDDs in the discovery dataset. Finally, the non-sparse methods, namely RIDGE and SVM, showed higher composite scores in the machine-learning process (Fig. 2D), and some of the markers constructed using the pipeline, including the random forest method, exhibited very high classification

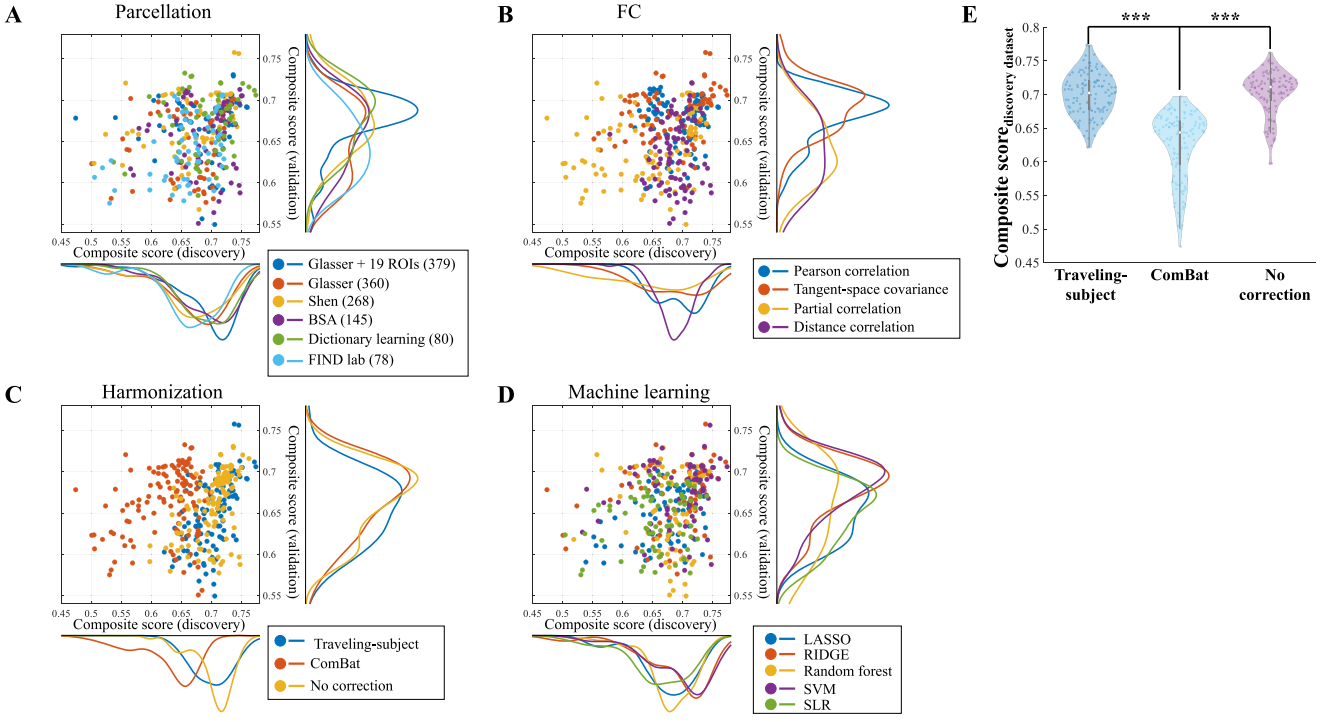


Fig. 2. Classification performance of markers constructed with 360 pipelines.
Distribution of composite scores in discovery and validation datasets for each process: (A) parcellation, (B) FC estimation, (C) site-difference harmonization, and (D) machine learning. (E) Distribution of composite scores of all classification biomarkers for each harmonization method on the discovery dataset. The composite scores of the biomarkers constructed using ComBat show significant lower values than those of the remaining biomarkers (Tukey–Kramer test, $^{***}p < 0.001$).

Table 3
Multi-way ANOVA for mean composite scores except for markers using ComBat. After eliminating markers constructed using ComBat options, four-way ANOVA was conducted on the average composite scores of all biomarkers, focusing on the four development processes.

Source	Sum sq.	df	Mean sq.	F	Sig.
Parcellation	0.03381	5	0.00676	37.91	$1.01*10^{-25}$
FC	0.05445	3	0.01815	101.74	$3.05*10^{-37}$
Harmonization	0.00211	1	0.00211	11.82	0.00074
Machine learning	0.04804	4	0.01201	67.32	$1.31*10^{-33}$
Parcellation× FC	0.02024	15	0.00135	7.56	$1.24*10^{-12}$
Parcellation× Harmonization	0.00167	5	0.00033	1.87	0.10246
Parcellation× Machine learning	0.01171	20	0.00059	3.28	$1.38*10^{-5}$
FC× Harmonization	0.00418	3	0.00139	7.81	$6.61*10^{-5}$
FC× Machine learning	0.0147	12	0.00122	6.87	$5.59*10^{-10}$
Harmonization× Machine learning	0.01013	4	0.00253	14.2	$5.70*10^{-10}$
Error	0.02926	164	0.00018		
Total	0.23376	236			

performance for Dataset II.

We investigated 240 pipelines that did not use ComBat, which showed significantly lower performance (Fig. 2E), to identify the effective options for the parcellation, FC computation, harmonization, and machine learning in detail. We used multi-way analysis of variance (ANOVA) to evaluate the effects of multiple factors on the average composite scores using the ANOVAN function in MATLAB (Table 3, MathWorks, R2018b). The average composite score was the mean value of the composite scores of both datasets and reflected the generalization capability of the marker. The average composite scores were standardized for four-way ANOVA. Except for the combination of parcellation and harmonization, all four main terms and five of the six interaction terms significantly affected the average composite scores. We determined the optimal method for each development process by comparing the composite scores of the markers of all pipelines, which could be calculated when the option in a certain process was fixed to one choice. Surface-based parcellation (Glasser + 19 ROIs (379)) and data-driven parcellation (dictionary learning (80)) led to the maximal performance (Fig. 3A). Tangent-space covariance and Pearson's full correlation as FC tended to outperform the partial and distance correlations (Fig. 3B). In terms of the harmonization approach, we found no significant difference between the no-correction option and traveling-subject harmonization (Fig. 3C). Regarding the machine-learning method, the results showed that non-sparse linear classifiers, including SVM and RIDGE, outperformed the other approaches (Fig. 3D). In summary, Glasser + 19 ROIs (379) or dictionary learning (80) in the parcellation process, Pearson's full correlation or tangent-space covariance in the FC calculation, traveling-subject or no correction in the harmonization, and non-sparse machine-learning methods were recommended for the development of classification biomarkers.

3.2. Evaluation of pipeline generalizability using dataset role swapping

We identified markers that showed high classification performance, even for the validation dataset, and the pipelines for constructing them. In view of the data dependency of machine learning, we constructed and verified markers using Dataset II as the discovery dataset to verify the generalizability of the results regarding the marker performance rankings and important factors (Fig. S1 and S2A). This attempt was made possible by swapping the roles of Dataset I, which was acquired using a unified protocol, and Dataset II, which contains data obtained with multiple imaging protocols (Table S1). We constructed MDD classification biomarkers using all pipelines, except those that included traveling-subject harmonization, because Dataset II had no traveling-subject dataset. We obtained the AUC, MCC, sensitivity, and specificity from both datasets and calculated the average composite scores and instabilities. For each subject, the classification probabilities were calculated using two biomarkers that were constructed using the same pipeline before and after dataset role swapping to evaluate the stability of the discrimination results. A comparison of the average composite scores before and after dataset role swapping showed that the effectiveness of the surface-based Glasser parcellation was remarkable, although a similar tendency was observed when using the datasets in the original role (Fig. 4A). For FC, Pearson's full correlation method commonly showed relatively high performance before and after the dataset role swapping. (Fig. 4B). In the harmonization process, the effectiveness of the no-correction option remained the same compared to ComBat, although the differences became substantially smaller (Fig. 4C). No change was observed in the effectiveness of the non-sparse method in the machine-learning process (Fig. 4D). Each pipeline was comprehensively ranked based on its average composite score, instability value of the classification biomarkers constructed before and after dataset role swapping, and correlation coefficient of the classification probability (Table 4). All five indicators used for the rankings were standardized for all pipelines, and the signs of the instability values were inverted.

