DOCTORAL THESIS

Functional magnetic resonance imaging-based methods for translational research of psychiatric disorders

Auther:
Ayumu YAMASHITA

Supervisor:
Dr. Shin ISHII

Co-supervisors:
Dr. Mitsuo KAWATO
Dr. Manabu KANO
Dr. Tetsuya MATSUDA

KYOTO UNIVERSITY

Abstract

Graduate School of Informatics
Department of Systems Science

Functional magnetic resonance imaging-based methods for translational research of psychiatric disorders

by Ayumu YAMASHTA

In the field of cognitive neuroscience, researchers have been studying to clarify the mechanism of the brain that causes various phenomena using functional magnetic resonance imaging (fMRI). With the development of new approaches have come attempts to apply fMRI to real-world problems, specifically in medical contexts. The approaches can be roughly divided into two types. One approach is prediction of outcomes (e.g. a diagnosis) from neuroimaging data. Growing studies of a data-driven approach point to the utility of resting-state fMRI can be used to interrogate a multitude of functional brain network (functional connectivity) simultaneously to discover the functional connectivity which associated with psychiatric disorder. This leads, for example, to assist in diagnosing whether participant is psychiatric disorder or not by observing functional connectivity pattern. The other approach is intervention for psychiatric disorders using fMRI neurofeedback in which real-time online fMRI signals are used to self-regulate brain function. FMRI neurofeedback is expected to become a next-generation therapy for psychiatric disorders, because this technique can non-invasively manipulate the brain activity. In the former, however, many previous studies have not been achieved to construct prediction model that can be truly useful for any imaging site because they used the dataset from few imaging sites and were mainly relying on the diagnosis which recently been known that the relationship with the neurobiological basis is weak. In the latter, since neurofeedback manipulating the local brain activity has not broad utility, improvement of technique is necessary to become a next-generation therapy for psychiatric disorders. In this thesis, we conducted three researches to solve these problems. In the first work, we developed a state-of-the-art harmonization method which enable us to analyze large-scale resting-state fMRI dataset from multiple imaging sites. In the second work, by using large-scale multi-site resting-state fMRI dataset we constructed a reliable prediction model of depressive symptoms which more directly related with biological basis than diagnosis. We found the functional connections which associated with major depressive disorder diagnosis and depressed symptoms simultaneously. These functional connections are likely to be a therapeutic target of intervention for psychiatric disorder. In the third work, we developed a connectivity neurofeedback which can induce an aimed direction of change in functional connectivity and a differential change in cognitive performance. This technique could be used for intervening the functional connectivity of therapeutic target. These works would provide possible framework of therapeutic intervention for psychiatric disorder using fMRI.

Acknowledgements

First of all, I would like to express my gratitude to Dr. Hiroshi Imamizu who has given me permission to study freely and continuous guidance. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. I deeply grateful the consistent support and advice of Dr. Mitsuo Kawato. He has given many insightful comments and suggestions and taught me how good research is done. I am greatly appreciating to Dr. Shin Ishii who is my supervisor in Kyoto University. He has given me a lot of continuous supports to improve my Ph.D. thesis. I would also like to thank my committee members, Dr. Tetsuya Matsuda and Dr. Manabu Kano for comments that greatly improved my Ph.D. thesis.

I would like to thank all previous and current Advanced Telecommunications Research Institute International (ATR) members, especially to Dr. Takeo Watanabe, Dr. Yuka Sasaki, Dr. Okito Yamashita, Dr. Jun Morimoto, Dr. Saori C. Tanaka, Dr. Yukiyasu Kamitani, Dr. Motoaki Kawanabe, Dr. Kazushi Ikeda, Dr. Tomohiro Shibata, Dr. Yuki Sakai, Dr. Takashi Yamada, Dr. Shunsuke Hayasaka, Dr. Atsunori Kanemura, Dr. Kazuhisa Shibata, Dr. Ai Koizumi, Dr. Tomohisa Asai, Dr. Cai Chang, Dr. Tomoyasu Horikawa, Dr. Kei Majima, Dr. Masahiro Yamashita, Dr. Makoto Fukushima, Dr. Lisi Giuseppe, Dr. Cortese Aurelio, Dr. Megumi Fukuda, Dr. Junichiro Furukawa, Dr. Ryu Ohata, Dr. Asuka Takai, Dr. Shunta Togo, Dr. Yu Takagi, Dr. Koji Ishihara, Mr. Shinya Chiyohara, Mr. Masashi Hamaya, Mr. Keita Suzuki, Mr. Yuto Okada, Mr. Kai Suwabe, and Mr. Takeshi Ito. They gave me a lot of useful and interesting knowledge about machine-learning and computational neuroscience. My long Ph.D. student days were supported by these collaborators.

I would also like to thank support ATR and Brain Activity Imaging Center staffs, especially to Mieko Namba, Ritsuko Mashimo, Mieko Hirata, Yoko Matsumoto, Kaori Nakamura, Kana Inoue, Kaori Tachi, Toshinori Yoshioka, Koujiro Fujii, Mitsutoshi Uchida, Yasuhiro Shimada, Akikazu Nishikido, Ichiro Fujimoto, Takanori Kochiyama, and all participants for my experiments. I could not conduct my research projects without their administrative help.

My sincere thanks also go to Dr. Hidehiko Takahashi in Kyoto University, Dr. Ryuichiro Hashimoto in Showa University, Dr. Noriaki Yahata in National Institutes for Quantum and Radiological Science and Technology, Dr. Yasumasa Okamoto in Hiroshima University, Dr. Kiyoto Kasai in University of Tokyo, Dr. Koji Matsuo in Yamaguchi University and their laboratory members for their help to collect data and many discussions. Without their precious support it would not be possible to conduct this research.

I was helped by many people in Kyoto University during pursuing my doctoral course study. I would like to thank the previous and current members of Ishii-laboratory. My significant interest to the informatics is founded during continuous discussion with them. Thanks to Dr. Shigeyuki Oba and Dr. Shin-ichi Maeda, Dr. Hidetoshi Urakubo, Dr. Henrik Skibbe, Dr. Ken Nakae, Dr. Naoki Honda, Dr. Yohei Kondo, Dr. Hiroshi Morioka, Dr. Yumi Shikauchi, Dr. Kourosh Meshgi, Dr. Kousuke Yoshida for giving valuable advice.

The short period of my belonging at Laboratory for Brain Connectomics Imaging at RIKEN Center for Biosystems Dynamics Research has given me the knowledge about the latest analysis pipeline of functional magnetic resonance imaging. I would like to thank Dr. Takuya Hayashi, Group leader of this laboratory, for his many supports.

The studies in this thesis were conducted under the “Development of BMI Technologies for Clinical Application” of the Strategic Research Program for Brain Sciences supported by the Japan Agency for Medical Research and Development (AMED). These studies were also partially supported by the ImPACT Program of the Council for Science, Technology and Innovation (Cabinet Office, Government of Japan). These studies were supported also by the Japan Society for the Promotion of Science through Grant-in-Aid for JSPS Fellows.

Special thanks also to the CiNet tennis club, ATR tennis club, and Kyoto University tennis club members for precious time to play tennis. Without their precious support it would not be possible to continue my PhD. student days.

Finally, I am deeply grateful to my beloved family: my mother, father, sister, and brother. Thank you for supporting me for everything.

Contents

Chapter 1 15

The history of functional magnetic resonance imaging studies 15

1.1 Functional magnetic resonance imaging 15

1.1.1 From a hypothesis-driven to a data-driven approach 16

1.1.2 From a brain-measurement to a brain-manipulation approach 16

1.2 Translational fMRI study for psychiatric disorder 17

1.2.1 Prediction of diagnosis and response to treatment 18

1.2.2 Intervention with fMRI neurofeedback training 18

1.3 Problems addressed in this thesis 19

1.3.1 Problem of using data collected from small number of imaging site 19

1.3.2 Problem of diagnosis-based analysis 20

1.3.3 Problem of controllability of neurofeedback training 20

1.4 Organization of this thesis 21

1.4.1 Development of a harmonization method of rs-fMRI data across multiple imaging sites 21

1.4.2 Investigation of common resting-state functional connectivity underlying MDD diagnosis and depressed symptoms 21

1.4.3 Development of functional connectivity neurofeedback 22

Chapter 2 23

Harmonization of rs-fMRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias 23

2.1 Materials and methods 23

2.1.1 Datasets 23

2.1.2 Preprocessing and calculation of the resting-state functional connectivity matrix . 28

2.1.3 Estimation of biases and factors 29

2.1.4 Quantification of the site differences 30

2.1.5 Spatial characteristics of measurement bias, sampling bias, and each factor in the brain 31

2.1.6 Hierarchical clustering analysis for measurement bias 31

2.1.7 Comparison of models for sampling bias 32

2.1.8 Traveling-subject harmonization procedures 34

2.1.9 Principal component analysis 34
2.1.10 Two-fold cross-validation evaluation procedure 34
2.2 Results 37
2.2.1 Quantification of site differences 37
2.2.2 Brain regions contributing most to biases and associated factors 39
2.2.3 Characteristics of measurement bias 40
2.2.4 Sampling bias is because of sampling from among a subpopulation 41
2.2.5 Visualization of the harmonization effect 42
2.2.6 Quantification of the effect of traveling-subject harmonization 45
2.3 Discussion 46
2.3.1 The effect sizes of measurement and sampling biases 46
2.3.2 Characteristics of measurement bias 46
2.3.3 Characteristics of sampling bias 47
2.3.4 The effect of harmonization method 47
2.3.5 Limitations 48
2.3.6 Summary 48

Chapter 3 51

A common brain network between major depressive disorder and symptoms of depression 51

3.1 Material and Methods 52
3.1.1 Participants 52
3.1.2 Datasets 53
3.1.3 Preprocessing and calculation of the resting-state FC matrix 57
3.1.4 MDD classifier in the training dataset 59
3.1.5 BDI score regression model in the training dataset 59
3.1.6 Generalization performance of the classifier and regression model 61
3.1.7 Identification of the FCs linked to diagnosis and symptoms 61
3.2 Results 61
3.2.1 MDD classifier in the discovery dataset 61
3.2.2 Regression models of BDI score in the discovery dataset 62
3.2.3 Generalization performance of the classifier and the regression model 63
3.2.4 Common FCs between major depressive disorder diagnosis and symptoms of depression 65
3.3 Discussion 67
3.3.1 Signatures of our classifier of MDD 67

3.3.2 Common FCs between diagnosis of MDD and symptoms of depression 68
3.3.3 Importance of symptom-based approach, rather than diagnosis-based approach ... 68
3.3.4 Candidate of theranostic biomarker 68
3.3.5 Summary 69

Chapter 4 71

Development of functional connectivity neurofeedback 71

4.1 Materials and Methods 71
4.1.1 Participants 71
4.1.2 Neurofeedback training 72
4.1.3 Resting-state fMRI (rs-fMRI) 76
4.1.4 Cognitive tasks 77
4.2 Results 78
4.2.1 Change in score 78
4.2.2 Change in functional connectivity during training 79
4.2.3 Change in resting-state functional connectivity 80
4.2.4 Change in cognitive performance 81
4.3 Discussion 82
4.3.1 Directions of change in reaction times dependent on the tasks 82
4.3.2 Difference in behaviors during training between subject groups 83
4.3.3 Difference in the activity of target ROIs during training between the groups 84
4.3.4 Change in resting-state functional connectivity 84
4.3.5 Effect of the initial functional connectivity on training 85
4.3.6 Associations among change in functional connectivity during training, change in resting-state functional connectivity, and change in cognitive performance 85
4.3.7 Application of connectivity neurofeedback training 85
4.3.8 Summary 86

Chapter 5 87

Conclusion and Future Directions 87

5.1 Main contributions of this thesis 87
5.2 Challenges for the future 88
5.2.1 Challenges in the data-driven approach 88
5.2.2 Challenges in the brain-manipulation approach 88

Appendix A 91

Appendix of Chapter 2 91

A.1 Magnitude distribution of both biases and each factor on functional connectivity 91
A.2 Field map correction 91
A.3 Selection of the regularization hyper-parameter lambda 92
A.4 Brain regions contributing the measurement bias of each site 92
A.5 Classifiers for MDD and SCZ, based on the four harmonization methods 92
A.6 Regression models for age based on the four harmonization methods 94

Appendix B 112

Appendix of Chapter 4 113

B.1 Other behavioral metrics in behavioral tasks 113
B.2 Difference in total score between subject groups 113
B.3 Strategies adopted by subjects to increase their score 114
B.4 Regional brain activity during the training 114
B.5 Relationship between online score and activity in each ROI 115
B.6 Effect of the initial functional connectivity on training 115
B.7 Moderation/mediation analysis 116

Bibliography 123

List of Figures

FIGURE 2.1 | Schematic examples illustrating the two main datasets. 25
FIGURE 2.2 | Statistics of magnitude distributions for each type of bias and each factor. 38
FIGURE 2.3 | Spatial distribution of each type of bias and each factor in various brain regions. 40
FIGURE 2.4 | Clustering dendrogram for measurement bias. 41
FIGURE 2.5 | Comparison of the two models of sampling bias. 42
FIGURE 2.6 | PCA dimension reduction in the SRPBS multi-disorder dataset after harmonization. 44
FIGURE 2.7 | Reduction of the measurement bias and improvement of signal to noise ratios for different harmonization methods. 45
FIGURE 3.1 | Schematic illustration of the study design. 52
FIGURE 3.2 | Schematic representation of the procedure for selecting FCs in the MDD classifier and BDI regression model, and assessing their predictive power. 60
FIGURE 3.3 | Classifier performance for MDD and regression performance for BDI score in the discovery dataset. 62
FIGURE 3.4 | Classifier performances for MDD and regression performance for BDI score in the independent validation dataset. 64
FIGURE 3.5 | Results of permutation test for MDD classifier. 65
FIGURE 3.6 | Common functional connectivity between diagnosis and symptoms. 66
FIGURE 3.7 | Similar FC values between training dataset and independent validation dataset in 7 common FCs. 67
FIGURE 4.1 | Neurofeedback training procedures. 73
FIGURE 4.2 | Change in score during neurofeedback training. 79
FIGURE 4.3 | Change in functional connectivity between the left primary motor area (lM1) and the left lateral parietal region (lLP) during neurofeedback training. 80
FIGURE 4.4 | Changes in cognitive performance from pre-neurofeedback to post-neurofeedback training 81

List of Tables

TABLE 2.1 | Demographic characteristics of patients included in the SRPBS multi-disorder dataset....26
TABLE 2.2 | Imaging protocols for resting-state fMRI in the SRPBS multi-disorder dataset 27
TABLE 3.1 | Demographic characteristics of participants in the discovery dataset..54
TABLE 3.2 | Unified imaging protocols for resting-state fMRI in the discovery dataset 54
TABLE 3.3 | Different imaging protocols among sites for resting-state fMRI in the discovery dataset....55
TABLE 3.4 | Demographic characteristics of participants in the independent validation dataset 55
TABLE 3.5 | Imaging protocols for resting-state fMRI in the independent validation dataset 56
TABLE 3.6 | Data availability statement 57
TABLE 3.7 | All common functional connections and weights in regression model of BDI score 66

Abbreviations Glossary

Abbreviations	Full Term
AAL	**A**natomical **A**utomatic **L**abeling
ABIDE	**A**utism **B**rain **I**maging **D**ata **E**xchange
ADHD	**A**ttention-**D**eficit **H**yperactivity **D**isorder
AIC	**A**kaike **I**nformation **C**riteria
AMED	the Japan **A**gency for **ME**dical research and **D**evelopment
ANOVA	**AN**alysis **O**f **VA**riance
ASD	**A**utism **S**pectrum **D**isorder
ATR	**A**dvanced **T**elecommunications **R**esearch Institute International
ATT	**A**dvanced **T**elecommunications Research Institute International **T**rio
ATV	**A**dvanced **T**elecommunications Research Institute International **V**erio
AUC	**A**rea **U**nder the **C**urve
BDI	**B**eck **D**epression **I**nventory-II
BIC	**B**ayesian **I**nformation **C**riteria
BOLD	**B**lood **O**xygen **L**evel **D**ependent
COI	**C**enter **O**f **I**nnovation
CRHD	**C**onnectomes **R**elated to **H**uman **D**isease
CSF	**C**erebro**S**pinal **F**luid
CV	**C**ross **V**alidation
CWST	**C**olor-**W**ord **S**troop **T**ask
DecNef	**Dec**oded **Ne**uro**f**eedback
DMN	**D**efault **M**ode **N**etwork
EFT	**E**riksen **F**lanker **T**ask
EPI	**E**cho **P**lanar **I**maging
FC	**F**unctional **C**onnectivity
FD	**F**rame-wise **D**isplacement
fMRI	**f**unctional **M**agnetic **R**esonance **I**maging
GE	**G**eneral **E**lectric Company
GLM	**G**eneral **L**inear **M**odel
HC	**H**ealthy **C**ontrol
HCP	**H**uman **C**onnectome **P**roject
HKH	**H**iroshima **K**ajikawa **H**ospital
HRC	**H**iroshima **R**ehabilitation **C**enter
HUH	**H**iroshima **U**niversity **H**ospital
KPM	**K**yoto **P**refectural University of **M**edicine

KUS	**K**yoto **U**niversity **S**kyra
KUT	**K**yoto **U**niversity **T**rio
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
lLP	**l**eft **L**ateral **P**arietal cortex
lM1	**l**eft primary **M**otor cortex
MAE	**M**ean **A**bsolute **E**rror
MCC	**M**atthews **C**orrelation **C**oefficient
MDD	**M**ajor **D**epressive **D**isorder
MNI	**M**ontreal **N**eurological **I**nstitute
MR	**M**agnetic **R**esonance
MVN	**M**otor/**V**isuospatial **N**etwork
OCD	**O**bsessive-**C**ompulsive **D**isorder
PC	**P**rincipal **C**omponent
PCA	**P**rincipal **C**omponent **A**nalysis
PVT	**P**sychomotor **V**igilance **T**ask
ROI	**R**egion **O**f **I**nterest
rs-fMRI	**r**esting-**s**tate **f**unctional **M**agnetic **R**esonance **I**maging
rs-fcMRI	**r**esting-**s**tate **f**unctional **c**onnectivity **M**agnetic **R**esonance **I**maging
SD	**S**tandard **D**eviation
SCZ	**SC**hi**Z**ophrenia
SRPBS	**S**trategic **R**esearch **P**rogram for **B**rain **S**ciences
SWA	**S**ho**WA** University
TD	**T**ypically **D**eveloped controls
tDCS	**t**ranscranial **D**irect **C**urrent **S**timulation
TE	**E**cho **T**ime
TMS	**T**ranscranial **M**agnetic **S**timulus
TR	**R**epetition **T**ime
UTO	**U**niversity of **TO**kyo
UYA	**YA**maguchi **U**niversity
WM	**W**hite **M**atter
YC1	**Y**aesu **C**linic scanner **1**
YC2	**Y**aesu **C**linic scanner **2**

Chapter 1

The history of functional magnetic resonance imaging studies

We are able to appreciate a beautiful view seen on the way to the laboratory, to reach for a cup of coffee to drink, and to be excited by watching sports. All these phenomena are caused by our brain. Brain further cause various dysfunction of our cognitive ability in psychiatric disorders due to its breakdown. To elucidate how our brain causes these phenomena and abnormality, cognitive neuroscientists have measured brain activity during various cognitive tasks and in many psychiatric disorders. Brain activity of a living human was invasively measured for the first time in 1924 (Adrian and Matthews, 1934; Berger, 1929; Compston, 2010). Seventy years later, with the discovery of the blood-oxygen level dependent (BOLD) response (Ogawa et al., 1990a; Ogawa et al., 1990b), it became possible to noninvasively measure human whole brain activity using functional magnetic resonance imaging (fMRI). The invention of fMRI was a break through that allowed basic cognitive neuroscience research to elucidate human brain mechanisms. Rapid advances of the technologies that measure and manipulate brain function have come attempts to apply fMRI to real-world problems (translational fMRI), specifically in medical contexts (Matthews et al., 2006; Poldrack and Farah, 2015). The 30 years adventure of fMRI studies generated great optimism about its potential for delivering clinically useful applications. While much progress has been made on the basic cognitive neuroscience research to elucidate human brain mechanisms, few results have been incorporated into clinical practice due to some problems. In this thesis, we conducted three researches to resolve these problems that are bottlenecks to develop the applications of fMRI resolving the real-world problems.

This chapter will first explain the basic principles of fMRI, summarize two recent major innovations in approach to research using fMRI in cognitive neuroscience field and introduce some recent translational fMRI studies accompanying with these two innovations. Finally, this chapter will summarize their bottlenecks to develop the applications of fMRI and explain how this thesis expands these bottlenecks.

1.1 Functional magnetic resonance imaging

MRI is a standard tool in Radiology that is used to capture high resolution images with good contrast between different body tissues such as in the brain. Using the phenomenon of nuclear magnetic resonance, the hydrogen nuclei generate a magnetic resonance (MR) signal that can be mapped and turned into an MR image. MR signal changes depending on the blood-oxygen level (Ogawa et al., 1990a; Ogawa et al., 1990b; Ogawa et al., 1992) in which deoxidized hemoglobin attenuates the MR signal and the MR signal changes dependent upon the amount of deoxidized hemoglobin in the blood. That is, when neural activities occur locally in a brain region, oxygen consumption in that brain region increases. To supply oxygen for that region, blood flow for that region increases (neurovascular coupling) (Roy and Sherrington, 1890). The total amount of deoxidized hemoglobin in that brain region decreases and the attenuated MR signal recovers. We must always be aware that fMRI does not directly measure electrical activity, such as the

firing of neurons, but rather that it indirectly measures brain activity by measuring the blood flow change accompanying the neural activity. In the last two decades, fMRI has routinely been used to answer questions about mind–brain relationships that go far beyond the simple localization of brain function (Bandettini, 2012; Norman et al., 2006; Poldrack and Farah, 2015).

1.1.1 From a hypothesis-driven to a data-driven approach

Since the development of fMRI, cognitive neuroscientists have traditionally focused on hypothesis-driven task-based approaches. However, data-driven discovery science, which has been very successful in the genetic research field (Wise, 2008), has not been conducted in the cognitive neuroscience field, because the accumulation and sharing of large-scale datasets for data mining is necessary for discovery science (Biswal et al., 2010).

Resting-state fMRI (rs-fMRI) has recently emerged as a powerful tool for the data-driven approach in the cognitive neuroscience field (Buckner et al., 2013; Smith et al., 2013). Imaging the brain during rest reveals large amplitude spontaneous low-frequency (<0.1Hz) fluctuations in fMRI signal (Fox and Raichle, 2007). For many years, researchers regarded theses spontaneous fluctuations as noise. However, the importance of rs-fMRI was eventually realized due to the facts that energy consumption is very large at rest and spontaneous fluctuations are temporally correlated across functionally related brain regions (Blamire et al., 1992; Fox and Raichle, 2007; Fox et al., 2005; Raichle, 2015a, b; Raichle et al., 2001). This temporal correlation between two brain regions is called “functional connectivity” (FC). Although there is still controversy regarding the relationship between the spontaneous fluctuations of the BOLD signal and neural activity at rest (Winder et al., 2017), electrophysiological studies in non-humans have revealed its neural basis (Arieli et al., 1996; Berkes et al., 2011; He et al., 2018; Kenet et al., 2003; Laufs et al., 2003; Lu et al., 2007; Ma et al., 2016; Magri et al., 2012; Mateo et al., 2017; Matsui et al., 2016; Scholvinck et al., 2010; Shmuel and Leopold, 2008; Vincent et al., 2007). Rs-fMRI can be used to interrogate a multitude of functional brain network simultaneously, without the requirement of selecting a priori hypotheses (Biswal et al., 2010; Smith et al., 2013). In recent years, it is becoming more and more possible to predict various personal characteristics such as sustained attention ability and age from individual brain networks estimated from rs-fMRI using machine learning techniques (Dosenbach et al., 2010; Rosenberg et al., 2016; Smith et al., 2013). There is also growing evidence that these networks may be important in psychiatric disorders (Deco and Kringelbach, 2014; Fornito et al., 2015).

1.1.2 From a brain-measurement to a brain-manipulation approach

In the cognitive neuroscience field, scientists have typically regarded physical variables, psychological variables, and human behavior as independent variables and brain activity as a dependent variable. That is, scientists have measured changes in brain activity, that occur as physical variables, psychological variables, and human behavior are varied (a brain-measurement approach). In such an experimental design, however, although we can measure how the brain activity *correlates* with these independent variables, we cannot clarify how the brain *causes* phenomena such as perception and emotion. Therefore, a different type of experimental design, in which brain activity is regarded as an independent variable and human behavior is regarded as a dependent variable, has

emerged in the cognitive neuroscience field (a brain-manipulation approach). For example, there are experiments where transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS), and fMRI neurofeedback training, are utilized. FMRI neurofeedback is a type of biofeedback in which real-time online fMRI signals are used to self-regulate brain function (Cox et al., 1995; deCharms, 2008; Sitaram et al., 2017; Sulzer et al., 2013). Some fMRI neurofeedback techniques are superior to brain stimulation methods in their ability to manipulate the brain, because these fMRI neurofeedback techniques can be integrated with advanced technology such as multi variate pattern analysis (Cohen et al., 2017; Kamitani and Tong, 2005; Norman et al., 2006) and network analysis (Bassett and Sporns, 2017), and can induce a specific activation pattern in the targeted brain region and alter the brain network, rather than simply increasing or decreasing the mean activation level (Megumi et al., 2015; Shibata et al., 2011; Sitaram et al., 2017; Watanabe et al., 2017). Furthermore, fMRI neurofeedback is also superior to brain stimulation in ethical aspects, because brain activation is voluntarily regulated through learning in fMRI neurofeedback training rather than via compulsorily changing brain activity by external stimuli such as TMS and tDCS.

Currently, there are three types of fMRI neurofeedback that differ based on the type of information used for feedback. The first, univariate neurofeedback, uses the average BOLD signal within a specific brain region of interest to increase or decrease the average activity in that region (deCharms, 2008; deCharms et al., 2005; Weiskopf et al., 2004). The second, decoded neurofeedback, uses the multi-variate activity pattern in a region to induce a specific piece of information in that region, such as orientation, color, or facial preference (Amano et al., 2016; deBettencourt et al., 2015; LaConte et al., 2007; Shibata et al., 2016; Shibata et al., 2011). The third, connectivity neurofeedback, uses the FC between regions to modulate connectivity between two targeted brain regions (Koush et al., 2015; Koush et al., 2013; Liew et al., 2016; Megumi et al., 2015). To our understanding, there are two mainstream sets of studies on connectivity neurofeedback. One uses dynamic causal modeling to modulate the state of a brain network and cognitive performance (Koush et al., 2015; Koush et al., 2013). The other uses Pearson's correlation coefficients of activity time courses between two targeted brain regions (Liew et al., 2016; Megumi et al., 2015).

1.2 Translational fMRI study for psychiatric disorder

With the development of these new approaches have come attempts to apply fMRI to real-world problems, specifically in medical contexts for psychiatric disorder (Matthews et al., 2006; Poldrack and Farah, 2015). Basic cognitive neuroscience is concerned with understanding brain mechanisms, whereas translational neuroscience is concerned with developing tools that are useful for clinical context (Castellanos et al., 2013). There are two mainstreams of translational fMRI study for psychiatric disorder. One is prediction of current state and future state such as diagnosis and response to treatment of psychiatric disorder using fMRI data (Woo et al., 2017). The other is fMRI neurofeedback as therapeutic interventions for psychiatric disorder (Sitaram et al., 2017; Stoeckel et al., 2014; Watanabe et al., 2017), in response to the urgent need for better treatments for psychiatric disorder. It is because that advanced fMRI neurofeedback could voluntarily regulate the brain function not just increase or decrease the brain activity through learning (Watanabe et al., 2017).

Recently, the functional connectome is in the foreground of cognitive neuroscience research aiming to achieve these clinical applications (Castellanos et al., 2013). The reason is that the view that the localization of psychological processes to specific areas of the brain provides only a partial account of brain function (Fornito et al., 2015). And now the brain is considered as a highly complex system in which interconnected network balances regional segregation and integration of function with strong specialization.

1.2.1 Prediction of diagnosis and response to treatment

With the development of the data-driven approach, current state of patients with psychiatric disorder (e.g. diagnosis) could be predicted from individual brain network (Woo et al., 2017; Xia and He, 2017). Rs-fMRI methods can lead to this purpose, because its relatively widespread availability (e.g. applicable at hospital without special equipment) and amenability (e.g. applicable for almost all psychiatric disorders without complex task) to large-scale aggregation across imaging sites and populations. Indeed, the rate of growth for studies incorporating rs-fMRI approaches has overtaken that of traditional task-based fMRI, with an increasing focus on clinical questions. Task-based imaging has struggled with marked variability in approaches and findings across laboratories, even when studying the same cognitive construct. Such variability is problematic for data aggregation (Milham, 2012).

In order to predict subject's state, variety of machine learning techniques have been applied to functional brain networks. For example, Rosenberg and colleagues constructed a prediction model of sustained attention ability based on functional connections (FCs) and this model can also predict symptoms of attention-deficit hyperactivity disorder (ADHD) (Rosenberg et al., 2016). Yahata and colleagues constructed a biomarker of autism spectrum disorder (ASD) which distinguishes between typically developed controls (TDs) and ASD patients based on a small number of resting-state FCs, using rs-fMRI data collected from imaging sites in Japan. Furthermore, this biomarker was found to generalize to data collected in the USA (Yahata et al., 2016). Drysdale and colleagues showed that patients with major depressive disorder (MDD) can be subdivided into four neurophysiological subtypes defined by distinct patterns of FCs. Clustering patients on this basis enabled the development of biomarkers with high performance for depression subtypes in multisite validation and independent validation datasets (Drysdale et al., 2017).

Some studies have begun to develop brain models for more difficult prediction problems that are future state of patients (e.g. response to treatment). For example, Reggente and colleagues collected rs-fMRI data from adults with obsessive-compulsive disorder (OCD) before and after 4 weeks of intensive daily cognitive behavioral therapy. They showed that pretreatment FC patterns could predict post treatment OCD severity (Reggente et al., 2018). The study reported above by Drysdale et al., also showed that biotypes of depression, defined based on FC, could predict responsiveness to TMS therapy (Drysdale et al., 2017).

1.2.2 Intervention with fMRI neurofeedback training

Whereas neuroscience applications of fMRI-based neurofeedback training have been used for investigating neural mechanisms (Amano et al., 2016; Shibata et al., 2016; Shibata et al., 2011), translational studies have attempted therapeutic interventions for psychiatric

disorders (Niv, 2013; Sitaram et al., 2017; Stoeckel et al., 2014; Sulzer et al., 2013; Watanabe et al., 2017). Using univariate neurofeedback, deCharms and colleagues showed that they could control pain perception by controlling activation in the rostral anterior cingulate cortex, a region putatively involved in pain perception (deCharms et al., 2005). Scheinost and colleagues demonstrated control of contamination anxiety by controlling activation in the orbitofrontal cortex (Scheinost et al., 2013). Young and Linden were able to control emotion by controlling activation in the amygdala and emotion network which included the ventrolateral prefrontal cortex and insula (Linden et al., 2012; Young et al., 2014). Using decoded neurofeedback, Koizumi and Taschereau-Dumouchel showed that they could reduce fear towards a target object (e.g. spider) by pairing reward with activation patterns in the visual cortex that representing the target object. Participants remained unaware of the content and purpose of the procedure (Koizumi et al., 2016; Taschereau-Dumouchel et al., 2018). However, although numerous studies have reported that psychiatric disorders are related to abnormal brain networks rather than local brain activity (Broyd et al., 2009; Fornito et al., 2015; Stam, 2014), there is no intervention using connectivity neurofeedbck due to its immature.

1.3 Problems addressed in this thesis

As this thesis have explained so far, although some progress has been made on the translational fMRI research for clinical application, fMRI has not thus far been truly useful in resolving the real-world problems. It is because that nobody constructed a reliable prediction model which generalize to data collected from any imaging site and that there is no intervention method which can apply for various psychiatric disorders. This section will explain some of the scientific reasons why this is the case.

1.3.1 Problem of using data collected from small number of imaging site

First problem is that almost all previous studies used neuroimaging dataset collected from small number of imaging site. The prediction model constructed from the dataset collected from small number of imaging site lacks the generalization ability to the data collected from new site not included when constructing the prediction model. To construct a reliable prediction model which generalize to dataset collected from any imaging site, it is essential to use a large sample size of data collected from multiple imaging sites. Therefore, acquiring and sharing large neuroimaging data have recently become critical to conduct reliable discovery science (Human Connectome Project (HCP) (Glasser et al., 2016b), [http://www.humanconnectomeproject.org/]; Human Brain Project [https://www.humanbrainproject.eu/en/]; UK Biobank [http://www.ukbiobank.ac.uk/]; and Strategic Research Program for Brain Sciences (SRPBS) (Yamada et al., 2017) [https://www.amed.go.jp/program/list/01/04/001_nopro.html]) (Biswal et al., 2010; Woo et al., 2017; Xia and He, 2017), especially for dataset including psychiatric disorders (Connectomes Related to Human Disease (CRHD), [https://www.humanconnectome.org/disease-studies]; Autism Brain Imaging Data Exchange (ABIDE); and SRPBS) (Di Martino et al., 2014; Pearlson, 2009; Yahata et al., 2017; Yamada et al., 2017). However, multisite datasets with multiple disorders raises difficult problems which are not present in a single-site based dataset of healthy population (e.g., HCP and UK Biobank). That is, even if a unified protocol is determined there exists difficulty in full control of scanner type, imaging protocol, and patient

demographics (Abraham et al., 2017; Nieuwenhuis et al., 2017; Orban et al., 2018; Yahata et al., 2016). Moreover, there often exists unpredictable differences in participant population among sites. Therefore, researchers must work with heterogeneous neuroimaging data when they analyze multisite dataset. In particular, site differences represent the greatest barrier when extracting disease factors by applying machine-learning techniques to such heterogeneous data (Dansereau et al., 2017) because disease factors tend to be confounded with site factors (Abraham et al., 2017; Nieuwenhuis et al., 2017; Orban et al., 2018; Watanabe et al., 2017; Yahata et al., 2017; Yahata et al., 2016; Yamada et al., 2017). This confounding occurs because a single site (or hospital) is apt to sample only a few types of psychiatric disorders (e.g., primarily schizophrenia from site A and primarily autism spectrum disorder from site B). Furthermore, site differences essentially consist of two types of biases: engineering bias (i.e., measurement bias) and biological bias (i.e., sampling bias). Measurement bias is completely noise, which includes differences in the properties of MRI scanners such as imaging parameters, field strength, MRI manufacturers, and scanner models, whereas sampling bias is biologically meaningful information which refers to differences in participant groups among sites. However, previous studies have not divided site differences into measurement bias and sampling bias. Therefore, existing methods to correct site-differences might fail to eliminate both biologically meaningless measurement bias and biologically meaningful sampling bias.

1.3.2 Problem of diagnosis-based analysis

Second problem is that most previous studies depend on clinical diagnosis. These previous studies identified resting-state FCs that characterized patients or sought to construct a biomarker which distinguishes between disordered patients and healthy controls (HCs) based on resting-state FCs (Woo et al., 2017). However, an increasing number of studies have highlighted the difficulty in finding a clear association between existing clinical diagnostic categories and neurobiological abnormalities (Clementz et al., 2016; Insel and Cuthbert, 2015; Singh and Rose, 2009). This is due to the fact that the diagnosis of patients is based on a complex mix of information, such as symptoms, epidemiological surveys, and clinical experience. The high co-morbidity of structural, functional, and genetic abnormalities across psychiatric disorders exacerbates this difficulty (Goodkind et al., 2015; Jacobi et al., 2004; Lee et al., 2013; McTeague et al., 2017). In order to understand the nature of mental health and psychiatric disorders in terms of varying degrees of dysfunctions in general psychological/biological systems, the importance of the research not dependent on diagnosis such as research domain criteria in which the goal is to understand the nature of mental health and illness in terms of varying degrees of dysfunctions in general psychological/biological systems has attracted attention.

1.3.3 Problem of controllability of neurofeedback training

Third problem is that most previous fMRI neurofeedback focus on manipulating brain activity within local brain region. Numerous studies have reported that psychiatric disorders are related to abnormal brain networks rather than local brain activity (Broyd et al., 2009; Fornito et al., 2015; Stam, 2014). Importantly, brain networks are also associated with certain cognitive functions (Barch et al., 2013; He et al., 2007; Kelly et al., 2008; Thompson et al., 2013). This indicates that using connectivity neurofeedback could be a promising approach for therapeutic intervention in psychiatric disorders and to

improve cognitive function. However, fewer studies have been conducted on connectivity neurofeedback. In particular, the controllability of connectivity neurofeedback is critical for applications aimed at psychiatric disorders. That is, it is important to examine whether connectivity neurofeedback can induce the aimed direction of change (i.e., an increase or a decrease) in FC and a change in cognitive performance. Previous studies using Pearson's correlation coefficients tested only increases in connectivity.

1.4 Organization of this thesis

1.4.1 Development of a harmonization method of rs-fMRI data across multiple imaging sites

In chapter 2, we first quantitatively investigated site-difference on rs-fMRI data and developed a novel harmonization method to make multi-site dataset available for analysis. To achieve this objective, we utilized a traveling-subject rs-fMRI dataset, wherein multiple participants travel to multiple sites for the assessment of measurement bias. This was used in conjunction with a multi-site multi-disorder rs-fMRI dataset to demonstrate that site differences are composed of biological sampling bias and engineering measurement bias. We found that effects on resting-state FC caused by both bias types were greater than or equal to those caused by psychiatric disorders. Furthermore, our findings indicated that each site can sample only from among a subpopulation of participants. This result suggests that it is essential to collect large neuroimaging data from as many sites as possible to appropriately estimate the distribution of the grand population. Finally, we developed a novel harmonization method that removed only the measurement bias by using the traveling-subject dataset. This achieved reduction of the measurement bias by 29% and improvement of the signal to noise ratios by 40%. Development of an accurate harmonization method promotes large-scale data analysis for the discovery approach.

1.4.2 Investigation of common resting-state functional connectivity underlying MDD diagnosis and depressed symptoms

In chapter 3, to break away from diagnosis-dependent analysis framework, we conducted a symptom-based approach, which describes the association between symptoms and neurobiological abnormalities. We constructed a reliable resting-state FC-based classifier for MDD and also constructed a regression model of Beck Depression Inventory-II (BDI) score, which is one of the most widely used test for measuring severity of depressed symptoms. The MDD classifier and the regression model of BDI generalized to an independent validation dataset obtained at completely different imaging sites. We found an overlap of about 30% between the connections related to depressed symptoms and those related to the diagnosis of MDD. These functional connections were particularly related to the salience network and the default mode network. Our study revealed a partially overlapping relationship between the biological basis of depressed symptoms and that of MDD diagnosis.

1.4.3 Development of functional connectivity neurofeedback

In chapter 4, in order to investigate whether the connectivity neurofeedback is a versatile tool that can be used as an intervention for psychiatric disorders, we investigated the hypothesis that connectivity neurofeedback can induce the aimed direction of change in FC, and the differential change in cognitive performance according to the direction of change in connectivity. We selected connectivity between the left primary motor cortex and the left lateral parietal cortex as the target. Subjects were divided into 2 groups, between which only the direction of change (an increase or a decrease in correlation) in the experimentally manipulated connectivity differed. Results showed that subjects were successfully able to induce the expected connectivity change in both directions. Furthermore, cognitive performance significantly and differentially changed from preneurofeedback to postneurofeedback training between the 2 groups. These findings indicate that connectivity neurofeedback can induce the aimed direction of change in connectivity and also a differential change in cognitive performance.

Chapter 2

Harmonization of rs-fMRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias

In chapter 2, we developed a novel harmonization method to make large-scale multi-site rs-fMRI data available for data-driven analysis.

When collecting large neuroimaging data associated with psychiatric disorders, images must be acquired from multiple sites because of the limited capacity of a single site. However, site differences represent the greatest barrier when acquiring multi-site neuroimaging data. We utilized a traveling-subject dataset in conjunction with a multi-site, multi-disorder dataset to demonstrate that site differences are composed of biological sampling bias and engineering measurement bias. Effects on resting-state functional connectivity (FC) based on pair-wise correlations because of both bias types were greater than or equal to those because of psychiatric disorders. Furthermore, our findings indicated that each site can sample only from among a subpopulation of participants. This result suggests that it is essential to collect large neuroimaging data from as many sites as possible to appropriately estimate the distribution of the grand population. Finally, we developed a novel harmonization method that removed only the measurement bias by using traveling-subject dataset and achieved the reduction of the measurement bias by 29% and the improvement of the signal to noise ratios by 40%. Our results provide fundamental knowledges on site-effects with future research using multi-site multi-disorder rs-fMRI data.