3.3. Evaluation of biomarker similarity

We investigated the similarities among the classification biomarkers constructed from the top 10 identified pipelines. The mean ± SEM of the composite scores in the discovery dataset for the top 10 markers was 0.722 ± 0.010 (Fig. 5A). The corresponding AUC, accuracy, sensitivity, specificity, and MCC scores were 0.774 ± 0.012, 0.680 ± 0.010, 0.756 ± 0.011, 0.661 ± 0.011, and 0.339 ± 0.016, respectively (Table S3). However, the mean ± SEM of the composite scores in the independent validation dataset was 0.711 ± 0.004. The corresponding AUC, accuracy, sensitivity, specificity, and MCC scores were 0.743 ± 0.005, 0.676 ± 0.005, 0.714 ± 0.018, 0.649 ± 0.015, and 0.358 ± 0.009, respectively. The classification results of any pair of top 10 classification biomarkers showed a very high concordance rate (Fig. 5B; Sorensen–Dice coefficient index: 0.789 ± 0.008, mean ± SEM). Each marker had 100 classifiers and 100 types of weights or importance values for FC. The summation of the absolute weights or importance of each biomarker was regarded as the degree of its contribution to the classification, and we confirmed whether the top 5% contributing FCs had network usage similarity among the markers. As each parcellation has its own ROIs for dividing the brain, simply comparing the high-contribution FC patterns among markers with different parcellations is impossible. Therefore, we placed the high-contribution FCs into a common brain map known as a brain network, as proposed by Yeo et al., and compared the similarity of the usage rates of the networks to which each FC belonged (Fig. S3). As the existence probability of all FCs in the brain network differed for each parcellation, we compared the network utilization rates between any two markers after correcting them. Significantly higher correlations were observed between the biomarkers that were constructed using the pipelines, including all parcellation and FC estimation patterns, except for the combination of dictionary learning parcellation and tangent-space covariance (Fig. 5C and D).

3.4. Application of pipelines to other disorders

Finally, we investigated whether high-performance biomarkers of other psychiatric disorders could be constructed with the higher-ranked pipelines identified using the MDD datasets (Fig. S1) because these psychiatric disorders also have no objective classification biomarkers. Using the top 10 pipelines shown in Table 4, we verified whether classification biomarkers for ASD and SCZ could be constructed using the same procedure as that for the MDD classification markers. In the top 10 pipelines, which did not contain a ComBat option, the classification

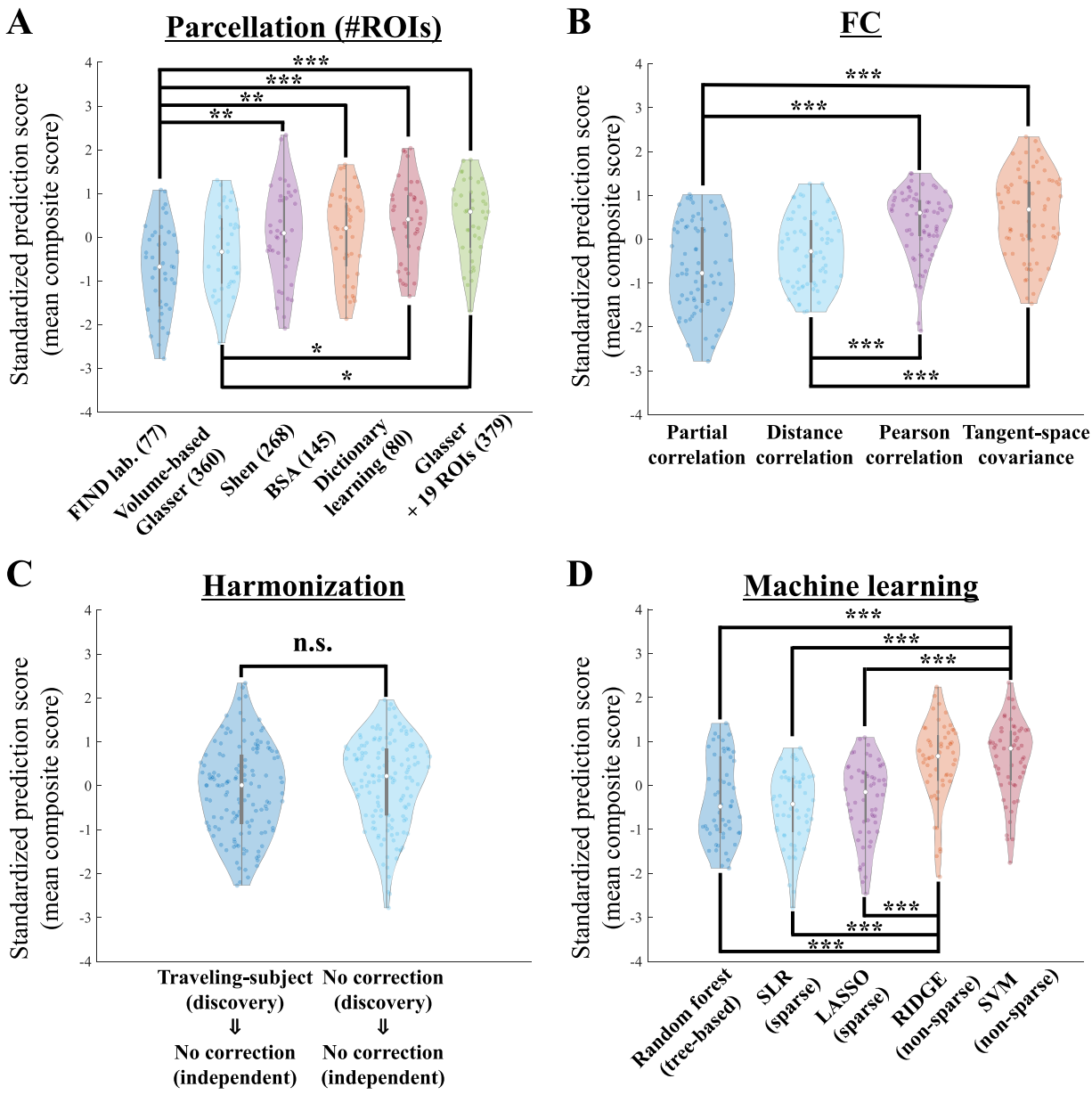


Fig. 3. Effect of choice of each method on prediction accuracy in each development process.
Distribution of classification performances for each method selected in each development process, after excluding biomarkers constructed using the ComBat method. (A) In the parcellation process, Glasser + 19 ROIs (379) and dictionary learning (80) performed better. (B) In the FC process, Pearson's full correlation and tangent-space covariance performed better. (C) There was no significant difference in the harmonization process. (D) In the machine-learning process, non-sparse methods outperformed sparse methods. In this analysis, standardized prediction scores were used to standardize the average composite scores of both datasets (Tukey–Kramer test, $***p < 0.001$, $**p < 0.01$, $*p < 0.05$).

biomarkers for ASD and SCZ were successfully constructed with classification performances that were equal to or higher than those of the MDD classification markers (Fig. 6).