2.1 Materials and methods

2.1.1 Datasets

We used two resting-state fMRI datasets for all analyses: (1) the SRPBS multi-disorder dataset, which encompasses multiple psychiatric disorders; (2) a traveling-subject dataset.

SRPBS multi-disorder dataset

This dataset included patients with five different disorders and healthy controls (HCs) who were examined at nine sites belonging to eight research institutions. A total of 805 participants were included: 482 HCs from nine sites, 161 patients with major depressive disorder (MDD) from five sites, 49 patients with autism spectrum disorder (ASD) from one site, 65 patients with obsessive-compulsive disorder (OCD) from one site, and 48 patients with schizophrenia (SCZ) from three sites (Table 2.1). Data were acquired using a Siemens TimTrio scanner at Advanced Telecommunications Research Institute International (ATT), a Siemens Verio scanner at Advanced Telecommunications Research Institute International (ATV), a Siemens Verio at the Center of Innovation in Hiroshima

University (COI), a GE Signa HDxt scanner at Hiroshima University Hospital (HUH), a Siemens Spectra scanner at Hiroshima Kajikawa Hospital (HKH), a Philips Achieva scanner at Kyoto Prefectural University of Medicine (KPM), a Siemens Verio scanner at Showa University (SWA), a Siemens TimTrio scanner at Kyoto University (KUT), and a GE MR750W scanner at the University of Tokyo (UTO). The rs-fMRI data were acquired using a unified imaging protocol at all but three sites (Table 2.2; https://bicr.atr.jp/rs-fmri-protocol-2/). Site differences in this dataset included both measurement and sampling biases (Fig. 2.1a). During the rs-fMRI scans, participants were instructed as follows, except at one site: "Please relax. Don't sleep. Fixate on the central crosshair mark and do not think about specific things." At the remaining site, participants were instructed to close their eyes rather than fixate on a central crosshair. For bias estimation, we only used data obtained using the unified protocol. (Patients with OCD were not scanned using this unified protocol; therefore, a disorder factor could not be estimated for OCD).

Traveling-subject dataset

We acquired a traveling-subject dataset to estimate measurement bias across sites in the SRPBS dataset. Nine healthy participants (all male participants; age range, 24–32 years; mean age, 27±2.6 years) were scanned at each of 12 sites in the SRPBS consortium, producing a total of 411 scan sessions. Data were acquired at the sites included in the SRPBS multi-disorder database (i.e., ATT, ATV, COI, HUH, HKH, KPM, SWA, KUT, and UTO) and three additional sites: Kyoto University (KUS; Siemens Skyra) and Yaesu Clinic 1 and 2 (YC1 and YC2; Philips Achieva) (Appendix Table A.1). 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 (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 underwent four rather than five sessions at the ATT site because of a poor physical condition. Thus, a total of 411 sessions were conducted [8 participants × (3×9＋2×2+5×3×1)＋1 participant × (3×9＋2×2+4×3×1)]. During each rs-fMRI session, participants were instructed to maintain a focus on a fixation point at the center of a screen, remain still and awake, and to think about nothing in particular. For sites that could not use a screen in conjunction with fMRI (HKH & KUS), a seal indicating the fixation point was placed on the inside wall of the MRI gantry. Although we had attempted to acquire this dataset using the same imaging protocol as that in the SRPBS multi-disorder dataset, there were some differences in the imaging protocol across sites because of limitations in parameter settings or the scanning conventions of each site (Appendix Table A.2). There were two phase-encoding directions (P→A and A→P), three MRI manufacturers (Siemens, GE, and Philips), four numbers of channels per coil (8, 12, 24, and 32), and seven scanner types (TimTrio, Verio, Skyra, Spectra, MR750W, SignaHDxt, and Achieva). Site differences in this dataset included measurement bias only as the same nine participants were scanned across the 12 sites (Fig. 2.1b).

All participants in all datasets provided written informed consent, and all recruitment procedures and experimental protocols were approved by the Institutional Review Boards of the principal investigators' respective institutions (Advanced Telecommunications Research Institute International (ATR), Hiroshima University, Kyoto Prefectural University of Medicine, Showa University, The University of Tokyo).

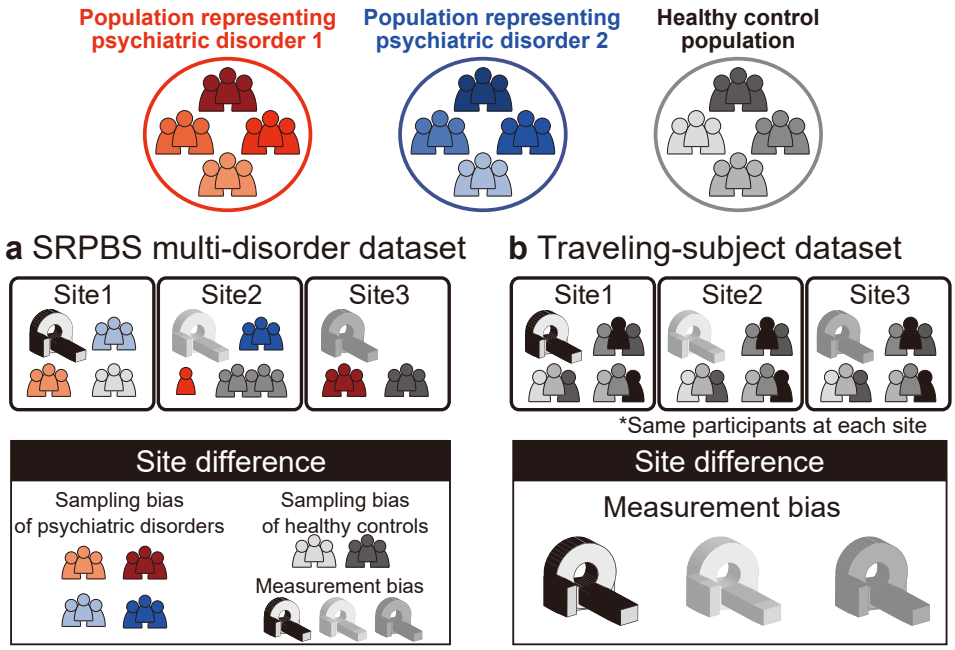

FIGURE 2.1 | Schematic examples illustrating the two main datasets.

(a) The SRPBS multi-disorder dataset includes patients with psychiatric disorders and healthy controls. The number of patients and scanner types differed among sites. Thus, site differences consist of sampling bias and measurement bias. (b) The traveling-subject dataset includes only healthy controls, and the participants were the same across all sites. Thus, site differences consist of measurement bias only. SRPBS: Strategic Research Program for Brain Sciences.

TABLE 2.1 | Demographic characteristics of patients included in the SRPBS multi-disorder dataset

Site	HC			MDD			ASD			OCD			SCZ			ALL				
	Number	Male /Female	Age (yr)	Number	Male /Female	Age (yr)	Number	Male /Female	Age (yr)	Number	Male /Female	Age (yr)	Number	Male /Female	Age (yr)	Number	Male /Female	Age (yr)	*1	*2
ATR TimTrio (ATT)	31	28/3	23.0±1.9	0	-	-	0	-	-	0	-	-	0	-	-	31	28/3	23.0±1.9	○	○
ATR Verio (ATV)	77	60/17	22.6±2.0	0	-	-	0	-	-	0	-	-	0	-	-	77	60/17	22.6±2.0	○	○
Hiroshima University Hospital (HUH)	66	37/29	34.6±13.0	57	32/25	43.3±12.2	0	-	-	0	-	-	0	-	-	123	69/54	38.6±13.3	-	○
Hiroshima Kajikawa Hospital (HKH)	29	17/12	45.4±9.5	23	13/10	43.6±11.6	0	-	-	0	-	-	0	-	-	52	30/22	44.6±10.5	-	○
Center of Innovation in Hiroshima University (COI)	10	5/5	43.5±13.5	38	20/18	44.0±11.0	0	-	-	0	-	-	0	-	-	48	25/23	43.9±11.4	○	○
Kyoto Prefectural University of Medicine (KPM)	52	28/24	29.1±7.3	0	-	-	0	-	-	65	30/35	31.9±9.8	0	-	-	117	58/59	30.6±8.8	-	○
Kyoto University (KUT)	35	18/17	36.3±8.9	9	5/4	45.2±15.9	0	-	-	0	-	-	22	11/11	40.4±8.4	66	34/32	38.9±10.2	○	○
Showa University (SWA)	40	8/32	30.9±8.5	0	-	-	49	45/4	32.9±8.1	0	-	-	12	11/1	41.8±9.2	101	64/37	33.2±8.9	○	○
University of Tokyo (UTO)	142	72/70	29.7±11.0	34	16/18	38.5±9.9	0	-	-	0	-	-	14	7/7	33.3±14.0	190	95/95	31.6±11.5	○	○
Summary	482	273/209	30.6±10.9	161	86/75	42.6±11.7	49	45/4	32.9±8.1	65	30/35	31.9±9.8	48	28/19	38.7±10.8	805	463/342	33.7±11.9	-	-

*1: Participants scanned using the unified protocol
*2: Sites used for constructing prediction models and principal component analysis
ATR: Advanced Telecommunications Research Institute International; HC: healthy control; MDD: major depressive disorder; ASD: autism spectrum disorder; OCD: obsessive-compulsive disorder; SCZ: schizophrenia; SRPBS: Strategic Research Program for Brain Sciences.

TABLE 2.2 | Imaging protocols for resting-state fMRI in the SRPBS multi-disorder 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**	**University of Tokyo**
Abbreviation	ATT	ATV	COI	HUH	HKH	KPM	SWA	KUT	UTO
MRI scanner	*Siemens TimTrio*	*Siemens Verio*	*Siemens Verio*	*GE Signa HDxt*	*Siemens Spectra*	*Philips Achieva*	*Siemens Verio*	*Siemens TimTrio*	*GE MR750w*
Magnetic field strength	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T
Number of channels per coil	12	12	12	8	12	8	12	32	24
Field-of-view (mm)	212 × 212	212 × 212	212 × 212	256 × 256	192 × 192	192 × 192	212 × 212	212 × 212	212 × 212
Matrix	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64
Number of slices	40 or 39	39	40	32	38	39	40	40	40
Number of volumes	240	240	240	143	107	194	240	240	240
In-plane resolution (mm)	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	4.0 × 4.0	3.0 × 3.0	3.0 × 3.0	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125
Slice thickness (mm)	3.2	3.2	3.2	3.2	3.0	3.0	3.2	3.2	3.2
Slice gap (mm)	0.8	0.8	0.8	0	0	0	0.8	0.8	0.8
TR (ms)	2,500	2,500	2,500	2,000	2,700	2,000	2,500	2,500	2,500
TE (ms)	30	30	30	27	31	30	30	30	30
Total scan time (min:s)	10:00	10:00	10:00	5:00	5:00	6:30	10:00	10:00	10:00
Flip angle (deg)	80	80	80	90	90	80	80	80	80
Slice acquisition order	Ascending	Ascending	Ascending	Ascending (Interleaved)	Ascending	Ascending	Ascending	Ascending	Ascending
Phase encoding	PA	PA	AP	PA	AP	AP	PA	PA	PA
Eye closed / fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Closed	Fixate	Fixate	Fixate

ATR: Advanced Telecommunications Research Institute International; fMRI: functional magnetic resonance imaging; SRPBS: Strategic Research Program for Brain Sciences; TR: repetition time; TE: echo time.

2.1.2 Preprocessing and calculation of the resting-state functional connectivity matrix

The rs-fMRI data were preprocessed using SPM8 implemented in MATLAB. The first 10 s of data were discarded to allow for T1 equilibration. Preprocessing steps included slice-timing correction, realignment, co-registration, segmentation of T1-weighted structural images, normalization to Montreal Neurological Institute (MNI) space, and spatial smoothing with an isotropic Gaussian kernel of 6 mm full-width at half-maximum. For the analysis of connectivity matrices, region of interests (ROIs) were delineated according to a 268-node gray matter atlas developed to cluster maximally similar voxels (Shen et al., 2013). The blood-oxygen-level dependent (BOLD) signal time courses were extracted from these 268 ROIs. To remove several sources of spurious variance, we used linear regression with 36 regression parameters (Ciric et al., 2017) such as six motion parameters, average signals over the whole brain, white matter, and cerebrospinal fluid. Derivatives and quadratic terms were also included for all parameters. A temporal band-pass filter was applied to the time series using a first-order Butterworth filter with a pass band between 0.01 Hz and 0.08 Hz to restrict the analysis to low-frequency fluctuations, which are characteristic of rs-fMRI BOLD activity (Ciric et al., 2017). Furthermore, to reduce spurious changes in FC because of head motion, we calculated frame-wise displacement (FD) and removed volumes with FD > 0.5 mm, as proposed in a previous study (Power et al., 2014). The FD represents head motion between two consecutive volumes as a scalar quantity (i.e., the summation of absolute displacements in translation and rotation). Using the aforementioned threshold, 5.4% ± 10.6% volumes (i.e., the average [approximately 13 volumes] ± 1 SD) were removed per 10 min of rs-fMRI scanning (240 volumes) in the traveling-subject dataset, 6.2% ± 13.2% volumes were removed per rs-fMRI session in the SRPBS multi-disorder dataset. If the number of volumes removed after scrubbing exceeded the average of –3 SD across participants in each dataset, the participants or sessions were excluded from the analysis. As a result, 14 sessions were removed from the traveling-subject dataset, 20 participants were removed from the SRPBS multi-disorder dataset. Furthermore, we excluded participants for whom we could not calculate FC at all 35,778 connections, primarily because of the lack of BOLD signals within an ROI. As a result, 99 participants were further removed from the SRPBS multi-disorder dataset.

We computed the ROI-based pairwise correlations as measure of FC. For each participant, the temporal correlations of rs-fMRI BOLD signals between pairs of ROIs were computed after averaging each voxelwise BOLD signal in each ROI. There are some candidates for the measure of FC such as the tangent method and partial correlation (Abraham et al., 2017; Ng et al., 2014); however, we used Pearson's correlation coefficients because they have been the most commonly used values in previous studies. FC was defined based on a functional brain atlas consisting of 268 nodes (regions) covering the whole brain, which has been widely utilized in previous studies (Finn et al., 2015; Noble et al., 2017; Rosenberg et al., 2016; Shen et al., 2013). The Fisher's *z*-transformed Pearson's correlation coefficients between the preprocessed BOLD signal time courses of each possible pair of nodes were calculated and used to construct 268 × 268 symmetrical connectivity matrices in which each element represents a connection strength, or edge, between two nodes. We used 35,778 connectivity values [(268 × 267)/2] of the lower triangular matrix of the connectivity matrix. To briefly investigate any site-effect on FC, we conducted one-way ANOVA with Site (9sites) as a factor to the

functional connections in the SRPBS multi-disorder dataset and recorded the number of significant differences between sites. We set the threshold to $p<0.05$, after Bonferroni correction. As a result, >30% of all connections (11,888/35,778) were significantly different between sites. Next, we briefly investigated the reproducibility of the resting state functional connectivity pattern due to the difference in site and the difference in day. We compared the reproducibility across days using ATR Tim Trio data (different 5 days) and the reproducibility among sites using all other sites data (11 sites). We calculated the reproducibility as a Pearson's correlation among 35,778 connectivity values. As a result, the average of the Pearson's correlation across days was 0.61 and the average of the Pearson's correlation among sites was 0.51. We found significant difference between these two correlation values (paired *t*-test, $df = 8$, $t = 6.72$, $p < 0.0005$). This result indicates that the reproducibility of resting state functional connectivity pattern decreased due to the difference of imaging sites.

2.1.3 Estimation of biases and factors

To quantitatively investigate the site differences in the rs-fcMRI data, we identified measurement biases, sampling biases, and disorder factors. We defined measurement bias for each site as a deviation of the correlation value for each functional connection from its average across all sites. We assumed that the sampling biases of the HCs and patients with psychiatric disorders differed from one another. Therefore, we calculated the sampling biases for each site separately for HCs and patients with each disorder. Disorder factors were defined as deviations from the HC values. Sampling biases were estimated for patients with MDD and SCZ because only these patients were sampled at multiple sites. Disorder factors were estimated for MDD, SCZ, and ASD because patients with OCD were not scanned using the unified protocol.

It is difficult to separate site differences into measurement and sampling biases using the SRPBS multi-disorder dataset alone because these two types of bias covaried across sites. Different samples (participants) were scanned using different parameters (scanners and imaging protocols). In contrast, the traveling-subject dataset included only measurement bias because the participants were fixed. By combining the traveling-subject dataset with the SRPBS multi-disorder dataset, we simultaneously estimated measurement bias and sampling bias as different factors affected by different sites. We utilized a constrained linear regression model to assess the effects of both types of bias and disorder factors on FC, as follows. *In the regression model for the SRPBS multi-disorder dataset,* the connectivity values of each participant in the SRPBS multi-disorder dataset were composed of the sum of the average connectivity values across all participants and all sites at baseline, measurement bias, sampling bias, and disorder factors. The combined effect of participant factors (individual difference) and scan-to-scan variations was regarded as noise. *In the regression model for the traveling-subject dataset,* the connectivity values of each participant for a specific scan in the traveling-subject dataset were composed of the sum of the average connectivity values across all participants and all sites, participant factors, and measurement bias. Scan-to-scan variation was regarded as noise. For each participant, we defined the participant factor as a deviation of connectivity values from the average across all participants. We estimated all biases and factors by simultaneously fitting the aforementioned two regression models to the FC values of the two different datasets. For this regression analysis, we used data from participants scanned using a unified imaging protocol in the SRPBS multi-disorder dataset and from all participants in the traveling-subject dataset. In summary, each bias or

each factor was estimated as a vector that included a dimension reflecting the number of connectivity values (35,778). Vectors included in our further analyses are those for measurement bias at 12 sites, sampling bias of HCs at six sites, sampling bias for patients with MDD at three sites, sampling bias for patients with SCZ at three sites, participant factors of nine traveling-subjects, and disorder factors for MDD, SCZ, and ASD. Although it seems difficult to separately estimate an effect of ASD and measurement bias because the data with ASD patients are acquired from one site, we could separately estimate by combining traveling subject dataset. For each connectivity, the regression model can be written as follows:

$$Connectivity = \mathbf{x}_{\boldsymbol{m}}{}^{\mathrm{T}}\boldsymbol{m} + \mathbf{x}_{\boldsymbol{s_{hc}}}{}^{\mathrm{T}}\boldsymbol{s_{hc}} + \mathbf{x}_{\boldsymbol{s_{mdd}}}{}^{\mathrm{T}}\boldsymbol{s_{mdd}} + \mathbf{x}_{\boldsymbol{s_{scz}}}{}^{\mathrm{T}}\boldsymbol{s_{scz}} + \mathbf{x}_{\boldsymbol{d}}{}^{\mathrm{T}}\boldsymbol{d} + \mathbf{x}_{\boldsymbol{p}}{}^{\mathrm{T}}\boldsymbol{p} + const + e,$$

$$\text{such that } \sum_j^9 p_j = 0, \sum_k^{12} m_k = 0, \sum_k^6 s_{hc_k} = 0, \sum_k^3 s_{mdd_k} = 0, \sum_k^3 s_{scz_k} = 0,\ d_1(\mathrm{HC}) = 0,$$

in which $\boldsymbol{m}$ represents the measurement bias $(12 \text{ sites} \times 1)$, $\boldsymbol{s_{hc}}$ represents the sampling bias of HCs $(6 \text{ sites} \times 1)$, $\boldsymbol{s_{mdd}}$ represents the sampling bias of patients with MDD $(3 \text{ sites} \times 1)$, $\boldsymbol{s_{scz}}$ represents the sampling bias of patients with SCZ $(3 \text{ sites} \times 1)$, $\boldsymbol{d}$ represents the disorder factor (3×1), $\boldsymbol{p}$ represents the participant factor $(\text{nine traveling subjects} \times 1)$, $const$ represents the average FC value across all participants from all sites, and $e \sim \mathcal{N}(0, \gamma^{-1})$ represents noise. $\mathbf{x}_{\boldsymbol{m}}, \mathbf{x}_{\boldsymbol{s_{hc}}}, \mathbf{x}_{\boldsymbol{s_{mdd}}}, \mathbf{x}_{\boldsymbol{s_{scz}}}, \mathbf{x}_{\boldsymbol{d}}, \mathbf{x}_{\boldsymbol{p}}$ are vectors represented by 1-of-K binary coding in which the target vector (e.g., $\mathbf{x}_{\boldsymbol{m}}$) for a measurement bias $\boldsymbol{m}$ belonging to site k is a binary vector with all elements equal to zero—except for element k, which equals 1. If a participant does not belong to any class, the target vector is a vector with all elements equal to zero. A superscript T denotes the transposition of a matrix or vector, such that $\mathbf{x}^{\mathrm{T}}$ represents a row vector. To eliminate the uncertainty of the constant term, we estimated measurement bias and each sampling bias by imposing constraints so that their average across sites would be 0. For each FC value, we estimated the respective parameters using regular ordinary least squares regression with L2 regularization, as the design matrix of the regression model is rank-deficient. When regularization was not applied, we observed spurious anticorrelation between the measurement bias and the sampling bias for HCs, and spurious correlation between the sampling bias for HCs and the sampling bias for patients with psychiatric disorders (Appendix Figure A.3a, left). These spurious correlations were also observed in the permutation data in which there were no associations between the site label and data (Appendix Figure A.3a, right). This finding suggests that the spurious correlations were caused by the rank-deficient property of the design matrix. We tuned the hyper-parameter lambda to minimize the absolute mean of these spurious correlations (Appendix Figure A.3c, left).

2.1.4 Quantification of the site differences

To quantitatively evaluate the magnitude of the effect of measurement and sampling biases on FC, we compared the magnitudes of both types of bias ($\boldsymbol{m}, \boldsymbol{s_{hc}}, \boldsymbol{s_{mdd}}$, **and** $\boldsymbol{s_{scz}}$) with the magnitudes of psychiatric disorders ($\boldsymbol{d}$) and participant factors ($\boldsymbol{p}$). For this purpose, we investigated the magnitude distribution of both biases, as well as the effects of psychiatric disorders and participant factors on FC overall 35,778 elements in a 35,778-dimensional vector to see how many functional connectivities were largely affected (Appendix Figures A.1ab). To quantitatively summarize the magnitude of the effect of each factor, we calculated the first, second, and third statistical moments of each

distribution. Furthermore, we calculated and compared the contribution size to determine the extent to which each bias type and factor explain the variance of the data in our linear model. After fitting the model, the *b*-th connectivity from subject *a* can be written, as follows:

$$Connectivity_{a,b} = \mathbf{x}_{\boldsymbol{m}}^{a\,\mathrm{T}}\boldsymbol{m^b} + \mathbf{x}_{\boldsymbol{s_{hc}}}^{a\,\mathrm{T}}\boldsymbol{s}_{hc}^{b} + \mathbf{x}_{\boldsymbol{s_{mdd}}}^{a\,\mathrm{T}}\boldsymbol{s}_{mdd}^{b} + \mathbf{x}_{\boldsymbol{s_{scz}}}^{a\,\mathrm{T}}\boldsymbol{s}_{scz}^{b} + \mathbf{x}_{\boldsymbol{d}}^{a\,\mathrm{T}}\boldsymbol{d^b} + \mathbf{x}_{\boldsymbol{p}}^{a\,\mathrm{T}}\boldsymbol{p^b} + const + e,$$

For example, the contribution size of measurement bias (i.e., the first term) in this model was calculated as

$$Contribution\ size_m = \frac{1}{N_m}\frac{1}{N_s * N}\sum_{a=1}^{N_s}\sum_{b=1}^{N}\frac{\left(\mathbf{x}_{\boldsymbol{m}}^{a\,\mathrm{T}}\boldsymbol{m^b}\right)^2}{\left(\mathbf{x}_{\boldsymbol{m}}^{a\,\mathrm{T}}\boldsymbol{m^b}\right)^2 + \left(\mathbf{x}_{\boldsymbol{s_{hc}}}^{a\,\mathrm{T}}\boldsymbol{s}_{hc}^{b}\right)^2 + \left(\mathbf{x}_{\boldsymbol{s_{mdd}}}^{a\,\mathrm{T}}\boldsymbol{s}_{mdd}^{b}\right)^2 + \left(\mathbf{x}_{\boldsymbol{s_{scz}}}^{a\,\mathrm{T}}\boldsymbol{s}_{scz}^{b}\right)^2 + \left(\mathbf{x}_{\boldsymbol{d}}^{a\,\mathrm{T}}\boldsymbol{d^b}\right)^2 + \left(\mathbf{x}_{\boldsymbol{p}}^{a\,\mathrm{T}}\boldsymbol{p^b}\right)^2 + \boldsymbol{e}^2},$$

in which N_m represents the number of components for each factor, N represents the number of connectivities, N_s represents the number of subjects, and $Contribution\ size_m$ represents the magnitude of the contribution size of measurement bias. These formulas were used to assess the contribution sizes of individual factors related to measurement bias (e.g., phase-encoding direction, scanner, coil, and fMRI manufacturer: Fig. 2.4b). We decomposed the measurement bias into these factors, after which the relevant parameters were estimated. Other parameters were fixed at the same values as previously estimated.

2.1.5 Spatial characteristics of measurement bias, sampling bias, and each factor in the brain

To evaluate the spatial characteristics of each type of bias and each factor in the brain, we calculated the magnitude of the effect on each ROI. First, we calculated the median absolute value of the effect on each functional connection among sites or participants for each bias and participant factor. We then calculated the absolute value of each connection for each disorder factor. The uppercase bold letters (e.g., $\boldsymbol{M}$) and subscript vectors (e.g., $\boldsymbol{m}_k$) represent the vectors for the number of functional connections:

$$\boldsymbol{M} = \underset{k}{\mathrm{median}}(|\boldsymbol{m}_k|), \boldsymbol{S_{hc}} = \underset{k}{\mathrm{median}}\left(\left|\boldsymbol{s}_{\boldsymbol{hc}_k}\right|\right), \boldsymbol{S_{mdd}} = \underset{k}{\mathrm{median}}\left(\left|\boldsymbol{s}_{\boldsymbol{mdd}_k}\right|\right),$$

$$\boldsymbol{S_{scz}} = \underset{k}{\mathrm{median}}\left(\left|\boldsymbol{s}_{\boldsymbol{scz}_k}\right|\right), \boldsymbol{D}_2 = |\boldsymbol{d}_2|, \boldsymbol{D}_3 = |\boldsymbol{d}_3|, \boldsymbol{P} = \underset{j}{\mathrm{median}}\left(\left|\boldsymbol{p}_j\right|\right).$$

We next calculated the magnitude of the effect on ROIs as the average connectivity value between all ROIs, except for themselves.

$$Effect\_on\_ROI_n = \frac{1}{N_{ROI} - 1}\sum_{v \neq n}^{N_{ROI}} Effect\_on\_connectivity_{n,v},$$

in which N_{ROI} represents the number of ROIs, $Effect\_on\_ROI_n$ represents the magnitude of the effect on the *n*-th ROI, and $Effect\_on\_connectivity_{n,v}$ represents the magnitude of the effect on connectivity between the *n*-th ROI and *v*-th ROI.

2.1.6 Hierarchical clustering analysis for measurement bias

We next investigated the characteristics of measurement bias. We first examined whether

similarities among the estimated measurement bias vectors for the 12 included sites reflect certain properties of MRI scanners such as phase-encoding direction, MRI manufacturer, coil type, and scanner type. We used hierarchical clustering analysis to discover clusters of similar patterns for measurement bias. This method has previously been used to distinguish subtypes of MDD, based on rs-fcMRI data (Drysdale et al., 2017). We calculated the Pearson's correlation coefficients among measurement biases $\boldsymbol{m}_k$ ($N \times 1$, where N is the number of functional connections) for each site k, and performed a hierarchical clustering analysis based on the correlation coefficients across measurement biases. To visualize the dendrogram, we used the "*dendrogram*", "*linkage*", and "*optimalleaforder*" functions in MATLAB (R2016b, Mathworks, USA).

2.1.7 Comparison of models for sampling bias

We investigated two alternative models for the mechanisms underlying sampling bias. We first theorized how the number of participants at each site affects the variance of sampling biases across connectivity values, as follows:

In the *single-population model*, we assumed that the FC values of each participant were generated from an independent Gaussian distribution, with a mean of 0 and a variance of ξ^2 for each connectivity value. Then, the FC vector for participant j at site k can be described as

$$\boldsymbol{c}_j^k \sim \mathcal{N}(\mathbf{0}, \xi^2 \mathbf{I}).$$

Let $\boldsymbol{c}_k$ be the vector of FC at site k averaged across participants. In this model, $\boldsymbol{c}_k$ represents the sampling bias and can be described as

$$\boldsymbol{c}_k = \frac{1}{N_k} \sum_{j=1}^{N_k} \boldsymbol{c}_j^k \sim \mathcal{N}\left(\mathbf{0}, \frac{\xi^2}{N_k} \mathbf{I}\right),$$

in which N_k represents the number of participants at site k. The variance across FC values for $\boldsymbol{c}_k$ is described as

$$V_k = \frac{1}{N} \sum_{i=1}^{N} (c_{ki} - \overline{\boldsymbol{c}_k})^2 = \frac{1}{N} \boldsymbol{c}_k{}^{\mathrm{T}} \left(\mathbf{I} - \frac{1}{N} \mathbf{1}\mathbf{1}'\right)^{\mathrm{T}} \left(\mathbf{I} - \frac{1}{N} \mathbf{1}\mathbf{1}'\right) \boldsymbol{c}_k \approx \frac{1}{N} \boldsymbol{c}_k{}^{\mathrm{T}} \boldsymbol{c}_k,$$

in which $\mathbf{1}$ represents the $N \times 1$ vector of ones and $\mathbf{I}$ represents the $N \times N$ identity matrix. Since N equals 35,778 and $\frac{1}{35778}$ is sufficiently smaller than 1, we can approximate

$$\mathbf{I} - \frac{1}{N} \mathbf{1}\mathbf{1}' \approx \mathbf{I}.$$

Then, the expected value of variance across FC values for sampling-bias can be described as

$$\mathbb{E}[V_k] \approx \frac{1}{N} \mathbb{E}[\boldsymbol{c}_k{}^{\mathrm{T}} \boldsymbol{c}_k] = \frac{1}{N} Tr\left(\frac{\xi^2}{N_k} \mathbf{I}\right) = \frac{\xi^2}{N_k}.$$

In the *different-population model*, we assumed that the FC values of each participant were generated from a different independent Gaussian distribution, with an average of $\boldsymbol{\beta_k}$ and a variance of ξ^2 depending on the population of each site. In this situation, the FC vector for participant *j* at site *k* can be described as

$$\boldsymbol{c}_j^k \sim \mathcal{N}(\boldsymbol{\beta}_k, \xi^2 \mathbf{I}).$$

Here, we assume that the average of the population $\boldsymbol{\beta_k}$ is sampled from an independent Gaussian distribution with an average of 0 and a variance of σ^2. That is, $\boldsymbol{\beta}_k$ is expressed as

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

The vector of FC for site *k* averaged across participants can then be described as

$$\boldsymbol{c}_k \sim \mathcal{N}\left(\mathbf{0}, \left(\frac{\xi^2}{N_k} + \sigma^2\right)\mathbf{I}\right).$$

The variance across FC values for $\boldsymbol{c}_k$ can be described as

$$\mathbb{E}[V_k] \approx \frac{\xi^2}{N_k} + \sigma^2.$$

In summary, the variance of sampling bias across FC values in each model is expressed by the number of participants at a given site, as follows:

$$\text{single-population model: } y_k = -x_k + 2\log_{10}\xi$$

$$\text{different-population model: } y_k = -\log_{10}(\xi^2 10^{-x_k} + \sigma^2),$$

in which $y_k = \log_{10}(v_k)$, v_k represents the variance across FC values for $\boldsymbol{s_{hc}}_k$, $\boldsymbol{s_{hc}}_k$ represents the sampling bias of HCs at site k ($N \times 1$: N is the number of FC), $x_k = \log_{10}(N_k)$, and N_k represents the number of participants at site *k*. We estimated the parameters ξ and σ using the MATLAB (R2016b, Mathworks, USA) optimization function "*fminunc*". To simplify statistical analyses, sampling bias was estimated based on FC in which the average across all participants was set to zero.

We aimed to determine which model provided the best explanation of sampling bias in our data by calculating the corrected Akaike information criterion (AICc; under the assumption of a Gaussian distribution) for small-sample data (Burnham and Anderson, 2003; Cortese et al., 2016), as well as BIC:

$$\text{AICc} = \sum_{k=1}^{6} \ln {\varphi_k}^2 + 2q + \frac{2q(q+1)}{(6-q-1)},$$

$$\text{BIC} = \sum_{k=1}^{6} \ln {\varphi_k}^2 + q * \log(6),$$

in which $\varphi_k = v_k - \widehat{v_k}$, $\widehat{v_k}$ represents the estimated variance, and q represents the

number of parameters in each model (1 or 2).

To investigate prediction performance, we used leave-one-site-out-cross-validation in which we estimated the parameters ξ and σ using data from five sites. The variance of sampling bias was predicted based on the number of participants at the remaining site. This procedure was repeated to predict variance values for sampling bias at all six sites. We then calculated the absolute errors between predicted and actual variances for all sites.

2.1.8 Traveling-subject harmonization procedures

We next developed a novel harmonization method that enabled us to subtract only the measurement bias using the traveling-subject dataset. Using a constrained linear model, we estimated the measurement bias separately from sampling bias (see 2.1.3 "Estimation of biases and factors"). Thus, measurement bias was removed by subtracting the estimated measurement biases. The harmonized FC values were set, as follows:

$$Connectivity^{Traveling-subject} = Connectivity - \mathbf{x_m}^{\mathrm{T}}\widehat{\boldsymbol{m}},$$

in which $\widehat{\boldsymbol{m}}$ represents the estimated measurement bias.

2.1.9 Principal component analysis

To visualize the site differences and disorder effects in the SRPBS multi-disorder dataset while maintaining its quantitative properties, we first visualized the site differences and disorder effects in the SRPBS multi-disorder rs-fcMRI dataset while maintaining its quantitative properties by using a principal component analysis (PCA)—an unsupervised dimension reduction method. To visualize whether most of the variation in the SRPBS multi-disorder dataset was still associated with imaging site after harmonization, we performed a PCA of FC values in the harmonized SRPBS multi-disorder dataset. We used the traveling-subject method for harmonization, as described in the following section. Finally, to visualize the measurement bias in the SRPBS multi-disorder dataset, we performed a PCA of FC values in the SRPBS multi-disorder data after subtracting only the sampling bias.

2.1.10 Two-fold cross-validation evaluation procedure

Existing harmonization methods estimate the site difference without separating site difference into the measurement bias and the sampling bias and subtract the site difference from data. Therefore, existing harmonization methods might have pitfall to eliminate not only biologically meaningless measurement bias but also eliminate biologically meaningful sampling bias. Here, we tested whether the traveling-subject harmonization method indeed removes only the measurement bias and whether the existing harmonization methods simultaneously remove the measurement and sampling biases. There are three commonly used harmonization methods: (1) a general linear model (GLM) harmonization method, site difference was estimated without adjusting for biological covariates (e.g., diagnosis) (Drysdale et al., 2017; Fortin et al., 2018; Rao et al., 2017). The GLM harmonization method adjusts the FC value for site difference using GLM. Site differences were estimated by fitting the regression model, which included site label only, to the SRPBS multi-disorder dataset only. The regression model can be written as

$$Connectivity = const + \mathbf{x}_{site}{}^{\mathrm{T}}\boldsymbol{site}^{GLM} + e, \quad (\mathbf{1})$$

in which $\boldsymbol{site}^{GLM}$ represents the site difference $(9 \text{ sites} \times 1)$. For each FC value, we estimated the parameters using regular ordinary least squares regression. Site differences were removed by subtracting the estimated site differences. Thus, the harmonized FC values were set, as follows:

$$Connectivity^{GLM} = Connectivity - \mathbf{x}_{site}{}^{\mathrm{T}}\widehat{\boldsymbol{site}^{GLM}},$$

in which $\widehat{\boldsymbol{site}^{GLM}}$ represents the estimated site difference.

(2) an adjusted GLM method, site difference was estimated while adjusting for biological covariates (Fortin et al., 2018; Rao et al., 2017). Site differences were estimated by fitting the regression model, which included site label and diagnosis label, to the SRPBS multi-disorder dataset. The regression model can be written as

$$Connectivity = const + \mathbf{x}_{site}{}^{\mathrm{T}}\boldsymbol{site}^{\boldsymbol{Adj}} + \mathbf{x}_{\boldsymbol{d}}{}^{\mathrm{T}}\boldsymbol{d}^{\boldsymbol{Adj}} + e, \quad (\mathbf{2})$$

In which $\boldsymbol{site}^{Adj}$ represents the site difference $(9 \text{ sites} \times 1)$. For each FC value, we estimated the parameters via regular ordinary least squares regression. Site differences were removed by subtracting the estimated site difference only. Thus, the harmonized FC values were set, as follows:

$$Connectivity^{Adj} = Connectivity - \mathbf{x}_{site}{}^{\mathrm{T}}\widehat{\boldsymbol{site}^{Adj}},$$

in which $\widehat{\boldsymbol{site}^{Adj}}$ represents the estimated site difference.

(3) a ComBat method (Fortin et al., 2018; Fortin et al., 2017; Johnson et al., 2007; Yu et al., 2018), a batch-effect correction tool commonly used in genomics, site difference was modeled and removed. The ComBat harmonization model extends the adjusted GLM harmonization method in two ways: (i) it models site-specific scaling factors and (ii) it uses empirical Bayesian criteria to improve the estimation of site parameters for small sample sizes. The model assumes that the expected connectivity value can be modeled as a linear combination of the biological variables and the site differences in which the error term is modulated by additional site-specific scaling factors.

$$Connectivity = const + \mathbf{x}_{site}{}^{\mathrm{T}}\boldsymbol{site}^{\boldsymbol{ComBat}} + \mathbf{x}_{\boldsymbol{d}}{}^{\mathrm{T}}\boldsymbol{d}^{\boldsymbol{ComBat}} + \delta_k e, \quad (\mathbf{3})$$

in which $\boldsymbol{site}^{ComBat}$ represents the site difference (9 sites × 1), and δ_k represents the scale parameter for site differences at site k for the respective connectivity value. The harmonized FC values were set, as follows:

$$Connectivity^{ComBat} = \frac{Connectivity - const - \mathbf{x}_{site}{}^{\mathrm{T}}\widehat{\boldsymbol{site}^{ComBat}} - \mathbf{x}_{\boldsymbol{d}}{}^{\mathrm{T}}\widehat{\boldsymbol{d}^{ComBat}}}{\widehat{\delta_k}} + const + \mathbf{x}_{\boldsymbol{d}}{}^{\mathrm{T}}\widehat{\boldsymbol{d}^{ComBat}},$$

in which $\widehat{\delta_k}$, $\widehat{\boldsymbol{d}^{ComBat}}$, and $\widehat{\boldsymbol{site}^{ComBat}}$ are the empirical Bayes estimates of δ_k, $\boldsymbol{d}^{ComBat}$, and $\boldsymbol{site}^{ComBat}$, respectively using "combat" function in https://github.com/Jfortin1/ComBatHarmonization. Thus, ComBat simultaneously models and estimates biological and nonbiological terms and algebraically removes the estimated additive and multiplicative site differences. Of note, in the ComBat model, we included diagnosis as covariates to preserve important biological trends in the data and avoid overcorrection.