4. Discussion and conclusions

In this study, we conducted a comprehensive evaluation of the analysis pipelines for resting-state FC biomarkers using a large-scale, multi-site fMRI dataset for MDD and HC. We explored option combinations in four sub-processes of the analysis pipeline: six types of brain parcellation, four types of FC computation, three types of site-difference harmonization, and five types of machine-learning methods. In total, 360 biomarkers were constructed using the SRPBS dataset acquired with a unified protocol (713 participants from four imaging sites) as the discovery dataset. Their classification performance was evaluated using datasets from other independent projects that were acquired with heterogeneous protocols (449 participants from four imaging sites) as the validation dataset. We evaluated the effect of each option on each of the four sub-processes based on the cross-validated classification performance on the discovery dataset and classification performance on the validation dataset. We repeated the same procedure after swapping the roles of the two datasets to determine the superior options shared by the resting-state FC biomarkers constructed using the dataset acquired with the unified protocol and using the dataset acquired with the heterogeneous protocol. We found that the pipelines tended to result in high classification performance, including Glasser's or dictionary learning parcellation, Pearson's full correlation or tangent covariance, no harmonization, and non-sparse classifiers, such as RIDGE and SVM.

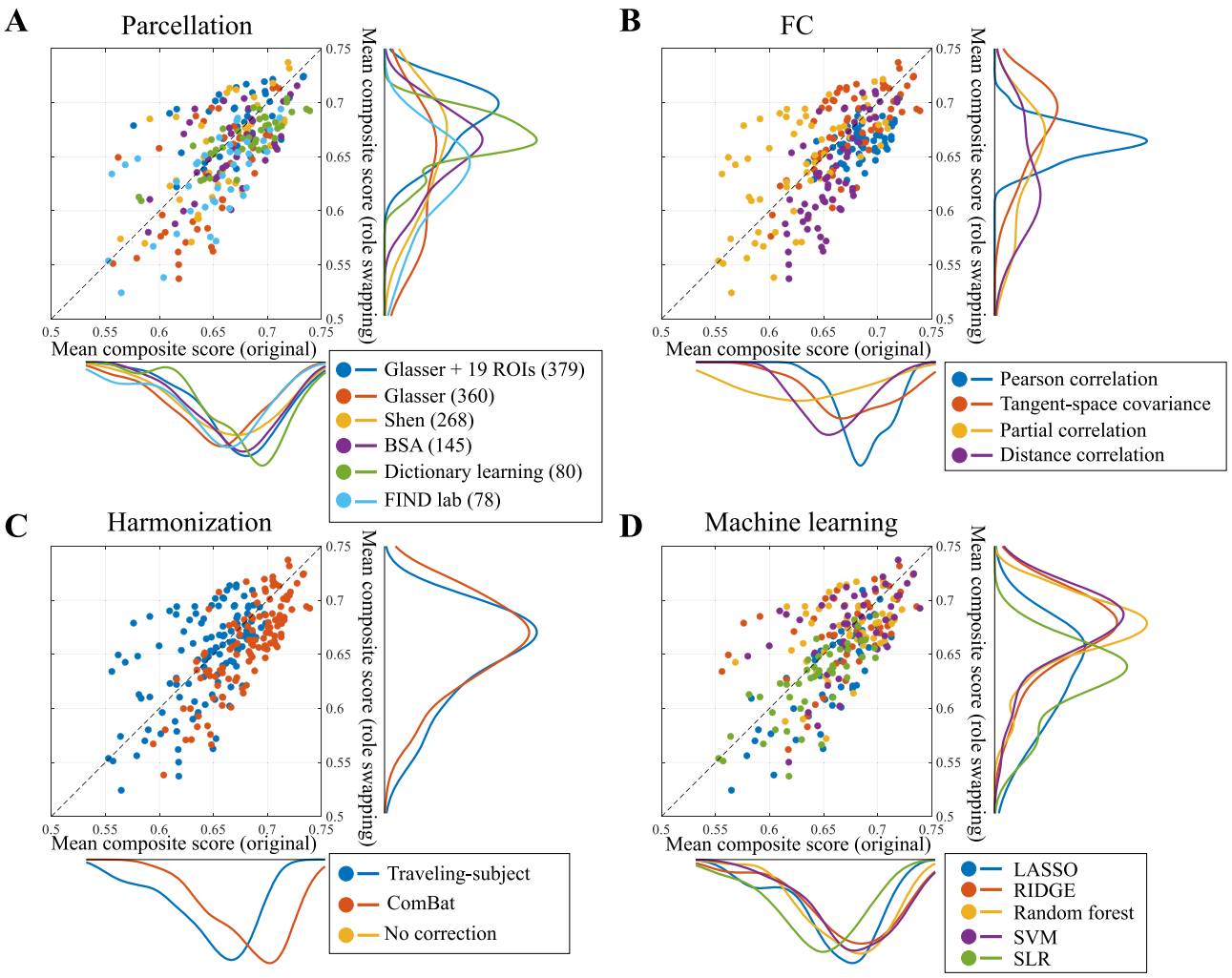


Fig. 4. Dataset role swapping to exclude discovery dataset-dependent pipelines.
Distribution of average composite scores for each process before and after dataset role swapping: (A) parcellation, (B) FC, (C) harmonization, and (D) machine learning.

Subsequently, we investigated the classification results and weight similarities between the top 10 biomarkers and observed commonalities, except in two biomarkers, using both data-driven parcellation and FC computation. Finally, we applied the top 10 pipelines to datasets of other psychiatric disorders (ASD and SCZ), and eight of the 10 biomarkers showed sufficient classification performance. Our results support the construction of standardized pipelines for multi-site and multi-disorder biomarkers. In terms of brain parcellation, Glasser parcellation was the most effective, followed by dictionary learning. Glasser parcellation is surface based. Other studies have reported that surface-based parcellation, which is superior for detecting brain activity with lower signal contamination than volume-based parcellation by 2D smoothing (Brodoehl et al., 2020), increases the statistical power, even in cases of lower spatial and temporal resolution (Anticevic et al., 2008; Coalson et al., 2018). This lower signal contamination may have improved the classification performance of the validation dataset, which consisted of fMRI data acquired using different protocols from those of the discovery dataset at various sites. In this study, 19 subcortical ROIs were added to the original surface-based Glasser parcellation, which may have contributed to further improvements in the classification performance (Fig. S4). The effectiveness of dictionary learning has been reported in a previous benchmarking study (Dadi et al., 2019). Although dictionary learning worked well with the dataset acquired using the unified protocol, its performance degraded significantly when learning was performed with the dataset acquired using heterogeneous protocols. Interestingly, the number of ROIs for the Glasser parcellation was 379, which was the highest in our exploration, whereas that for the dictionary learning was only 80, which was the lowest.

We confirmed that Pearson's full correlation and tangent-space covariance were better than the other options for FC. The latter exhibited slightly higher performance than the former, although the difference was not statistically significant. The effectiveness of the tangent-space covariance was also consistent with previous reports (Dadi et al., 2019; Yang et al., 2022). Because Pearson's full correlation uses the average fMRI signals within each ROI, this approach overlooks the spatial patterns of the voxel- or vertex-wise signals within individual ROIs. We tested the recently proposed distance correlation (Yoo et al., 2019) to exploit the information of voxel patterns within each ROI. However, the average performance was not good, indicating high sensitivity of the voxel patterns to protocol and scanner differences.

In terms of the harmonization method, we found no performance differences between the traveling-subject harmonization and no-correction methods, although the performance of ComBat degraded when the SRPBS dataset was used as the discovery dataset. When we trained the biomarkers with heterogeneous datasets, using only ComBat as the harmonization method, we did not observe any difference between no correction and ComBat (Fig. S2B). Additionally, an advanced harmonization method derived from ComBat, known as CovBat (Chen

Table 4
Top 10 superior pipelines for development of MDD classification biomarker.
The 240 MDD classification biomarkers were ranked based on the sum of five values: two standardized composite scores, two standardized and sign-swapped instability scores, and the standardized consistency of classifier outputs.

Pipeline ranking	Parcellation	FC	Harmonization	Machine learning	Dataset I → Dataset II		Dataset II → Datatset I		Standardized consistency of classifier outputs	Total score
					Standardized composite score	Standardized instability	Standardized composite score	Standardized instability		
#1	Glasser + 19 ROIs (379)	Tangent	No correction	SVM	1.824	1.644	1.317	0.320	0.765	5.870
#2	Glasser + 19 ROIs (379)	Tangent	No correction	RIDGE	1.808	1.624	1.407	0.258	0.662	5.758
#3	Glasser + 19 ROIs (379)	Pearson	ComBat	Random forest	0.501	0.918	1.385	1.077	1.730	5.610
#4	Shen (268)	Pearson	No correction	Random forest	1.360	0.649	0.804	1.153	1.578	5.544
#5	Dictionary learning (80)	Tangent	No correction	SVM	1.971	0.901	0.947	1.131	0.362	5.312
#6	FIND lab (78)	Tangent	No correction	SVM	1.281	0.939	0.907	1.055	1.091	5.273
#7	Dictionary learning (80)	Tangent	No correction	RIDGE	1.892	0.939	1.013	1.244	0.179	5.268
#8	Glasser + 19 ROIs (379)	Pearson	No correction	Random forest	1.077	0.786	0.784	1.162	1.269	5.078
#9	Glasser + 19 ROIs (379)	Pearson	No correction	SVM	1.482	1.177	1.084	0.504	0.723	4.970
#10	Dictionary learning (80)	Pearson	ComBat	Random forest	0.687	0.279	1.192	1.229	1.540	4.927

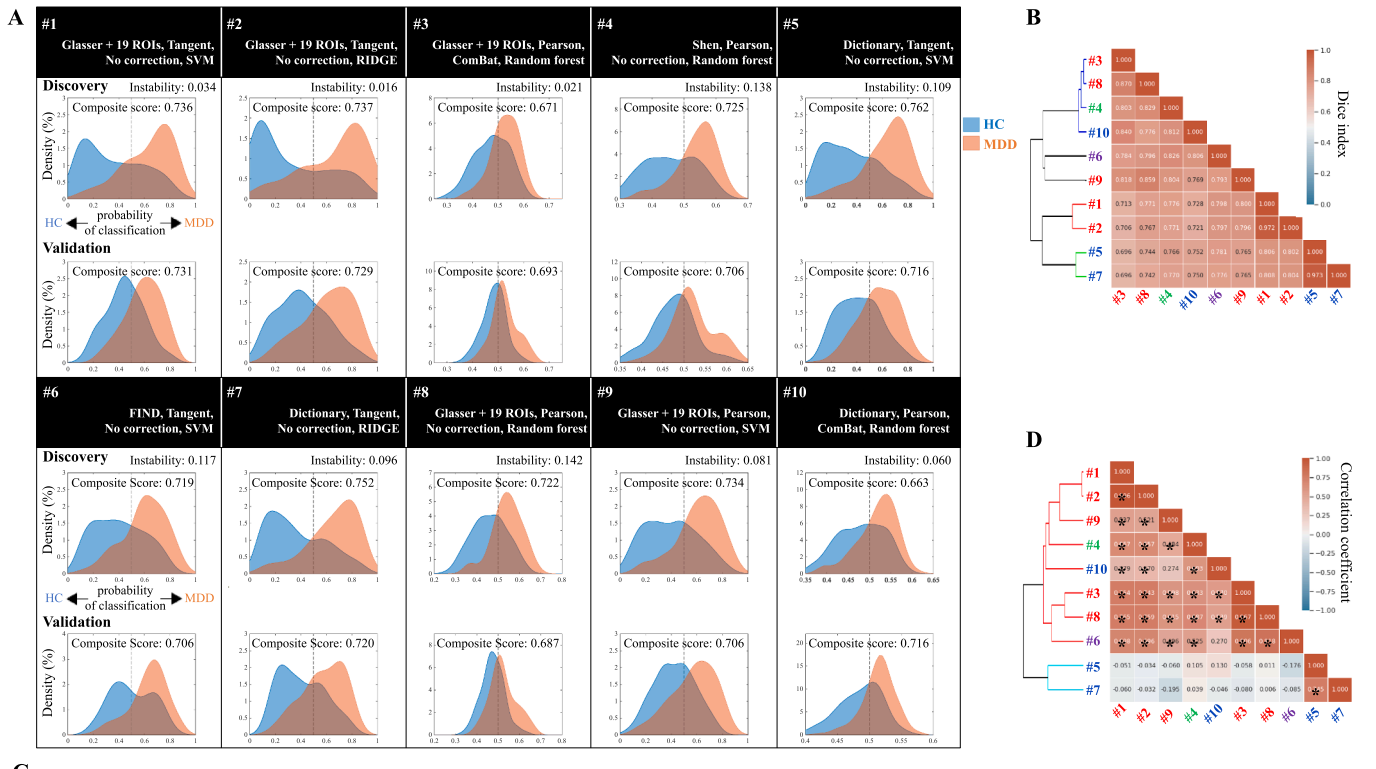
et al., 2022), exists. CovBat is a technique that more comprehensively accounts for the interaction between covariates and batch effects. Although not included in the comparison in this study, we performed marker construction for the top 10 pipelines using Dataset I as the discovery dataset and Dataset II as the validation dataset, replacing the harmonization method with CovBat. This analysis demonstrated that the application of CovBat does not necessarily improve the classification performance of the markers (discovery dataset AUC in 10-fold CV: 0.774 ± 0.012 before CovBat, 0.679 ± 0.013 after CovBat: 0.774 ± 0.012, mean ± SEM, Fig. S6). In summary, we did not identify a clear effect of harmonization on the classification performance. One possible reason for this is that the pattern of disease factors is sufficiently different from the pattern owing to site differences, and machine learning automatically weighs the disease factors (Abraham et al., 2017; Yamashita et al., 2019). However, harmonization remains important when interpreting constructed biomarkers closely. For example, a previous study using sparse classifiers to construct MDD biomarkers showed that the number of relevant FCs increased after traveling-subject harmonization compared to the no-correction method (Yamashita et al., 2020).

Our results demonstrated that non-sparse machine learning methods were preferred, which is consistent with a previous report (Dadi et al., 2019). This implies that the FC required for MDD diagnoses is widespread throughout the brain. Although the non-sparse methods worked better provided that the classification performance was considered, sparse classifiers, which can select a small number of important FCs (Yamashita et al., 2020), are advantageous if applications target specific FCs, such as FC neurofeedback (Megumi et al., 2015; Yamashita et al., 2017) and the identification of drug discovery target brain areas (Ichikawa et al., 2020). Although deep-learning methods are known to be effective in constructing classification biomarkers, higher performance is achieved with a larger sample size; thus, the performance will vary depending on the sample size (Quaak et al., 2021). Deep learning was not included in our pipeline comparison to avoid the effects of the sample size.

A similarity analysis of the classification results and FC weights of the top 10 pipelines showed that the classification patterns were highly similar among the top 10 biomarkers, and the weight patterns were highly similar among eight biomarkers. Two markers that showed low weight similarity to the remaining markers used a combination of dictionary learning parcellation and tangent-space covariance, both of which are data-driven methods. We can discriminate the MDD characteristics that differ from those of biomarkers using other pipelines with multiple data-driven methods. An interesting future study could take advantage of the different characteristics of multiple biomarkers using an ensemble learning framework to provide more robust discrimination results.