For evaluation, we performed 2-fold cross-validation evaluations in which the SRPBS multi-disorder dataset was partitioned into two equal-size subsamples (fold1 data and fold2 data) with the same proportions of sites. Between these two subsamples, the measurement bias is common, but the sampling bias is different (because the scanners are common, and participants are different). We estimated the measurement bias (or site difference including the measurement bias and the sampling bias for the existing methods) by applying the harmonization methods to the fold1 data and subtracted the measurement bias or site difference from the fold2 data. We then estimated the measurement bias in the fold2 data. For the existing harmonization methods, if the site difference estimated by using fold1 contains only the measurement bias, the measurement bias estimated in fold2 data after subtracting the site difference should be smaller than that of without subtracting the site difference (Raw). To separately estimate measurement bias and sampling bias in both subsamples while avoiding information leak, we also divided the traveling-subject dataset into two equal-size subsamples with the same proportions of sites and subjects. We concatenated one subsample of traveling-subject dataset to fold1 data to estimate the measurement bias for traveling-subject method (estimating dataset) and concatenated the other subsample of traveling-subject dataset to fold2 data for testing (testing dataset). That is, in the traveling-subject harmonization method, we estimated the measurement bias using the estimating dataset and removed the measurement bias from the testing dataset. By contrast, in the other harmonization methods, we estimated the site difference using the fold1 data (not including the subsample of traveling-subject dataset) and removed the site difference from the testing dataset. We then estimated the measurement bias using the testing dataset and evaluated the standard deviation of the magnitude distribution of measurement bias calculated in the same way as described in "2.1.4 Analysis of contribution size" section. To verify whether important information such as participant factors and disorder factors are kept in the testing dataset, we also estimated the disorder factors and participant factors and calculated the ratio of the standard deviation of measurement bias to the standard deviation of participant factor and disorder factor as signal to noise ratios. This procedure was done again by exchanging the estimating dataset and the testing dataset. In the traveling-subject harmonization method, we estimated the measurement bias by applying the regression model. Thus, the harmonized FC values in testing dataset were set, as follows:

$$connectivity_{testing\ \ dataset}^{Traveling-subject} = Connectivity_{testing\ \ dataset} - \mathbf{x_m}^{\mathrm{T}}\widehat{\boldsymbol{m}}_{estimating\ \ dataset},$$

in which $\widehat{\boldsymbol{m}}_{estimating\ \ dataset}$ represents the estimated measurement bias using the estimating dataset.

By contrast, in the other harmonization methods, we estimated the site differences by applying the regression models written in equations (1)–(3) to the estimating dataset (fold1 data). Thus, the harmonized FC values in testing dataset were set, as follows:

$$connectivity_{testing\ \ dataset}^{GLM} = Connectivity_{testing\ \ dataset} - \mathbf{x_{site}}^{\mathrm{T}}\widehat{\boldsymbol{site}^{GLM}}_{fold1},$$

$$connectivity_{testing\ \ dataset}^{Adj} = Connectivity_{testing\ \ dataset} - \mathbf{x_{site}}^{\mathrm{T}}\widehat{\boldsymbol{site}^{Adj}}_{fold1},$$

$$connectivity_{testing\ \ dataset}^{ComBat} = Connectivity_{testing\ \ dataset} - \mathbf{x_{site}}^{\mathrm{T}}\widehat{\boldsymbol{site}^{ComBat}}_{fold1},$$

in which $\widehat{site^{GLM}}_{fold1}, \widehat{site^{Adj}}_{fold1}, \widehat{site^{ComBat}}_{fold1}$ represents the estimated site differences using fold1 data.

We then estimated the measurement bias, participant factor, and disorder factors by applying the regression model written in equation (1) to the harmonized FC values in the testing dataset. Finally, we evaluated the standard deviation of the magnitude distribution of measurement bias calculated in the same way as described in "Quantification of site differences" section among the harmonization methods. This procedure was done again by exchanging the estimating dataset and the testing dataset.

2.2 Results

2.2.1 Quantification of site differences

To quantitatively evaluate the magnitude of the effect of measurement and sampling biases on FC, we investigated the magnitude distribution of both biases, as well as the effects of psychiatric disorders and participant factors on FC and calculated the first, second, and third statistical moments of each distribution. Based on the mean values and cube roots of the third moments (Fig. 2.2a), all distributions could be approximated as bilaterally symmetric with a mean of zero. Thus, distributions with larger squared roots of the second moments (standard deviations) affect more connections with larger effect sizes (Fig. 2.2b). The value of the standard deviation was largest for the participant factor (0.0662), followed by these values for the measurement bias (0.0411), the SCZ factor (0.0377), the MDD factor (0.0328), the ASD factor (0.0297), the sampling bias for HCs (0.0267), sampling bias for patients with SCZ (0.0217), and sampling bias for patients with MDD (0.0214). To compare the sizes of the standard deviation of the magnitude distribution between participant factors and measurement bias, we evaluated the variance of each distribution. All pairs of variances were analyzed using Ansari–Bradley tests. Our findings indicated that the variances of magnitude distributions in 10 of 12 measurement biases were significantly larger than in the MDD factor; the variances of magnitude distributions in seven of 12 measurement biases were significantly larger than in the SCZ factor; and the variances of all magnitude distributions in measurement biases were significantly larger than the variance of the MDD factor (Appendix Table A.6). The largest variance of magnitude distribution in the sampling bias was significantly larger than in the MDD factor (Appendix Table A.7). Variances of magnitude distributions in all participant factors were significantly larger than that in all measurement biases (nine participant factors × 12 measurement biases = 108 pairs; W^*: mean = –59.80, max = –116.81, min = –3.69; p value after Bonferroni correction: max = 0.011, min = 0, n = 35,778). The standard deviation of the magnitude distribution in the participant factor was approximately twice that in the SCZ, MDD, and ASD factors. Furthermore, we plotted fractions of the data variance determined using the aforementioned factors (contribution size) in our linear model (Fig. 2.2d). The results were consistent with the analysis of the standard deviation (Fig. 2.2c, middle).

These results indicated that the effect size of the measurement bias on FC is smaller than that of the participant factor but is mostly larger than the disorder factors, which suggested that measurement bias represents a serious limitation in research regarding psychiatric disorders. Furthermore, the effect sizes of the sampling biases were comparable with those of the disorder factors. This finding indicates that sampling bias

also represents a major limitation in psychiatric research. In addition, the effect size of the participant factor was much greater than that among patients with SCZ, MDD, or ASD. Such relationships make the development of rs-fcMRI-based classifiers of psychiatric or developmental disorders very challenging. If the disorder factor and site factor are confounded in functional connections, to develop robust and generalizable classifiers across multiple sites, we have to select disorder-specific and site-independent abnormal functional connections (Takagi et al., 2017; Watanabe et al., 2017; Yahata et al., 2017; Yahata et al., 2016; Yamada et al., 2017).

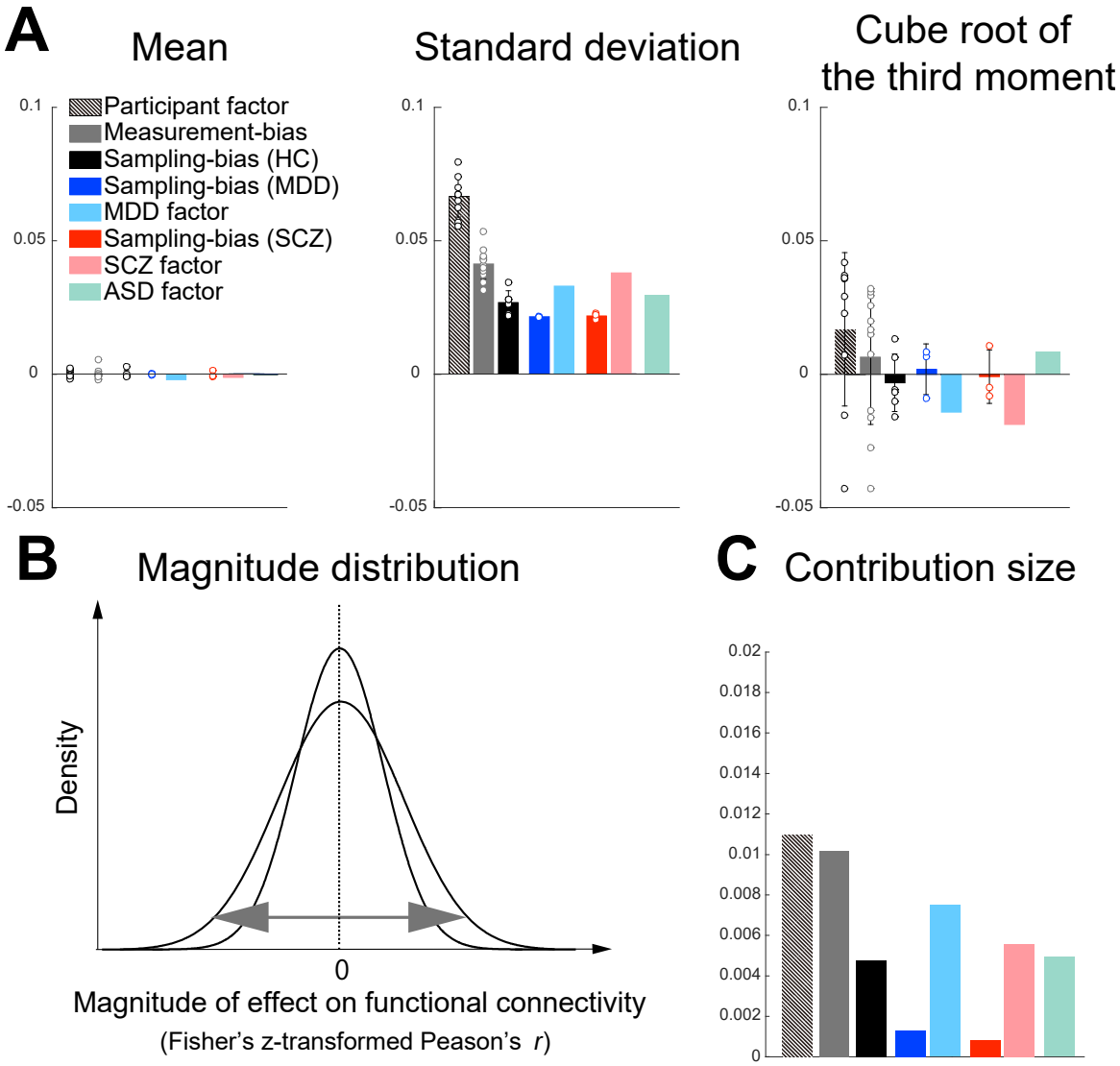

FIGURE 2.2 | Statistics of magnitude distributions for each type of bias and each factor.

(a) The means, standard deviations, and third moments standardized to the same scale on the vertical axis (i.e., cube root) for each type of bias and each factor. Bars represent the average value, while the error bars represent the standard deviation across sites or participants. Each data point represents one participant or one site. (b) Schematic examples illustrating the magnitude distribution. (c) Contribution size of each bias and each factor. HC: healthy controls; SCZ: schizophrenia; MDD: major depressive disorder; ASD: autism spectrum disorder.

2.2.2 Brain regions contributing most to biases and associated factors

To evaluate the spatial characteristics of each type of bias and each factor in the brain, we were able to visualize the relative contribution of individual ROIs to each bias or factor in the whole brain (Fig. 2.3). Consistent with the findings of previous studies, the effect of the participant factor was large for several ROIs in the cerebral cortex, especially in the prefrontal cortex, but small in the cerebellum and visual cortex (Finn et al., 2015). The effect of measurement bias was large in inferior brain regions where functional images are differentially distorted depending on the phase-encoding direction (Jezzard and Balaban, 1995; Weiskopf et al., 2006). Connections involving the medial dorsal nucleus of the thalamus were also heavily affected by both MDD, SCZ and ASD. Effects of the MDD factor were observed in the dorsomedial prefrontal cortex and the superior temporal gyrus in which abnormalities have also been reported in previous studies (Drysdale et al., 2017; Kaiser et al., 2015; Mulders et al., 2015). Effects of the SCZ factor were observed in the left inferior parietal lobule, bilateral anterior cingulate cortices, and left middle frontal gyrus in which abnormalities have been reported in previous studies (Kuhn and Gallinat, 2013; Li et al., 2017; Minzenberg et al., 2009). Effects of the ASD factor were observed in the putamen, the medial prefrontal cortex, and the right middle temporal gyrus in which abnormalities have also been reported in previous studies (Abraham et al., 2017; Anderson et al., 2011; Yahata et al., 2016). The effect of sampling bias for HCs was large in the inferior parietal lobule and the precuneus, both of which are involved in the default mode network and the middle frontal gyrus. Sampling bias for disorders was large in the medial dorsal nucleus of the thalamus, left dorsolateral prefrontal cortex, dorsomedial prefrontal cortex, and cerebellum for MDD (Drysdale et al., 2017); and in the prefrontal cortex, cuneus, and cerebellum for SCZ (Li et al., 2017).

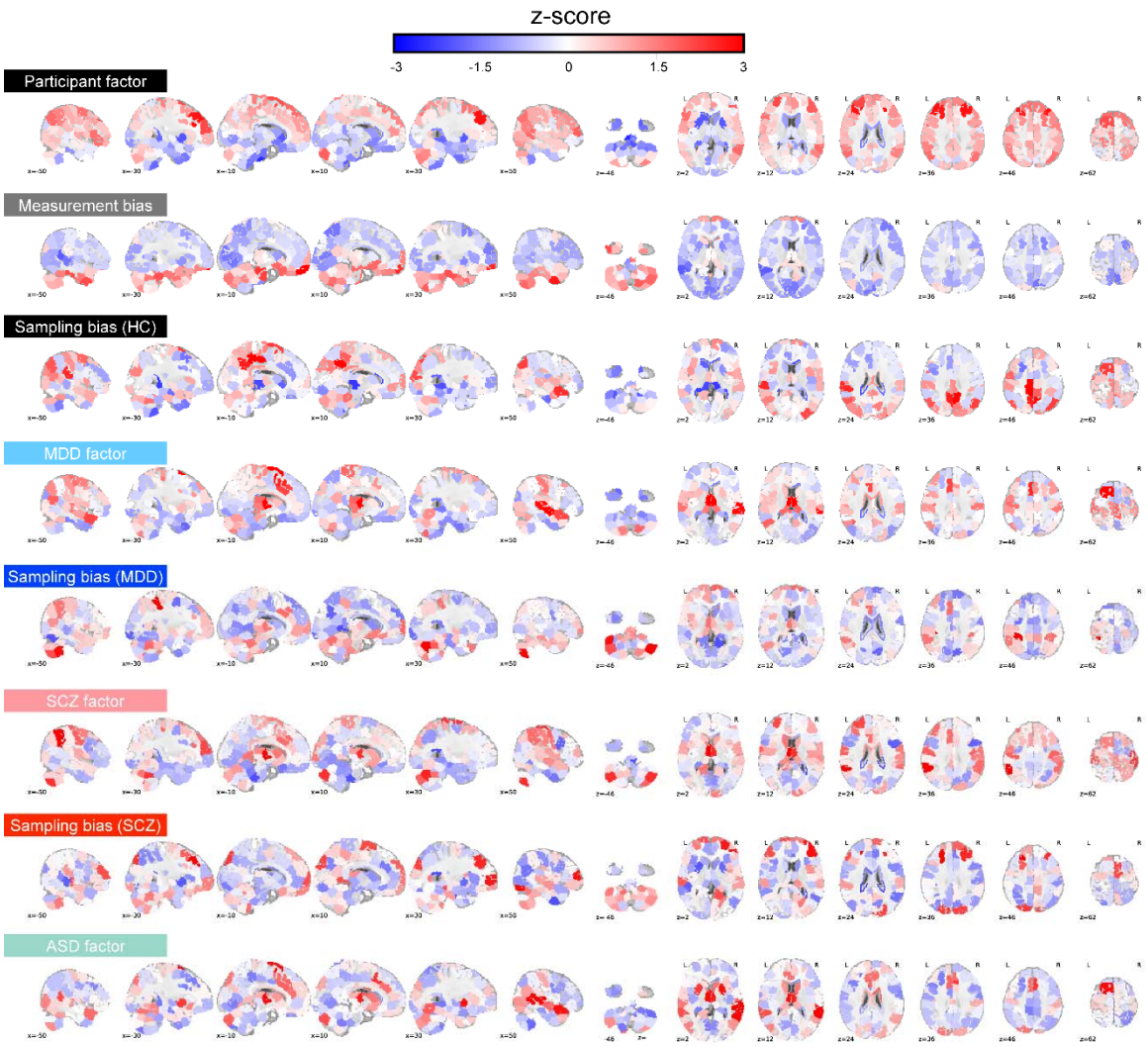

FIGURE 2.3 | Spatial distribution of each type of bias and each factor in various brain regions.

Mean effects of connectivity for all 268 ROIs. For each ROI, the mean effects of all functional connections associated with that ROI were calculated for each bias and each factor. Warmer (red) and cooler (blue) colors correspond to large and small effects, respectively. The magnitudes of the effects are normalized within each bias or each factor (*z*-score). ROI: region of interest; HC: healthy control; SCZ: schizophrenia; MDD: major depressive disorder; ASD: autism spectrum disorder.

2.2.3 Characteristics of measurement bias

We next investigated the characteristics of measurement bias. We first examined whether similarities among the estimated measurement bias vectors for the 12 included sites reflect certain properties of MRI scanners such as phase-encoding direction, MRI manufacturer, coil type, and scanner type. As a result, the measurement biases of the 12 sites were divided into phase-encoding direction clusters at the first level (Fig. 2.4a). They were divided into fMRI manufacturer clusters at the second level, and further divided into coil type clusters, followed by scanner model clusters. Furthermore, we quantitatively verified the magnitude relationship among factors by using the same model to assess the contribution of each factor (Fig. 2.4b; “2.1.4 Quantification of the site

differences"). The contribution size was largest for the phase-encoding direction (0.0391), followed by the contribution sized for fMRI manufacturer (0.0318), coil type (0.0239), and scanner model (0.0152). These findings indicate that the main factor influencing measurement bias is the difference in the phase-encoding direction, followed by fMRI manufacturer, coil type, and scanner model, respectively.

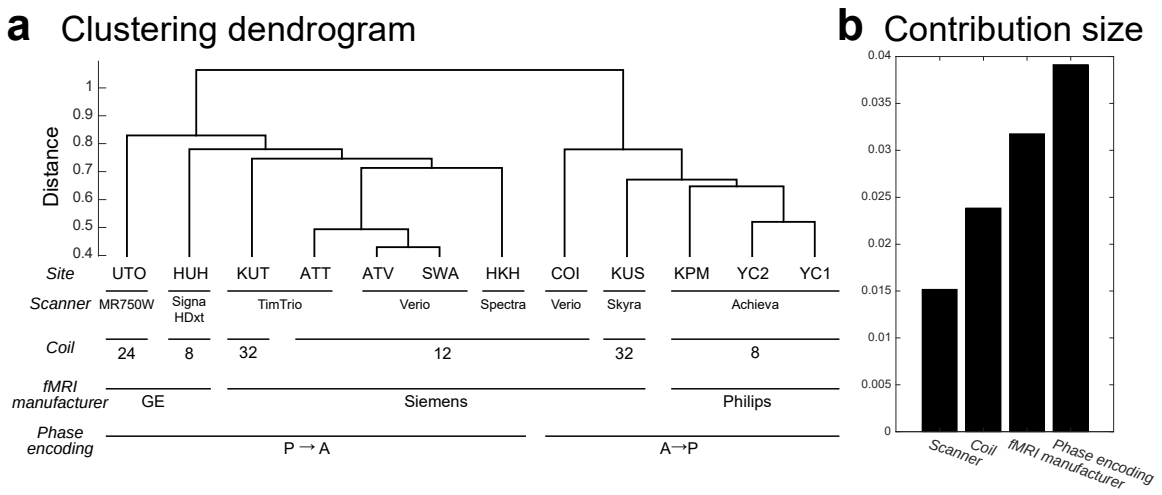

FIGURE 2.4 | Clustering dendrogram for measurement bias.

(a) The height of each linkage in the dendrogram represents the dissimilarity (1 - *r*) between the clusters joined by that link. (b) Contribution size of each factor. UTO: University of Tokyo; HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KUS: Siemens Skyra scanner at Kyoto University; KPM: Kyoto Prefectural University of Medicine; YC1: Yaesu Clinic 1; YC2: Yaesu Clinic 2.

2.2.4 Sampling bias is because of sampling from among a subpopulation

To investigate the characteristics of sampling bias, we investigated two alternative models for the mechanisms underlying sampling bias. Our results indicated that the different-subpopulation model provided a better fit for our data than the single-population model (Fig. 2.5c; different-subpopulation model: AICc = -108.80 and BIC = -113.22; single-population model: AICc = -96.71 and BIC = -97.92). Furthermore, the predictive performance was significantly higher for the different-subpopulation model than for the single-population model (one-tailed Wilcoxon signed-rank test applied to absolute errors: $Z = 1.67$, $p = .0469$, $n = 6$; Figs. 2.5d and 2.5e). This result indicates that sampling bias is not only caused by random sampling from a single grand population, depending on the number of participants among sites, but also by sampling from among different subpopulations. Sampling biases thus represent a major limitation in attempting to estimate a true single distribution of HC or patient data based on measurements obtained from a finite number of sites and participants.

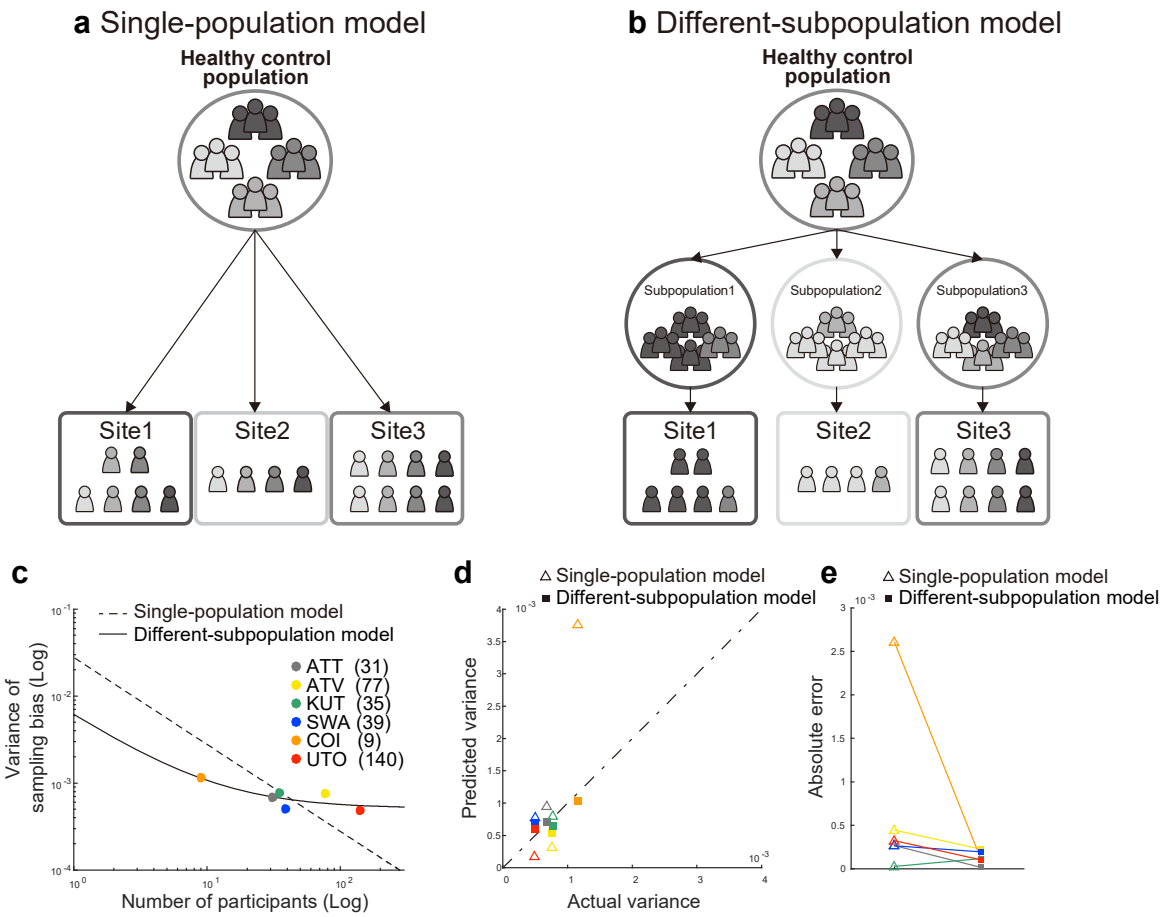

FIGURE 2.5 | Comparison of the two models of sampling bias.

Schematic examples illustrating the single-population (a) and different-subpopulation models (b) and the results of model fitting (c). The *x*-axis represents the number of participants on a logarithmic scale, while the *y*-axis represents the variance of sampling bias on a logarithmic scale. The broken line represents the prediction of the single-population model, while the solid line represents the prediction of the different-subpopulation model. Each data point represents one site. (d) Results of the predictions determined by using each model. The *x*-axis represents the actual variance, while the *y*-axis represents the predicted variance. Open triangles correspond to the single-population model, while filled squares correspond to the different-subpopulation model. (e) Performance of prediction using the two models, based on the absolute error between the actual and predicted variance. UTO: University of Tokyo; COI: Center of Innovation in Hiroshima University; SWA: Showa University; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International.

2.2.5 Visualization of the harmonization effect

To visualize the site differences and disorder effects in the SRPBS multi-disorder dataset while maintaining its quantitative properties, we first performed an unsupervised dimension reduction of the raw rs-fcMRI data using a PCA. All participant data in the SRPBS multi-disorder dataset were plotted on two axes consisting of the first two PCs (Fig. 2.6a, small, light-colored symbols). First two PCs could explain about 6% of the

variance in the whole data (Fig 2.6b, 3.5% for 1st PC and 2.5% for 2nd PC). Dark-colored markers indicated the averages of projected data across healthy controls in each site and the average within each psychiatric disorder in the subspace spanned by the 2 components. For the raw data, there was a clear separation of the Hiroshima University Hospital (HUH) site for PC1, which explained most of the variance in the data. To visualize the effects of the harmonization process, we plotted the data after subtracting only the measurement bias from the SRPBS multi-disorder dataset (Fig. 2.6c). In Fig. 2.6b, the differences among sites represent the sampling bias. Relative to the result of raw data, which reflects the data before harmonization, the HUH site moved much closer to the origin (i.e., grand average) and showed no marked separation from the other sites. This result indicates that the separation of the HUH site observed in Fig. 2.6a was caused by measurement bias, which was removed following harmonization. Furthermore, harmonization was effective in distinguishing patients and HCs scanned at the same site. Since patients with ASD were only scanned at the Showa University (SWA) site, the averages for patients with ASD (▲) and HCs (blue ●) scanned at this site were projected to nearly identical positions (Fig. 2.6a). However, the two symbols are clearly separated from one another in Fig. 2.6c. The effect of a psychiatric disorder (ASD) could not be observed in the first two PCs without harmonization but became detectable following the removal of measurement bias. Finally, to visualize the measurement bias in the SRPBS multi-disorder dataset, we plotted the data after subtracting only the sampling bias from the SRPBS multi-disorder dataset (Fig. 2.6d). Relative to the result of harmonized data, the HUH site showed marked separation from the other sites as same as in raw data (Fig. 2.6a).

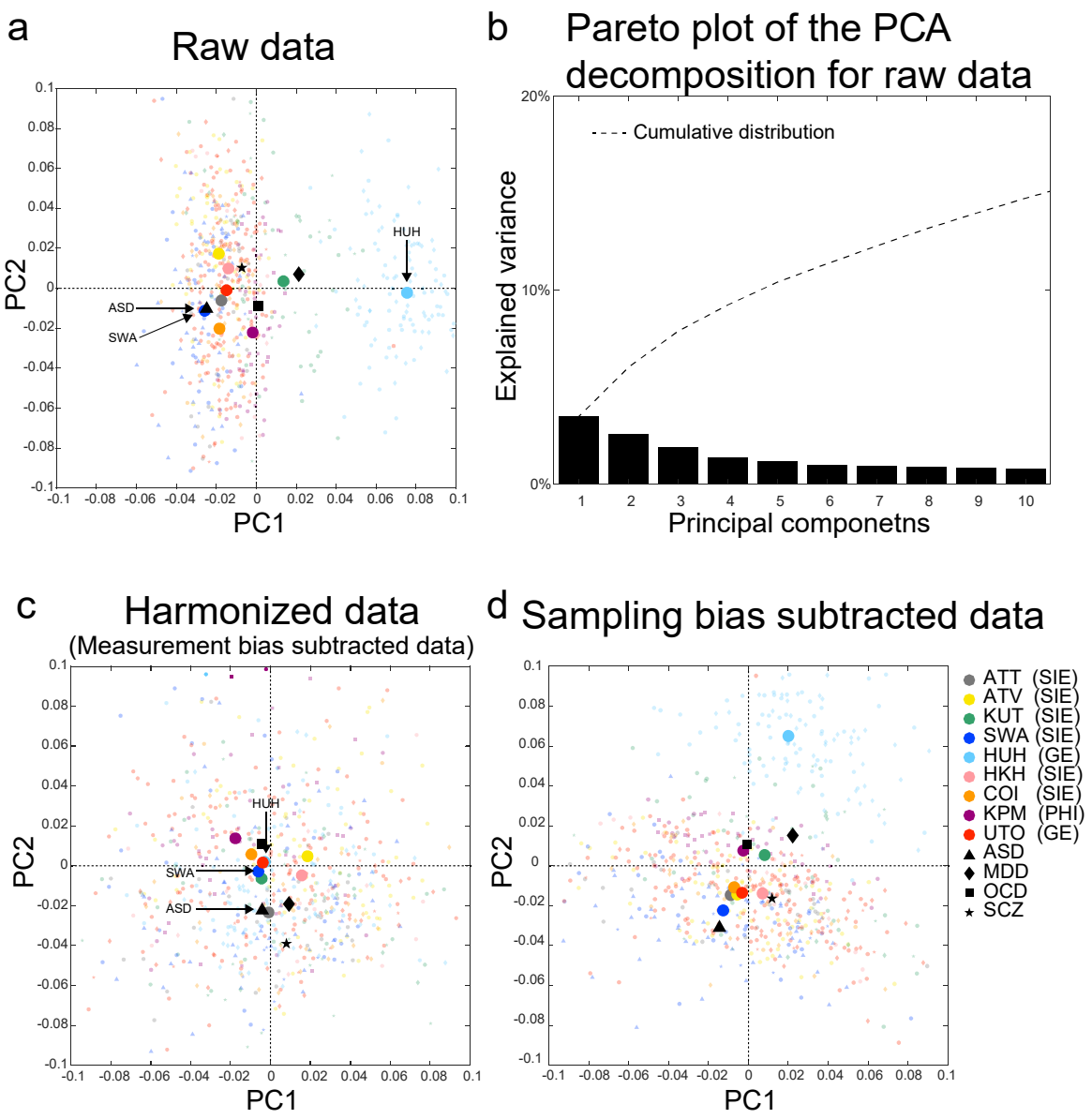

FIGURE 2.6 | PCA dimension reduction in the SRPBS multi-disorder dataset after harmonization.

Comparison among (a) raw data, (c) harmonized data (measurement bias subtracted data), and (d) sampling bias subtracted data. All participants in the SRPBS multi-disorder dataset projected into the first two principal components (PCs), as indicated by small, light-colored markers. Dark-colored markers indicated the averages of the projected data across healthy controls at each site and the average within each psychiatric disorder in the subspace spanned by the 2 components. The color of the marker represents the site, while the shape represents the psychiatric disorder. (b) The pareto plot of the PCA decomposition for raw data. The pareto plot shows how much variance is explained by each principal component. PCA: principal component analysis; SRPBS: Strategic Research Program for Brain Sciences; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; KUT: Siemens TimTrio scanner at Kyoto University; SWA: Showa University; HUH: Hiroshima University Hospital; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima

University; KPM: Kyoto Prefectural University of Medicine; UTO: University of Tokyo; ASD: Autism Spectrum Disorder. MDD: Major Depressive Disorder. OCD: Obsessive Compulsive Disorder. SCZ: Schizophrenia. SIE: Siemens fMRI. GE: GE fMRI. PHI: Philips fMRI.

2.2.6 Quantification of the effect of traveling-subject harmonization

We tested whether the traveling-subject harmonization method indeed removes only the measurement bias and whether the existing harmonization methods simultaneously remove the measurement and sampling biases. Fig. 2.7 shows that the standard deviation of the magnitude distribution of the measurement bias and the ratio of the standard deviation of the magnitude distribution of the measurement bias to that of participant factor and disorder factor in the both fold data for the four harmonization methods and without harmonization (Raw). Our result shows that the reduction of the standard deviation of the magnitude distribution of the measurement bias from the Raw was highest in the traveling-subject method among all methods (29% reduction compared to 3% in the second highest value for ComBat method). Moreover, improvement in the signal to noise ratios were also highest in our method for participant factor (41% improvement compared to 3% in the second highest value for ComBat method) and for disorder factor (39% improvement compared to 3% in the second highest value for ComBat method). These results indicate that the traveling-subject harmonization method indeed removed the measurement bias and improved the signal to noise ratios.

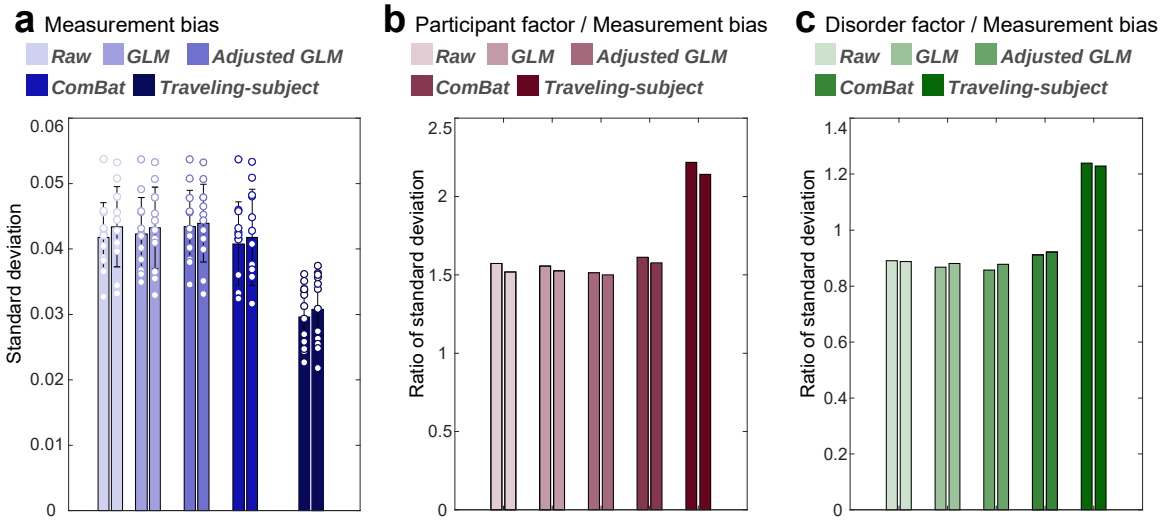

FIGURE 2.7 | Reduction of the measurement bias and improvement of signal to noise ratios for different harmonization methods.

(a) Standard deviation of the magnitude distribution of the measurement bias. The error bars represent the standard deviation across sites. Each data point represents one site. (b) Ratio of standard deviation of the magnitude distribution of the measurement bias to that of the participant factor. (c) Ratio of standard deviation of the magnitude distribution of the measurement bias to that of the disorder factor. Different colored columns show the results from different harmonization method. Two columns of the same color show the results of the two folds. GLM: generalized linear model.

2.3 Discussion

In the present study, by acquiring a separate traveling-subject dataset and the SRPBS multi-disorder dataset, we separately estimated measurement and sampling biases for multiple sites, which we then compared with the magnitude of disorder factors. Furthermore, we investigated the origin of each bias in multi-site datasets. Finally, to overcome the problem of site difference, we developed a novel harmonization method that enabled us to subtract the measurement bias by using a traveling-subject dataset and achieved the reduction of the measurement bias by 29% and the improvement of the signal to noise ratios by 40%.

2.3.1 The effect sizes of measurement and sampling biases

Previous studies have focused on measurement bias and compared its magnitude to the participant factor by using a traveling subject design in a finger tapping task-fMRI (Gountouna et al., 2010) and resting-state fMRI (Noble et al., 2017). These studies revealed the magnitude of measurement bias is smaller than the participant factor. Although such a result was also obtained in this study, the novelty of this study exists in that we separately estimated measurement and sampling biases and then compared them with the magnitude of disorder factors. We assessed the effect sizes of measurement and sampling biases in comparison with the effects of psychiatric disorders on resting-state FC. Our findings indicated that measurement bias exerted significantly greater effects than disorder factors, whereas sampling bias was comparable to (or even larger than) the disorder effects (Fig. 2.2). This result is very important finding to collect resting-state fMRI data from multiple sites and to construct biomarkers of psychiatric disorder based on multi-site data in the clinical field. However, we did not control for variations in disease stage and treatment in our dataset. Although controlling for such heterogeneity may increase the effect size of disorder factors, such control is not feasible when collecting big data from multiple sites. Therefore, it is important to appropriately remove measurement bias from heterogeneous patient data to identify relatively small disorder effects. This issue is essential for investigating the relationships among different psychiatric disorders because disease factors are often confounded by site differences. As previously mentioned, it is common for a single site to sample only a few types of psychiatric disorders (SCZ and ASD from sites A and B, respectively). In this situation, it is critical to dissociate disease factors from site differences. This dissociation can be accomplished by subtracting only the measurement bias, which is estimated using the traveling subject dataset.

2.3.2 Characteristics of measurement bias

Our results indicated that measurement bias is primarily influenced by differences in the phase-encoding direction, followed by differences in fMRI manufacturer, coil type, and scanner model (Fig. 2.4). These results are consistent with our finding of large measurement biases in the inferior brain regions (Fig. 2.3), the functional imaging of which is known to be influenced by the phase-encoding direction (Jezzard and Clare, 1999; Weiskopf et al., 2006). Previous studies have reported that the effect because of the difference in the phase-encoding direction can be corrected using the field map obtained at the time of imaging (Hutton et al., 2002; Jenkinson, 2003; Jezzard and Balaban, 1995; Jezzard and Clare, 1999). The field map was acquired in parts of the traveling-subject

dataset; therefore, we investigated the effectiveness of field map correction by comparing the effect size of the measurement bias and the participant factor between functional images with and without field map correction. Our prediction was as follows: if field map correction is effective, the effect of measurement bias will decrease, while that of the participant factor will increase following field map correction. Field map correction using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12) reduced the effect of measurement bias in the inferior brain regions (whole brain: 3% reduction in the standard deviation of the magnitude distribution of the measurement bias) and increased the effect of the participant factor in the whole brain (3% increase in the standard deviation of the magnitude distribution of the participant factor; Appendix Figures A.2a and A.2b). However, the effect of measurement bias remained large in inferior brain regions (Appendix Figure A.2a), and hierarchical clustering analysis revealed that the clusters of the phase-encoding direction remained dominant (Appendix Figure A.2c). These results indicate that, even with field map correction, it is largely impossible to remove the influence of differences in phase-encoding direction on FC. Thus, harmonization methods are still necessary to remove the effect of these differences and other measurement-related factors. However, some distortion correction methods have been developed, such as top-up method and symmetric normalization (Andersson et al., 2003; Gorgolewski et al., 2017), and further studies are required to verify the efficacy of these methods.

2.3.3 Characteristics of sampling bias

Our data supported the different-subpopulation model rather than the single-population model (Fig. 2.5), which indicates that sampling bias is caused by sampling from among different subpopulations. Furthermore, these findings suggest that, during big data collection, it is better to sample participants from several imaging sites than to sample many participants from a few imaging sites. These results were obtained only by combining the SRPBS multi-disorder database with a traveling-subject dataset (https://bicr.atr.jp/decnefpro/). To the best of our knowledge, the present study is the first to demonstrate the presence of sampling bias in rs-fcMRI data, the mechanisms underlying this sampling bias, and the effect size of sampling bias on resting-state FC, which was comparable to that of psychiatric disorders. We analyzed sampling bias among HCs only, because the number of sites was too small to conduct an analysis of patients with psychiatric diseases.