Regarding the networks identified as important among eight of the top 10 markers, connectivity within the somatomotor network (SMN) was consistently the most significant for classification (Fig. 5). Abnormalities in SMN connectivity in MDD patients were also reported by Javaheripour et al. (2021). Considering that their analysis was based on large-scale data collected from multiple institutions, it can be regarded as a reliable finding that reflects common brain functional abnormalities in MDD patients. Furthermore, the lack of commonality in the utilization of the frontoparietal network (FPN) or default-mode network (DMN) across our eight markers was consistent with their view that it is difficult to detect these abnormalities in MDD patients using rs-fMRI data.

Although this study conducted a search for superior classification biomarkers for construction pipelines by developing MDD classification biomarkers, other psychiatric disorders also require the establishment of objective biomarkers. We demonstrated that eight of the top 10 pipelines might also be effective in developing classification biomarkers for ASD and SCZ in addition to MDD. As classification biomarkers have been constructed for these three psychiatric disorders, the classification probability of each subject for each disorder can be obtained from the three classification biomarkers in a single fMRI acquisition. Combining multiple classification probabilities based on biophysiological backgrounds is expected to fuel patient classification regardless of conventional categorical diagnoses. We will also attempt to study patient-clustering methods using a multiple-disorder dataset, which is consistent with the Research Domain Criteria concept that seeks precision medicine for psychiatry (Insel, 2014). Both ASD classification biomarkers were constructed using pipelines including a combination of Pearson's full correlation, ComBat harmonization, and the random forest method in a discovery dataset. As the two discovery dataset sites consisted of only TD subjects and the ASD patients were concentrated in one site, excessive correction was caused by ComBat, and the characteristics of the ASD patients in the FC appeared to be lost. In fact, if



Parcellation	Glasser + 19 ROIs (379)	Glasser + 19 ROIs (379)	Glasser + 19 ROIs (379)	Shen (268)	Dictionary learning (80)	Glasser + 19 ROIs (379)	Glasser + 19 ROIs (379)	FIND lab (78)	Dictionary learning (80)	Dictionary learning (80)
FC	Tangent	Tangent	Pearson	Pearson	Pearson	Pearson	Pearson	Tangent	Tangent	Tangent
Harmonization	No correction	No correction	ComBat	No correction	ComBat	ComBat	No correction	No correction	No correction	No correction
Machine learning	SVM	RIDGE	SVM	Random forest	Random forest	Random forest	Random forest	SVM	SVM	RIDGE

Fig. 5. Classification and weight similarity among top 10 high-performance biomarkers.
(A) Classification performance of top 10 MDD classification biomarkers. The upper row shows the correlation coefficients between the top markers in the classification probability. The probability distributions for the classification results of MDD in Dataset I (discovery, 10-fold CV test, upper) and Dataset II (validation test, bottom) are shown. The MDD and HC distributions are depicted in red and blue, respectively. (B) Correlation coefficients between the top markers in the classification probability. (C, D) Similarity of corrected network usage amounts of important FCs amount among the top 10 markers. A two-thirds combination showed a significantly higher correlation with the corrected network usage amount (permutation test, $*p < 0.05$).

classification biomarkers are created in a pipeline containing ComBat for only two sites, including ASD, their classification performance will be improved (Fig. S5).

This study has several limitations, which are outlined below. First, although this study examined generalization, it did not address prospective generalization. Our recent study demonstrated real prospective generalization using independent validation cohort data that did not exist at the time of the biomarker development (Okada et al., 2023). In this study, as both datasets were prepared prior to developing the pipeline, prospective generalization could not be validated. Future validation should include the BMB dataset (Koike et al., 2021) upon its release to examine prospective generalization properly. Second, the superiority of the pipeline exhibited only minor performance differences using a relatively small sample size compared with image processing and a large language model. Although these differences were statistically significant, the small sample size weakened the robustness of the findings, suggesting that caution should be exercised before dismissing any method. However, the SRPBS dataset (Tanaka et al., 2021) was the largest multi-center, multi-disorder fMRI database globally at the time of its release (Tanaka et al., 2024). Future work should replicate the validation of this study using the BMB dataset. Third, the validation of the retrospective harmonization method was limited by the asymmetry of the current datasets; traveling-subject data were only available for the discovery dataset and not for the validation cohort in our datasets. When traveling-subject data are used in medical practice, it is generally expected that the data will be collected for the intended devices and protocols. The generality of our assessment of pipelines depends on the availability of traveling-subject data. Fourth, this study focused only on classification-based diagnostic markers. It is important to recognize that biomarkers can also be stratification, state, or prognostic markers, and our conclusions are specific to diagnostic markers using classification algorithms. Fifth, the feature selection algorithm was not tested within

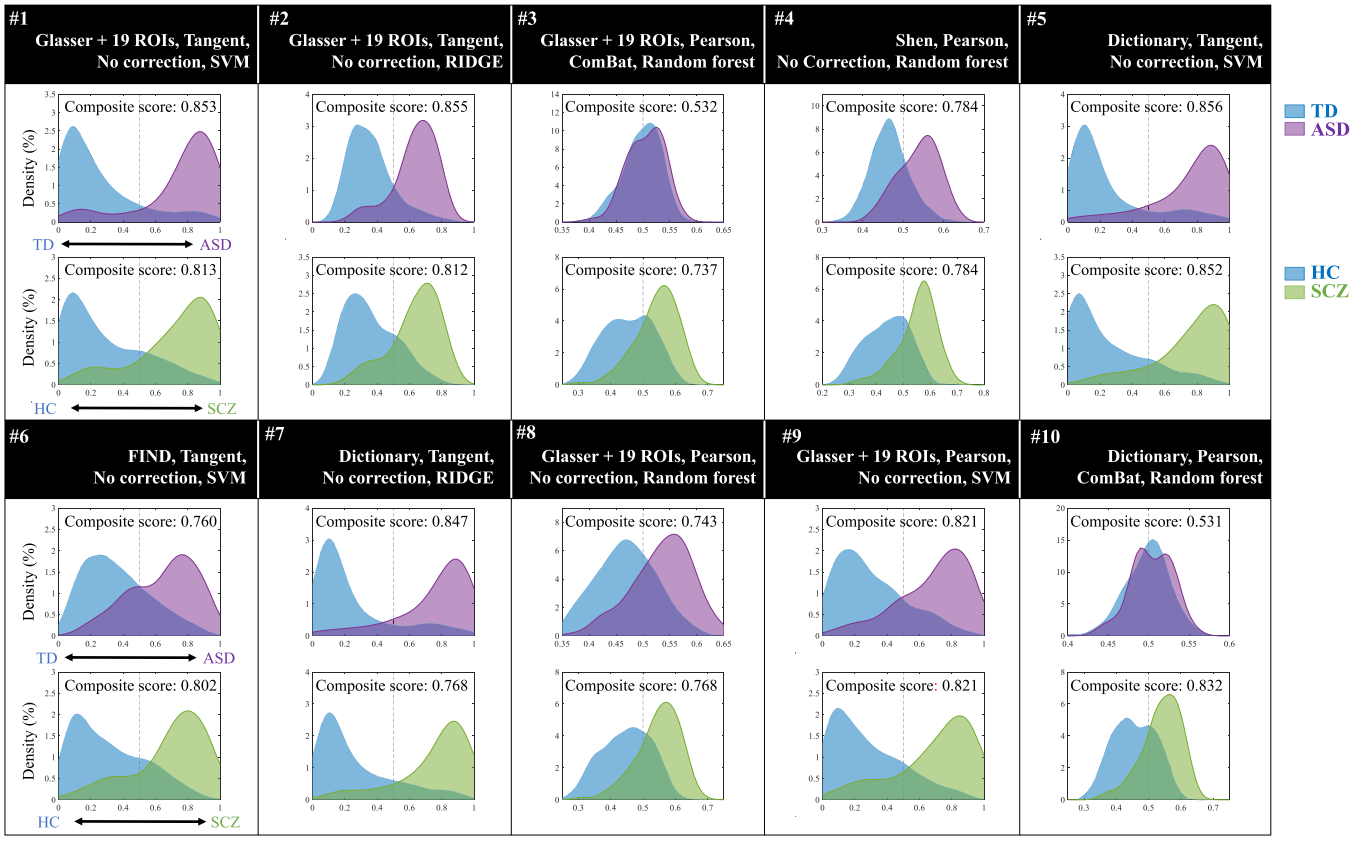


Fig. 6. Development of ASD and SCZ classification biomarkers using top 10 analysis pipelines.
Constructing ASD and SCZ classification biomarkers using the top 10 pipelines showed high classification and generalization performance. Eight of the top 10 biomarkers showed high classification performance in the composite scores. The TD or HC, ASD, and SCZ distributions are depicted in blue, purple, and green, respectively.