2.3.4 The effect of harmonization method

We developed a novel harmonization method using a traveling-subject dataset (i.e., traveling-subject method), which was then compared with existing harmonization methods. Our results demonstrated that the traveling-subject method outperformed other conventional GLM-based harmonization methods and ComBat method. The traveling-subject method achieved reduction of the measurement bias by 29% compared to 3% in the second highest value for ComBat method and improvement of the signal to noise ratios by 40% compared to 3% in the second highest value for ComBat method. This result indicates that the traveling-subject dataset helps to properly estimate the measurement bias and also helps to harmonize the rs-fMRI data across imaging sites towards development of a wide range of final applications. As one example of final application, we constructed biomarkers for psychiatric disorders based on rs-fcMRI data, which distinguishes between HCs and patients, and a regression model to predict participants' age based on rs-fcMRI data using SRPBS multi-disorder dataset (see

"Classifiers for MDD and SCZ, based on the four harmonization methods" and "Regression models of participant age based on the four harmonization methods" in Appendix A). We quantitatively evaluated the harmonization method to investigate the generalization performance to independent validation dataset, which was not included in SRPBS multi-disorder dataset. Although the ComBat method achieved the highest performance for the MDD classifier and regression model of age, it was inferior to the raw method for the SCZ classifier. By contrast, the traveling-subject harmonization method always improved the generalization performance as compared with the case where harmonization was not performed. These results indicate that the traveling-subject dataset also helps the constructing a prediction model based on multi-site rs-fMRI data. As a future work, it is necessary to improve traveling-subject method by incorporating hierarchy like ComBat.

2.3.5 Limitations

The present study possesses some limitations of note. The accuracy of measurement bias estimation may be improved by further expanding the traveling-subject dataset. This can be achieved by increasing the number of traveling participants or sessions per site. However, as mentioned in a previous traveling-subject study (Noble et al., 2017), it is costly and time-consuming to ensure that numerous participants travel to every site involved in big database projects. Thus, the cost-performance tradeoff must be evaluated in practical settings. The numbers of traveling participants and MRI sites used in this study (nine and twelve, respectively) were larger than those used in a previous study (eight and eight, respectively) (Noble et al., 2017), and the number of total sessions in this study (411) was more than three times larger than that used in the previous study (128) (Noble et al., 2017). Furthermore, although we estimated the measurement bias for each connectivity, hierarchical models of the brain (e.g., ComBat) may be more appropriate for improving the estimates of measurement bias. Regarding the number of sites in the data with psychiatric disorders, we believe that uniqueness of our study exists in the datasets of multiple disorders and multiple sites with traveling subject data rather than the number of sites for a single disorder. For example, although ABIDE (Abraham et al., 2017; Di Martino et al., 2014) collected the data from patients with ASD from 17 sites, it significantly differs from our study because it does not use a unified protocol for data collection and does not include a traveling subject dataset. In this study, we have collected the data using a unified protocol with healthy controls from 6 sites, patients with MDD from 3 sites, patients with ASD from one site, patients with SCZ from 3 sites, patients with OCD from one site, and a traveling subject dataset from 12 sites. These datasets enabled us to compare the magnitude of the effect between site differences (measurement or sampling bias) and multiple disorder factors, which is the key point of our study. To the best of our knowledge, such multi-site multi-disorder resting-state fMRI dataset has not existed so far.

2.3.6 Summary

In this chapter, by acquiring a separate traveling-subject dataset and the SRPBS multi-disorder database, we revealed that site differences were composed of biological sampling bias and engineering measurement bias. The effect sizes of these biases on FC were greater than or equal to the effect sizes of psychiatric disorders, highlighting the importance of controlling for site differences when investigating psychiatric disorders. Furthermore, using the traveling-subject dataset, we developed a novel traveling-subject

method that harmonizes the measurement bias only by separating sampling bias from site differences. Our findings verified that the traveling-subject method outperformed conventional GLM-based harmonization methods and ComBat method. These results suggest that a traveling-subject dataset can help to harmonize the rs-fMRI data across imaging sites.

Chapter 3

A common brain network between major depressive disorder and symptoms of depression

In chapter 3, we constructed a reliable resting-state functional connectivity (FC)-based classifier of psychiatric disorder and also constructed a regression model of disorder symptoms to resolve the problem of diagnosis-based analysis.

In our study, we focused on major depressive disorder (MDD). MDD is diagnosed when depressed symptoms persist for more than two weeks and causes the greatest worldwide social loss (Ferrari et al., 2013; Hay et al., 2017; Kassebaum et al., 2016; Vos et al., 2015). Despite the significance of this disorder, its neurobiological basis remains poorly understood. Therefore, it is important to investigate resting-state FC that is associated with the depressed symptoms and compare this to FCs associated with MDD to strengthen neuroscientific understanding, and aid future diagnosis and treatment of MDD. In this study, we harnessed a machine-learning algorithm to automatically and objectively identify resting-state FC related to diagnosis and symptoms. We constructed an MDD classifier that distinguished between HCs and MDD patients based on FC patterns. We examined the FCs that were identified as important for constructing the MDD classifier. In addition, we constructed a regression model of Beck Depression Inventory-II (BDI) scores (Dozois and Covin, 2004). This scale is widely used for measuring the severity of depressed symptoms. We examined the FCs that were identified as important for constructing this regression model. Furthermore, we investigated the common FCs between MDD diagnosis and depressed symptoms (Fig. 3.1).

We have developed a regression model of Beck Depression Inventory-II (BDI) scores based on resting-state functional connectivity, and a classifier for major depressive disorder (MDD) that could distinguish between healthy controls (HCs) and MDD patients based on resting-state functional connectivity. Next, we investigated the common connectivity patterns between MDD diagnosis and depressed symptoms.

To construct a reliable classifier and regression model using a machine-learning algorithm, it is essential to use a large sample size of data collected from multiple imaging sites for training. We used a discovery rs-fMRI dataset of 713 participants, which included 149 MDD patients, collected from 4 imaging sites in different regions of Japan. Furthermore, to ensure the reproducibility of the identified FCs, it is critical to demonstrate the generalizability of the classifier and regression model with an independent validation dataset (Munafò et al., 2017; Nosek and Errington, 2017; Poldrack et al., 2017; Whelan and Garavan, 2014). Our independent validation dataset consisted of 449 participants, which included 185 MDD patients, collected from 4 imaging sites that were not included in the discovery dataset.

As a result, a reliable and reproducible neuroimaging-based classifier for MDD and regression model for BDI scores, respectively, were developed. Furthermore, this classifier and regression model could be generalized to the independent validation dataset.

We found an overlap of approximately 30% FCs between depressed symptoms and MDD diagnosis. Interestingly, common FCs were associated with the salience and default mode networks. Taken together, our study has revealed some neurobiological underpinnings of the depressed symptoms in MDD patients. This result leads to the elucidation of the neurological basis of MDD and development of a theranostic biomarker that can contribute to MDD diagnosis and the determination of therapeutic targets for depressed symptoms (Watanabe et al., 2017; Yahata et al., 2017; Yamada et al., 2017).

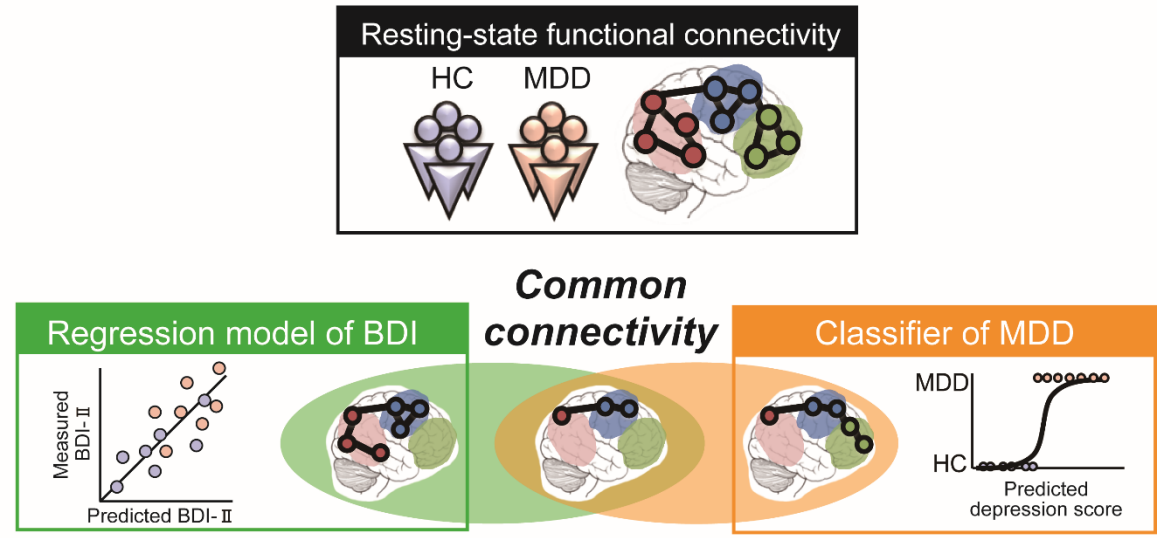

FIGURE 3.1 | Schematic illustration of the study design.

We have developed a regression model of Beck Depression Inventory-II (BDI) scores based on resting-state FC, and a classifier for major depressive disorder (MDD) that could distinguish between healthy controls (HCs) and MDD patients based on resting-state FC. Next, we investigated the common connectivity patterns between MDD diagnosis and depressed symptoms.

3.1 Material and Methods

3.1.1 Participants

We used two rs-fMRI datasets for the analyses: (1) Dataset 1 contained data from 713 participants (564 HCs from 4 sites, 149 MDD patients from 3 sites; Table 1); (2) Dataset 2 contained data from 449 participants (264 HCs from independent 4 sites, 185 MDD patients from independent 4 sites; Table 3). Depressed symptoms were evaluated using the BDI-II score obtained from most participants in each dataset. This study was carried out in accordance with the recommendations of the institutional review boards of the principal investigators' respective institutions (Hiroshima University, Kyoto University, Showa University, University of Tokyo, and Yamaguchi University) with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the institutional review boards of the principal investigators' respective institutions (Hiroshima University, Kyoto University, Showa University, University of Tokyo, and Yamaguchi University).

3.1.2 Datasets

Dataset 1—the "discovery dataset"—was used to construct an MDD classifier and BDI linear regression model. Data were acquired using parameters shown in Table 2. Each participant underwent a single rs-fMRI session for 10 min. Within the Japanese SRPBS DecNef project, we planned to acquire the rs-fMRI data using a unified imaging protocol (Table 3.2; http://bicr.atr.jp/rs-fmri-protocol-2/). However, there were erroneously two phase-encoding directions (P→A and A→P). In addition, different sites had different MRI hardware (Table 3.3). During the rs-fMRI scans, participants were instructed to "Relax. Stay awake. Fixate on the central crosshair mark, and do not concentrate on specific things."

Dataset 2—the "independent validation dataset"—was used to test the MDD classifier and BDI regression model. Data were acquired following protocols reported in Table 3.5. The sites used were different from Dataset 1. Each participant underwent a single rs-fMRI session of 5 or 8 min.

Most data utilized in this study can be downloaded publicly from the DecNef Project Brain Data Repository at https://bicr-resource.atr.jp/decnefpro/. The data availability statement of each site were described in Table 3.6.

TABLE 3.1 | Demographic characteristics of participants in the discovery dataset

Site	HC				MDD				ALL			
	Number	Male/ Female	Age (yr)	BDI	Number	Male/ Female	Age (yr)	BDI	Number	Male/ Female	Age (yr)	BDI
Center of Innovation in Hiroshima University (COI)	124 (123)	46/78	51.9±13.4	8.2±6.3	70 (70)	31/39	45.0±12.5	26.2±9.9	194 (193)	77/117	49.4±13.5	14.7±11.7
Kyoto University (KUT)	169 (139)	100/69	35.9±13.6	6.0±5.4	17 (17)	11/6	43.9±13.3	27.7±10.1	186 (156)	111/75	36.7±13.7	8.3±9.1
Showa University (SWA)	101 (97)	86/15	28.4±7.9	4.4±3.8	0	-	-	-	101 (97)	86/15	28.4±7.9	4.4±3.8
University of Tokyo (UTO)	170 (24)	78/92	35.6±17.5	6.7±6.5	62 (32)	36/26	38.7±11.6	20.4±11.4	232 (56)	114/118	36.4±16.2	14.5±11.8
Summary	564 (383)	310/254	38.0±16.1	6.3±5.6	149 (119)	78/71	42.3±12.5	24.9±10.7	713 (502)	388/325	38.9±15.5	10.7±10.6

The number in parentheses indicate the number of participants with BDI score. All demographic distributions are matched between the MDD and HC populations in the discovery dataset ($p > 0.05$). BDI: Beck Depression Inventory-Ⅱ; HC: Healthy Control; MDD: Major Depressive Disorder.

TABLE 3.2 | Unified imaging protocols for resting-state fMRI in the discovery dataset

Magnetic field strength	Field of view (mm)	Matrix	Number of slices	Number of volumes	In-plane resolution (mm)	Slice thickness (mm)	Slice gap (mm)	TR (ms)	TE (ms)	Total scan time (min:s)	Flip angle (deg)	Slice acquisition order	Eye closed / fixate
3.0 T	212 × 212	64 × 64	40	240	3.3125 × 3.3125	3.2	0.8	2,500	30	10:00	80	Ascending	Fixate

TABLE 3.3 | Different imaging protocols among sites for resting-state fMRI in the discovery dataset

Site	Center of Innovation in Hiroshima University	Kyoto University TimTrio	Showa University	University of Tokyo
Abbreviation	COI	KUT	SWA	UTO
MRI scanner	*Siemens Verio*	*Siemens TimTrio*	*Siemens Verio*	*GE MR750w*
Channels per coil	12	32	12	24
Phase encoding	AP	PA	PA	PA

TABLE 3.4 | Demographic characteristics of participants in the independent validation dataset

Site	HC				MDD				ALL			
	Number	Male/ Female	Age (yr)	BDI	Number	Male/ Female	Age (yr)	BDI	Number	Male/ Female	Age (yr)	BDI
Hiroshima Kajikawa Hospital (HKH)	29 (29)	12/17	45.4±9.5	5.1±4.6	33 (33)	20/13	44.8±11.5	28.5±8.7	62 (62)	32/30	45.1±10.5	17.6±13.7
Hiroshima Rehabilitation Center (HRC)	49 (49)	13/36	41.7±11.7	9.1±8.5	16 (16)	6/10	40.5±11.5	35.3±9.5	65 (65)	19/46	41.4±11.5	15.6±14.3
Hiroshima University Hospital (HUH)	66 (66)	29/37	34.6±13.0	6.9±5.9	57 (57)	32/25	43.3±12.2	30.9±9.0	123 (123)	61/62	38.6±13.3	18.0±14.1
Yamaguchi University (UYA)	120 (120)	50/70	45.9±19.5	7.1±5.6	79 (78)	36/43	50.3±13.6	29.7±10.7	199 (198)	86/113	47.6±17.5	16.0±13.6
Summary	264 (264)	104/160	42.2±16.5	7.2±6.3	185 (184)	94/91	46.3±13.0	30.3±9.9	449 (448)	198/251	43.9±15.3	16.7±13.9

The number in parentheses indicate the number of participants with BDI score. Demographic distribution of age is matched between the MDD and HC populations in the independent validation dataset ($p > 0.05$). Demographic distribution of sex ratio is 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.

TABLE 3.5 | Imaging protocols for resting-state fMRI in the independent validation dataset

Site	**Hiroshima Kajikawa Hospital**	**Hiroshima Rehabilitation Center**	**Hiroshima University Hospital**	**Yamaguchi University**
Abbreviation	HKH	HRC	HUH	UYA
MRI scanner	*Siemens Spectra*	*GE Signa HDxt*	*GE Signa HDxt*	*Siemens Skyra*
Magnetic field strength	3.0 T	3.0 T	3.0 T	3.0 T
Channels per coil	12	8	8	20
Field-of-view (mm)	192 × 192	256 × 256	256 × 256	220 × 220
Matrix	64 × 64	64 × 64	64 × 64	64 × 64
Number of slices	38	32	32	34
Number of volumes	107	143	143	200
In-plane resolution (mm)	3.0 × 3.0	4.0 × 4.0	4.0 × 4.0	3.4 × 3.4
Slice thickness (mm)	3.0	4	3.2	4.0
Slice gap (mm)	0	0	0	1.0
TR (ms)	2,700	2,000	2,000	2,500
TE (ms)	31	27	27	30
Total scan time (min:s)	5:00	4:46	5:00	8:28
Flip angle (deg)	90	90	90	80
Slice acquisition order	Ascending	Ascending (Interleaved)	Ascending (Interleaved)	Ascending
Phase encoding	AP	AP	PA	PA
Eyes closed / open / fixate	Fixate	Fixate	Fixate	Closed

TABLE 3.6 | **Data availability statement**

Site	Number of subjects	Type of data availablity
Center of Innovation in Hiroshima University (COI)	194	1
Kyoto University (KUT)	186	2
Showa University (SWA)	101	2
University of Tokyo (UTO)	232	2
Hiroshima University Hospital (HUH)	123	1
Hiroshima Kajikawa Hospital (HKH)	62	1
Hiroshima Rehabilitation Center (HRC)	65	1
Yamaguchi University (UYA)	199	4
Traveling subject	9	3
Summary	1171	

Note: Type of data availability

1) freely available without restriction allowing commercial re-use

2) freely available but not allowing commercial re-use

3) available after registration to our record but not allowing commercial re-use

4) available only to our research group

3.1.3 Preprocessing and calculation of the resting-state FC matrix

We performed preprocessing of the rs-fMRI data using FMRIPREP version 1.0.8 (Esteban et al., 2018). The first 10 s of the data were discarded to allow for T1 equilibration. Preprocessing steps included slice-timing correction, realignment, co-registration, distortion correction, segmentation of T1-weighted structural images, normalization to Montreal Neurological Institute (MNI) space, and spatial smoothing with an isotropic Gaussian kernel of 6 mm full-width at half-maximum. The distortion correction was not performed for dataset 2 due to the lack of fieldmap data. For more details of the pipeline see http://fmriprep.readthedocs.io/en/latest/workflows.html.

Parcellation of brain regions: To analyze the data using Human Connectome Project (HCP) style surface-based methods we used the ciftify toolbox version 2.0.2 (https://edickie.github.io/ciftify/#/). This allowed us to analyze our data, which lacked the T2-weighted image required for HCP pipelines, using an HCP-like surface-based pipeline. Next, we used the Glasser's 379 surface-based parcellations (cortical 360 parcellations + subcortical 19 parcellations) as regions of interests (ROIs), which are considered reliable brain parcellations (Glasser et al., 2016a). The BOLD signal time courses were extracted from these 379 ROIs. To facilitate the comparison of our results with previous studies, we identified the anatomical names of important ROIs and the names of intrinsic brain networks that included the ROIs using anatomical automatic labeling (AAL) (Tzourio-Mazoyer et al., 2002) and Neurosynth (http://neurosynth.org/locations/).

Physiological noise regression: Physiological noise regressors were extracted by applying CompCor (Behzadi et al., 2007). Principal components were estimated for the anatomical

CompCor (aCompCor). A mask to exclude signals with a cortical origin was obtained by eroding the brain mask and ensuring that it contained subcortical structures only. Five aCompCor components were calculated within the intersection of the subcortical mask and union of the CSF and WM masks calculated in T1-weighted image space, after their projection to the native space of functional images in each session. To remove several sources of spurious variance, we used linear regression with twelve regression parameters, such as six motion parameters, average signals over the whole brain, and five aCompCor components.

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

Head motion: Frame-wise displacement (FD) (Power et al., 2014) was calculated for each functional session using Nipype (https://nipype.readthedocs.io/en/latest/). FD was used in the subsequent scrubbing procedure. To reduce spurious changes in FC from head motion, we removed volumes with FD >0.5 mm, as proposed in a previous study (Power et al., 2014). The FD represents head motion between two consecutive volumes as a scalar quantity (i.e., the summation of absolute displacements in translation and rotation). Using the aforementioned threshold, 6.3% ± 13.5 volumes (mean ± SD) were removed per one rs-fMRI session in all datasets. If the ratio of the excluded volumes after scrubbing exceeded the mean + 3 SD, participants were excluded from the analysis. As a result, 35 participants were removed from all datasets. Thus, we included 683 participants (545 HCs, 138 MDD patients) in the discovery dataset and 444 participants (263 HCs, 181 MDD patients) in the independent validation dataset for further analysis.

Calculation of FC matrix: FC was calculated as the temporal correlation of rs-fMRI BOLD signals across 379 ROIs for each participant. There are some candidates to measure FC, such as the tangent method and partial correlation; however, we used Pearson's correlation coefficients because they are the most commonly used values in previous studies. The Fisher's z-transformed Pearson's correlation coefficients were calculated between the preprocessed BOLD signal time courses of each possible pair of ROIs and used to construct 379 × 379 symmetrical connectivity matrices in which each element represented a connection strength between two ROIs. We used 71,631 FC values [(379 × 378)/2] of the lower triangular matrix of the connectivity matrix for further analysis.

Control of site differences: Next, we used a traveling-subject harmonization method to control for site differences in FC in the discovery dataset. This method enabled us to subtract pure site differences (measurement bias), which are estimated from the traveling-subject dataset wherein multiple participants travel to multiple sites to assess measurement bias. More detailed information has been described in Chapter 2 (see 2.1.8 "Traveling-subject harmonization procedures"). We used the ComBat harmonization method (Fortin et al., 2018; Fortin et al., 2017; Johnson et al., 2007; Yu et al., 2018) to control for site differences in FC in the independent validation dataset because we did not have a traveling-subject dataset for those sites. Note that the ComBat method is a more advanced method to control for site effects when compared with the conventional regression method.

3.1.4 MDD classifier in the training dataset

We constructed a biomarker for MDD, which distinguished between HCs and MDD patients, using the discovery dataset based on 71,631 FC values. To construct the classifier, a machine-learning technique was applied. Based on our previous study (Yahata et al., 2016), we have assumed that psychiatric disorder factors were not associated with whole-brain connectivity, but with a specific subset of connections. Therefore, we conducted logistic regression analyses using the least absolute shrinkage and selection operator (LASSO) method to select the optimal subset of FCs (Tibshirani, 1996). A logistic function was used to define the probability of a participant belonging to the MDD class, as follows:

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

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

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

in which $\|\boldsymbol{w}\|_1 = \sum_i^N |w_i|$ and λ represent hyper-parameters that control the amount of shrinkage applied to the estimates. To estimate λ, we conducted a nested cross-validation procedure by using the "*lassoglm*" function in MATLAB (R2016b, Mathworks, USA) and set "NumLambda" = 25 and "CV" = 10. Classification accuracy was evaluated using a 10-fold cross validation (CV) procedure (Fig. 3.2). At each CV, we used the under-sampling (Wallace et al., 2011) method to construct the classifier, because the training dataset was unbalanced with regard to the numbers of MDD patients and HCs. Almost 130 MDD patients and 130 HCs were randomly sampled from the 10-fold training set, and classifier performance was tested using the 10-fold test set (Fig. 3.2). Under-sampling is disadvantageous because it does not allow the classifier to learn using the excluded data; therefore, we repeated the aforementioned random sampling procedure 10 times (i.e., subsampling), and the mean classifier output value (diagnostic probability) was considered indicative of the classifier output (Fig. 3.2). Diagnostic probability values >0.5 were considered indicative of MDD patients. We calculated the area under the curve (AUC) using the "*perfcurve*" function in MATLAB (R2016b, Mathworks, USA). In addition, we calculated the accuracy, sensitivity, and specificity. Furthermore, we evaluated classifier performance for the unbalanced dataset using the Matthews correlation coefficient (MCC) (Chicco, 2017; Matthews, 1975a), which takes into account the ratio of the confusion matrix size.

3.1.5 BDI score regression model in the training dataset

We constructed a linear regression model to predict the BDI score using the discovery dataset based on 71,631 FC values. To construct the linear regression model, a machine-learning technique was applied to participants with BDI scores in the discovery dataset. Next, we employed linear regression using the LASSO method, as follows:

$$Predicted\ \ BDI_{sub} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{c}_{sub},$$

in which $Predicted\ \ BDI_{sub}$ represents the BDI score of a participant; $\boldsymbol{c}_{sub}$ represents a FC vector for a given participant; and $\boldsymbol{w}$ represents the weight vector of the linear regression. We also conducted a 10-fold CV procedure for this regression model (Fig. 3.2); however, no FC values were selected by LASSO in 7 out of 10 folds. Therefore, we constructed a regression model using the combination of FC values selected in all 10 folds in the training dataset (Fig. 3.2). This caused information leakage across the folds; therefore, the training dataset may be overfitting. This meant that it was important to confirm generalization performance by applying this regression model to an independent validation dataset, as described below. Finally, we calculated the mean absolute error (MAE) and Pearson's correlation coefficients between the predicted and measured BDI scores.

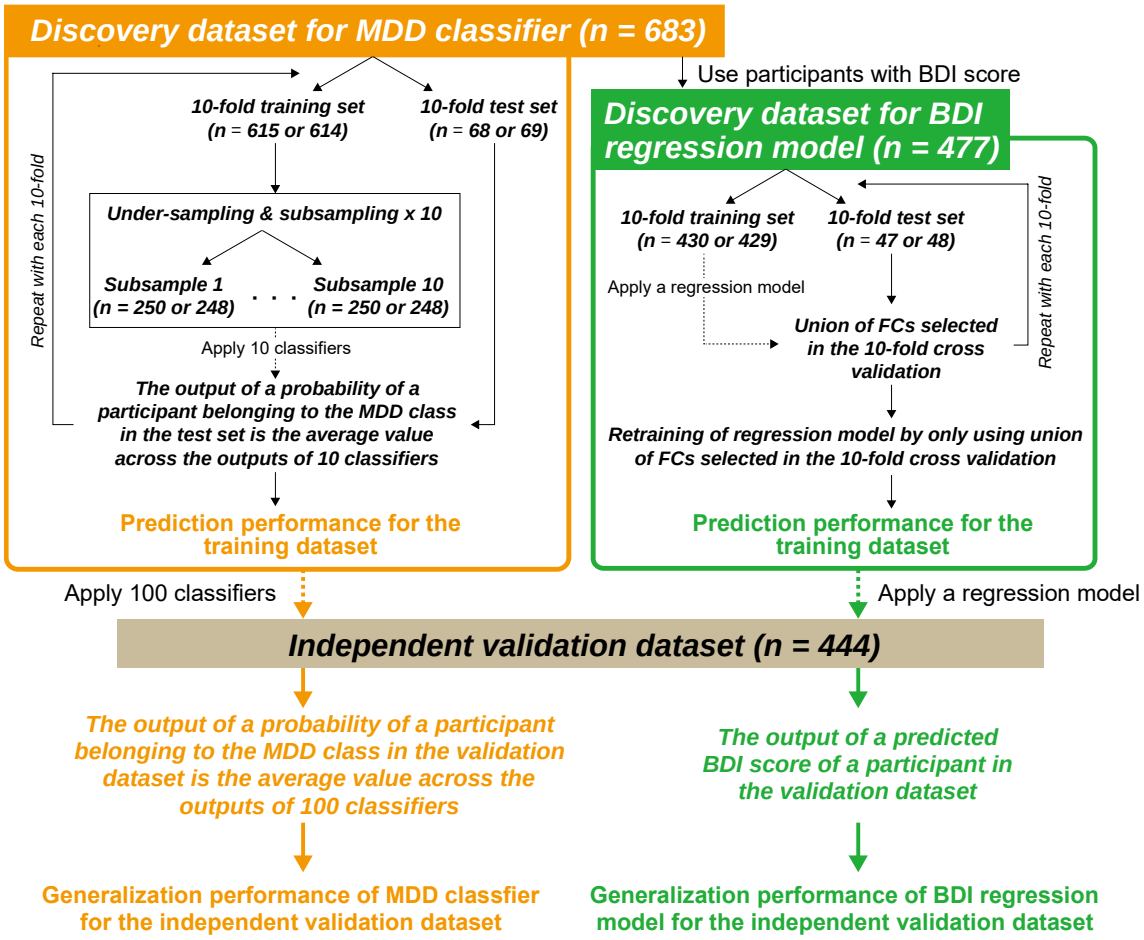

FIGURE 3.2 | Schematic representation of the procedure for selecting FCs in the MDD classifier and BDI regression model, and assessing their predictive power.

Both the MDD classifier and the BDI regression model were constructed using 10-fold cross validation in the discovery dataset. The number of participants in the 10-fold cross validation (68 or 69 for the MDD classifier, and 47 or 48 for the BDI regression model) and subsampling (250 or 248) changed according to the folds. Generalization performances were evaluated by applying the constructed model to the independent validation dataset. BDI: Beck Depression Inventory-II; MDD: Major Depressive Disorder; FC: functional connection.

3.1.6 Generalization performance of the classifier and regression model

We tested the generalizability of the classifier and regression model using an independent validation dataset. We had created 100 classifiers of MDD (10-fold × 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 cross validation 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 cross validation, and repeated this permutation procedure 100 times.

3.1.7 Identification of the FCs linked to diagnosis and symptoms

We examined resting-state FC for MDD diagnosis and depressed symptoms by extracting the important FCs that related to 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 binomial test. The mean number of FCs included in the MDD classifier per 1 CV was 329.1 (the number of FCs selected at least once in 10 subsampling); therefore, we assumed that the binomial distribution was $B(10, 329/71631)$. We set the significance level as 0.05/71631, correcting for multiple comparisons (Bonferroni correction). Therefore, FCs selected ≥3 times during the 10-fold CV were regarded as diagnostically important. Similarly, the mean number of FCs included in the BDI regression model per 1 CV was 3.4; therefore, we assumed that the binomial distribution was $B(10, 3/71631)$. In this case, FCs that were selected ≥1 time during the 10-fold CV were regarded as relevant to depressed symptoms.

3.2 Results

3.2.1 MDD classifier in the discovery dataset

The classifier separated MDD- from HC-populations with an accuracy of 66%. The corresponding AUC was 0.77, indicating acceptable discriminatory ability. Figure 3.3a shows that the two diagnostic-probability distributions of the MDD and HC populations were clearly separated by the threshold of 0.5, to the right (MDD) and to the left (HCs) for the discovery dataset. The sensitivity was 77% and specificity was 64%. This led to an acceptable MCC of 0.33. We found that acceptable classification accuracy was achieved for not only the entire dataset but also individual datasets of the three imaging sites (Fig.

3.3b) almost to the same degree. We have only HC population in SWA dataset but note that its probability distribution is comparable to those of HC populations in the other sites.

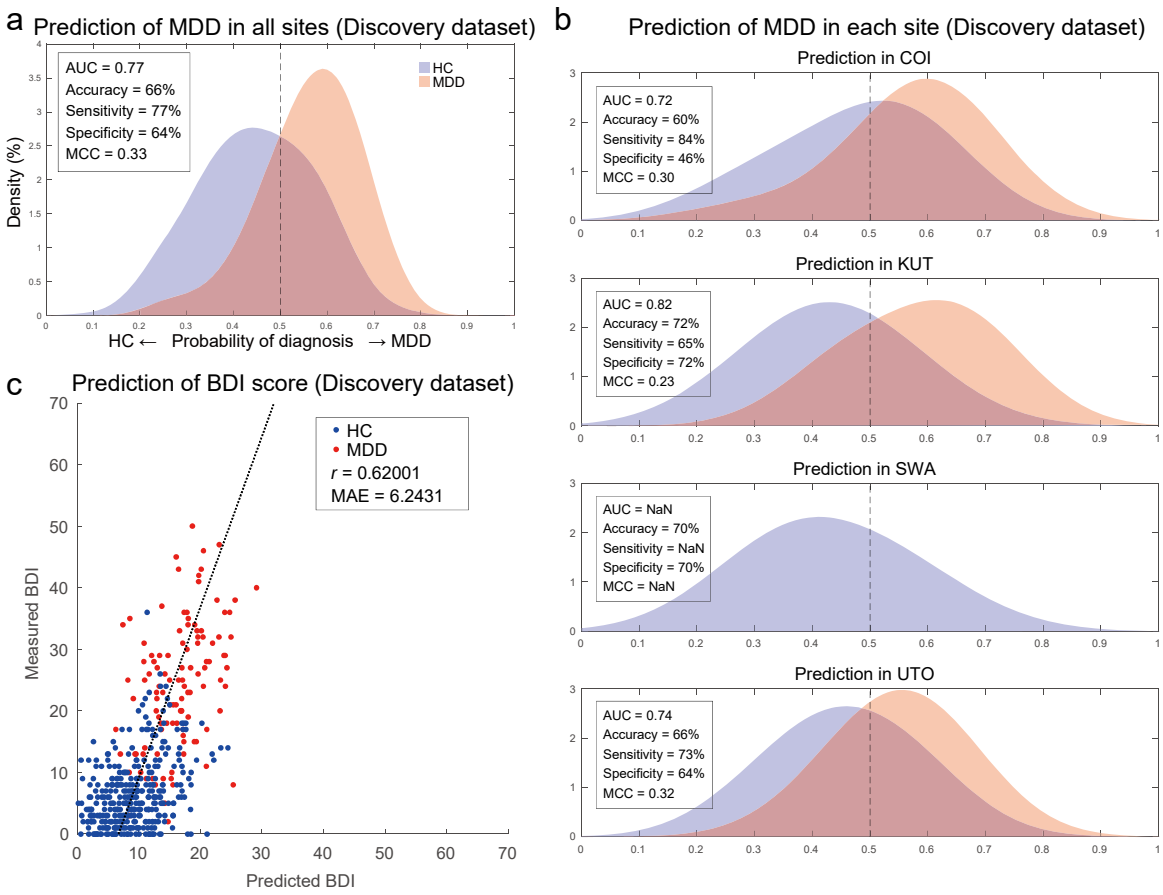

FIGURE 3.3 | Classifier performance for MDD and regression performance for BDI score in the discovery dataset.

(a) The probability distribution for the diagnosis of MDD in the discovery dataset and (b) the probability distributions for each imaging sites. The MDD and HC distributions are depicted in red and blue, respectively. (c) Scatter plots of measured BDI and predicted BDI. The solid line indicates the linear regression of the measured BDI from the predicted BDI. The correlation coefficient (r) and the 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; MCC: Matthews correlation coefficient; COI: Center of Innovation in Hiroshima University; KUT: Kyoto University; SWA: Showa University; UTO: University of Tokyo.

3.2.2 Regression models of BDI score in the discovery dataset

We found that the BDI score was well predicted with statistically significant correlation (r = 0.62, $p = 5.3 \times 10^{-52}$; mean absolute error = 6.24, Fig. 3.3c). Furthermore, we found that significant correlation was achieved not only for the entire data set but also separately for HC population and MDD population (HC population, r = 0.38, $p = 5.3 \times 10^{-14}$; MDD population; r = 0.42, $p = 4.5 \times 10^{-6}$).

3.2.3 Generalization performance of the classifier and the regression model

The classifier separated MDD- from HC-populations with an accuracy of 70% in the independent validation dataset. The corresponding AUC was 0.75 (Permutation test, $p < 0.01$; see below), indicating acceptable discriminatory ability. Figure 3.4a shows that the two diagnostic-probability distributions of the MDD and HC populations were clearly separated by the threshold of 0.5, to the right (MDD) and to the left (HCs). The sensitivity was 68% and specificity was 71%. This led to an acceptable MCC of 0.38 (Permutation test, $p < 0.01$; see below). We found that acceptable classification accuracy was achieved for not only the entire data set but also individual dataset of the four imaging sites (Fig. 3.4b). Furthermore, we also found that the BDI score was moderately well predicted with statistically significant correlation ($r = 0.21$, $p = 9.1 \times 10^{-6}$; mean absolute error = 11.8; Fig. 3.4c; Permutation test, $p < 0.01$; see below).

Figure 3.5 shows the histograms of the AUC and the MCC in the permutation test (100 repetitions) for MDD classifier. The vertical red lines indicate the AUC and MCC of the MDD classifier in the independent validation data without permutation. These results indicate that both AUC and MCC were significant at $p = 0.01$. For the BDI classifier, we could not construct any regression model in the whole permutation procedure, because no FC was selected at the nested cross validation in the LASSO procedure. This result indicates that the performance of the BDI regression model in the independent validation data without permutation was significant at $p = 0.01$.

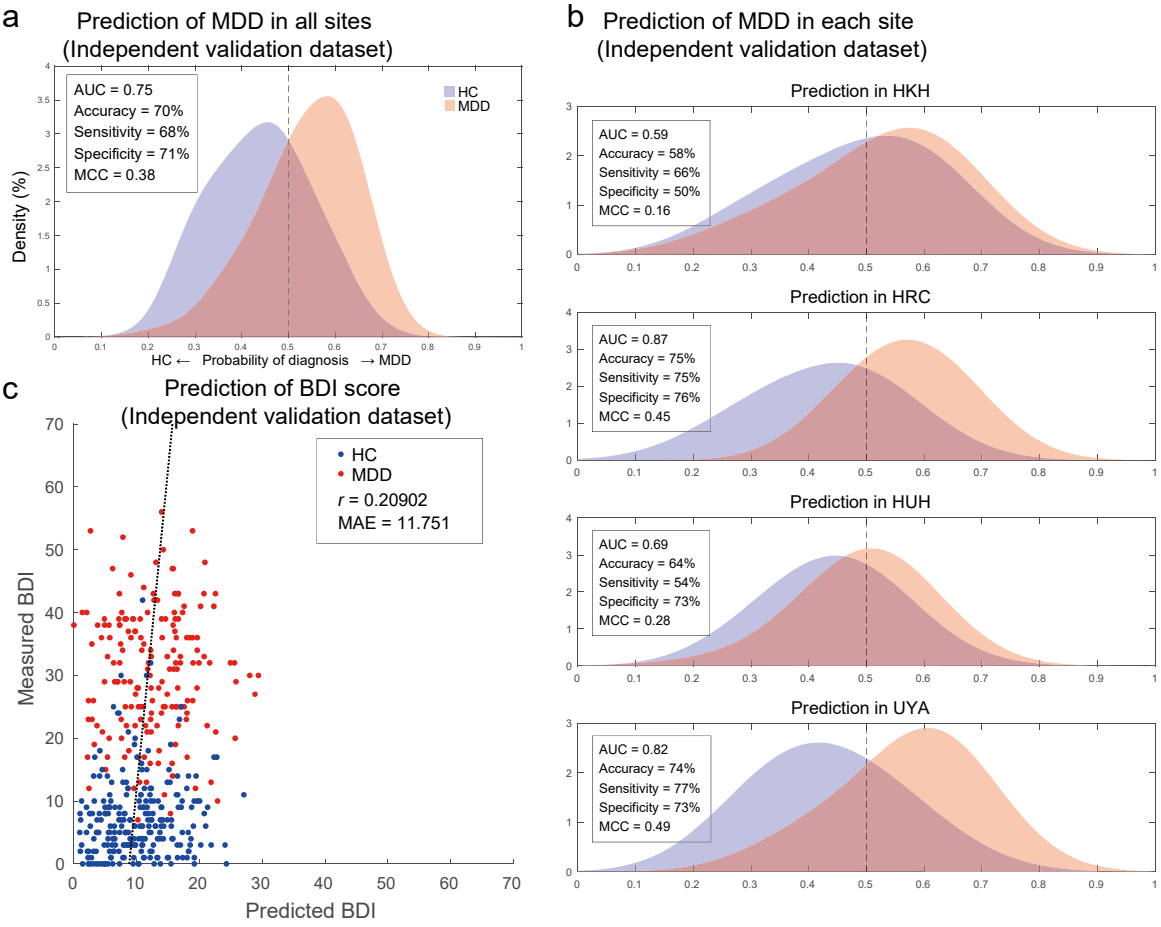

FIGURE 3.4 | Classifier performances for MDD and regression performance for BDI score in the independent validation dataset.

(a) The probability distribution for the diagnosis of MDD in the independent validation dataset and (b) the probability distributions for each imaging sites. The MDD and HC distributions are depicted in red and blue, respectively. (c) Scatter plots of measured BDI and predicted BDI. The solid line indicates the linear regression of the measured BDI from the predicted BDI. The correlation coefficient (r) and the 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; MCC: Matthews correlation coefficient; HKH: Hiroshima Kajikawa Hospital; HRC: Hiroshima Rehabilitation Center; HUH: Hiroshima University Hospital; UYA: Yamaguchi University.

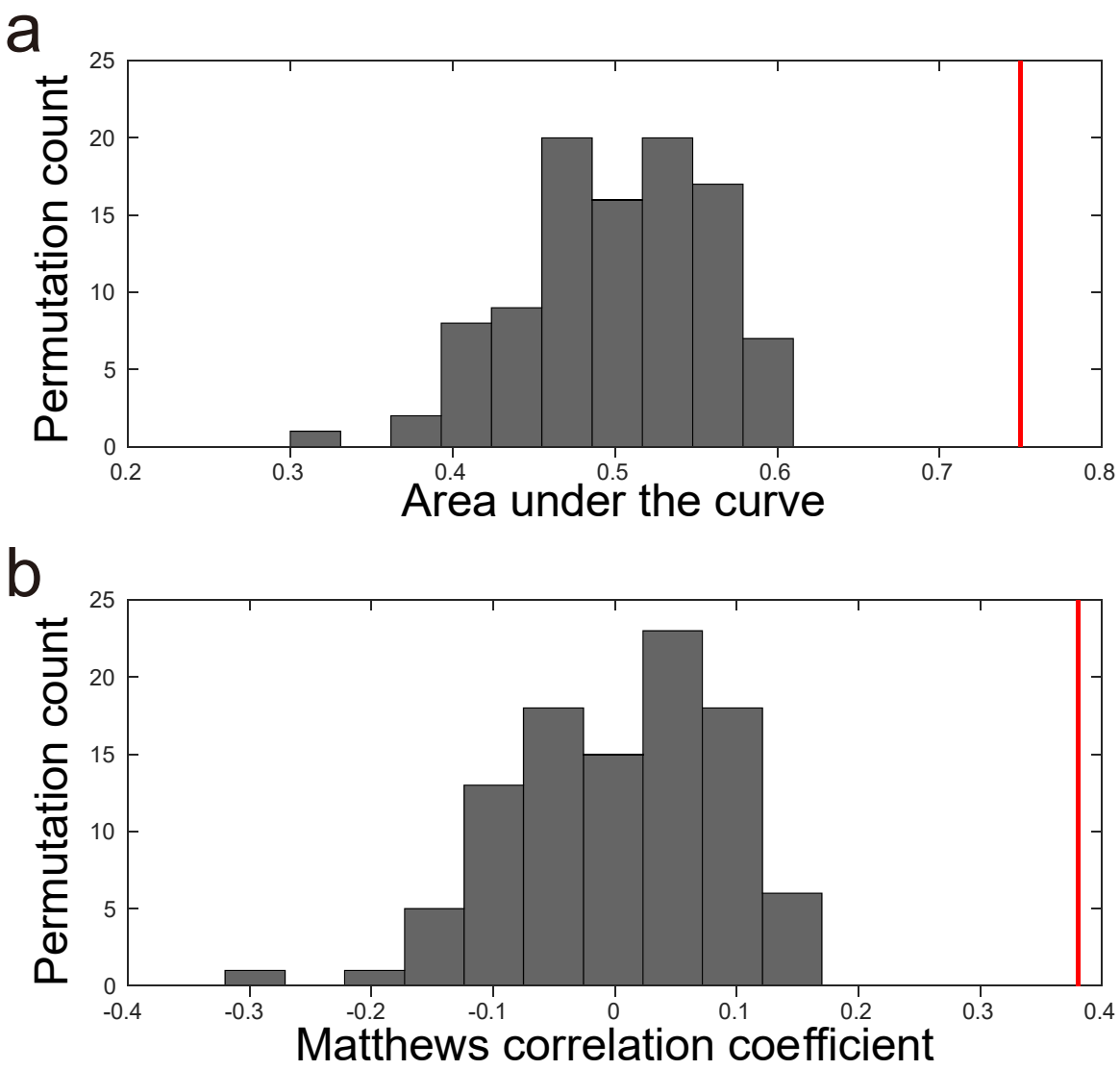

FIGURE 3.5 | Results of permutation test for MDD classifier.

Panels (a) and (b) show the histograms of the performances in the permutation test (100 repetitions) for the independent validation data, area under the curve (AUC) and Matthews correlation coefficient (MCC) respectively. The vertical red lines indicate the AUC and MCC of the MDD classifier in the independent validation data without permutation. Both AUC and MCC were significant at $p = 0.01$, as demonstrated by the two panels.

3.2.4 Common FCs between major depressive disorder diagnosis and symptoms of depression

We identified the 340 FCs related to the diagnosis and the 21 FCs related to the symptoms that were automatically and objectively identified from the data for the reliable classification of MDD and HCs and the regression model of BDI score by the machine-learning algorithms. Seven FCs were common between the diagnosis 340 FCs and the symptoms 21 FCs. Figure 3.6 shows the spatial distribution of the common 7 FCs. A detailed list of FC properties is provided in Table 6. Furthermore, the mean FC values

of these 7 FCs were very similar between in the discovery dataset and in the independent validation dataset (Fig. 3.7). This result suggests that the intersected 7 FCs are trustworthy in characterizing neural substrates of MDD and depressed symptoms.

To test statistical significance of the possibility of 7 FCs overlap, we tried to compare the number of FCs in permutation test. However, as stated in the previous subsection, no BDI regression model could be meaningfully constructed in the permutation test. This result indicates that the overlap of 7 FCs with MDD classifier and BDI regression model were significant. We also checked the significance of overlap by calculating the probability of 7 overlap FCs when we randomly selected 340 FCs from 71,631 FCs and 21 FCs from 71,631 FCs independently. We repeated this random sample 10,000 times. As a result, the number of 1 overlap FC was 891 times, the number of 2 overlap FCs was 48 times, and the number of 3 overlap FCs was 1 time. That is, even the probability of 2 FCs overlap was $p < 0.005$. This result also indicates that the overlap of 7 FCs with MDD classifier and BDI regression model were significant.

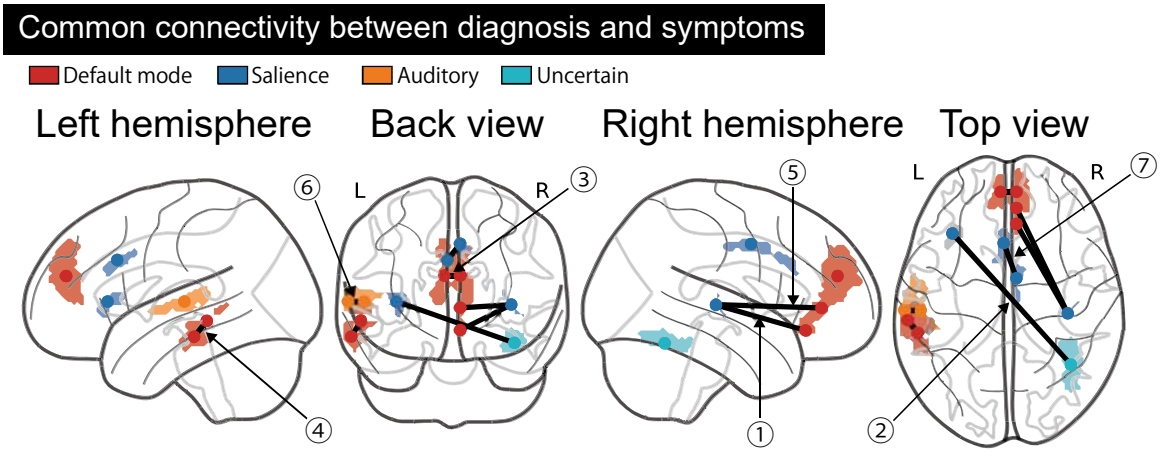

FIGURE 3.6 | Common functional connectivity between diagnosis and symptoms.

The 7 common functional connectivity viewed from left, back, right, and top. Inter-hemisphere connections are shown only in back and top views. Regions are color-coded by intrinsic network.

TABLE 3.7 | All common functional connections and weights in regression model of BDI score

ID	ROI1			ROI2			Weight
	Glasser's area name	AAL Label	Network	Glasser's area name	AAL Label	Network	
1	R.52	R.Insula	Salience	R.s32	R.Frontal_Med_Orb	Default mode	-4.84
2	L.FOP5	L.Insula	Salience	R.FFC	R.Fusiform	Uncertain	4.12
3	L.9m	L.Frontal_Sup_Medial	Default mode	R.9m	R.Frontal_Sup_Medial	Default mode	-4.02
4	L.STSvp	L.Temporal_Mid	Default mode	L.TE1m	L.Temporal_Mid	Default mode	-3.80
5	R.52	R.Insula	Salience	R.a24	R.Cingulum_Ant	Default mode	-1.71
6	L.PBelt	L.Temporal_Sup	Auditory	L.A4	L.Temporal_Sup	Auditory	-0.78
7	L.a24pr	L.Cingulum_Mid	Salience	R.p24pr	R.Cingulum_Mid	Salience	-0.17

Labels of ROIs were determined by referring to AAL and Neurosynth (http://neurosynth.org/locations/)

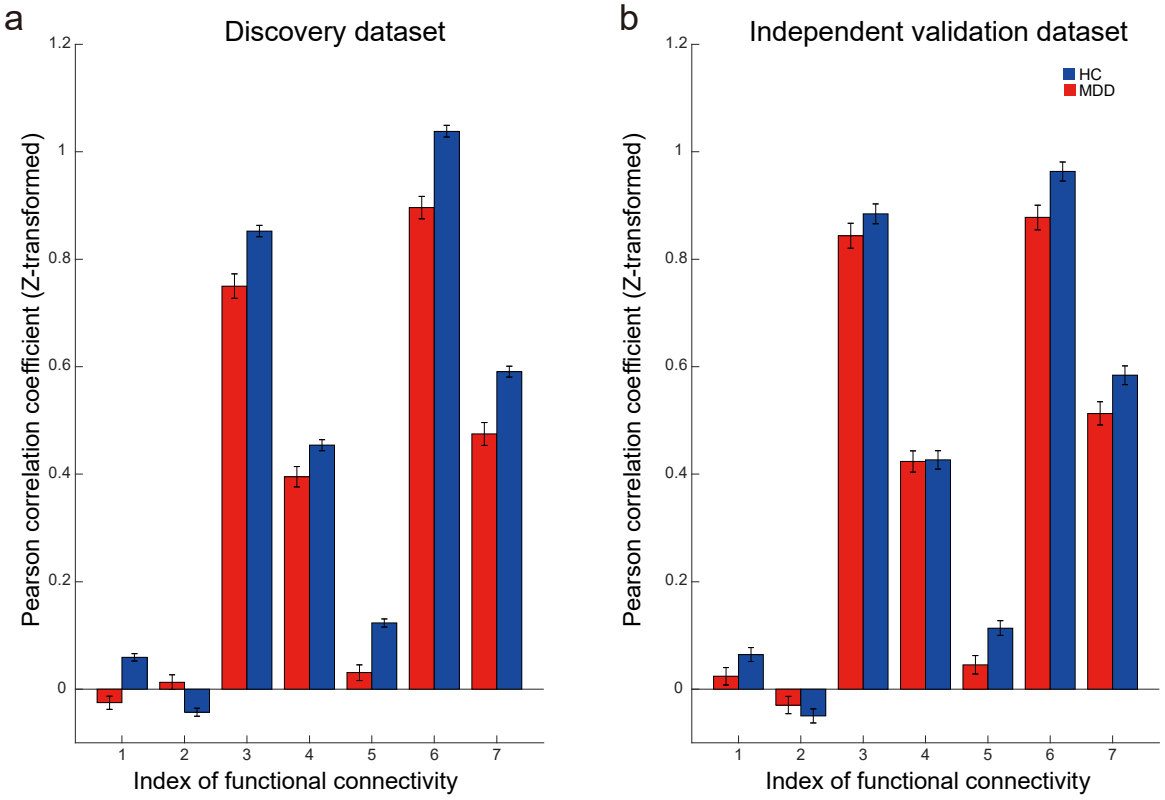

FIGURE 3.7 | Similar FC values between discovery dataset and independent validation dataset in 7 common FCs.

(a) The functional connectivity (FC) values of 7 common FC for both healthy controls (HCs; blue bar) and major depressive disorders (MDD; red bar) in the discovery dataset. (b) The FC values of 7 common FC for both HCs and MDD in the independent validation dataset. Error bar shows the standard error.

3.3 Discussion

3.3.1 Signatures of our classifier of MDD

In the present study, we constructed the reliable neuroimaging-based classifier for MDD and the regression model of BDI by investigating the whole-brain patterns of FCs using the rs-fMRI data of 138 MDD patients and 545 HCs collected at multiple sites. The constructed MDD classifier achieved acceptable generalized prediction performance AUC of 0.75 and MCC of 0.38 for the independent validation dataset. Furthermore, acceptable generalized prediction performance was achieved for not only the entire data set but also datasets of the four imaging sites (Fig. 3.4b), from where no data were included in the discovery dataset. This generalization was achieved even though the imaging protocols in the independent validation datasets were different from those of the discovery dataset, that is the unified SRPBS DecNef protocol. Successful construction of FC-based MDD classifiers, which generalized to independent validation data, have been reported previously (Drysdale et al., 2017; Ichikawa et al., 2017). However, our work presents the first achievement, to our knowledge, of successful generalized classification for MDD without a restriction on the MDD subtype: Drysdale concentrated the MDD patients with treatment resistance and Ichikawa restricted the MDD patients only with melancholic

type.

For comparison, we constructed the BDI-based classifier for MDD using discovery dataset. The constructed MDD classifier achieved high generalized prediction performance AUC of 0.97 and MCC of 0.80 for the independent validation dataset. In this classifier, the subjects with BDI score of 14 or more were diagnosed as MDD patients.

3.3.2 Common FCs between diagnosis of MDD and symptoms of depression

The machine-learning algorithms reliably identified the common 7 FCs as important FCs for both the classifier of MDD and the regression model of BDI score (Figs. 3.6, 3.7 and Table 3.7). We could summarize characteristics of the 7 FCs as following two points. First, regarding the strength of FCs, all of the 7 FCs exhibited the hypoconnectivity in the MDD populations (that is, strength, or an absolute value, of FC is closer to 0 in MDD than HC individuals; Fig. 3.7). Second, the common 7 FCs are closely related to the default mode network and the salience network regarding their attribution to intrinsic functional networks. Out of the 13 brain regions comprising these 7 FCs, 5 regions belonged to the default mode network and 5 regions belonged to the salience network. Furthermore, 6 FCs have nodes (ROIs) in either the default mode network or the salience network. 2 FCs out of 7 FCs were connectivity between the two networks. 2 FCs were connectivity within the default mode network, and 1 FC was connectivity within the salience network. The default mode network and the salience network have been repeatedly implicated as essential neural correlates of depression (Dutta et al., 2014; Greicius et al., 2007; Kaiser et al., 2015; Menon, 2015; Mulders et al., 2015; Peng et al., 2018; Wang et al., 2012).

3.3.3 Importance of symptom-based approach, rather than diagnosis-based approach

As we mentioned in Chapter 1 (see 1.3.2 “Problem of diagnosis-based analysis”), most of previous studies (75%) on predictive models of psychiatric disorders have focused on diagnosis-based approach (Woo et al., 2017). However, research interests have recently moved to symptom-based approach that describes how symptoms are related to neurobiological abnormality. To the best of our knowledge, our work presents the first achievement of the regression model of BDI score based on resting-state FC that generalized across discovery and independent validation datasets all in a data-driven manner with a large sample size of about 1000 participants (Connolly et al., 2013; Davey et al., 2012; Furman et al., 2011; Peng et al., 2018; Salomons et al., 2014; Sheline et al., 2010; Strikwerda-Brown et al., 2015). Depressed symptoms can be seen in other diseases such as bipolar disorder and schizophrenia (Takizawa et al., 2014). Therefore, it may be possible to identify relationship in neurobiological abnormalities among such diseases by investigating the co-morbidity of the neurological basis of the depressed symptoms, which were revealed in this study, in the future work. Such approach would lead to an integrated understanding of mental illness (Ayuso-Mateos et al., 2010; Xia et al., 2018).

3.3.4 Candidate of theranostic biomarker

Although biomarkers have been developed with the aim of diagnosing patients, the focus has shifted to the identification of biomarkers that determine therapeutic targets (i.e.

theranostic biomarker), which would allow for more personalized treatment approaches. The seven FCs discovered in this study are promising candidates of theranostic biomarker for MDD, because these FCs are related to not only diagnosis of MDD but also depressed symptoms. It is a future work to investigate if modulation of these FCs would truly change depression symptoms or the effect of treatment by using an intervention method on FC such as a functional connectivity neurofeedback training (Koush et al., 2017; Megumi et al., 2015; Yahata et al., 2017; Yamada et al., 2017; Yamashita et al., 2017).

3.3.5 Summary

In this chapter, we constructed a reliable neuroimaging-based classifier for MDD and a regression model of BDI by investigating the whole-brain patterns of FCs using the rs-fMRI data. The MDD classifier and the regression model of BDI achieved acceptable generalization performance: Their predictions successfully generalized to the independent validation dataset collected with different imaging protocols at different imaging sites from those of the discovery dataset. We found overlap of about 30% between the resting-state FCs (7 FCs out of 21 FCs) related to depressed symptoms and those related to diagnosis of MDD. These common 7 FCs were particularly related to the salience network and the default mode network. Our study revealed the biological basis of depressed symptoms, which is one of heterogeneous symptoms in MDD. This result would contribute to the elucidation of the neurological basis of MDD.

Chapter 4

Development of functional connectivity neurofeedback

In chapter 4, we investigated the hypothesis that connectivity neurofeedback can induce the aimed direction of change in functional connectivity (FC), and the differential change in cognitive performance according to the direction of change in connectivity to resolve the problem of controllability of neurofeedback training in order to develop applications aimed at treatment for psychiatric disorders. Our connectivity neurofeedback training developed previously can control FC by rewarding spontaneous changes in FC (Megumi et al., 2015). Subjects underwent training with intermittent feedback of the temporal correlation (FC) between BOLD signals in two brain regions immediately after each trial. Subjects learned to control connectivity in a trial-and-error manner through training. To investigate our hypothesis, we separated subjects into two groups, in which we aimed to increase or decrease the FC and compared the resulting changes in cognitive performance from pre-neurofeedback to post-neurofeedback training between the two groups.

Because our previous study (Megumi et al., 2015) already successfully increased the connectivity between the left primary motor cortex (lM1), which belongs to the motor/visuospatial network (MVN), and the left lateral parietal cortex (lLP), which belongs to the default mode network (DMN) (Raichle, 2010, 2015a), we selected this connectivity as the target for neurofeedback training. Furthermore, we conducted a psychomotor vigilance task (PVT), the Eriksen flanker task (EFT), and the color-word Stroop task (CWST) before and after the neurofeedback training, since previous studies have shown that these three tasks are associated with the MVN, the DMN, or both (Hinds et al., 2013; Kelly et al., 2008; Liu et al., 2015; Thompson et al., 2013). In the current study, FC between lM1 and lLP was normally negative (e.g., r = -0.4). Therefore, we hereafter refer to a change in FC from r = -0.4 to -0.1, for instance, as an "increase," while we refer to a change such as that from r = -0.4 to -0.7 as a "decrease."

4.1 Materials and Methods

4.1.1 Participants

Thirty healthy subjects (4 women; mean age [mean ± standard deviation: SD], 22.7 ± 1.7 years; age range, 20–27 years) participated in the neurofeedback experiment. We randomly assigned subjects to an "increased FC" group (n = 18: 1 woman; mean age, 22.6 ± 1.8 years; age range, 20–27 years) or a "decreased FC" group (n = 12: 3 women; mean age, 22.8 ± 1.6 years; age range, 21–26 years). Although there are fewer female participants, there was no significant difference in the male:female ratio between the two groups (Fisher's exact test: p = 0.27). Twenty-five (13 in the "increased FC" group and 12 in the "decreased FC" group) out of 30 subjects completed behavioral testing sessions before and after the neurofeedback training. The other five subjects did not participate in the behavioral testing sessions. We excluded one subject in the "decreased FC" group from cognitive performance analysis because that subject did not follow the instructions. We also excluded another subject in the "decreased FC" group from the EFT analysis

because that subject misunderstood the instructions. All subjects were right-handed according to the Edinburgh inventory (Oldfield, 1971). The Institutional Review Board of Advanced Telecommunications Research Institute International (ATR) approved this study, which was performed in accordance with the tenets of the Declaration of Helsinki. All subjects provided written informed consent.

4.1.2 Neurofeedback training

Brain imaging and region of interest (ROI) definition

MR images were obtained using a 3-T Siemens MAGNETOM Verio scanner (Kyoto, Japan). BOLD signals were measured using an echo planar imaging (EPI) sequence (repetition time [TR], 2000 ms; echo time [TE], 26 ms; flip angle, 80°). The entire brain was covered in 33 axial slices (3.5 mm of thickness, no gap), voxel size was 3 × 3 × 3.5 mm, and field of view was 192 × 192 mm. T1-weighted structural images were acquired with a resolution of 1 × 1 × 1 mm. T2-weighted structural images were acquired with a resolution of 0.75 × 0.75 × 3.5 mm.

Following our previous study (Megumi et al., 2015), we selected the lM1, which is included in the MVN, and the lLP, which is included in the DMN, as the two target ROIs for calculating a feedback score in the connectivity neurofeedback training. lM1 was defined as Brodmann area 4 according to the anatomical map given in PickAtlas (http://fmri.wfubmc.edu/software/PickAtlas) (Lancaster et al., 1997; Maldjian et al., 2003). lLP was defined as a sphere with a 7.5-mm radius centered at (x, y, z) = (-45, -67, 36) in the Montreal Neurological Institute standard brain coordinates (MNI; Montreal, QC) according to a previous study of brain networks (Fox et al., 2005). We adopted the spherical ROI for lLP because we could not find an anatomical definition of LP as a part of DMN in the literature. By contrast, we adopted the anatomical ROI for lM1 because 1) it is well defined in anatomical maps including PickAtlas, and 2) the spherical ROI centered at M1 may include the somatosensory cortex.

Because these ROIs were defined in the standard brain, we identified corresponding voxels in the functional images of each individual subject's brain using a deformation module in SPM8 (Wellcome Trust Center for Neuroimaging, London, UK; www.fil.ion.ucl.ac.uk/spm/). We obtained several volumes of functional images for this purpose at the beginning of the experiment on each day and used the identified voxels as ROIs for calculating scores in the subsequent training blocks. Furthermore, the position and orientation of scan images on every training day were carefully matched those on the first day.

Neurofeedback training procedure

Subjects received neurofeedback training to increase or decrease FC between the two target ROIs. Each subject received training for 4 days (Fig. 4.1a: DAY1–DAY4). On each training day, subjects performed 6 blocks, each of which was composed of 10 trials. Prior to training, subjects underwent four baseline-estimation blocks to measure a subject-specific baseline correlation between the two target ROIs (Fig. 4.1a: BASE). The baseline-estimation block was identical to the neurofeedback training block, except that the score was randomly determined (see "*Online calculation of feedback score"*).

Our training procedure in each trial followed our previous study (Megumi et al., 2015). A trial in each block began with a rest period of 14 s, during which the "=" cue was presented on the screen (Fig. 4.1b: Rest). When the cue changed to "+," subjects performed the tapping motor imagery task for 14 s (Fig. 4.1b: Motor Imagery). Subjects were instructed to imagine tapping their thumbs with their fingers randomly as fast as possible. Furthermore, they were asked to produce kinesthetic imagery, rather than attempt visual imagery, and to not overtly move their hands during the task. If no task instruction was provided, cognitive states during learning were expected to differ largely among subjects, thereby making data analysis difficult. Therefore, we administered a motor imagery task to the subjects to constraint the subjects' cognitive states. After the motor imagery period, a feedback score calculated by the online MRI system (see "*Online calculation of feedback score"*) was presented on the screen as a green disc (Fig. 4.1b: Feedback). Subjects were instructed that the disc becomes bigger as they improve at producing tapping imagery; however, the disc size actually corresponded to the score determined by the temporal correlation (FC) between BOLD signals in lM1 and lLP (see "*Online calculation of feedback score"*). Subjects were instructed to make the disc size as large as possible so that it would become larger than the red circle (a baseline) and reach the outer green circle (a target). The calculations of the baseline and target are described in section "*Online calculation of feedback score"*. Subjects were informed that additional monetary reward (up to JPY 3,000) would be paid in proportion to their total score, and they received this at the end of the experiments on each day.

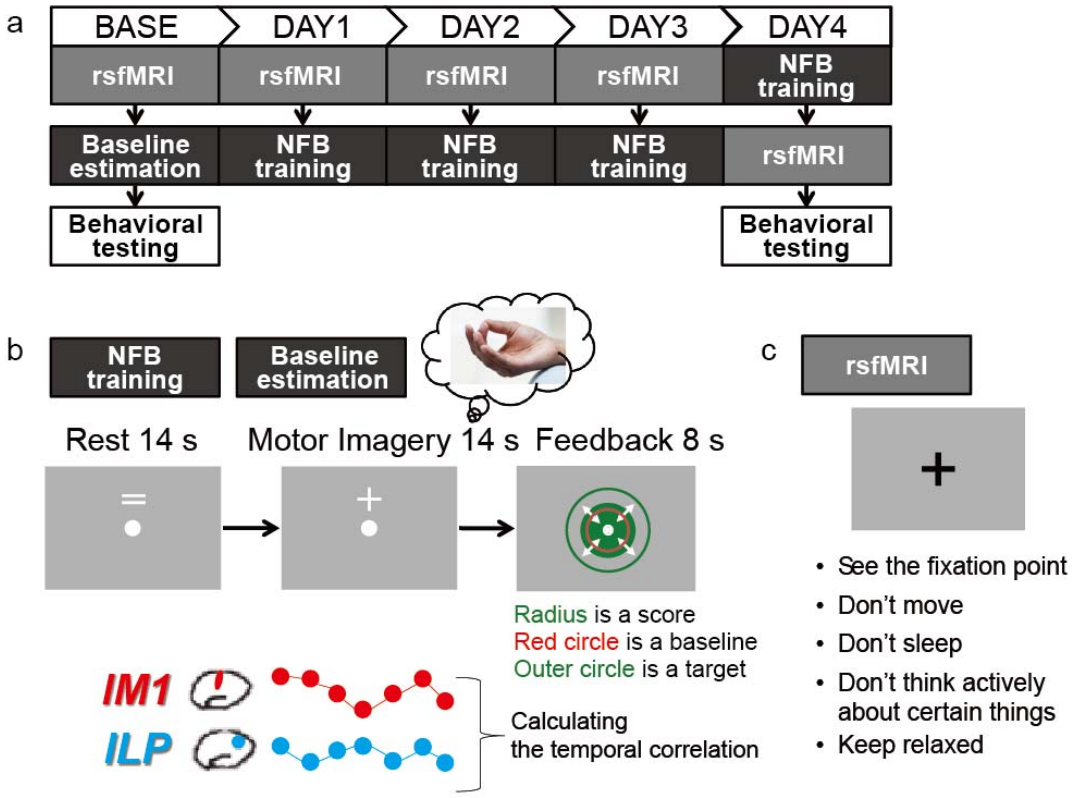

FIGURE 4.1 | Neurofeedback training procedures.

(a) Neurofeedback (NFB) training schedule. Experiments lasted for 5 days, including baseline estimation of temporal correlation between BOLD signals in the target regions

(BASE). Resting-state activity (rs-fMRI) was measured daily before NFB training from BASE to DAY3. On DAY4, rs-fMRI was measured after NFB training. Subjects performed a cognitive task on BASE and DAY4. (b) Timeline and displays for subjects in a training trial. After a rest period while the "=" cue was presented on the screen, subjects were instructed to imagine finger tapping during the motor imagery period while the "+" cue was presented. A solid green disc was presented to the subjects in the feedback period: the disc size was proportional to the correlation between BOLD signals in two target regions (left primary motor area [lM1] and left lateral parietal region [lLP]) during the motor imagery period. (c) Display for subjects in rs-fMRI. Subjects were instructed to keep looking at a fixation point at the screen center, to keep still, to stay awake, and to not think about specific things.

Online calculation of feedback score

We used in-house MATLAB software (Mathworks Inc., Natick, MA), including realignment modules of SPM8, for online processing. This software ran on a connected computer and accessed data files in the MRI system. Each volume of the functional image was realigned in real time to the first volume obtained on each day.

Seven volumes were obtained during the motor imagery period in each trial, but the first volume was discarded and one volume from the feedback period was added as compensation for hemodynamic delay. One may argue that a one-volume shift (2 s) is not enough to fully compensate for the hemodynamic delay (4–8 s); however, we followed two previous neurofeedback studies (Bray et al., 2007; Megumi et al., 2015) that used a 2-s shift and succeeded in changing brain activity. We followed these studies to minimize the delay of feedback to participants.

BOLD signal time courses were extracted from the lM1 and lLP ROIs (averaged across voxels) in these volumes. To remove several sources of spurious artifacts in BOLD signals, we conducted an online linear regression, including (i) six motion parameters, in addition to averaged signals over (ii) gray matter, (iii) white matter, and (iv) cerebrospinal fluid (Fox et al., 2005). To completely remove global signals that may be related to instrumental, motion-related, and physiological fluctuations (Caballero-Gaudes and Reynolds, 2017), we included signals averaged over the gray matter in the regression model. However, this may have removed neuronal signals in the ROIs if their activity strongly affected the average signal. In our post-hoc analysis, we calculated signals averaged over the gray matter excluding the ROIs, but they were similar to those including the ROIs (temporal correlation: $r > 0.999$), suggesting that the activity of the ROI unlikely affected the average signal.

We estimated coefficients for these parameters from the preceding 180 volumes (a moving window), which corresponded to one neurofeedback block, and regressed out the signals correlated with the parameters from a newly acquired volume. To maintain a constant number of moving volumes (180), we used the volumes acquired in the preceding block for the online regression in the early part of each block. Because there was no preceding block for the first block in the neurofeedback training, we conducted the 6-min resting condition block just before the training, which was not included in the offline analysis. Furthermore, to remove low-frequency trends from BOLD signals, a high-pass temporal filter (cutoff frequency of 0.0075 Hz) was applied to the time courses within each block (using only volumes from the same block).

Using the time courses after the noise reduction, the feedback score of the *i*-th trial was calculated as

$$Score_i = \frac{50(Correlation_i + Correlation_{Target} - 2Correlation_{Base})}{Correlation_{Target} - Correlation_{Base}}. \quad (1)$$

$$0 \leq Score_i \leq 100$$

Here, $Correlation_i$ represents the correlation between the BOLD signals averaged in each of the two ROIs. We developed this score for an intuitive feedback to participants: their baseline performance corresponded to 50 while their better performance was rewarded by the increase in the score from 50. Specifically, $Correlation_{Base}$ was the median correlation in the baseline-estimation block (40 trials) on the first day (BASE). SD was also calculated in the baseline-estimation block. $Correlation_{Target}$ was determined to restrict the appearance of a score of 100 to one time per block on average: $Correlation_{Target}$ was ($Correlation_{Base}$ + 1.28 SD) in the "increased FC" group and ($Correlation_{Base}$ – 1.28 SD) in the "decreased FC" group. Therefore, if $Correlation_i$ is equal to $Correlation_{Base}$, the score is 50 in both groups. If $Correlation_i$ increases in the "increased FC" group (or decreases in the "decreased FC" group) from $Correlation_{Base}$ to $Correlation_{Target}$, the score rises from 50 to 100. If $Correlation_i$ decreases in the "increased FC" group from $Correlation_{Base}$ to $Correlation_{Base} - 1.28$ SD (or if $Correlation_i$ increases in the "decreased FC" group from $Correlation_{Base}$to $Correlation_{Base} + 1.28$ SD), the score decreases from 50 to 0. Any score below 0 or above 100 was maintained at 0 or 100, respectively. The score was calculated immediately after the acquisition of the first volume in the feedback period (2 s). Preprocessing and score calculation were completed within 2 s. Thus, subjects received the score within 4 s after the end of the imagery periods.

To prevent learning in the baseline-estimation block, we gave the subjects a pseudo-random score, which was generated from a normal distribution having a mean of 50 and SD of 30.3. The SD was determined to restrict the appearance of a score of 0 or 100 to one time per block. At the beginning of the first training day (DAY1), we told subjects that the feedback score had been randomly determined on the previous day (BASE).

Change in score during training

We investigated the daily changes in score during the neurofeedback training. In total, each subject had 280 scores (BASE = 40 scores, DAY1–DAY4 = 60 × 4 scores). To investigate the daily changes in the score, we applied a mixed-effects model based on linear regression (Aarts et al., 2014) to the scores adopting training day (a continuous value) as a fixed effect and subject as a random effect. We used a maximum likelihood method for estimation of coefficients as implemented in the *lme4* package (https://github.com/lme4/lme4) of R version 3.2.1 (https://www.r-project.org). We calculated *p*-values using the *lmeTest* package of R.

Change in functional connectivity during training

We investigated the daily changes in FC between lM1 and lLP during the neurofeedback training in our offline analysis. The fMRI data were preprocessed with SPM8 on MATLAB. Preprocessing steps included slice-timing correction, realignment,

coregistration, segmentation of T1-weighted structural image, normalization into MNI space, and spatial smoothing with an isotropic Gaussian kernel of 8 mm full width at half maximum. BOLD signal time courses were extracted from the two ROIs (averaged across voxels). Sources of spurious variance were removed as described in "*Online calculation of feedback score"* in this section. Then, we calculated the FC as Fisher's z-transformed Pearson's correlation coefficients between the BOLD signals in the two ROIs using seven volumes during the motor imagery period in each trial, in the same fashion as the online calculation of the feedback score, i.e., the first volume was discarded and one volume from the feedback period was added as compensation for hemodynamic delay. In total, each subject had 280 functional connectivities (BASE = 40 connectivities, DAY1–DAY4 = 60 × 4 connectivities).

To compare the daily changes in FC between groups, we applied a mixed-effects model to the functional connectivities including training day (a continuous value), group, and the interaction between group and day as fixed effects, and subject as a random effect. Further, to investigate whether the changes of FC were induced in the aimed direction in each group, we applied the mixed-effects model, as a post-hoc analysis of the effect of training day on FC, separately to each group.

4.1.3 Resting-state fMRI (rs-fMRI)

To investigate the daily changes in resting-state FC between the target ROIs (lM1 and lLP) as well as in connectivity between the network-level ROIs (DMN and MVN), we measured rs-fMRI every day (Fig. 4.1A). The rs-fMRIs were measured before the neurofeedback training except for the last day (after the neurofeedback on DAY4).

Brain imaging and calculation of the resting-state functional connectivity

During the rs-fMRI measurements, subjects were instructed to keep looking at a fixation point at the center of a screen, to keep still, to stay awake, and to not think about specific things. MRI scans were obtained using a 3-T Siemens MAGNETOM Verio scanner. BOLD signals were measured using an EPI sequence (time, 10 min; TR, 2500 ms; TE, 30 ms; flip angle, 80°). The entire brain was covered in 40 axial slices (3.5 mm of thickness, no gap), voxel size was 3.3 × 3.3 × 3.5 mm, and field of view was 212 × 212 mm.

The rs-fMRI data were preprocessed with SPM8 on MATLAB. The first four volumes were discarded to allow for T1 equilibration. Preprocessing steps were same as those listed in *Change in Functional Connectivity during Training* in the section 4.1.2 "Neurofeedback training". BOLD signal time courses were extracted from the four ROIs (lM1, lLP, MVN, and DMN) and averaged across voxels in each ROI. To determine network-level ROIs (DMN and MVN), we applied a spatial independent component analysis (Calhoun et al., 2001) to rs-fMRI data from 66 subjects (12 women; mean age, 23.2 ± 2.3; age range, 20–31 years). We visually inspected MVN and DMN ROIs based on the following criteria: MVN includes bilateral primary motor cortex and supplementary motor area (Biswal et al., 1995), while DMN includes the medial prefrontal cortex, medial parietal cortex, and lateral parietal cortex (Raichle, 2010). These network ROIs correspond to ICN8 (MVN) and ICN13 (DMN) as shown in Fig. 2 of a previous study (Laird et al., 2011). To remove several sources of spurious variance, linear regression was performed, including (i) six motion parameters in addition to averaged signals over (ii) whole brain, (iii) white matter, and (iv) cerebrospinal fluid. A temporal

band-pass filter of 0.009–0.08 Hz was applied to the time series to restrict the analysis to low-frequency fluctuations that characterize rs-fMRI BOLD activity (Fox et al., 2005). Furthermore, to reduce spurious changes in FC by head motion, we calculated frame-wise displacement (FD) and removed volumes with FD > 0.5 mm, as proposed by the original article on scrubbing (Power et al., 2014). FD represents head motion between two consecutive volumes as a scalar quantity (summation of absolute displacements in translation and rotation). According to the above threshold, 3.8% (almost 9 volumes) ± 7.0% (1 SD) volumes were removed per 10 min of rs-fMRI session (240 volumes). Then, we computed the resting-state FC as Fisher's z-transformed Pearson correlation coefficients between the preprocessed BOLD signals in two target ROIs (lM1 and lLP) and in two network-level ROIs (MVN and DMN).

Change in resting-state functional connectivity

To statistically evaluate the daily changes in resting-state FC and to compare the changes between groups, we applied a mixed-effects model to the resting-state functional connectivities. We included group, training day, and the interaction between group and day as fixed effects and subject as a random effect. Further, as a post-hoc analysis of the effect of day on the resting-state functional connectivities in each group, we applied a mixed-effects model, as a post-hoc analysis of the effect of day on resting-state FC, separately to each group.

4.1.4 Cognitive tasks

To investigate the effect of the neurofeedback training on cognitive performance, subjects carried out a PVT, EFT, and CWST outside the MRI using a personal computer and keyboard before and after the entire neurofeedback training. Here, we know the direct relationship between cognitive performance and strength of the FC of MVN and DMN only in the vigilance task, in which the more the FC decreases the faster subjects react to a target and vice versa.

Cognitive task procedures

Task procedures in the current study followed those in the previous studies.

1) PVT: PVT is a task that measures the ability to sustain attentional focus. Subjects pressed a key in response to a stimulus that occasionally appeared on a screen. Subjects fixated on a centrally presented white cross on a gray background. When the cross was changed to black, subjects pressed the left arrow key with their right index finger as quickly as possible. Then, the cross changed to white again. If subjects failed to respond within 9 s, the cross automatically returned to white. Subjects performed four blocks (about 5 min × 4), each block containing 5 trials. The inter-trial interval varied from 10 to 90 s. We measured reaction time as the time from the change of the cross color to the key press. If reaction time was over 2 SD from average in each subject, the trial was excluded from further analysis. This definition of reaction time and exclusion criterion was also used in the following tasks.

2) EFT: EFT is a response inhibition test that measures the ability to suppress inappropriate responses in a particular context. Subjects fixated on a centrally presented black cross on a gray background. When five arrows (arrow direction was right or left) appeared on screen, subjects pressed the right or left arrow key as quickly as possible,

which corresponded to the direction of the central arrow in the array of 5 arrows. Subjects used their right index or middle finger to press the right or left arrow key, respectively. The task included incongruent and congruent conditions. Under the congruent condition, five arrows pointed in the same direction (e.g. <<<<<), whereas under the incongruent condition, the central arrow pointed to the opposite direction from the others (e.g. <<><<). If subjects failed to respond within 3 s, the arrows disappeared. Subjects performed two blocks (6 min × 2), each block containing 24 trials for each condition, presented in a pseudorandom order. Inter trial interval was 4.5 s. We calculated the reaction time for each condition.

3) CWST: CWST is also a response inhibition test, and it measures the ability to suppress inappropriate responses in a particular context. Subjects fixated on a centrally presented black cross on a gray background. According to the pre-task cue ("+" or "-") presented before presenting stimuli, subjects pressed the key that corresponded to the meaning or color of the presented stimulus with their right index, middle or ring finger as quickly as possible. The stimulus was word (red, blue, or yellow, in Japanese) with color (red, blue, or yellow). For example, if the pre-task cue was "+," subjects pressed the key corresponding to the meaning of the word; if the pre-task cue was "-," subjects pressed the key corresponding to the color of the word. The task included incongruent and congruent conditions. Under the congruent condition, the color of the word was the same as the meaning of the word, whereas under the incongruent condition, the color of the word was different from the meaning of the word. If subjects failed to respond within 3 s, the stimulus disappeared. Subjects performed two blocks (6 min × 2), each block containing 24 trials for each condition presented in a pseudorandom order. Inter trial interval was again 4.5 s. We calculated the reaction time for each condition.

Change in cognitive performance

To compare the changes in the reaction time of each task from pre-neurofeedback to post-neurofeedback training between the two groups, we applied a mixed-effects model to the all reaction times separately for each task. We included group, day (pre-neurofeedback and post-neurofeedback training), and the interaction between group and day as fixed effects and subject as a random effect. Further, as a post-hoc analysis of the effect of day on reaction times in each group, we applied a mixed-effects model including day as a fixed effect and subject as a random effect separately to each group in each task.

4.2 Results

4.2.1 Change in score

Figure 4.2 shows the change in score during the neurofeedback training, averaged across the blocks and subjects as a function of training day. Here, the score increases when the connectivity changes in the aimed direction for each subject group (Equation 1). We applied a mixed-effects model to the scores and examined whether a regression coefficient for the day was greater than zero. As a result, we found a significantly positive effect of day on the score (DAY: $t = 1.70$, $p = 0.044$ [one-side]). This result indicates that subjects increased their score during the neurofeedback training.