the pipeline owing to the significant increase in computational complexity. However, the sparse algorithms examined in this study implicitly select features. Sixth, the comparison was based on the classification accuracy, rather than on the clinical interpretability of the marker. In the future, we should add the criterion that the biomarkers should be easy to interpret. Seventh, the influence of age bias on FC was mitigated using an undersampling method; however, the impact of gender was not considered. Although no gender bias was observed in the classification results using markers built from the top 10 pipelines (data not shown), we believe that a gender correction process should be incorporated into the pipeline in the future.

In summary, we searched for a pipeline with excellent generalization performance and marker construction from a combination of comprehensive methods based on fMRI data acquired from multiple large-scale facilities and obtained multiple candidates. Our identified pipelines are likely to apply to other psychiatric disorders, and we anticipate that a combination of multiple fMRI markers will contribute to a deeper understanding of such disorders and facilitate the advancement of diagnostics and personalized medicine, unconstrained by current disease classifications. We hope that the widespread use of fMRI markers in clinical practice will reduce the treatment period for patients and lead to the development of new treatment methods for patients for whom no effective treatment methods are currently available.

Data and code availability statement

We are a registered member of the DecNef Project Brain Data Repository (https://bicr.atr.jp/decnefpro/data), and the data for this study are available from its website upon reasonable request by qualified researchers. See Tanaka et al. (2021) for more detailed information on the datasets.

CRediT authorship contribution statement

Yuji Takahara: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Yuto Kashiwagi:** Writing – review & editing, Investigation, Conceptualization. **Tomoki Tokuda:** Writing – review & editing, Investigation, Conceptualization. **Junichiro Yoshimoto:** Writing – review & editing, Investigation, Conceptualization. **Yuki Sakai:** Writing – review & editing, Investigation, Conceptualization. **Ayumu Yamashita:** Writing – review & editing, Resources. **Toshinori Yoshioka:** Writing – review & editing. **Hidehiko Takahashi:** Writing – review & editing, Data curation. **Hiroto Mizuta:** Writing – review & editing, Data curation. **Kiyoto Kasai:** Writing – review & editing, Data curation. **Akira Kunimitsu:** Writing – review & editing, Data curation. **Naohiro Okada:** Writing – review & editing, Data curation. **Eri Itai:** Writing – review & editing, Data curation. **Hotaka Shinzato:** Writing – review & editing, Data curation. **Satoshi Yokoyama:** Writing – review & editing, Data curation. **Yoshikazu Masuda:** Writing – review & editing, Data curation. **Yuki Mitsuyama:** Writing – review & editing, Data curation. **Go Okada:** Writing – review & editing, Data curation. **Yasumasa Okamoto:** Writing – review & editing, Data curation. **Takashi Itahashi:** Writing – review & editing, Data curation. **Haruhisa Ohta:** Writing – review & editing, Data curation. **Ryu-ichiro Hashimoto:** Writing – review & editing, Data curation. **Kenichiro Harada:** Writing – review & editing, Data curation. **Hirotaka Yamagata:** Writing – review & editing, Data curation. **Toshio Matsubara:** Writing – review

& editing, Data curation. **Koji Matsuo:** Writing – review & editing, Data curation. **Saori C. Tanaka:** Writing – review & editing, Data curation. **Hiroshi Imamizu:** Writing – review & editing, Data curation. **Koichi Ogawa:** Writing – review & editing, Conceptualization. **Sotaro Momosaki:** Writing – review & editing, Conceptualization. **Mitsuo Kawato:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Okito Yamashita:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Okito Yamashita reports financial support was provided by XNef, Inc. Okito Yamashita reports financial support was provided by Shionogi & Co., Ltd. Okito Yamashita has patent Brain functional connectivity correlation value adjustment method, brain functional connectivity correlation value adjustment system, brain activity classifier harmonization method, brain activity classifier harmonization system, and brain activity biomarker system pending to WO2020075737A1. Ayumu Yamashita has patent Brain functional connectivity correlation value adjustment method, brain functional connectivity correlation value adjustment system, brain activity classifier harmonization method, brain activity classifier harmonization system, and brain activity biomarker system pending to WO2020075737A1. Hiroshi Imamizu has patent Brain functional connectivity correlation value adjustment method, brain functional connectivity correlation value adjustment system, brain activity classifier harmonization method, brain activity classifier harmonization system, and brain activity biomarker system pending to WO2020075737A1. Mitsuo Kawato has patent Brain functional connectivity correlation value adjustment method, brain functional connectivity correlation value adjustment system, brain activity classifier harmonization method, brain activity classifier harmonization system, and brain activity biomarker system pending to WO2020075737A1. Yuji Takahara has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Yuto Kawashiwagi has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Tomoki Tokuda has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Mitsuo Kawato has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Ayumu Yamashita has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Okito Yamashita has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Yuki Sakai has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Junichiro Yoshiomoto has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Go Okada has patent Brain functional connectivity correlation value clustering device, brain functional connectivity correlation value clustering system, brain functional connectivity correlation value clustering method, brain functional connectivity correlation value classifier program, and brain activity marker classification system pending to WO2022014682A1. Co-authors are employees of XNef, Inc.. - Y.S., T.Y., and M.K. Corresponding author and co-authors are employees of Shionogi & Co., Ltd.. - Y.T., Y.K., K.O. and S.M. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was mainly supported by the collaborative funds of XNef, Inc. and Shionogi & Co., Ltd. The data resource was supported by JSPS KAKENHI (Grant Numbers JP20H03605, JP21H05174, and JP21H05171), AMED (Grant Numbers JP22dm0307008, JP22dm0307009, JP19dm0207069, JP18dm0307001, and JP18dm0307004, JP23wm0625001, JP24wm0625502), Moonshot R&D (Grant Number JPMJMS2021), and UTokyo Institute for Diversity and Adaptation of Human Mind (UTIDAHM) and the International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study (UTIAS). This work was partially supported by Innovative Science and Technology Initiative for Security Grant Number JPJ004596, ATLA, Japan. We would like to thank Dr. Wenjun Bai for his helpful comments. We would like to thank Editage (www.editage.jp) for English language editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neunet.2025.107335.

References

Abraham, A., Milham, M. P., di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage, 147*, 736–745. https://doi.org/10.1016/j.neuroimage.2016.10.045

Anticevic, A., Dierker, D. L., Gillespie, S. K., Repovs, G., Csernansky, J. G., van Essen, D. C., & Barch, D. M. (2008). Comparing surface-based and volume-based analyses of functional neuroimaging data in patients with schizophrenia. *NeuroImage, 41*(3), 835–848. https://doi.org/10.1016/J.NEUROIMAGE.2008.02.052

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage, 145*(Pt B), 137–165. https://doi.org/10.1016/J.NEUROIMAGE.2016.02.079

Arnow, B. A., Blasey, C., Williams, L. M., Palmer, D. M., Rekshan, W., Schatzberg, A. F., Etkin, A., Kulkarni, J., Luther, J. F., & Rush, A. J. (2015). Depression subtypes in predicting antidepressant response: A report from the iSPOT-D trial. *American Journal of Psychiatry, 172*(8), 743–750. https://doi.org/10.1176/APPI.AJP.2015.14020181/SUPPL_FILE/AUG2015_DEPRESSION.MP3

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage, 37*(1), 90–101. https://doi.org/10.1016/J.NEUROIMAGE.2007.04.042

Brodoehl, S., Gaser, C., Dahnke, R., Witte, O. W., & Klingner, C. M. (2020). Surface-based analysis increases the specificity of cortical activation patterns and connectivity results. *Scientific Reports, 10*(1), 1–13. https://doi.org/10.1038/S41598-020-62832-Z

Brown, C.J., & Hamarneh, G. (2016). *Machine learning on human connectome data from MRI*. http://arxiv.org/abs/1611.08699.

Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C., & Thomas Yeo, B. T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 106*(5), 2322–2345. https://doi.org/10.1152/JN.00339.2011

Calhoun, V. D., Sui, J., Kiehl, K., Turner, J., Allen, E., & Pearlson, G. (2012). Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder. *Frontiers in Psychiatry, 2*, 1–13. https://doi.org/10.3389/FPSYT.2011.00075

Castellanos, F. X., di Martino, A., Craddock, R. C., Mehta, A. D., & Milham, M. P. (2013). Clinical applications of the functional connectome. *NeuroImage, 80*, 527–540. https://doi.org/10.1016/J.NEUROIMAGE.2013.04.083

Chen, G., Ward, B. D., Xie, C., Li, W., Wu, Z., Jones, J. L., Franczak, M., Antuono, P., & Li, S. J. (2011). Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging. *Radiology, 259*(1), 213–221. https://doi.org/10.1148/RADIOL.10100734

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., Shou, H., & Alzheimer's Disease Neuroimaging Initiative. (2022). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping, 43*(4), 1179–1195.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining, 10*(1), 1–17. https://doi.org/10.1186/s13040-017-0155-3

Choi, E. Y., Thomas Yeo, B. T., & Buckner, R. L. (2012). The organization of the human striatum estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 108*(8), 2242–2263. https://doi.org/10.1152/JN.00270.2012

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage, 154*, 174–187. https://doi.org/10.1016/J.NEUROIMAGE.2017.03.020

Clare Kelly, A. M., Uddin, L. Q., Biswal, B. B., Castellanos, F. X., & Milham, M. P (2008). Competition between functional brain networks mediates behavioral variability. *NeuroImage, 39*(1), 527–537. https://doi.org/10.1016/J.NEUROIMAGE.2007.08.008

Coalson, T. S., van Essen, D. C., & Glasser, M. F. (2018). The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proceedings of the National Academy of Sciences of the United States of America, 115*(27), E6356–E6365. https://doi.org/10.1073/PNAS.1801582115/-/DCSUPPLEMENTAL

Craddock, C., Jbabdi, S., Yan, C.-G., & Vogelstein, J. T. (2013). Article in nature methods. *Nature Methods, 10*(6), 524–539. https://doi.org/10.1038/nmeth.2482

Craddock, R. C., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine : Official Journal of the Society ofMagnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine, 62*(6), 1619. https://doi.org/10.1002/MRM.22159

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., & Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage, 192*, 115–134. https://doi.org/10.1016/J.NEUROIMAGE.2019.02.062

Dansereau, C., Benhajali, Y., Risterucci, C., Pich, E. M., Orban, P., Arnold, D., & Bellec, P. (2017). Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. *NeuroImage, 149*, 220–232. https://doi.org/10.1016/j.neuroimage.2017.01.072

Dickie, E. W., Anticevic, A., Smith, D. E., Coalson, T. S., Manogaran, M., Calarco, N., Viviano, J. D., Glasser, M. F., van Essen, D. C., Voineskos, A. N., Luke, S., & Usa, C. M. (2019). ciftify: A framework for surface-based analysis of legacy MR acquisitions HHS Public Access. *Neuroimage, 197*, 818–826. https://doi.org/10.5281/zenodo.2651201

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods, 16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage, 167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage, 161*, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Garrity, A. G., Pearlson, G. D., McKiernan, K., Lloyd, D., Kiehl, K. A., & Calhoun, V. D. (2007). Aberrant "default mode" functional connectivity in schizophrenia. *The American Journal of Psychiatry, 164*(3), 450–457. https://doi.org/10.1176/AJP.2007.164.3.450

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex Europe PMC funders group. *Nature, 536*(7615), 171–178. https://doi.org/10.1038/nature18933

Graña, M., & Silva, M. (2021). Impact of machine learning pipeline choices in autism prediction from functional connectivity data. *International Journal of Neural Systems, 31*(4), 1–20. https://doi.org/10.1142/S012906572150009X

Greicius, M. D., Srivastava, G., Reiss, A. L., & Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America, 101*(13), 4637. https://doi.org/10.1073/PNAS.0308627101

Ichikawa, N., Lisi, G., Yahata, N., Okada, G., Takamura, M., Hashimoto, R. I., Yamada, T., Yamada, M., Suhara, T., Moriguchi, S., Mimura, M., Yoshihara, Y., Takahashi, H., Kasai, K., Kato, N., Yamawaki, S., Seymour, B., Kawato, M., Morimoto, J., & Okamoto, Y. (2020). Primary functional brain connections associated with melancholic major depressive disorder and modulation by antidepressants. *Scientific Reports, 10*(1), 1–12. https://doi.org/10.1038/s41598-020-60527-z

Insel, T. R. (2014). The nimh research domain criteria (rdoc) project: Precision medicine for psychiatry. *American Journal of Psychiatry, 171*(4), 395–397. https://doi.org/10.1176/APPI.AJP.2014.14020138

Jafri, M. J., Pearlson, G. D., Stevens, M., & Calhoun, V. D. (2008). A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *NeuroImage, 39*(4), 1666. https://doi.org/10.1016/J.NEUROIMAGE.2007.11.001

Javaheripour, N., Li, M., Chand, T., Krug, A., Kircher, T., Dannlowski, U., Nenadić, I., Hamilton, JP., Sacchet, MD., Gotlib, IH., Walter, H., Frodl, T., Grimm, S., Harrison, BJ., Wolf, CR., Olbrich, S., van Wingen, G., Pezawas, L., Parker, G., Hyett, MP., Sämann, PG., Hahn, T., Steinsträter, O., Jansen, A., Yuksel, D., Kämpe, R., Davey, CG., Meyer, B., Bartova, L., Croy, I., Walter, M., & Wagner, G. (2021). Altered resting-state functional connectome in major depressive disorder: a mega-analysis from the PsyMRI consortium. *Translational Psychiatry, 11*(1), 511. https://doi.org/10.1038/S41398-021-01619-W

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it. *Molecular Psychiatry, 17*(12), 1174–1179. https://doi.org/10.1038/MP.2012.105

Koike, S., Tanaka, S. C., Okada, T., Aso, T., Yamashita, A., Yamashita, O., Asano, M., Maikusa, N., Morita, K., Okada, N., Fukunaga, M., Uematsu, A., Togo, H., Miyazaki, A., Murata, K., Urushibata, Y., Autio, J., Ose, T., Yoshimoto, J., … Hayashi, T. (2021). Brain/MINDS beyond human brain MRI project: A protocol for multi-level harmonization across brain disorders throughout the lifespan. *NeuroImage: Clinical, 30*, Article 102600. https://doi.org/10.1016/J.NICL.2021.102600

Martino, A. D., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., & Alaerts, K. (2013). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *JS Verhoeven, 10*, 659–667. https://doi.org/10.1038/mp.2013.78

Megumi, F., Yamashita, A., Kawato, M., & Imamizu, H. (2015). Functional MRI neurofeedback training on connectivity between two regions induces long-lasting changes in intrinsic functional network. *Frontiers in Human Neuroscience, 9*, 160. https://doi.org/10.3389/fnhum.2015.00160. Www.Frontiersin.Org

Mellema, C. J., Nguyen, K. P., Treacher, A., & Montillo, A. (2022). Reproducible neuroimaging features for diagnosis of autism spectrum disorder with machine learning. *Scientific Reports |, 12*, 3057. https://doi.org/10.1038/s41598-022-06459-2

Mensch, A., Mairal, J., Thirion, B., & Varoquaux, G. (2016). Dictionary learning for massive matrix factorization. In *, 4. 33rd international conference on machine learning, ICML 2016* (pp. 2601–2610).