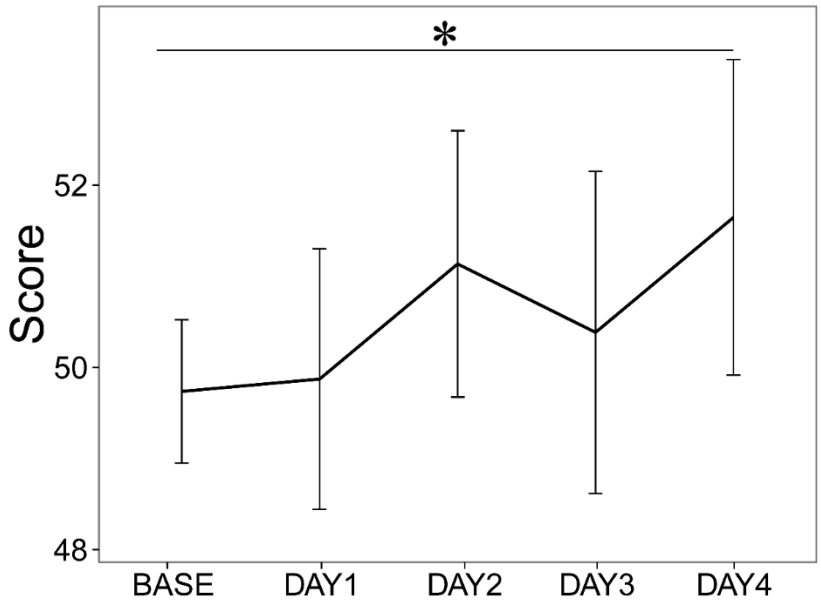

FIGURE 4.2 | Change in score during neurofeedback training.
Score averaged across subjects (n = 30) as a function of training day (error bars: standard error). A mixed-effects model identified a significant main effect of training day (t = 1.73, p = 0.044 [one-side]). *: $p < 0.05$

4.2.2 Change in functional connectivity during training

Figure 4.3 shows the change in FC between lM1 and lLP during the neurofeedback training, averaged across the blocks and subjects as a function of training day in each group. To compare the daily changes in FC between groups, we applied a mixed-effects model to the FC. As a result, we found significant effects for the day and the interaction between group and day (DAY: t = -3.22, p = 0.0012; DAY × Group: t = 3.86, p = 0.00011) but not for the group (Group: t = -0.627, p = 0.53). This suggests that the change in FC across days was different between the groups.

Since we defined the lLP ROI as a sphere with a 7.5-mm radius, one may argue that the size was so large it could include confounding noise. In our post-hoc analysis, we reduced the radius from 7.5 to 4.0 mm, and recalculated the connectivity between the ROIs. We still found significant effects for the day and the interaction between group and day (DAY: t = -3.34, p = 0.00082; DAY × Group: t = 3.82, p = 0.00013), but not for the group (Group: t = -0.834, p = 0.40).

Further, to investigate whether the changes in FC were induced in the aimed direction during the training in each group, we applied a mixed-effects model separately for each group. We found a significant effect of training day in both groups (“increased FC” group: t = 2.17, p = 0.029; “decreased FC” group: t = -3.18, p = 0.0014). Mean FC increased from -0.12 ± 0.029 at BASE to -0.061 ± 0.046 on DAY4 in the “increased FC” group and decreased from -0.11 ± 0.034 at BASE to -0.22 ± 0.044 on DAY4 in the “decreased FC” group. These results indicate that FC between lM1 and lLP during the training changed from pre-neurofeedback to post-neurofeedback training in the aimed direction in each group, i.e., the FC increased in the “increased FC” group and decreased in the “decreased FC” group.

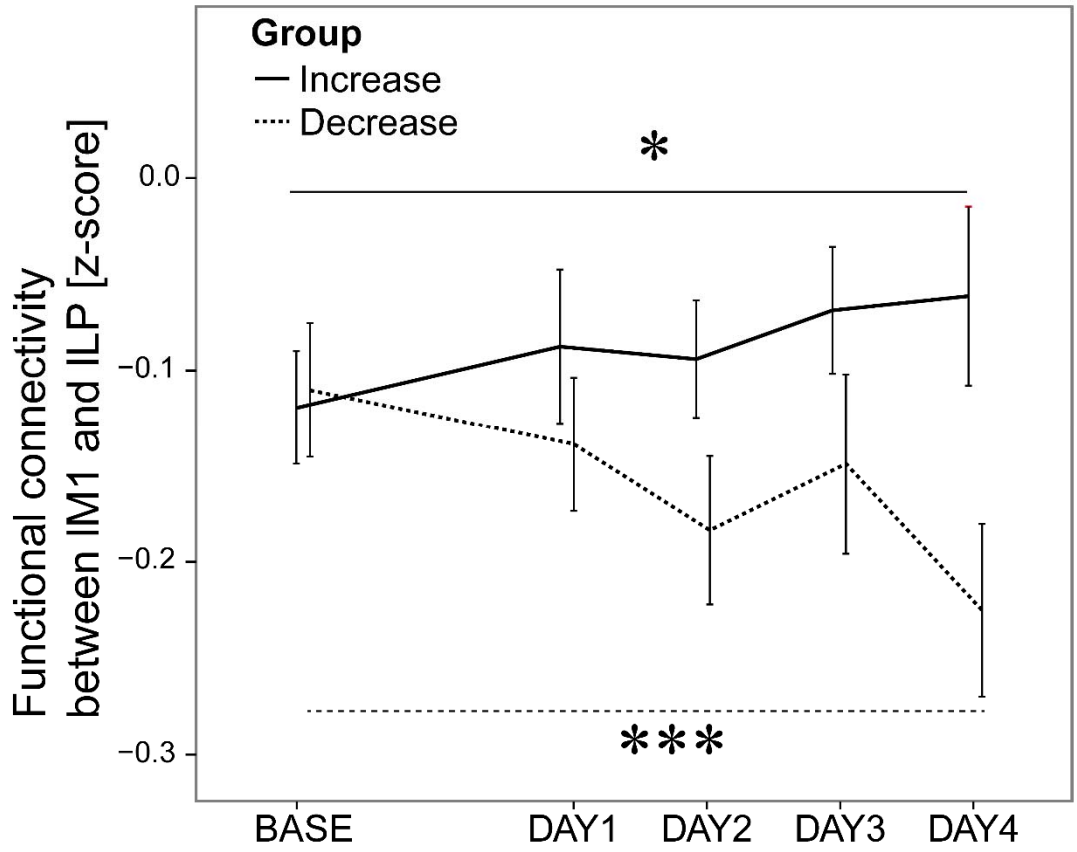

FIGURE 4.3 | Change in functional connectivity between the left primary motor area (lM1) and the left lateral parietal region (lLP) during neurofeedback training.

Functional connectivity (FC) averaged across subjects in each group (solid line: increased FC group, $n = 18$; broken line: decreased FC group, $n = 12$) as a function of training day (error bars: standard error). A mixed-effects model identified a significant interaction between group and day ($t = 3.86$, $p = 0.0014$). ***: $p < 0.005$, *: $p < 0.05$ according to post-hoc analysis of the main effect of day for each group.

4.2.3 Change in resting-state functional connectivity

To compare the daily changes in resting-state FC between the two groups, we applied a mixed-effects model to the resting-state functional connectivities between lM1 and lLP and to the connectivity between MVN and DMN. We did not find any significant effect in any connectivity (for lM1–lLP, DAY: $t = 0.41$, $p = 0.67$; Group: $t = 1.28$, $p = 0.20$; DAY × Group: $t = -0.78$, $p = 0.43$; for MVN-DMN, DAY: $t = 1.20$, $p = 0.23$; Group: $t = -0.11$, $p = 0.91$; DAY × Group: $t = 0.89$, $p = 0.37$). Further, to investigate the change in connectivity across days in each group, we applied a mixed-effects model separately for each group. Consequently, we found a significant effect of training day in the "increased FC" group for the connectivity between MVN and DMN ("increased FC" group: $t = 2.93$, $p = 0.0045$; "decreased FC" group: $t = 1.18$, $p = 0.24$), but not for the connectivity between lM1 and lLP ("increased FC" group: $t = -0.70$, $p = 0.48$; "decreased FC" group: $t = 0.43$, $p = 0.66$). Specifically, FC between MVN and DMN increased from -0.26 ± 0.058 at BASE to -0.13 ± 0.064 on DAY4 in the "increased FC" group. Therefore, the direction of change in FC between MVN and DMN in the "increased FC" group was consistent with our aimed direction, the connectivity change in the neurofeedback sessions (Fig. 4.3), and

that in a previous study (Megumi et al. 2015).

4.2.4 Change in cognitive performance

Figure 4.4 shows the changes in reaction time from the pre-neurofeedback to post-neurofeedback training stage averaged across subjects in each group. Note that there was no significant difference in reaction time or error rate between the two groups for the pre-neurofeedback training stage (Appendix Table B.1). Owing to the absence of significant differences in task performance before the training, we show only the changes in reaction time in Fig. 4.4. However, the following statistical analyses were applied to raw reaction-time data without subtraction or averaging. To compare the changes in reaction time from the pre-neurofeedback to post-neurofeedback training stage between the two groups, we applied a mixed-effects model to the reaction times in each task. As a result, the interaction effect between group and day was significant in PVT, EFT *congruent*, and CWST *congruent* (PVT: t = -2.72, p = 0.0065; EFT *congruent*: t = 2.41, p = 0.016; CWST *congruent*: t = -2.67, p = 0.0075), but not in EFT *incongruent* or CWST *incongruent* (EFT *incongruent*: t = 1.18, p = 0.23; CWST *incongruent*: t = -0.50, p = 0.61). These significant interactions suggest that the changes in reaction time from the pre-neurofeedback to post-neurofeedback training stage were different between the groups.

Further, we applied a mixed-effects model to the reaction times separately for each group in PVT, EFT *congruent*, and CWST *congruent*. The main effect of training day was significant in the "increased FC" group in PVT and CWST *congruent* (PVT: t = -3.85, p = 0.00013; EFT *congruent*: t = 0.58, p = 0.56; CWST *congruent*: t = 6.93, p < 0.0001) and in the "decreased FC" group in EFT *congruent* and CWST *congruent* (PVT: t = 0.12, p = 0.90; EFT *congruent*: t = -2.52, p = 0.011; CWST *congruent*: t = 8.53, p < 0.0001). These results indicate that the change in the reaction time from the pre-neurofeedback to post-neurofeedback training stage could be identified in the "increased FC" group in PVT and CWST *congruent* and in the "decreased FC" group in EFT *congruent* and CWST *congruent*. Although some of these main effects of day might have been affected by repetition of the same task, the interaction effects between the groups could not be explained by such repetition.

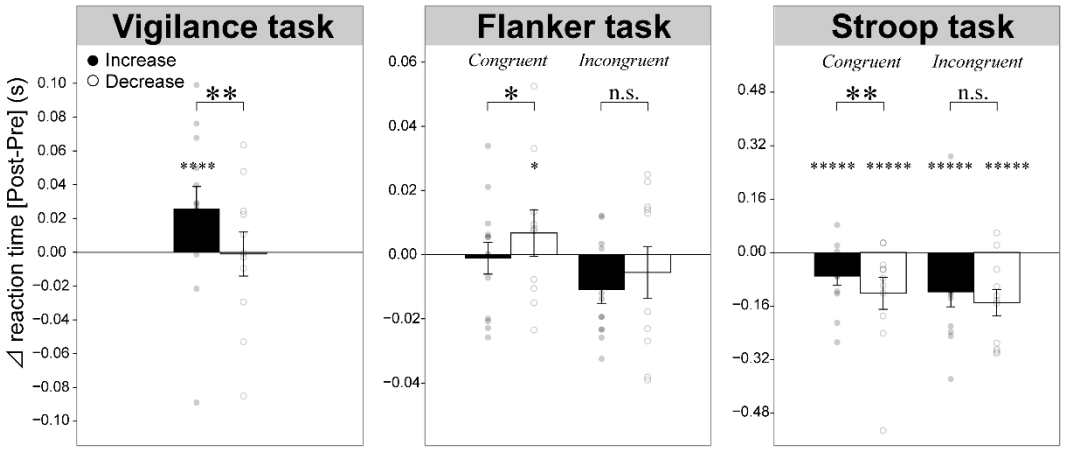

FIGURE 4.4 | Changes in cognitive performance from pre-neurofeedback to post-neurofeedback training.

All panels show changes in reaction time of the increased (n = 13) and decreased (n = 11) FC groups. Each marker (black or white circle) represents one subject; each bar represents each group's average (error bars: standard error). A mixed-effects model identified a significant interaction between group and training day (Vigilance: t = -2.72, p = 0.0065; Flanker *congruent*: t = 2.41, p = 0.016; Stroop *congruent*: t = -2.67, p = 0.0075). Among these tasks, in which interaction effects were significant, the main effect of day was significant in the increased FC group in Vigilance and Stroop *congruent* (Vigilance: t = -3.85, p = 0.00013; Stroop *congruent*: t = 6.93, p < 0.0001) and in the decreased FC group in Flanker *congruent* and Stroop *congruent* (Flanker *congruent*: t = -2.52, p = 0.011; Stroop *congruent*: t = 8.53, p < 0.0001). *****: p < 0.0001, ****: p < 0.0005, **: p < 0.01, *: p < 0.05, n.s.: not significant.

4.3 Discussion

Using the connectivity neurofeedback training method, we experimentally manipulated the FC between lM1 and lLP and examined the change in performance from the pre-neurofeedback to post-neurofeedback training stage. The FC in each group indeed changed in the aimed direction during the training (Fig. 4.3). Furthermore, we identified significant change in some cognitive performances between the groups (Fig. 4.4). These findings indicate that connectivity neurofeedback can induce the aimed direction of change in FC as well as induce a differential change in cognitive performance.

4.3.1 Directions of change in reaction times dependent on the tasks

We found significant change in some cognitive performances, but the directions of change in reaction time were different for each task. For example, the reaction times of the vigilance task increased in the "increased FC" group but the reaction times of the flanker task (*congruent*) increased in the "decreased FC" group.

Regarding the EFT, a previous study (Kelly et al., 2008) on brain–behavior relationships investigated the coefficient of variation (standard deviation divided the by mean) as an index of task performance. Thus, we could not predict the direction of change in reaction time for EFT. However, Kelly et al. (2008) showed that the FC between MVN and DMN is positively correlated with coefficient of variation. Our additional analysis indicated that the directions of change in coefficient of variation were consistent with those in the previous study, i.e., coefficient of variation increased from pre-neurofeedback to post-neurofeedback training in the "increased FC" group and decreased in the "decreased FC" group, although their interaction effect did not reach a statistically significant level (Appendix B.1).

Regarding the CWST, a previous study (Liu et al., 2015) on brain–behavior relationships investigated the Stroop effect (mean reaction time of incongruent condition – mean reaction time of congruent condition) as an index of task performance. However, because they investigated the relationship between the regional homogeneity and Stroop effect, we also could not predict the direction of change in reaction time. Our additional analysis indicated that the change in Stroop effect was not significant (Appendix B.1). Furthermore, from Fig. 4.4, we can easily assume that there are considerable learning effects on the reaction time in both groups, since the Stroop task may be more difficult than the other two tasks. However, even if there were learning effects, the significant

interaction between group and day suggests that the reaction times were significantly and differentially changed from pre-neurofeedback to post-neurofeedback training between the two groups. However, we could not conclude that neurofeedback training influenced the reaction time itself or the learning effect in the Stroop task (*congruent*).

Unlike the other two tasks, previous studies found a concrete relationship between reaction time in the vigilance task and FC (Thompson et al., 2013). Thompson et al. (2013) divided their subjects into two groups according to fast or slow reaction time for PVT. They reported that the fast reaction time group showed more greatly decreased negative resting-state FC between MVN and DMN than did the slow reaction time group. This finding is consistent with our results. Hinds et al. (2013) examined fMRI activities of MVN and DMN during PVT. Their subjects could rapidly respond to a stimulus when activity in a part of MVN (the supplementary motor area) increased and activity in DMN decreased (Hinds et al., 2013). It is assumed that DMN is more active than MVN when subjects are waiting for a stimulus in PVT, whereas MVN becomes more active than DMN when subjects respond to the stimulus. Based on these studies, as well as our own, one could hypothesize that subjects who have a more negative resting-state FC could more rapidly enhance MVN activity and suppress DMN activity, which leads to a shorter reaction time. However, further evidence is needed to verify this hypothesis.

4.3.2 Difference in behaviors during training between subject groups

In the neurofeedback experiment, only the rewarded direction of change in FC was flipped during the neurofeedback training between the "increased FC" and "decreased FC" groups. Nevertheless, the changes in the reaction time of some tasks from the pre-neurofeedback to post-neurofeedback training stage differed between the two groups (Fig. 4.4). This suggests that the change in FC influences the change in cognitive performance. However, factors other than the rewarded direction, which cannot be experimentally controlled, may have differed between the two groups and caused the difference in the change in cognitive performance. We examined these factors as follows.

Total score during training: The score was calculated according to the equivalent formula for the two groups (see "*Online calculation of feedback score"*). The resulting total score may have differed between the groups, and this difference may have caused a difference in their motivation during training and thus a change in cognitive performance. Therefore, we compared the total score over the training in the increase group with the score in the decrease group (Appendix B.2). The total score averaged across subjects was 14379 ± 809 (mean ± 95% confidence interval) for the "increased FC" group and 13860 ± 1448 for the "decreased FC" group. We used a two-sample *t*-test to compare the total score but did not obtain a significant difference between the groups ($t = 0.68$, $p = 0.50$).

Strategies adopted by subjects: We provided identical instructions to both groups, telling them that the disc (score) becomes bigger as subjects improve at producing the tapping imagery during the training. However, the actual strategies that were adopted by the subjects may have differed between the groups through their trial-and-error learning. The difference in strategy may have caused a difference in regional brain activity and thus affected subsequent cognitive performance. We conducted a post-experiment debriefing with 25 of the 30 subjects (13 subjects in the increase group and 12 subjects in the decrease group) and examined the differences in their strategy for the motor imagery

(Appendix B.3 and Appendix Table B.2). We analyzed the reported strategies in five aspects: image category (items: kinesthetic, visual, or both), hand laterality (left, right, or both hands), tapping sequence (fixed or random), imagery with or without manipulated object (e.g., a subject imagined typing on a computer keyboard), and imagery with or without a rhythm (Appendix Table B.3). We counted the numbers of items across subjects and compared these numbers between the groups. We calculated the *p* value as the probability that these results would be obtained if we separated subjects randomly (Appendix Table B.4). As a result, there was no significant difference in the numbers between the groups.

4.3.3 Difference in the activity of target ROIs during training between the groups

Because the aim of the current study is to control the FC but not to control the averaged activity in a specific ROI, it is important to check the changes in activities in the target regions. We applied a mixed effects model to averaged activity in each target region in the same manner as our analysis of FC (see 4.2.2 "Change in functional connectivity during training" and Appendix B.4). As a result, we found a significant interaction effect of day on the lLP activation. This result may indicate that subjects altered only the activity in the lLP and that the change in temporal correlation between the target regions is an epiphenomenon. However, this was not the case owing to the following reason. We investigated whether subjects could get the information about the activity in the lLP from the feedback score to calculate the correlation between the feedback score and the activity in lLP (Appendix B.5). If there is no correlation between the feedback score and the brain activity in lLP, subjects could not have directly altered the activity in lLP through the training. As a result, we did not find a significant correlation between the feedback score and activity in the lLP ("increased FC" group: $r = 0.016$, $p = 0.23$; "decreased FC" group: $r = -0.01$, $p = 0.54$). These results indicate that subjects could not get information about activity in the lLP from the feedback score. Therefore, subjects altered the FC between the lLP and lM1, and the activity in the lLP might have been collaterally altered.

4.3.4 Change in resting-state functional connectivity

Our previous study (Megumi et al., 2015) showed the significant increase in the resting-state FC between the target ROIs (lM1 and lLP) from pre-neurofeedback to post-neurofeedback training. However, our current study failed to observe a significant change in the resting-state FC between the target ROIs. A possible reason is that the effect of neurofeedback training may have been smaller than in our previous study. In fact, our previous study showed an increase of about 0.2 in correlation between the two target ROIs during the training in comparison to about 0.1 in our current study. This change in correlation between ROIs might have been insufficient for generalization of the training effect from the training to rest periods.

By contrast, at the network level, we found a significant increase in resting-state FC between MVN and DMN from pre-neurofeedback to post-neurofeedback training despite the smaller effect of neurofeedback training than that in our previous study. A possible reason is the difference in the number of voxels between ROI and network analyses: network-level ROIs have more voxels (about 5000 voxels) than target ROIs (lM1 and lLP: about 100 voxels). Correlation calculated from signal time courses averaged over the larger number of voxels is more reliable than that from smaller number of voxels in most

cases. This may have helped us find a significant increase in resting-state FC between the two network-level ROIs. However, we did not observe a significant decrease in the resting-state FC between MVN and DMN in the "decreased FC" group. Because the connectivity is negative between MVN and DMN in nature, further decreasing the negative connectivity may be difficult (e.g., changing correlation from r = -0.4 to -0.6) in comparison to increasing it (e.g., from r = -0.4 to -0.2). In fact, we confirmed that the distribution of the FC between DMN and MVN is positively skewed (skewness = 0.68), suggesting that probability of a decrease is less than that of an increase in correlation.

4.3.5 Effect of the initial functional connectivity on training

We examined whether the differential changes in FC and cognitive performances between the two groups were induced by the difference in initial FC (Appendix B.6). At group level, we did not observe a significant difference in the initial FC (lM1-lLP) between the two groups (t-test, t = 0.20, p = 0.84). Thus, the initial difference unlikely explains the differential changes between the two groups. At individual level, we did not find any significant correlations between the initial FC and the change in FC and cognitive performances (see Appendix B.6 for details). These results indicate that the change in FC and cognitive performances were not induced by the difference in the initial FC between the two groups.

4.3.6 Associations among change in functional connectivity during training, change in resting-state functional connectivity, and change in cognitive performance

We examined the associations among 1) the changes in FC of lM1-lLP during neurofeedback training, 2) the changes in resting-state FC of MVN-DMN, and 3) the changes in cognitive performance of the three tasks, in which the interaction between groups and days yielded significant effects. We analyzed data of the "increased FC" group, in which a significant change in resting-state connectivity of MVN-DMN was observed. Using linear regression, we conducted a moderation/mediation analysis. This displayed a significant effect of the change in FC during training on the change in reaction time of CWST *congruent* (β = -1.41, SE = 0.61, t = -2.29, p = 0.044, adjusted R^2 = −0.22) (Appendix B.7 and Appendix B.1). This result suggests that the change in reaction time of CWST *congruent* was directly affected by changes in the FC during training rather than by changes in the resting-state FC. However, our moderation/mediation analysis shed light on only a fraction of many factors related to the connectivity neurofeedback. Further studies are required to verify the robust relationship between cognitive function and FC.

4.3.7 Application of connectivity neurofeedback training

Disturbances in regional or brain-wide FC have been reported for numerous neurological and psychiatric diseases (Broyd et al., 2009; Fornito et al., 2015; Fox and Raichle, 2007; Stam, 2014). These pathological disturbances have been related to the severity of cognitive dysfunctions in individual patients (Hawellek et al., 2011; He et al., 2007; Yahata et al., 2016). From this perspective, online fMRI neurofeedback (Sulzer et al., 2013) is expected to become a next-generation therapeutic tool (Esmail and Linden, 2014; Stoeckel et al., 2014)(Decoded Neurofeedback Project within the Strategic Research Program for Brain Sciences [SRPBS]: https://bicr.atr.jp/decnefpro/). In the future,

connectivity neurofeedback training methods may contribute to a remedy for such disturbances and to improvement of impaired cognitive functions by regulating the FC rather than only the level of regional brain activity, as traditionally implemented by most neuromodulation techniques such as single-ROI-based neurofeedback, transcranial magnetic stimulation, and deep brain stimulation.

Our current study shows that connectivity neurofeedback can not only increase but also decrease FC. Therefore, connectivity neurofeedback shows potential for future therapeutic interventions against psychiatric and neurological disorders caused by not only hyper-connectivity but also hypo-connectivity. For example, in patients with Alzheimer's disease, FC is reduced between the right hippocampus and many component regions of the DMN, while connectivity increases between the left hippocampus and the right dorsolateral prefrontal cortex (Broyd et al., 2009). In patients with depression, FC is increased between the subgenual cingulate cortex and the DMN (Broyd et al., 2009; Greicius et al., 2007). In autism spectrum disorders, connectivity is reduced between the anterior and posterior DMN regions (Broyd et al., 2009).

Furthermore, our study suggests the possibility of developing a technique of neurofeedback manipulation to cancel out the behavioral change induced by previous methods of neurofeedback manipulation. This is important for ensuring safeguards in clinical applications of connectivity neurofeedback.

4.3.8 Summary

In this chapter, using the connectivity neurofeedback training method, we tested the hypothesis that connectivity neurofeedback can induce the aimed direction of change in FC and cognitive performance. As a result, subjects could increase or decrease the FC between two brain regions, and cognitive performance was significantly and differentially changed from pre-neurofeedback to post-neurofeedback training between the two groups. We did not find a significant difference in behaviors between the groups during the training, except for the rewarded direction of change in FC between the two regions. These findings suggest that connectivity neurofeedback can induce the aimed direction of change in FC as well as a change in cognitive performance.

Chapter 5

Conclusion and Future Directions

This thesis demonstrated three types of research to expand the bottlenecks that prevent us to develop the application of fMRI resolving the clinical and cognitive problems. This thesis is here summed up by describing its contributions and the remaining challenges for future translational fMRI study.

5.1 Main contributions of this thesis

The chapter 2 demonstrated that site differences are composed of biological sampling bias and engineering measurement bias by utilizing a traveling-subject dataset in conjunction with a multi-site, multi-disorder dataset. The effects on resting-state functional connectivity because of both bias types were greater than or equal to psychiatric disorder effects. Furthermore, our findings indicated that each site can sample only from a subpopulation of participants. This result suggests that it is essential to collect large neuroimaging data from as many sites as possible to appropriately estimate the distribution of the grand population. And also, we developed a state-of-the-art harmonization method for multi-site rs-fMRI data by using traveling-subject dataset and achieved the reduction of the measurement bias. Since development of an accurate harmonization method enable us to analyze large multi-site multi-disorder dataset, it promotes discovery science in cognitive neuroscience field.

In the chapter 3, we constructed a reliable neuroimaging-based classifier and regression model for MDD and BDI score, respectively, by investigating whole-brain resting-state functional connectivity patterns. The MDD classifier and BDI regression model generalized to an independent validation dataset obtained at different imaging sites. We found an approximately 30% overlap in functional connections related to depressed symptoms and MDD diagnosis. These functional connections were associated with the salience and default mode networks. Our study revealed a partially overlapping relationship between the biological basis of depressed symptoms and MDD diagnosis. Our study would make a significant contribution to the elucidation of the neurological basis of MDD and future development of a theranostic biomarker that contribute to not only diagnosis, but also determination of therapeutic targets in depressive symptoms. That is, since overlap connections related to MDD diagnosis and depressed symptoms, we can diagnose MDD by observing these connections and these connections may express disease state and intervention to these connections would cause improvement of depressed symptoms.

In the chapter 4, we investigated the hypothesis that connectivity neurofeedback can induce the aimed direction of change in functional connectivity, and the differential change in cognitive performance according to the direction of change in connectivity. We showed evidence that connectivity neurofeedback can induce the aimed direction of change in connectivity and a differential change in cognitive performance depending on the direction of the change in connectivity. Our results provide experimental evidence to the theory that manipulating brain networks lead a change in cognitive function.

As a translational fMRI study, these results would provide one possible framework of therapeutic intervention for psychiatric disorder using fMRI. First, a theranostic biomarker of functional connectivity is discovered by analyzing large-scale multisite rs-fMRI data which can be enabled by harmonization method. Therapeutic intervention for patients with disorder would be done by modify that functional connectivity by using connectivity neurofeedback.

5.2 Challenges for the future

Although this thesis would contribute to translational fMRI study, we have several important challenges to make fMRI truly useful in real-world.

5.2.1 Challenges in the data-driven approach

In the data-driven approach, if we can achieve the following things, it could be truly useful for clinical application of psychiatric disorder.

1. *Redefinition of psychiatric disorder based on biological neural basis.* An increasing number of studies have pointed out difficulty in finding a clear association between existing clinical diagnostic categories and neurobiological abnormalities (Clementz et al., 2016; Insel and Cuthbert, 2015; Singh and Rose, 2009). Therefore, redefinition of psychiatric disorder based on biological neural bases have been focused. This may be feasible by applying unsupervised learning techniques such as clustering analysis to large multi-site multi-disorder neuroimaging dataset.

2. *Achievement of precision medicine.* Precision medicine is treatment targeted to the needs of individual patients on the basis of genetic, or biological neural bases that distinguish a given patient from other patients with similar clinical presentations. This may be feasible by acquiring longitudinal data which include neuroimaging data before and after treatment such as medication or cognitive behavioral therapy.

5.2.2 Challenges in the brain-manipulation approach

In the brain-manipulation approach, if we can achieve the following things, fMRI neurofeedback training could have truly useful clinical intervention for psychiatric disorder.

3. *Placebo-controlled experiment.* The effectiveness of fMRI neurofeedback training is still under discussion; we must therefore conduct more double-blind, placebo-controlled, randomized neurofeedback studies. There are thus far still only two papers which show the effectiveness of neurofeedback training in double-blind, placebo-controlled, randomized conditions (Vincent et al., 2007; Young et al., 2017; Young et al., 2018).

4. *Optimization of experiment protocol.* Since the current experiment protocol often requires one full week of neurofeedback training, the burden on the patients is large. As it is yet unclear which approach is appropriate for clinical application, more work is needed to provide evidence-based guidelines (Stoeckel et al., 2014; Sulzer et al., 2013; Yahata et al., 2017; Yamada et al., 2017).

5. *Elucidation of neural mechanism in neurofeedback.* Why can fMRI neurofeedback training cause changes in the brain, despite the fact that fMRI neurofeedback manipulates the BOLD signal rather than neural activity itself? (Sitaram et al., 2017; Watanabe et al., 2017) This question may be answered in non-human electrophysiological studies by investigating the change in neural activity when the BOLD signal is altered.

6. *Elucidation of the relationship between brain networks and cognitive function.* Even if neurofeedback training can change the connectivity within brain networks, it is difficult to apply this knowledge to the real-world without better understanding of how brain networks give rise to cognitive functions. This may be clarified by constructing and verifying a mathematical model of cognitive function based on brain networks (Anzellotti and Coutanche, 2018; Ezaki et al., 2017; Ito et al., 2017; Mill et al., 2017; Tompson et al., 2018).

Finally, in translational studies, we tend to conduct research from the viewpoint of "being useful for the real-world". However, I would like to continue my research without forgetting that scientists should also be responsible for elucidating the mechanisms behind phenomena, in order to prevent a repeat of past mistakes such as the era of the lobotomy.

Appendix A

Appendix of Chapter 2

A.1 Magnitude distribution of both biases and each factor on functional connectivity

To quantitatively evaluate the effect of measurement and sampling biases on functional connectivity (FC), we compared the magnitudes of both types of bias with the magnitudes of psychiatric disorders and participant factors. For this purpose, we investigated the magnitude distribution of both biases, as well as the effects of psychiatric disorders and participant factors on FC overall 35,778 elements in a 35,778-dimensional vector to see how many functional connectivities were largely affected. Figure A.1a: the *x*-axis shows the magnitude as Fisher's *z*-transformed Pearson's correlation coefficients, while the *y*-axis shows the density of the number of connectivities. Figure A.1b shows the same data, except the *y*-axis represents the log-transformed number of connectivities for better visualization of small values. There were significant differences among biases and factors for larger magnitudes near the tails of their distributions. For example, the number of connectivities, which was largely affected (i.e., a magnitude larger than 0.2), was more than 100 for the participant factor, approximately 100 for measurement bias, and nearly 0 for all sampling biases, as well as all disorder factors.

A.2 Field map correction

We investigated the effect of field-map correction on data harmonization (Hutton et al., 2002; Jenkinson, 2003; Jezzard and Balaban, 1995). We used SPM12 for field-map correction, in accordance with the SPM protocol. A total of 35,778 functional connections were calculated from echo-planar images (EPIs) following field-map correction. Participant factors and measurement biases were estimated by fitting the regression model to the traveling-subject dataset only. The regression model can be described as follows:

$$Connectivity = const + \mathbf{x_p}^{\mathrm{T}}\boldsymbol{p} + \mathbf{x_m}^{\mathrm{T}}\boldsymbol{m} + e,$$

$$\text{such that } \sum_j^9 p_j = 0, \sum_k^{12} m_k = 0.$$

To evaluate the spatial effect of field-map correction on various brain regions, we visualized the difference in the effect on each ROI between datasets with and without field-map correction. We also visualized the effect of measurement bias on each ROI using data subjected to field-map correction (Figure A.2a). We calculated the standard deviation of the measurement bias and the participant factor and compared the results between datasets with and without field-map correction (Figure A.2b). Furthermore, we performed hierarchical clustering analysis of measurement bias in the dataset subjected to field-map correction (Figure A.2c). Figures A.1a and A.1b demonstrate that field-map correction remarkably reduced the effect of measurement bias in the cerebellum and lower regions of the frontal cortex, while also increasing the effect of the participant

factor. However, the presence of the cluster for phase-encoding direction in Figure A.2c indicates that field-map correction did not completely eliminate the influence of the difference in the phase-encoding direction.

A.3 Selection of the regularization hyper-parameter lambda

Because the design matrix of the regression model was rank-deficient, *L2* regularization was applied when estimating each type of bias and factor. When regularization was not applied, we observed spurious anti-correlation between measurement bias and sampling bias for healthy controls, as well as spurious correlation between sampling bias for healthy controls and sampling bias for patients with psychiatric disorders (Figure A.3a, left). These spurious correlations can even be observed in the permutation data, in which there were no associations between site label and data (Figure A.3a, right). This suggests that the spurious correlations were caused by the rank-deficient property of the design matrix. We utilized the hyper-parameter lambda to minimize the absolute mean of these spurious correlations (Figure A.3c, left). We confirmed that the values of lambda for the real data were almost identical to those for the permutation data (Figure A.3c, right). Furthermore, although the ability to explain the data decreases when using regularization, we confirmed that the degree of decrease due to regularization was less than 1% (Figure A.3d).

A.4 Brain regions contributing the measurement bias of each site

To evaluate the spatial distribution of the measurement bias of each site in the whole brain, we utilized the same method in figure 3.3 in the main text. We projected connectivity information to anatomical regions of interest (ROIs). Figure A.4 shows the relative contribution of individual ROIs to measurement bias of each site in the whole brain.

A.5 Classifiers for MDD and SCZ, based on the four harmonization methods

To quantitatively evaluate the harmonization method, we constructed biomarkers for psychiatric disorders using the SRPBS multi-disorder dataset, which distinguishes between HCs and patients, based on resting-state FC. We compared four different harmonization methods for the removal of site difference from the SRPBS multi-disorder dataset: (1) by using a traveling-subject method; (2) by using a ComBat method; (3) by using a GLM method; and (4) by using an Adjusted GLM. We also compared these four methods to the non-harmonization method (Raw method).

We aimed to focus on multi-site data; therefore, we targeted data from patients with MDD and SCZ who were sampled from multiple sites. To construct each classifier, a machine-learning technique was applied to (1) whole brain FCs for HCs and patients with MDD from the SRPBS multi-disorder dataset (425 HCs from nine sites and 135 patients with MDD from five sites; Table 1 in the main text) or (2) whole brain FCs for HCs and

patients with SCZ from the SRPBS multi-disorder dataset (425 HCs from nine sites and 44 patients with SCZ from three sites; Table 1 in the main text). Based on our previous results, we assumed that disorder factors were not associated with whole-brain connectivity, but with a specific subset of connections (Yahata et al., 2016). Therefore, we conducted logistic regression analyses using the least absolute shrinkage and selection operator (LASSO) method to select the optimal subset of functional connections from among 35,778 connections. A logistic function was used to define the probability of a participant belonging to the MDD (or SCZ) class, as follows:

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

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

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

in which $\|\boldsymbol{w}\|_1 = \sum_i^N |w_i|$ and λ represent hyper-parameters that controls the amount of shrinkage applied to the estimates. To estimate λ properly, we used the "*lassoglm*" function in MATLAB (R2016b, Mathworks, USA) and set "NumLambda" = 25 and "CV" = 10.

Because the SRPBS multi-disorder dataset is unbalanced with regard to the numbers of patients and HCs, we used the under-sampling (Wallace et al., 2011) and resampling method to develop and evaluate classifiers. To train the MDD classifier, 130 patients with MDD and 130 HCs were randomly sampled from the dataset, and classifier performance was tested among the remaining participants. For the SCZ classifier, data were randomly sampled for 40 patients with SCZ and 40 HCs. We calculated the area under the curve (AUC), accuracy, sensitivity, and specificity. Furthermore, to properly evaluate classifier performance even for the unbalanced dataset, we calculated the Matthews correlation coefficients (MCC) (Chicco, 2017; Matthews, 1975b) as indicators of classifier performance. MCC correctly takes into account the ratio of the confusion matrix size. Especially in unbalanced datasets, the MCC is able to identify whether prediction is proceeding appropriately, whereas accuracy is not. Under-sampling is disadvantageous in that it does not allow the classifier to learn using the excluded data. Therefore, to ensure all participants were used during classifier training, we repeated the aforementioned procedure 100 times (i.e., resampling), and the average value of classifier performance was considered indicative of classifier performance in the training dataset. Figs. A.5a and A.5b show the classifier performances and the left panels of Figs. A.6a and A.6b depict the distribution of the average probability for a given disorder. The generalizability of the models was tested using parts of the dataset for the completely independent validation cohort obtained from the following sites: Hiroshima Rehabilitation Center (HRC) and Yamaguchi University (UYA) (47 HCs and 12 patients with MDD from HRC; 117 HCs and 76 patients with MDD from UYA) for the MDD classifier; Kyoto University Trio (KTT) for the SCZ classifier (61 HCs and 36 patients with SCZ from KTT) (see Tables A.2, A.3 and A.5). Since we created 100 classifiers using the training data, we entered the independent cohort data into all trained classifiers and averaged the resultant probability

values. For each participant, when the average probability was greater than 0.5, the diagnostic class label *y* of this participant was set equal to 1 (MDD or SCZ, $y = 1$; HC, $y = 0$). Classifier performance in the independent cohort is shown in Figs. A.6c and A.6d and the right panels of Figs. A.6a and A.6b depict the distribution of the average probability for a given disorder. Further details of classifier performance are presented in Tables A.8– A.11.

Classifier output was defined as the probability of a participant being categorized into the MDD or SCZ class. Diagnostic probability values greater than 0.5 were considered indicative of a psychiatric disorder. The distribution of diagnostic probability in the training dataset (left panels in Figs. A.6a and A.6b) revealed that patients with psychiatric disorders and HCs were clearly separated by a threshold of 0.5 (the middle line in each panel) for all methods. By contrast, the distribution of diagnostic probability in the independent cohort (right panels in Figs. A.6a and A.6b) revealed that patients with psychiatric disorders and HCs were separated by a threshold of 0.5 only for the traveling-subject and ComBat methods. No such separation was observed for the other methods because of the leftward shift of the distributions for HCs and patients. Thus, in the independent cohort, very low sensitivity (below 0.5) and unduly high specificity were observed for the GLM and adjusted GLM methods, whereas medium sensitivity (approximately 0.5) and unduly high specificity were observed for the raw methods for the MDD and SCZ classifiers (Figs. A.6c and A.6). Unduly high specificity was achieved because patients and HCs were indifferently classified as HCs. This result indicates that the GLM and adjusted GLM methods are unable to remove site differences and may even negatively impact classification (Fortin et al., 2018; Rao et al., 2017).

We next compared the generalizability of the traveling-subject, ComBat, and raw methods. Classifier performance was evaluated using MCC and the traveling-subject method was superior to the raw method for MDD and SCZ classifiers. The ComBat method was also superior to the raw method for the MDD classifier but inferior to the raw method for the SCZ classifier. The index values for the MDD classifier were as follows for MCC: 0.376 (ComBat method) > 0.348 (traveling-subject method) > 0.267 (raw method). The values for the SCZ classifier were as follows for the MCC: 0.520 (traveling-subject method) > 0.474 (raw method) > 0.400 (ComBat method). In the traveling-subject and ComBat methods, the threshold of 0.5 was nearly correctly set at the approximate intersection between the HC and patient distributions, whereas the threshold was shifted slightly rightward in the raw method. These results indicate that harmonization of the SRPBS multi-disorder dataset, based on the traveling-subject and ComBat methods, outperformed other harmonization methods with regard to classifier generalizability. However, because the ComBat method was inferior to even the raw method for the SCZ classifier, ComBat may not be appropriate for certain datasets.

A.6 Regression models for age based on the four harmonization methods

To further investigate the effectiveness of the harmonization methods, we constructed regression models to predict participant age using four different harmonization methods, as well as the raw method, and compared prediction performance among the models. To construct each regression model, a machine-learning technique was applied to whole brain FCs from HCs (425 HCs from nine sites; Table 1 in the main text). We then

employed linear regression using the LASSO method, as follows:

$$y_{sub} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{c}_{sub},$$

in which y_{sub} represents the age of the participant, $\boldsymbol{c}_{sub}$ represents a FC vector of the participant, and $\boldsymbol{w}$ represents the weight vector of the linear regression. The performance of linear regression in the training data was evaluated via 10-fold cross validation. We calculated the MAE and Pearson's correlation coefficients between predicted age and actual age. The generalizability of the models was examined using parts of the completely independent validation cohort dataset obtained from the ATR TimTrio, ATR Verio, and ATR Prisma sites (223 HCs, Tables A.4 and A.5). This dataset is collected for different purpose from this study and contains subjects from a wide range of ages (20–69). We developed 10 linear regression models using the training data (each cross validation); therefore, we entered the independent cohort data into all trained linear regression models. The average output was regarded as the predicted age. We also calculated the MAE and Pearson's correlation coefficients between the predicted age and the actual age. Figure A.7 shows the scatter plot of the actual age and the predicted age in the independent cohort. Figure A.5c shows the same plot in the training dataset.

The ComBat method achieved the lowest mean absolute error (MAE) value and the highest *r* value, whereas the traveling-subject method achieved the second lowest MAE value and the second highest *r* value (Fig. A.7). Furthermore, the MAE values of the ComBat and traveling-subject methods were significantly lower than that of the raw method (two-tailed paired *t*-test; ComBat: $p = 3.1 \times 10^{-20}$, $t = -10.18$, $df = 222$; traveling-subject: $p = 6.3 \times 10^{-8}$, $t = -5.5$, $df = 222$). These results indicate that the ComBat and traveling-subject methods outperformed the other harmonization methods for constructing a regression model to predict a participant's age with regard to generalizability.

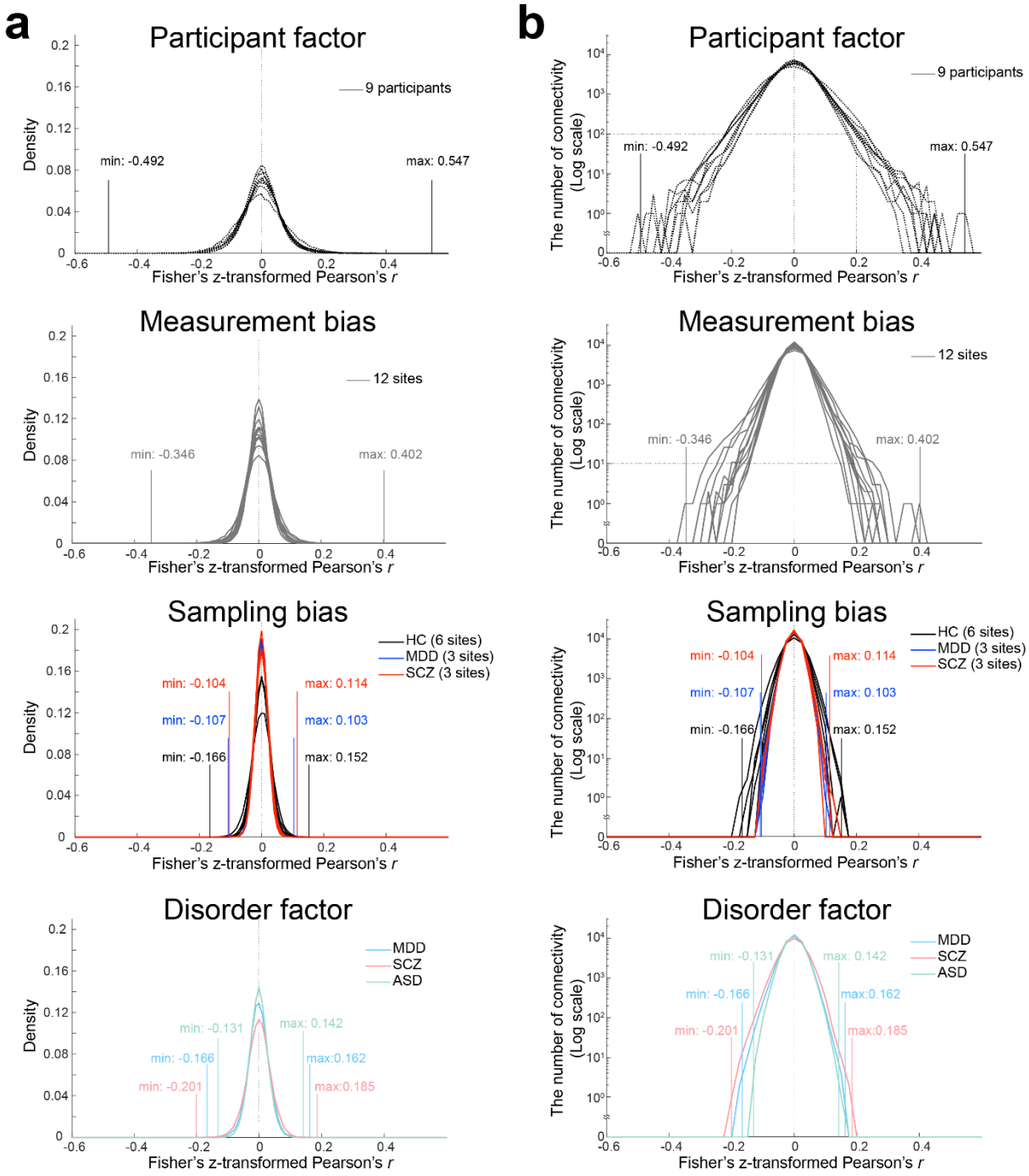

FIGURE A.1 | Distributions and statistics for each type of bias and each factor. (a, b) The distribution of the effects of each bias and each factor on functional connectivity vectors. Functional connectivity was measured based on Fisher's z-transformed Pearson's correlation coefficients. The *x*-axis represents the effect size of the Fisher's z-transformed Pearson's correlation coefficients. In (a) and (b), the *y*-axis represents the density of connectivity and the log-transformed the number of connections, respectively. Each line represents one participant or one site. HC: healthy controls; SCZ: schizophrenia; MDD: major depressive disorder; ASD: autism spectrum disorder.

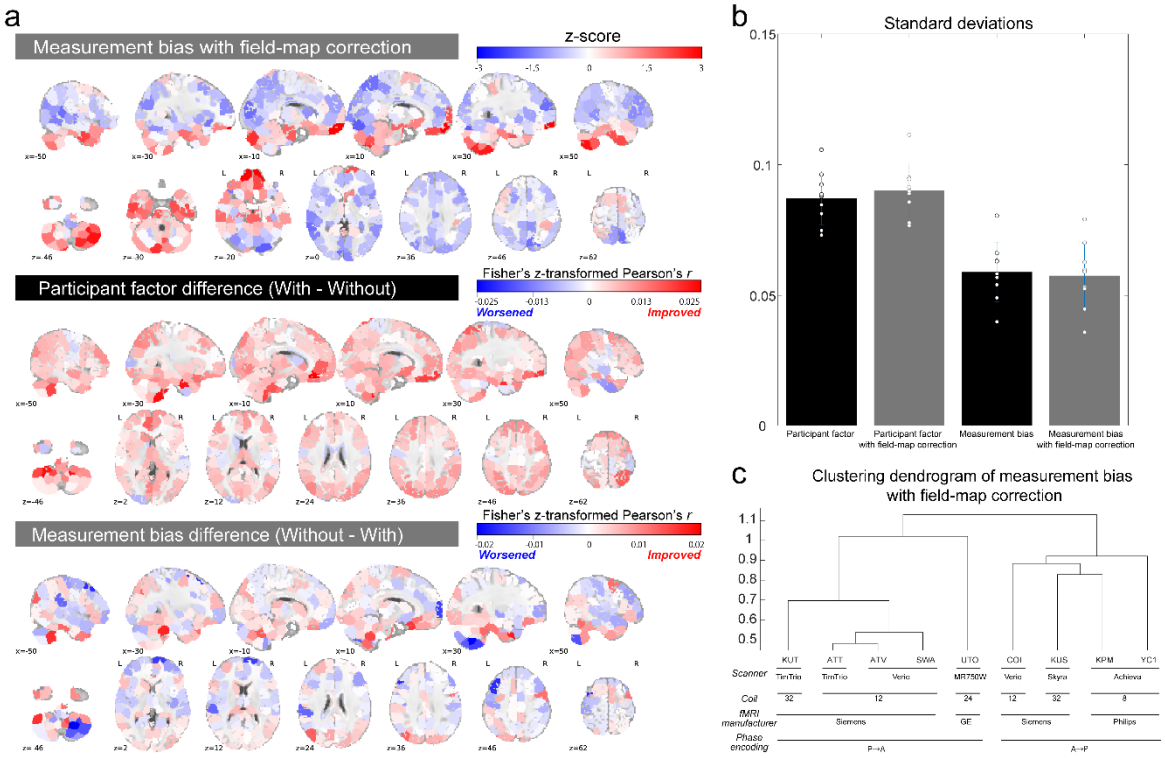

FIGURE A.2 | Effects of field-map correction. (a) Top: Mean effects of connectivity at all 268 ROIs with field-map correction. Color-coding follows that for Figure 4 in the main text. Difference between field-map-corrected and -uncorrected datasets for participant factor (middle) and measurement bias (bottom). Red represents positive effects due to correction (i.e., increase in participant factor and decrease in measurement bias). Blue represents negative effects (i.e., decrease in participant factor and increase in measurement bias). (b) The standard deviations of participant factor and measurement bias after field-map correction. Bars represent the average, while error bars represent the standard deviation across sites or participants. Each data point represents one participant or one site. (c) Clustering dendrogram for measurement bias after field-map correction. The height of each linkage in the dendrogram represents the distance between the clusters joined by that link. ROI: region of interest; UTO: University of Tokyo; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; COI: Center of Innovation in Hiroshima University; KUS: Siemens Skyra scanner at Kyoto University; KPM: Kyoto Prefectural University of Medicine; YC1: Yaesu Clinic 1.

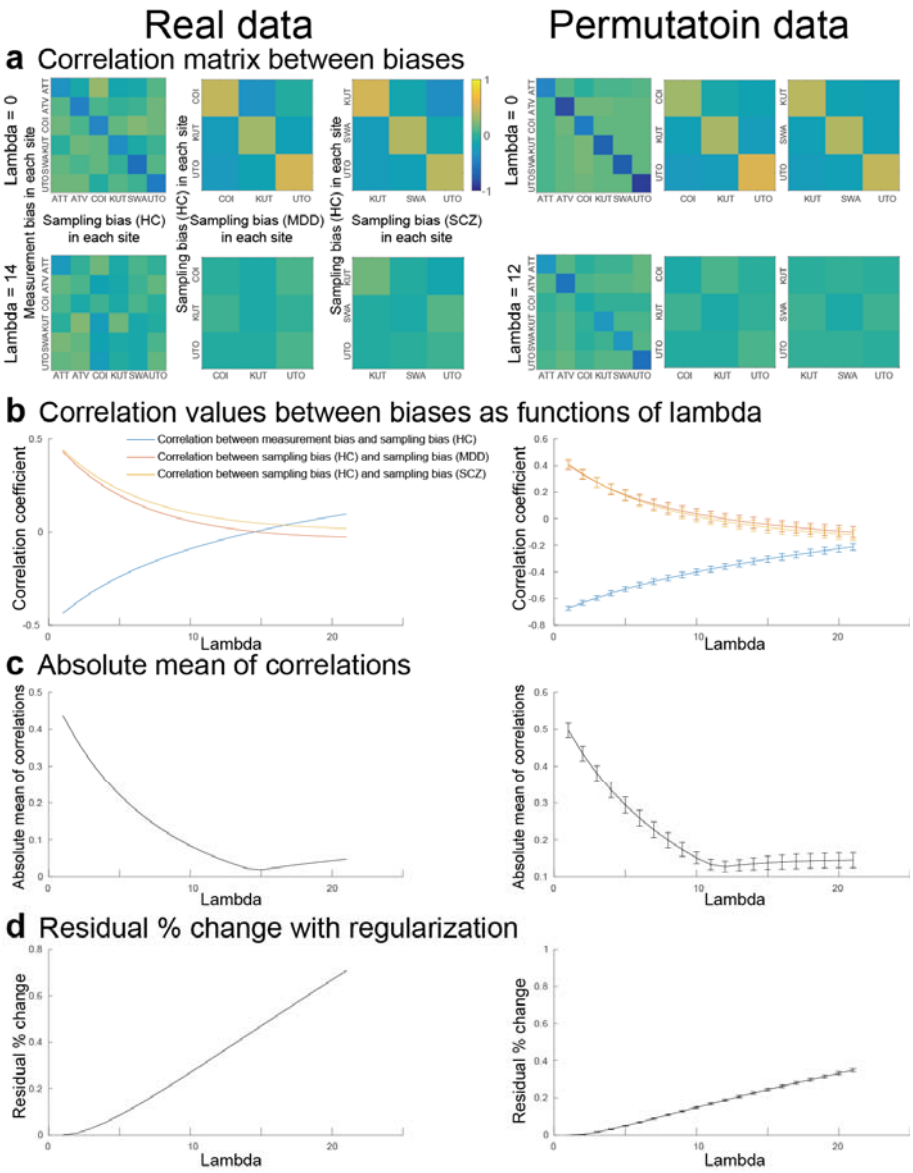

FIGURE A.3 | Selection of the regularization hyper-parameter lambda. (a) Correlation matrix between measurement biases and sampling biases in healthy controls (HCs), and matrices between sampling biases of HCs and sampling biases of patients with psychiatric disorders at lambda = 0 and lambda = 14, 12 (left: real data, right: permutation data). (b) Correlation values between the two types of bias as functions of lambda from 0 to 20 (left: real data, right: permutation data). Correlations were calculated between the measurement and sampling biases of HCs, and between the sampling biases of HCs and sampling biases of patients with psychiatric disorders. (c) Absolute mean of three correlations as a function of lambda. (d) Percentage change in the residual error between model and real data as a function of lambda. UTO: University of Tokyo; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; COI: Center of Innovation in Hiroshima University; MDD: major depressive disorder; SCZ: schizophrenia.

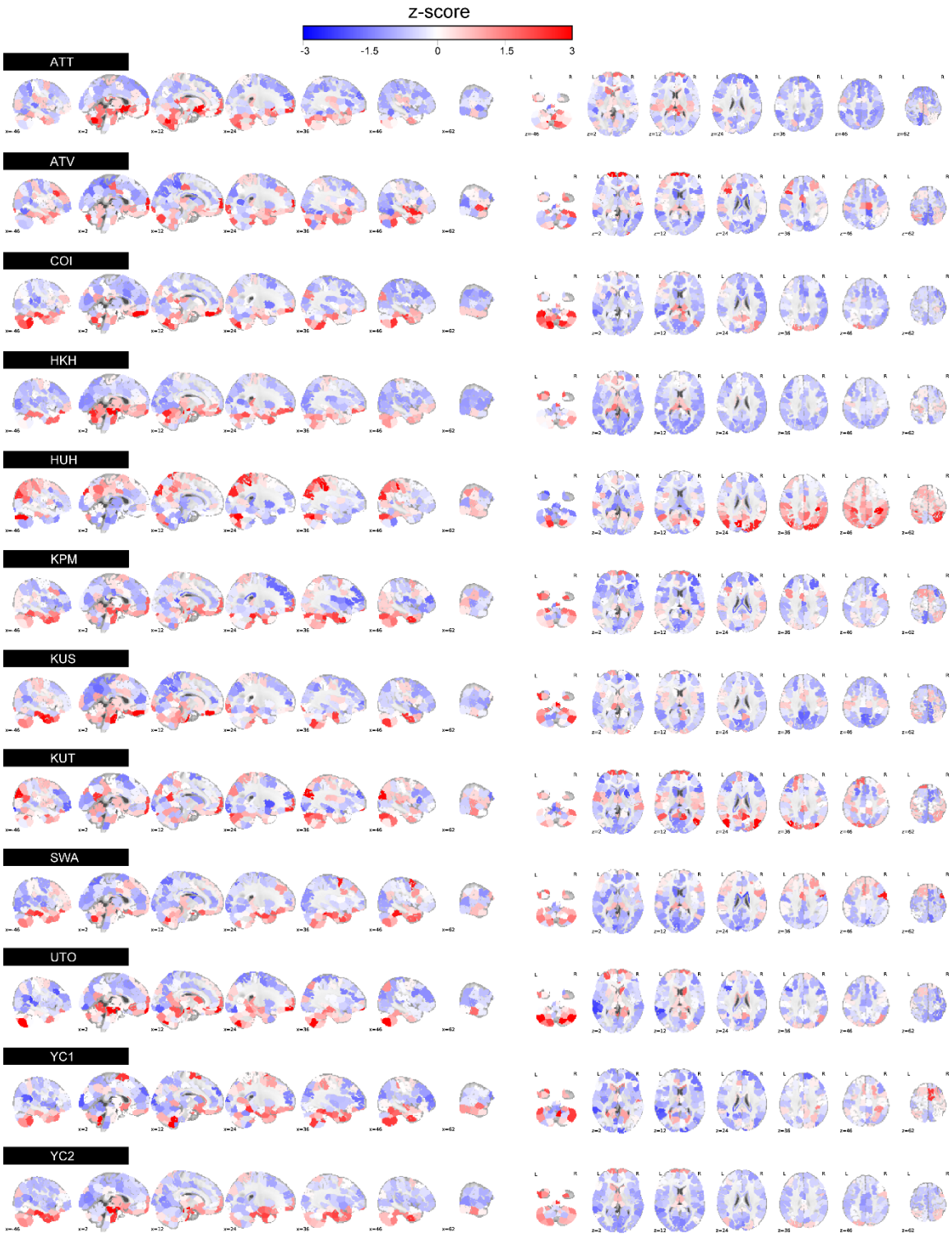

FIGURE A.4 | Spatial distribution of the measurement bias of each site in various brain regions. Mean effects of connectivity for all 268 ROIs. For each ROI, the mean effects of all functional connections associated with that ROI were calculated for the measurement bias of each site. Warmer (red) and cooler (blue) colors correspond to large and small effects, respectively. The magnitudes of the effects are normalized within each site (*z*-score). ROI: region of interest; UTO: University of Tokyo; HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International;

SWA: Showa University; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KUS: Siemens Skyra scanner at Kyoto University; KPM: Kyoto Prefectural University of Medicine; YC1: Yaesu Clinic 1; YC2: Yaesu Clinic 2.

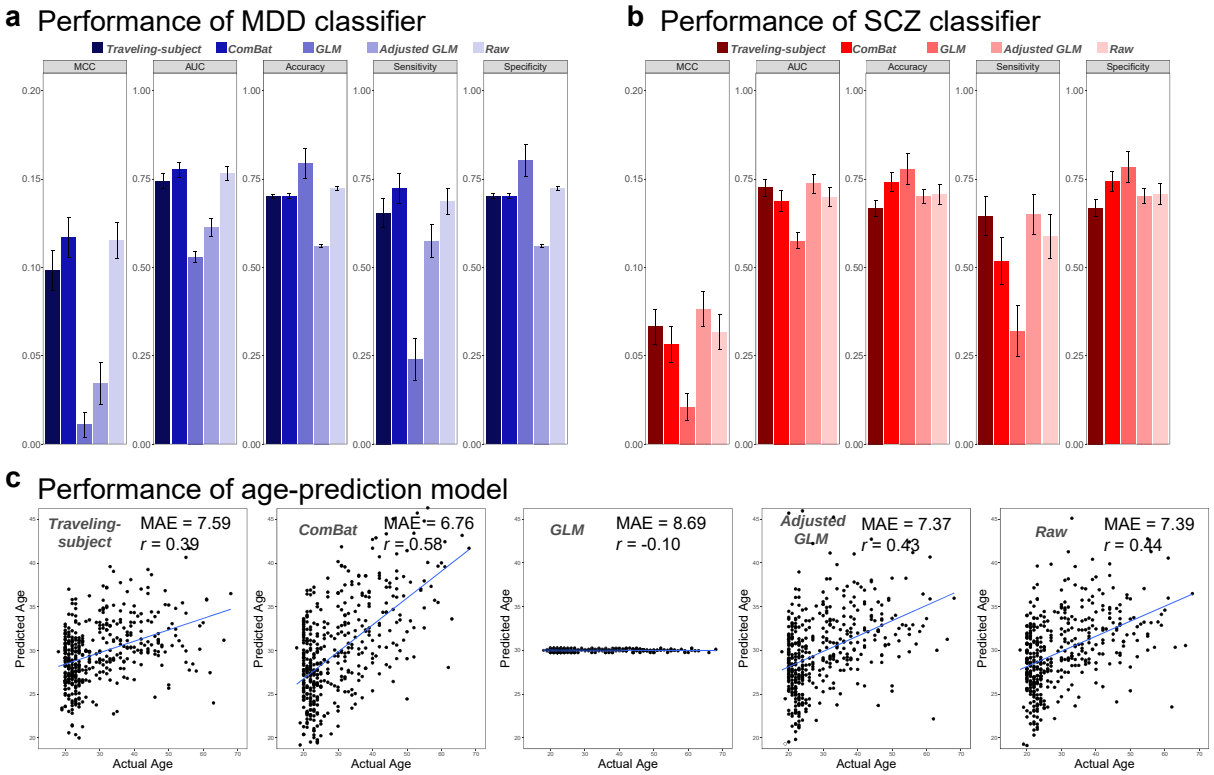

FIGURE A.5 | Performance of disorder classifiers and age regression model in the training dataset. (a, b) Performance of each classifier in the training dataset for each harmonization method (blue for MDD, red for SCZ). Bars represent the average, while error bars represent the standard deviation across 100 re-samplings. (c) Scatter plot of actual age and predicted age for each harmonization method. The solid line represents the linear regression of the actual age from the predicted age. The mean absolute error (MAE) and correlation coefficient (*r*) are also shown. Each data point represents one participant. MDD: major depressive disorder; SCZ: schizophrenia; MCC: Matthews correlation coefficient; AUC: area under the curve.

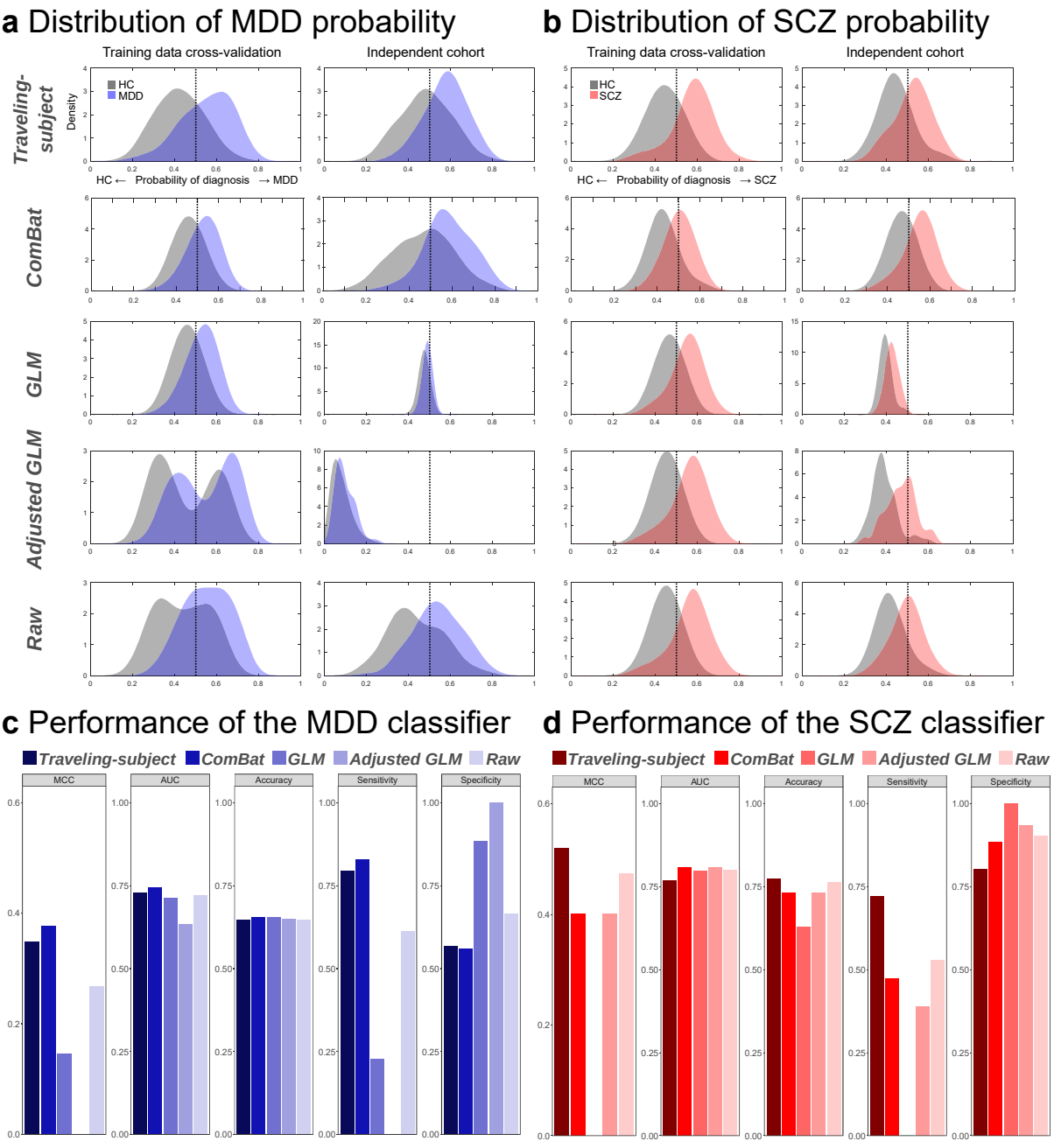

FIGURE A.6 | Classifier performances for MDD and SCZ for different harmonization methods. (a) The probability distribution for the diagnosis of MDD in the training dataset (left) and independent cohort (right) for each harmonization method. The MDD and HC distributions are depicted in blue and gray, respectively. (b) The probability distribution for the diagnosis of SCZ in the training dataset (left) and independent cohort (right) for each harmonization method. The SCZ and HC distributions are depicted in red and gray, respectively. (c, d) Classifier performance in the independent cohort for each harmonization method and each classifier (blue for MDD, red for SCZ). MCC: Matthews correlation coefficient; AUC: area under the curve; MDD: major depressive disorder; SCZ: schizophrenia; HC: healthy control.

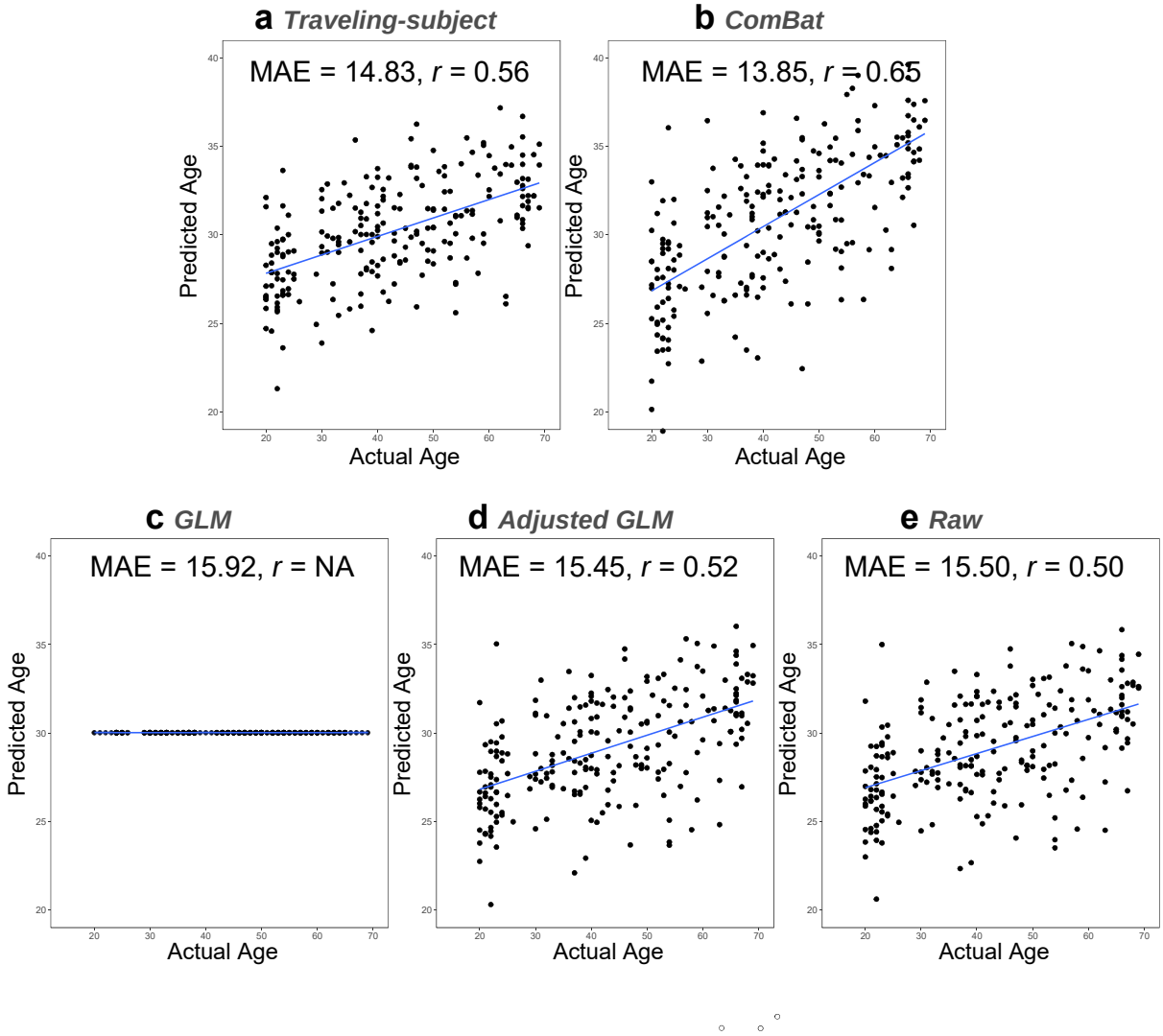

FIGURE A.7 | Performance of a regression model for the prediction of a participant's age for different harmonization methods. Scatter plots of actual age and predicted age. The solid line indicates the linear regression of the actual age from the predicted age. The mean absolute error (MAE) and correlation coefficient (*r*) are shown in each panel. Each data point represents one participant. Each panel shows the results for the (a) traveling-subject method, (b) ComBat method, (c) GLM method, (d) adjusted GLM method, or (e) raw method (i.e., the data were not harmonized across sites). GLM: general linear model.

TABLE A.1 | 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***
Magnetic field strength	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T
Number of channels per coil	12	12	12	8	12	8	12	32	32	24	8	8
Field-of-view (mm)	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212	212 x 212
Matrix	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64
Number of slices	40	39	40	35	35	40	40	40	40	40	40	40
Number of volumes	240	240	240	240	240	240	240	240	240	240	240	240
In-plane resolution (mm)	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125
Slice thickness (mm)	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2
Slice gap (mm)	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
TR (ms)	2,500	2,500	2,500	2,500	2,500	2,500	2,500	2,500	2,500	2,500	2,500	2,500
TE (ms)	30	30	30	30	30	30	30	30	30	30	30	30
Total scan time (min:s)	10:00	10:00	10:00	10:00	10:00	10:00	10:00	10:00	10:00	10:00	10:00	10:00
Flip angle (deg)	80	80	80	80	80	80	80	80	80	80	80	80
Slice acquisition order	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending
Phase encoding	PA	PA	AP	PA	PA	AP	PA	PA	AP	PA	AP	AP
Eye closed / fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
Field map	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔	-

UTO: University of Tokyo; HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KUS: Siemens Skyra scanner at Kyoto University; KPM: Kyoto Prefectural University of Medicine; YC1: Yaesu Clinic 1; YC2: Yaesu Clinic 2; TR: repetition time; TE: echo time.

TABLE A.2 | Demographic characteristics of patients in the independent validation cohort dataset for MDD prediction

Site	HC			MDD			ALL		
	Number	Male /Female	Age (y)	Number	Male /Female	Age (y)	Number	Male /Female	Age (y)
HRC	47	12/35	42.4±11.3	14	4/10	38.2±9.1	61	16/45	41.5±10.9
UYA	120	50/70	45.9±19.5	79	36/43	50.3±13.6	199	86/113	47.6±17.5
Summary	167	62/105	44.9±17.6	93	40/53	42.6±11.7	260	102/158	46.2±16.4

HRC: Hiroshima Rehabilitation Center; UYA: Yamaguchi University; HC: healthy control; MDD: major depressive disorder

TABLE A.3 | Demographic characteristics of patients in the independent validation cohort dataset for SCZ prediction

Site	HC			SCZ			ALL		
	Number	Male /Female	Age (y)	Number	Male /Female	Age (y)	Number	Male /Female	Age (y)
KTT	77	50/27	28.8±8.98	56	29/27	37.8±9.42	133	79/54	32.6±10.2

KTT: Kyoto University (Trio scanner); HC: healthy control; SCZ: schizophrenia

TABLE A.4 | Demographic characteristics of patients in the independent validation cohort dataset for age prediction

Site	HC		
	Number	Male /Female	Age (y)
ATT	40	16/24	42.3±15.4
ATV	134	77/57	43.2±14.6
ATP	48	21/27	41.8±16.6
Summary	232	124/108	42.7±15.1

ATP: Prisma scanner at Advanced Telecommunications Research Institute International; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; HC: healthy control

TABLE A.5 | Imaging protocols for resting-state fMRI in the independent validation cohort dataset

Site	**ATR TimTrio**	**ATR Verio**	**ATR Prisma**	**Hiroshima Rehabilitation Center**	**Yamaguchi University**	**Kyoto University Trio**
Abbreviation	**ATT**	**ATV**	**ATP**	**HRC**	**UYA**	**KTT**
MRI scanner	***Siemens TimTrio***	***Siemens Verio***	***Siemens Prisma***	***GE Signa HDxt***	***Siemens Skyra***	***Siemens Trio***
Magnetic field strength (T)	3.0	3.0	3.0	3.0	3.0	3.0
Number of channels per coil	12	12	12	8	20	8
Field-of-view (mm)	212 × 212	212 × 212	212 × 212	256 × 256	220 × 220	256 × 192
Matrix	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 48
Number of slices	40 or 39	39	40	32	34	30
Number of volumes	240	240	240	143	200	180
In-plane resolution (mm)	3.3125 × 3.3125	3.3125 × 3.3125	3.3125 × 3.3125	4.0 × 4.0	3.4 × 3.4	4.0 × 4.0
Slice thickness (mm)	3.2	3.2	3.2	4	4.0	4.0
Slice gap (mm)	0.8	0.8	0.8	0	1.0	0
TR (ms)	2,500	2,500	2,500	2,000	2,500	2,000
TE (ms)	30	30	30	27	30	30
Total scan time (min:s)	10:00	10:00	10:00	4:46	8:28	6:00
Flip angle (deg)	80	80	80	90	80	90
Slice acquisition order	Ascending	Ascending	Ascending	Ascending (Interleaved)	Ascending	Ascending (Interleaved)
Phase encoding	PA	PA	PA	AP	PA	AP
Eyes closed / open / fixate	Fixate	Fixate	Fixate	Fixate	Closed	Fixate

ATR: Advanced Telecommunications Research Institute International

TABLE A.6 | Comparison between the variances of distributions in measurement bias and disorder factor

Disorder factors \ Measurement bias	ATT	ATV	COI	HKH	HUH	KPM	KUS	KUT	SWA	UTO	YC1	YC2
MDD	-	MDD > M	M > MDD	M > MDD	M > MDD	M > MDD	M > MDD	M > MDD	M > MDD	M > MDD	M > MDD	M > MDD
	W^* = 1.31	W^* = -9.55	W^* = 30.51	W^* = 25.25	W^* = 59.76	W^* = 33.72	W^* = 31.96	W^* = 13.92	W^* = 19.64	W^* = 21.46	W^* = 42.94	W^* = 34.85
SCZ	SCZ > M	SCZ > M	M > SCZ	M > SCZ	M > SCZ	M > SCZ	M > SCZ	SCZ > M	-	-	M > SCZ	M > SCZ
	W^* = -17.95	W^* = -28.89	W^* = 11.38	W^* = 6.48	W^* = 41.53	W^* = 14.40	W^* = 12.55	W^* = -5.36	W^* = 0.23	W^* = 2.00	W^* = 23.46	W^* = 15.75
ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD	M > ASD
	W^* = 14.61	W^* = 3.76	W^* = 43.53	W^* = 38.66	W^* = 71.98	W^* = 46.62	W^* = 45.13	W^* = 27.21	W^* = 32.9	W^* = 34.6	W^* = 55.74	W^* = 47.61

M > MDD (SCZ, ASD): Measurement-bias is larger than MDD (SCZ, ASD) factor; MDD (SCZ, ASD) > M: MDD (SCZ, ASD) factor is larger than measurement-bias

* if W^* = 2.63, then p = 0.05 after Bonferroni correction (12), n = 35,778

UTO: University of Tokyo; HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KUS: Siemens Skyra scanner at Kyoto University; KPM: Kyoto Prefectural University of Medicine; YC1: Yaesu Clinic 1; YC2: Yaesu Clinic 2; MDD: major depressive disorder; SCZ: schizophrenia.

TABLE A.7 | Comparison between variances of the distributions in sampling bias for healthy control and disorder factors

Sampling bias / Disorder factors	ATT	ATV	COI	KUT	SWA	UTO
MDD	MDD > S W^* = -26.52	MDD > S W^* = -22.77	S > MDD W^* = 8.28	MDD > S W^* = -24.01	MDD > S W^* = -49.66	MDD > S W^* = -54.40
SCZ	SCZ > S W^* = -45.87	SCZ > S W^* = -41.95	SCZ > S W^* = -11.53	SCZ > S W^* = -42.85	SCZ > S W^* = -67.90	SCZ > S W^* = -72.35
ASD	ASD > S W^* = -13.02	ASD > S W^* = -9.51	S > ASD W^* = 21.66	ASD > S W^* = -10.51	ASD > S W^* = -36.66	ASD > S W^* = -41.64

S > MDD (SCZ, ASD): Sampling-bias is larger than MDD (SCZ, ASD) factor, MDD (SCZ, ASD) > S: MDD (SCZ, ASD) factor is larger than sampling-bias

* if W^* = 2.4, then p = 0.05 after Bonferroni correction (6), n = 35,778

UTO: University of Tokyo; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; COI: Center of Innovation in Hiroshima University; MDD: major depressive disorder; SCZ: schizophrenia.