Milham, P. M., Damien, F., Maarten, M., & Stewart, H. M. (2012). The ADHD-200 Consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience, 6*, 1–5. https://doi.org/10.3389/fnsys.2012.00062

Okada, G., Yoshioka, T., Yamashita, A., Itai, E., Yokoyama, S., Kamishikiryo, T., Shinzato, H., Masuda, Y., Mitsuyama, Y., Kan, S., Kurata, A., Takamura, M., Yoshino, A., Mantani, A., Yamamoto, O., Yokota, N., Tamura, T., Jitsuiki, H., Kawato, M., … Okamoto, Y. (2023). Verification of the brain network marker of major depressive disorder: Test-retest reliability and anterograde generalization performance for newly acquired data. *Journal of Affective Disorders, 326*, 262–266. https://doi.org/10.1016/j.jad.2023.01.087

Perrot, M., Rivière, D., & Mangin, J. F. (2011). Cortical sulci recognition and spatial normalization. *Medical Image Analysis, 15*(4), 529–550. https://doi.org/10.1016/J.MEDIA.2011.02.008

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage, 211*, 1–24. https://doi.org/10.1016/J.NEUROIMAGE.2020.116604

Plitt, M., Barnes, K. A., & Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage: Clinical, 7*, 359–366. https://doi.org/10.1016/J.NICL.2014.12.013

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage, 84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Quaak, M., Van De Mortel, L., Thomas, M., & Van Wingen, G. (2021). Deep learning applications for the classification of psychiatric disorders using neuroimaging data: Systematic review and meta-analysis. *NeuroImage Clinical, 30*, 1–21. https://doi.org/10.1016/j.nicl.2021.102584

Shen, X., Tokoglu, F., Papademetris, X., & Constable, R. T. (2013). Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage, 82*, 403–415. https://doi.org/10.1016/j.neuroimage.2013.05.081

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., & Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral Cortex, 22*(1), 158–165. https://doi.org/10.1093/cercor/bhr099

Tanaka, S. C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., … Imamizu, H. (2021). A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data, 8*(1), 1–15. https://doi.org/10.1038/s41597-021-01004-8

Tanaka, S. C, Kasai, K., Okamoto, Y., Koike, S., Hayashi, T., Yamashita, A., … Hanakawa, T. (2024). The status of MRI databases across the world focused on psychiatric and neurological disorders. *Psychiatry and Clinical Neuroscience, 78*(10), 563–579. https://doi.org/10.1111/PCN.13717

Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fisch, B., Liu, H., & Buckner, R. L (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 106*(3), 1125–1165. https://doi.org/10.1152/JN.00338.2011

Tokuda, T., Yoshimoto, J., Shimizu, Y., Okada, G., Takamura, M., Okamoto, Y., … Doya, K. (2018). Identification of depression subtypes and relevant brain regions using a data-driven approach. *Scientific Reports, 8*(1), 1–13. https://doi.org/10.1038/S41598-018-32521-Z

van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn human connectome project: An overview. *NeuroImage, 80*, 62–79. https://doi.org/10.1016/J.NEUROIMAGE.2013.05.041

van Essen, D. C., & Ugurbil, K. (2012). The future of the Human connectome. *Neuroimage, 62*(2), 1299. https://doi.org/10.1016/J.NEUROIMAGE.2012.01.032

Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., & Thirion, B. (2010). Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In *, 13. MICCAI: international conference on medical image computing and computer-assisted intervention* (pp. 200–208). https://doi.org/10.1007/978-3-642-15705-9_25

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2011). Class imbalance, redux. In *Proceedings - IEEE international conference on data mining, ICDM* (pp. 754–763). https://doi.org/10.1109/ICDM.2011.33

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews, 57*, 328–349. https://doi.org/10.1016/J.NEUBIOREV.2015.08.001

Yahata, N., Morimoto, J., Hashimoto, R., Lisi, G., Shibata, K., Kawakubo, Y., … Kawato, M. (2016). ARTICLE A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature Communications, 7*, 1–12. https://doi.org/10.1038/ncomms11254

Yamashita, A., Hayasaka, S., Kawato, M., & Imamizu, H. (2017). Connectivity neurofeedback training can differentially change functional connectivity and cognitive performance. *Cerebral Cortex, 27*, 4960–4970. https://doi.org/10.1093/cercor/bhx177

Yamashita, A., Sakai, Y., Yamada, T., Yahata, N., Kunimatsu, A., Okada, N., Itahashi, T., Hashimoto, R., Mizuta, H., Ichikawa, N., Takamura, M., Okada, G., Yamagata, H., Harada, K., Matsuo, K., Tanaka, S. C., Kawato, M., Kasai, K., Kato, N., … Imamizu, H. (2020). Generalizable brain network markers of major depressive disorder across multiple imaging sites. *PLoS Biology, 18*(12), 1–26. https://doi.org/10.1371/journal.pbio.3000966

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., … Imamizu, H. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biology, 17*(4). https://doi.org/10.1371/journal.pbio.3000042

Yamashita, O., Sato, M. A., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage, 42*(4), 1414–1429. https://doi.org/10.1016/j.neuroimage.2008.05.050

Yang, X., Zhang, N., & Schrader, P. (2022). A study of brain networks for autism spectrum disorder classification using resting-state functional connectivity. *Machine Learning with Applications, 8*, Article 100290. https://doi.org/10.1016/J.MLWA.2022.100290

Yoo, K., Rosenberg, M. D., Hsu, W.-T., Zhang, S., Li, C.-S. R., Scheinost, D., Todd Constable, R., & Chun, M. M. (2018). Connectome-based predictive modeling of attention: Comparing different functional connectivity features and prediction methods across datasets HHS public access. *Neuroimage, 167*, 11–22. https://doi.org/10.1016/j.neuroimage.2017.11.010

Yoo, K., Rosenberg, M. D., Noble, S., Scheinost, D., Constable, R. T., & Chun, M. M. (2019). Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *NeuroImage, 197*, 212–223. https://doi.org/10.1016/J.NEUROIMAGE.2019.04.060

Yoshihara, Y., Lisi, G., Yahata, N., Fujino, J., Matsumoto, Y., Miyata, J., Sugihara, G.-I., Urayama, S.-I., Kubota, M., Yamashita, M., Hashimoto, R., Ichikawa, N., Cahn, W., van Haren, N. E. M., Mori, S., Okamoto, Y., Kasai, K., Kato, N., Imamizu, H., … Takahashi, H. (2020). Overlapping but asymmetrical relationships between schizophrenia and autism revealed by brain connectivity. *Schizophrenia Bulletin, 46*(5), 1210–1218. https://doi.org/10.1093/schbul/sbaa021

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., & Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping, 39*(11), 4213–4227. https://doi.org/10.1002/HBM.24241

Zhou, Y., Liang, M., Tian, L., Wang, K., Hao, Y., Liu, H., Liu, Z., & Jiang, T. (2007). Functional disintegration in paranoid schizophrenia using resting-state fMRI. *Schizophrenia Research, 97*(1–3), 194–205. https://doi.org/10.1016/J.SCHRES.2007.05.029