TABLE A.8 | Results of MDD prediction

Training dataset	**Matthews correlation coefficient (MCC)**	**Diagnostic odds ratio (DOR)**	**F-value**	**Area under the curve (AUC)**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Traveling-subject	0.1	-	0.07	0.74	0.7	0.04	0.65	0.7
ComBat	0.12	-	0.07	0.78	0.7	0.04	0.72	0.7
GLM	0.01	-	0.02	0.53	0.79	0.01	0.24	0.8
Adjusted GLM	0.03	-	0.04	0.61	0.56	0.02	0.57	0.56
Raw	0.12	-	0.07	0.77	0.72	0.04	0.69	0.72
Independent cohort	**Matthews correlation coefficient (MCC)**	**Diagnostic odds ratio (DOR)**	**F-value**	**Area under the curve (AUC)**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Traveling-subject	0.35	5.09	0.61	0.73	0.65	0.50	0.80	0.57
ComBat	0.38	6.21	0.63	0.75	0.65	0.50	0.83	0.56
GLM	0.15	2.24	0.31	0.71	0.65	0.51	0.23	0.88
Adjusted GLM	NA	NA	NA	0.63	0.65	NA	0	1
Raw	0.27	3.15	0.55	0.72	0.65	0.50	0.61	0.66

MDD: major depressive disorder; GLM: general linear model

TP: true positive, TN: true negative, FP: false positive, FN: false negative

$$\mathrm{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

$$\mathrm{DOR} = \frac{TP * TN}{FP * FN},$$

$$\mathrm{F\ \ value} = 2\frac{\mathrm{precision} * \mathrm{sensitivity}}{\mathrm{precision} + \mathrm{recall}}, \mathrm{precision} = \frac{TP}{TP + FP}, \mathrm{sensitivity} = \frac{TP}{TP + FN}$$

TABLE A.9 | All results for MDD prediction at each site

Training dataset	**Area under the curve (AUC)**					**Accuracy**					**Sensitivity**					**Specificity**				
	Traveling-subject	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*
ATT	-	-	-	-	-	0.77	0.68	0.76	0.99	0.9	-	-	-	-	-	0.77	0.68	0.76	0.99	0.9
ATV	-	-	-	-	-	0.82	0.70	0.78	1	0.91	-	-	-	-	-	0.82	0.70	0.78	1	0.91
COI	0.74	0.71	0.64	0.5	0.69	0.66	0.67	0.84	0.15	0.49	0.67	0.64	0.26	0.99	0.79	0.66	0.67	0.94	0.01	0.44
HUH	0.64	0.87	0.52	0.55	0.59	0.51	0.77	0.81	0.87	0.34	0.68	0.80	0.18	0.14	0.85	0.5	0.77	0.84	0.9	0.31
HKH	0.47	0.72	0.48	0.56	0.54	0.47	0.71	0.79	0.24	0.44	0.59	0.65	0.17	0.78	0.59	0.47	0.71	0.8	0.22	0.44
KPM	-	-	-	-	-	0.68	0.71	0.78	0.06	0.75	-	-	-	-	-	0.68	0.71	0.78	0.06	0.75
SWA	-	-	-	-	-	0.74	0.65	0.77	0.99	0.9	-	-	-	-	-	0.74	0.65	0.77	0.99	0.9
KUT	0.83	0.81	0.6	0.49	0.72	0.76	0.76	0.83	0.8	0.82	0.7	0.73	0.36	0.18	0.48	0.77	0.76	0.83	0.8	0.82
UTO	0.71	0.68	0.58	0.54	0.66	0.72	0.65	0.8	0.05	0.74	0.55	0.61	0.27	1	0.41	0.72	0.65	0.81	0.04	0.75
Independent cohort	**AUC**					**Accuracy**					**Sensitivity**					**Specificity**				
	Traveling-subject	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*
HRC	0.73	0.76	0.64	0.74	0.73	0.54	0.68	0.8	0.8	0.47	0.83	0.75	0	0	0.83	0.47	0.66	1	1	0.38
UYA	0.76	0.74	0.74	0.7	0.78	0.68	0.65	0.61	0.61	0.7	0.76	0.82	0	0	0.57	0.62	0.55	1	1	0.79

MDD: major depressive disorder; GLM: general linear model; UTO: University of Tokyo; HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KPM: Kyoto Prefectural University of Medicine; HRC: Hiroshima Rehabilitation Center; UYA: Yamaguchi University.

TABLE A.10 | Results of SCZ prediction

Training dataset	**Matthews correlation coefficient (MCC)**	**Diagnostic odds ratio (DOR)**	**F-value**	**Area under the curve (AUC)**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Traveling-subject	0.07	-	0.04	0.73	0.67	0.02	0.64	0.67
ComBat	0.06	-	0.03	0.69	0.74	0.02	0.52	0.74
GLM	0.02	-	0.01	0.58	0.78	0.01	0.32	0.78
Adjusted GLM	0.08	-	0.04	0.74	0.7	0.02	0.65	0.7
Raw	0.06	-	0.03	0.7	0.71	0.02	0.59	0.71
Independent cohort	**Matthews correlation coefficient (MCC)**	**Diagnostic odds ratio (DOR)**	**F-value**	**Area under the curve (AUC)**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Traveling-subject	0.52	10.62	0.70	0.77	0.77	0.68	0.72	0.80
ComBat	0.40	6.90	0.57	0.81	0.73	0.71	0.47	0.89
GLM	NA	NA	NA	0.80	0.63	NA	0	1
Adjusted GLM	0.4	9.07	0.52	0.81	0.73	0.78	0.39	0.93
Raw	0.47	10.24	0.62	0.80	0.76	0.76	0.53	0.9

SCZ: schizophrenia; GLM: general linear model

TABLE A.11 | All results for SCZ prediction at each site

Training dataset	**Area under the curve (AUC)**					**Accuracy**					**Sensitivity**					**Specificity**				
	Traveling-subject	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*
ATT	-	-	-	-	-	0.6	0.72	0.76	0.74	0.71	-	-	-	-	-	0.6	0.72	0.76	0.74	0.71
ATV	-	-	-	-	-	0.7	0.73	0.76	0.78	0.74	-	-	-	-	-	0.7	0.73	0.76	0.78	0.74
COI	-	-	-	-	-	0.74	0.77	0.75	0.81	0.78	-	-	-	-	-	0.74	0.77	0.75	0.81	0.78
HUH	-	-	-	-	-	0.64	0.74	0.76	0.72	0.69	-	-	-	-	-	0.64	0.74	0.76	0.72	0.69
HKH	-	-	-	-	-	0.6	0.75	0.74	0.76	0.75	-	-	-	-	-	0.6	0.75	0.74	0.76	0.75
KPM	-	-	-	-	-	0.7	0.77	0.76	0.84	0.81	-	-	-	-	-	0.7	0.77	0.76	0.84	0.81
SWA	0.78	0.73	0.62	0.78	0.78	0.65	0.73	0.82	0.61	0.68	0.78	0.57	0.34	0.81	0.79	0.65	0.73	0.83	0.6	0.67
KUT	0.75	0.73	0.63	0.7	0.7	0.64	0.75	0.86	0.53	0.63	0.67	0.57	0.31	0.7	0.62	0.64	0.76	0.89	0.52	0.63
UTO	0.64	0.61	0.56	0.58	0.59	0.69	0.72	0.78	0.68	0.7	0.52	0.42	0.27	0.45	0.39	0.69	0.73	0.78	0.68	0.71
Independent cohort	**AUC**					**Accuracy**					**Sensitivity**					**Specificity**				
	Traveling-subject	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*	*Traveling-subject*	*ComBat*	*GLM*	*Adjusted GLM*	*Raw*
KTT	0.77	0.81	0.8	0.81	0.8	0.77	0.73	0.63	0.73	0.76	0.72	0.47	0	0.39	0.53	0.8	0.89	1	0.93	0.9

SCZ: schizophrenia; GLM: general linear model; UTO: University of Tokyo; HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KPM: Kyoto Prefectural University of Medicine; KTT: Siemens Trio scanner at Kyoto University.

TABLE A.12 | Data availability statement

Site	Number of subjects	Type of data availablity
ATR TimTrio (ATT)	31	3
ATR Verio (ATV)	77	3
Hiroshima University Hospital (HUH)	123	1
Hiroshima Kajikawa Hospital (HKH)	52	1
Center of Innovation in Hiroshima University (COI)	48	1
Kyoto Prefectural University of Medicine (KPM)	117	4
Kyoto University (KUT)	66	2
Showa University (SWA)	101	2
University of Tokyo (UTO)	190	2
Traveling subject	9	3
Summary	814	

Note: Type of data availability

1) freely available without restriction allowing commercial re-use

2) freely available but not allowing commercial re-use

3) available after registration to our record but not allowing commercial re-use

4) available only to our research group

Appendix B

Appendix of Chapter 4

B.1 Other behavioral metrics in behavioral tasks

In previous studies, Kelly et al. (2008) showed the relationship between the functional connectivity (FC) of the default mode network (DMN) and task positive network (TPN) and the reaction time coefficient of variation (standard deviation divided by the mean) in the Flanker task (Kelly et al., 2008), and Liu et al. (2015) showed the relationship between the FC of DMN and Stroop effect measured by the difference in mean reaction time between conditions (mean reaction time of incongruent condition − mean reaction time of congruent condition) (Liu et al., 2015). Therefore, we investigated the change in these behavioral metrics from pre- to post-neurofeedback training. Furthermore, we also investigated the changes in error rate in all tasks.

We applied a mixed-effects model to each metric in the same manner as in our analysis of mean reaction times (main text, section 3.4). We did not find any interaction effect between group and day in any metric (Table A.1). From these results, we could not conclude that the error rate, the coefficient of variation in the Flanker task, or the Stroop effect changed from pre- to post-neurofeedback training in any group. However, the direction of change in coefficient of variation is consistent with previous studies. The coefficient of variation increased from 0.084 ± 0.0084 (mean ± 95% confidence interval) for pre-neurofeedback training to 0.087 ± 0.010 for post-neurofeedback training in the "increased FC" group and decreased from 0.097 ± 0.020 for pre- to 0.87 ± 0.012 for post-neurofeedback training in the "decreased FC" group.

B.2 Difference in total score between subject groups

The total score during the training was associated with monetary reward received by the subjects. Thus, the difference in the score between the groups may have caused differences in motivation for involvement in the experiment and thus performance in the cognitive task. Therefore, we compared the difference in total score between the groups using a two-sample *t*-test. There was no difference between the groups ($t = 0.68$, $p = 0.50$). The total score averaged across subjects was 14379 ± 809 (mean ± 95% confidence interval) for the "increased FC" group and 13860 ± 1448 for the "decreased FC" group. Thus, the difference in score was only 3.7% of the average score across the groups. This result indicates that there was no significant difference in the amount of reward between the groups and little possibility that the difference in score caused different changes in cognitive performance.

B.3 Strategies adopted by subjects to increase their score

Because the difference in subjects' strategies for getting a high score may have affected the cognitive task performance, we examined whether the strategies differed between the groups. We conducted a post-experiment questionnaire with 25 of the 30 subjects ("increased FC" group: n = 13; "decreased FC" group: n = 12). The experimenter told subjects, "Please let me know how you imagined the finger tapping during neurofeedback training. Also, if you tried any strategy other than the one that I instructed, please let me know what you did." The free answers of the individual subjects are listed in Table A.2.

We tabulated their answers with respect to five categories: 1) modality of imagery (items: kinesthetic*, visual, or both), 2) moved finger (right*, left, or both), 3) movement sequence (random* or fixed), 4) presence of an object (present or absent*), and 5) presence of a rhythm (present or absent*). Our instruction was to imagine tapping the thumbs with the fingers randomly as fast as possible and to produce kinesthetic imagery related to tapping. Thus, the items marked with asterisks (default items) should be mentioned if subjects had followed our instructions and if they made any comment on each category. We assumed that subjects adopted a strategy including the default items when they did not make a comment on the category (Table A.3). We counted the numbers of items across subjects and compared the numbers between the groups. We calculated the p value as the probability that these results would have been obtained if we had separated subjects randomly. As a result, there were no significant differences in the numbers between the groups (Table A.4). These results indicate that there was no significant difference in the strategy adopted for increasing the score between the groups.

B.4 Regional brain activity during the training

We investigated the daily changes in the regional brain activity in two target ROIs (lM1 and lLP) during neurofeedback training in our offline analysis in the same manner as in our analysis of FC between the two ROIs.

We used the mean BOLD signals in each ROI instead of the connectivity. We averaged the signals in seven volumes during a motor imagery period in each trial (the first volume was discarded and one volume from the feedback period was added as compensation for hemodynamic delay following the online calculation of the feedback score). In total, each subject had 280 mean BOLD signals in each ROI (BASE = 40, DAY1–DAY4 = 60 * 4). Then, to compare the daily changes in mean BOLD signals between the two groups, we applied a mixed-effects model to the mean BOLD signals in each ROI.

As a result, we found a significant interaction effect in lLP (DAY: $t = 4.87$, $p = 1.1 \times 10^{-6}$; Group: $t = 0.17$, $p = 0.86$; DAY × Group: $t = -2.83$, $p = 0.0045$) but not in lM1 (DAY: $t = -4.41$, $p = 1.0 \times 10^{-5}$; Group: $t = 0.49$, $p = 0.62$; DAY × Group: $t = 0.17$, $p = 0.86$). The significant interaction suggests that the change in the regional brain activity in lLP across days was different between the groups.

B.5 Relationship between online score and activity in each ROI

To investigate the relationship between the online score and regional activities in the two target ROIs, we calculated the correlation between the scores and activity averaged within each ROI using the offline data for each group. Because each subject underwent 280 trials (BASE = 40, DAY1–DAY4 = 60 × 4), we used 280 × 18 pairs of score and activity in each ROI to calculate the correlation in the "increased FC" group and 280 × 12 pairs in the "decreased FC" group.

As a result, we did not find any significant correlation between the feedback score and the regional brain activity in lLP ("increased FC" group: $r = 0.016$, $p = 0.23$; "decreased FC" group: $r = -0.01$, $p = 0.54$) or in lM1 ("increased FC" group: $r = 0.0071$, $p = 0.61$; "decreased FC" group: $r = 0.028$, $p = 0.10$). This result indicates that subjects could not get information about regional brain activities in either of the target ROIs from the online score.

B.6 Effect of the initial functional connectivity on training

To investigate whether the differential changes in FC and cognitive performances between the two groups were induced by the difference in initial resting-state FC, we examined whether 1) the average initial resting-state FC of lM1 and lLP in this study was different from that in a healthy population and 2) the observed differences in changes in FC during training and cognitive performances were induced by the differences in the initial FC during training between the two groups.

Average resting-state FC in a healthy population was estimated using another data set collected in our laboratory (232 healthy subjects [108 women]; mean age, 42.7 ± 15.1 years; age range, 20–69 years). In order to match subjects' ages to those in this study, we only included subjects aged ≤ 30 years (61 healthy subjects [27 women]; mean age, 23.4 ± 3.1 years; age range, 20–30 years) in the dataset. We calculated the average resting-state FC of lM1 and lLP in this data set and compared it to the average resting-state FC in this study. As a result, the average resting-state FC of lM1 and lLP in this study (mean ± SD = -0.0568 ± 0.1936) was not different from that in the healthy population (mean ± SD = -0.0605 ± 0.228).

Further, we investigated whether the differences in changes in FC during training and cognitive performances were affected by the difference in initial FC during training between the two groups. First, we defined the initial FC as that measured on the first day (BASE) and confirmed that the average was not different between the two groups (*t*-test, $t = 0.20$, $p = 0.84$). This suggests that the initial FC probably did not explain the differential changes between the two groups.

Moreover, we conducted correlation analyses between an individual's initial FC and the effect of training in the following two aspects. In the first aspect, we tested whether the individual's initial FC was correlated with the slope of FC across training days. We used a linear model to assess FC across training days for each subject and used

the slope as an index of individual differences in training effects. As a result, we did not find any significant correlation between an individual's initial FC and the effect of training ("increased FC" group: r = -0.13, p = 0.58; "decreased FC" group: r = -0.068, p = 0.83; all subjects: r = 0.11, p = 0.55). These results indicate that the subject-specific initial FC could not affect the effect of neurofeedback training.

In the second aspect, we investigated the relationship between an individual's initial FC and the changes in reaction time. As a result, we did not find any significant correlations ("increased FC" group: PVT, r = -0.14, p = 0.63; EFT *congruent*, r = 0.30, p = 0.30; CWST *congruent*, r = -0.056, p = 0.85; "decreased FC" group: PVT, r = -0.29, p = 0.37; EFT *congruent*, r = -0.22, p = 0.53; CWST *congruent*, r = -0.43, p = 0.18; all subjects: PVT, r = -0.16, p = 0.43; EFT *congruent*, r = -0.28, p = 0.18; CWST *congruent*, r = -0.26, p = 0.21). This result also indicates that subject-specific initial FC cannot explain the differential changes in reaction time between the two groups.

B.7 Moderation/mediation analysis

We examined the associations among 1) the changes in FC of lM1-lLP during neurofeedback training, 2) the changes in resting-state FC of MVN-DMN, and 3) the changes in cognitive performance of the three tasks, in which the interaction between group and day yielded significant effects. We analyzed data of the "increased FC" group, in which a significant change in resting-state connectivity of MVN-DMN was observed (t = 2.93, p = 0.0045; see 4.2.3 "Change in resting-state functional connectivity"). First, using a linear regression, we examined the effect of the change in FC during training on the changes in reaction time of the three tasks. We found a significant effect on the change in reaction time of CWST *congruent* (β = -1.42, SE = 0.58, t = -2.42, p = 0.033, adjusted R^2 = 0.28) but not on PVT and EFT *congruent* (PVT: β = -0.091, SE = 0.37, t = -0.24, p = 0.81, adjusted R^2 = 0.084; EFT *congruent*: β = 0.026, SE = 0.14, t = 0.18, p = 0.85, adjusted R^2 = −0.087; Figure B.1C). Second, we examined the effects of the change in FC during training on the change in resting-state FC. We did not find any significant effect (β = -0.11, SE = 0.54, t = -0.20, p = 0.84, adjusted R^2 = 0.086; Figure B.1A). Finally, we examined the effects of the change in resting-state FC on the changes in reaction time of the three tasks. We did not find any significant effects (PVT: β = 0.33, SE = 0.18, t = 1.83, p = 0.094, adjusted R^2 = 0.16; EFT *congruent*: β = 0.11, SE = 0.069, t = 1.68, p = 0.12, adjusted R^2 = 0.13; CWST *congruent*: β = 0.15, SE = 0.39, t = 0.38, p = 0.70, adjusted R^2 = 0.076; Figure B.1B). Additionally, a regression analysis was performed while including both the change in FC during training and the change in resting-state FC in the model. We found a significant effect on the change in reaction time of CWST *congruent* (β = -1.41, SE = 0.61, t = -2.29, p = 0.044, adjusted R^2 = 0.22) but not on PVT and EFT *congruent* (PVT: β = -0.055, SE = 0.34, t = -0.16, p = 0.87, adjusted R^2 = 0.082; EFT *congruent*: β = 0.039, SE = 0.13, t = 0.29, p = 0.77, adjusted R^2 = 0.053; Figure B.1C').

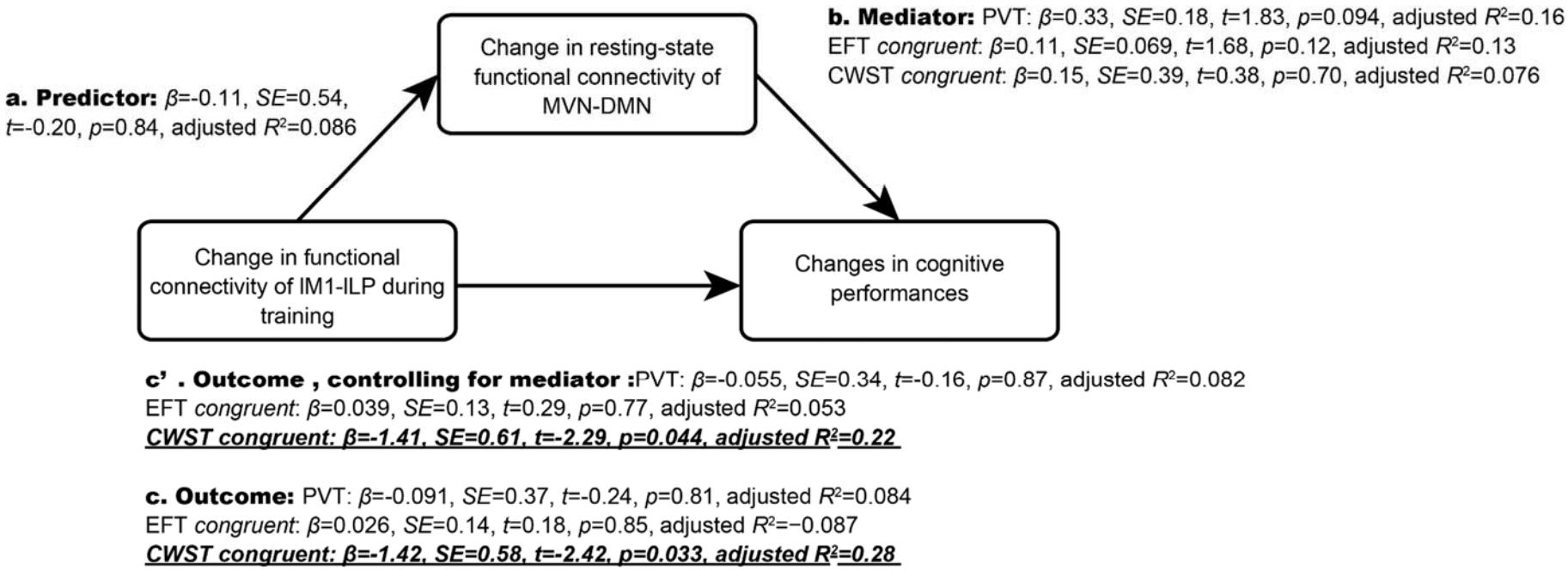

FIGURE B.1 | (a) Predictor is the change in functional connectivity between lM1-lLP during neurofeedback training. (b) Mediator is the change in resting-state functional connectivity between MVN-DMN. (c) Outcomes are the changes in cognitive performances. c denotes the relationship between predictor and outcomes, and c' denotes the same relationship after controlling for the effect of the mediator.

TABLE B.1 | All results of the cognitive performances tasks

	Pre-training		**Two-sample *t*-test for pre-training (Inc. vs. Dec.)**		**Post-training**		**Interaction effect of day and group**		**Main effect of day**			
									Inc.		Dec.	
Reaction time (s)	Inc.	Dec.	*t*-value	*p*-value	Inc.	Dec.	*t*-value	*p*-value	*t*-value	*p*-value	*t*-value	*p*-value
Vigilance	0.39	0.43	-1.28	0.21	0.42	0.42	**-2.72**	**0.0065**	**-3.85**	**0.00013**	0.12	0.90
Flanker *congruent*	0.38	0.39	-0.77	0.44	0.39	0.40	**2.41**	**0.016**	0.58	0.56	**-2.52**	**0.011**
Flanker *incongruent*	0.48	0.49	-0.45	0.65	0.47	0.48	1.18	0.23	**4.49**	**<0.0001**	1.87	0.060
Stroop *congruent*	0.69	0.73	-0.56	0.57	0.60	0.59	**-2.67**	**0.0075**	**6.93**	**<0.0001**	**8.53**	**<0.0001**
Stroop *incongruent*	0.96	1.01	-0.44	0.66	0.80	0.87	-0.50	0.61	**6.67**	**<0.0001**	**6.25**	**<0.0001**
Error rate (%)												
Vigilance	5.41	4.00	1.39	0.17	5.83	5.50	1.20	0.24	-0.59	0.55	-1.78	0.10
Flanker *congruent*	3.99	3.95	0.046	0.96	4.16	4.37	0.43	0.66	-0.24	0.80	-0.80	0.43
Flanker *incongruent*	7.29	9.16	-0.79	0.43	6.07	8.12	0.18	0.85	1.45	0.17	0.55	0.59
Stroop *congruent*	6.59	7.29	-0.59	0.55	6.42	6.25	-0.55	0.58	0.20	0.83	1.00	0.34
Stroop *incongruent*	18.4	22.2	-1.02	0.31	13.5	15.8	-0.13	0.89	2.02	0.065	1.89	0.087
Coefficient of variation												
Flanker *congruent*	0.081	0.090	-1.04	0.30	0.085	0.094	-0.049	0.96	0.68	0.50	0.49	0.63
Flanker *incongruent*	0.084	0.97	-1.31	0.21	0.087	0.87	-1.22	0.23	0.44	0.66	-1.13	0.28
Stroop effect (s)												
Stroop	0.28	0.30	-0.22	0.82	0.23	0.27	0.34	0.73	-1.48	0.16	-0.61	0.55

Inc., “increased functional connectivity” group; Dec., “decreased functional connectivity” group

TABLE B.2 | Free answers given by individual subjects

Participants	Strategy
	"Increased functional connectivity" group
1	I performed kinesthetic motor imagery.
2	I visualized three-dimensional images.
3	I performed kinesthetic motor imagery. I imagined the haptic sensation of two fingers touching.
4	I visualized an image while trying to alter the location to which the image was projected.
5	I performed motor imagery, and the imagined movement was rhythmic.
6	I performed kinesthetic motor imagery of my hand.
7	I vividly imagined my hand and fingers moving slowly.
8	I imagined moving my fingers, as well as the movements of my hands and feet as if I were walking.
9	I imagined the thumb rapidly tapping with every other finger, from index to little finger, for both hands. I continued to look at the fixation point without any thought during rest.
10	I tried to imagine moving hands without making actual movements.
11	I imagined running at the same time I imagined the thumb tapping with every other finger, from index to little finger, for both hands.
12	I imagined tapping the thumb with every other finger, randomly. I also imagined tapping a keyboard or buttons on a calculator, and scratching objects.
13	I performed kinesthetic motor imagery of my right hand and visualized it in front of my face. I imagined the thumb tapping with every other finger, from index to little finger, and then backwards.
	"Decreased functional connectivity" group
14	I performed motor imagery of all fingers, excluding the thumb, and the imagined movement was rhythmic. I performed kinesthetic motor imagery of my right hand.
15	I performed kinesthetic and visual motor imagery.
16	I performed kinesthetic motor imagery of my right hand.
17	I performed visual motor imagery. I imagined the thumb tapping with every other finger precisely.
18	I kinesthetically imagined the thumb having prolonged contact with every other finger. I also imagined the same movement at a higher speed.
19	I performed motor imagery, without particularly focusing on my hand.
20	I imagined tapping my thumb with my little finger.
21	I performed motor imagery and the imagined movement was in sync with the rhythm of music in my head.
22	I imagined the thumb tapping with every other finger, from index to little finger, mainly with the right hand. Occasionally, the order of tapping was random.
23	At first, I imagined the movement of either the right or the left hand and discovered that using the left hand led to a higher score. From then on, I kept using the left hand. I shaped my right hand as if it formed the Japanese character Tsu (つ, similar to alphabet 'C'), and imagined the thumb tapping with every other finger, from index to little finger. I also performed motor imagery according to the instructions. I also imagined the same movement, while shaping my hand as if holding a hamburger. I again imagined the same movement, while holding my hand with only the thumb sticking out.
24	I performed motor imagery using my right hand, which is my dominant hand. I also visualized my hand from different angles.
25	I performed motor imagery using my right hand.

TABLE B.3 | Subjects' answers with respect to five categories

Participants	Image	Hand	Sequence	Object	Rhythm
	"Increased functional connectivity" group				
1	Kinesthetic	*	*	*	*
2	Visual	*	*	*	*
3	Kinesthetic	*	*	*	*
4	Visual	*	*	*	*
5	*	*	*	*	Present
6	Kinesthetic	*	*	*	*
7	*	*	*	*	Present
8	*	*	*	*	*
9	*	Both	Fixed	*	*
10	Kinesthetic	*	*	*	*
11	*	Both	Fixed	*	*
12	*	*	Random	Present	*
13	Kinesthetic	Right	Fixed	*	*
	"Decreased functional connectivity" group				
14	Kinesthetic	Right	*	*	Present
15	Both	*	*	*	*
16	Kinesthetic	Right	*	*	*
17	Visual	*	*	*	*
18	Kinesthetic	*	*	*	*
19	*	*	*	*	*
20	*	*	Fixed	*	*
21	*	*	*	*	Present
22	*	Right	Fixed	*	*
23	*	Both	*	*	*
24	Visual	Right	*	*	*
25	*	Right	*	*	*

* When subjects made no comments on image type, hand, sequence or object, default items were filled in each cell (image: kinesthetic, hand: right, order: absent, object: absent, rhythm: absent)

TABLE B.4 | The number of items in free answers

Summary	Image			Hand			Sequence		Object		Rhythm	
	Kinesthetic	Visual	Both	Right	Left	Both	Fixed	Random	Present	Absent	Present	Absent
Increase	11	2	0	11	0	2	3	10	1	12	2	11
Decrease	9	2	1	11	0	1	2	10	0	12	2	10
Probability	$p = 0.19$			$p = 0.40$			$p = 0.35$		$p = 0.52$		$p = 0.40$	

Bibliography

Aarts, E., et al., 2014. A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience* 17, 491-496.

Abraham, A., et al., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *Neuroimage* 147, 736-745.

Adrian, E.D., Matthews, B.H.C., 1934. THE BERGER RHYTHM: POTENTIAL CHANGES FROM THE OCCIPITAL LOBES IN MAN. *Brain* 57, 355-385.

Amano, K., et al., 2016. Learning to Associate Orientation with Color in Early Visual Areas by Associative Decoded fMRI Neurofeedback. *Curr Biol* 26, 1861-1866.

Anderson, J.S., et al., 2011. Functional connectivity magnetic resonance imaging classification of autism. *Brain* 134, 3742-3754.

Andersson, J.L.R., et al., 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 20, 870-888.

Anzellotti, S., Coutanche, M.N., 2018. Beyond Functional Connectivity: Investigating Networks of Multivariate Representations. *Trends Cogn Sci* 22, 258-269.

Arieli, A., et al., 1996. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science* 273, 1868-1871.

Ayuso-Mateos, J.L., et al., 2010. From depressive symptoms to depressive disorders: the relevance of thresholds. *Br J Psychiatry* 196, 365-371.

Bandettini, P.A., 2012. Twenty years of functional MRI: the science and the stories. *Neuroimage* 62, 575-588.

Barch, D.M., et al., 2013. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* 80, 169-189.

Bassett, D.S., Sporns, O., 2017. Network neuroscience. *Nat Neurosci* 20, 353-364.

Behzadi, Y., et al., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* 37, 90-101.

Berger, H., 1929. Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten* 87, 527-570.

Berkes, P., et al., 2011. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* 331, 83-87.

Biswal, B., et al., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med* 34, 537-541.

Biswal, B.B., et al., 2010. Toward discovery science of human brain function. *Proc Natl Acad Sci U S A* 107, 4734-4739.

Blamire, A.M., et al., 1992. Dynamic mapping of the human visual cortex by high-speed magnetic resonance imaging. *Proc Natl Acad Sci U S A* 89, 11069-11073.

Bray, S., et al., 2007. Direct instrumental conditioning of neural activity using functional magnetic resonance imaging-derived reward feedback. *J Neurosci* 27, 7498-7507.

Broyd, S.J., et al., 2009. Default-mode brain dysfunction in mental disorders: a systematic review. *Neurosci Biobehav Rev* 33, 279-296.

Buckner, R.L., et al., 2013. Opportunities and limitations of intrinsic functional connectivity MRI. *Nat Neurosci* 16, 832-837.

Burnham, K.P., Anderson, D.R., 2003. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

Caballero-Gaudes, C., Reynolds, R.C., 2017. Methods for cleaning the BOLD fMRI signal. *Neuroimage* 154, 128-149.

Calhoun, V.D., et al., 2001. A method for making group inferences from functional

MRI data using independent component analysis. *Hum Brain Mapp* 14, 140-151.

Castellanos, F.X., et al., 2013. Clinical applications of the functional connectome. *Neuroimage* 80, 527-540.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. *BioData Min* 10, 35.

Ciric, R., et al., 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* 154, 174-187.

Clementz, B.A., et al., 2016. Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *Am J Psychiatry* 173, 373-384.

Cohen, J.D., et al., 2017. Computational approaches to fMRI analysis. *Nat Neurosci* 20, 304-313.

Compston, A., 2010. The Berger rhythm: potential changes from the occipital lobes in man. *Brain* 133, 3-6.

Connolly, C.G., et al., 2013. Resting-state functional connectivity of subgenual anterior cingulate cortex in depressed adolescents. *Biol Psychiatry* 74, 898-907.

Cortese, A., et al., 2016. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat Commun* 7, 13669.

Cox, R.W., et al., 1995. Real‐time functional magnetic resonance imaging. *Magn Reson Med* 33, 230-236.

Dansereau, C., et al., 2017. Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. *Neuroimage* 149, 220-232.

Davey, C.G., et al., 2012. Regionally specific alterations in functional connectivity of the anterior cingulate cortex in major depressive disorder. *Psychol Med* 42, 2071-2081.

deBettencourt, M.T., et al., 2015. Closed-loop training of attention with real-time brain imaging. *Nat Neurosci* 18, 470-475.

deCharms, R.C., 2008. Applications of real-time fMRI. *Nat Rev Neurosci* 9, 720-729.

deCharms, R.C., et al., 2005. Control over brain activation and pain learned by using real-time functional MRI. *Proc Natl Acad Sci U S A* 102, 18626-18631.

Deco, G., Kringelbach, M.L., 2014. Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron* 84, 892-905.

Di Martino, A., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry* 19, 659-667.

Dosenbach, N.U., et al., 2010. Prediction of individual brain maturity using fMRI. *Science* 329, 1358-1361.

Dozois, D.J.A., Covin, R., 2004. The Beck Depression Inventory-II (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS). *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment.* John Wiley & Sons Inc, Hoboken, NJ, US, pp. 50-69.

Drysdale, A.T., et al., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med* 23, 28-38.

Dutta, A., et al., 2014. Resting state networks in major depressive disorder. *Psychiatry Res* 224, 139-151.

Esmail, S., Linden, D., 2014. Neural Networks and Neurofeedback in Parkinson's Disease. *NeuroRegulation* 1, 240-272.

Esteban, O., et al., 2018. FMRIPrep: a robust preprocessing pipeline for functional MRI. *bioRxiv*.

Ezaki, T., et al., 2017. Energy landscape analysis of neuroimaging data. *Philos Trans A Math Phys Eng Sci* 375.

Ferrari, A.J., et al., 2013. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med* 10, e1001547.

Finn, E.S., et al., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci* 18, 1664-1671.

Fornito, A., et al., 2015. The connectomics of brain disorders. *Nat Rev Neurosci* 16, 159-172.

Fortin, J.P., et al., 2018. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167, 104-120.

Fortin, J.P., et al., 2017. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* 161, 149-170.

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci* 8, 700-711.

Fox, M.D., et al., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci U S A* 102, 9673-9678.

Furman, D.J., et al., 2011. Frontostriatal functional connectivity in major depressive disorder. *Biology of Mood & Anxiety Disorders* 1, 11.

Glasser, M.F., et al., 2016a. A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171-178.

Glasser, M.F., et al., 2016b. The Human Connectome Project's neuroimaging approach. *Nat Neurosci* 19, 1175-1187.

Goodkind, M., et al., 2015. Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry* 72, 305-315.

Gorgolewski, K.J., et al., 2017. BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Comput Biol* 13, e1005209.

Gountouna, V.E., et al., 2010. Functional Magnetic Resonance Imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. *Neuroimage* 49, 552-560.

Greicius, M.D., et al., 2007. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol Psychiatry* 62, 429-437.

Hawellek, D.J., et al., 2011. Increased functional connectivity indicates the severity of cognitive impairment in multiple sclerosis. *Proc Natl Acad Sci U S A* 108, 19066-19071.

Hay, S.I., et al., 2017. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* 390, 1260-1344.

He, B.J., et al., 2007. Breakdown of functional connectivity in frontoparietal networks underlies behavioral deficits in spatial neglect. *Neuron* 53, 905-918.

He, Y., et al., 2018. Ultra-Slow Single-Vessel BOLD and CBV-Based fMRI Spatiotemporal Dynamics and Their Correlation with Neuronal Intracellular Calcium Signals. *Neuron* 97, 925-939 e925.

Hinds, O., et al., 2013. Roles of default-mode network and supplementary motor area in human vigilance performance: evidence from real-time fMRI. *J Neurophysiol* 109, 1250-1258.

Hutton, C., et al., 2002. Image distortion correction in fMRI: A quantitative evaluation. *Neuroimage* 16, 217-240.

Ichikawa, N., et al., 2017. Identifying melancholic depression biomarker using whole-brain functional connectivity. *arXiv preprint arXiv:1704.01039*.

Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? Precisely. *Science* 348, 499-500.

Ito, T., et al., 2017. Cognitive task information is transferred between brain regions via resting-state network topology. *Nat Commun* 8, 1027.

Jacobi, F., et al., 2004. 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.

Jenkinson, M., 2003. Fast, automated, N-dimensional phase-unwrapping algorithm. *Magn Reson Med* 49, 193-197.

Jezzard, P., Balaban, R.S., 1995. Correction for geometric distortion in echo planar images from B0 field variations. *Magn Reson Med* 34, 65-73.

Jezzard, P., Clare, S., 1999. Sources of distortion in functional MRI data. *Hum Brain Mapp* 8, 80-85.

Johnson, W.E., et al., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118-127.

Kaiser, R.H., et al., 2015. Large-Scale Network Dysfunction in Major Depressive Disorder: A Meta-analysis of Resting-State Functional Connectivity. *JAMA Psychiatry* 72, 603-611.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8, 679-685.

Kassebaum, N.J., et al., 2016. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* 388, 1603-1658.

Kelly, A.M., et al., 2008. Competition between functional brain networks mediates behavioral variability. *Neuroimage* 39, 527-537.

Kenet, T., et al., 2003. Spontaneously emerging cortical representations of visual attributes. *Nature* 425, 954-956.

Koizumi, A., et al., 2016. Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nat Hum Behav* 1.

Koush, Y., et al., 2015. Learning Control Over Emotion Networks Through Connectivity-Based Neurofeedback. *Cereb Cortex*.

Koush, Y., et al., 2017. Learning Control Over Emotion Networks Through Connectivity-Based Neurofeedback. *Cereb Cortex* 27, 1193-1202.

Koush, Y., et al., 2013. Connectivity-based neurofeedback: dynamic causal modeling for real-time fMRI. *Neuroimage* 81, 422-430.

Kuhn, S., Gallinat, J., 2013. Resting-state brain activity in schizophrenia and major depression: a quantitative meta-analysis. *Schizophr Bull* 39, 358-365.

LaConte, S.M., et al., 2007. Real-time fMRI using brain-state classification. *Hum Brain Mapp* 28, 1033-1044.

Laird, A.R., et al., 2011. Behavioral interpretations of intrinsic connectivity networks. *J Cogn Neurosci* 23, 4022-4037.

Lancaster, J.L., et al., 1997. Automated labeling of the human brain: a preliminary report on the development and evaluation of a forward-transform method. *Hum Brain Mapp* 5, 238-242.

Laufs, H., et al., 2003. Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. *Proc Natl Acad Sci U S A* 100, 11053-11058.

Lee, S.H., et al., 2013. Genetic relationship between five psychiatric disorders

estimated from genome-wide SNPs. *Nat Genet* 45, 984-994.

Li, T., et al., 2017. Brain-Wide Analysis of Functional Connectivity in First-Episode and Chronic Stages of Schizophrenia. *Schizophr Bull* 43, 436-448.

Liew, S.L., et al., 2016. Improving Motor Corticothalamic Communication After Stroke Using Real-Time fMRI Connectivity-Based Neurofeedback. *Neurorehabil Neural Repair* 30, 671-675.

Linden, D.E., et al., 2012. Real-time self-regulation of emotion networks in patients with depression. *PLoS One* 7, e38115.

Liu, C., et al., 2015. Predicting stroop effect from spontaneous neuronal activity: a study of regional homogeneity. *PLoS One* 10, e0124405.

Lu, H., et al., 2007. Synchronized delta oscillations correlate with the resting-state functional MRI signal. *Proc Natl Acad Sci U S A* 104, 18265-18269.

Ma, Y., et al., 2016. Resting-state hemodynamics are spatiotemporally coupled to synchronized and symmetric neural activity in excitatory neurons. *Proc Natl Acad Sci U S A* 113, E8463-E8471.

Magri, C., et al., 2012. The amplitude and timing of the BOLD signal reflects the relationship between local field potential power at different frequencies. *J Neurosci* 32, 1395-1407.

Maldjian, J.A., et al., 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* 19, 1233-1239.

Mateo, C., et al., 2017. Entrainment of Arteriole Vasomotor Fluctuations by Neural Activity Is a Basis of Blood-Oxygenation-Level-Dependent "Resting-State" Connectivity. *Neuron* 96, 936-948 e933.

Matsui, T., et al., 2016. Transient neuronal coactivations embedded in globally propagating waves underlie resting-state functional connectivity. *Proc Natl Acad Sci U S A* 113, 6556-6561.

Matthews, B.W., 1975a. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405, 442-451.

Matthews, B.W., 1975b. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405, 442-451.

Matthews, P.M., et al., 2006. Applications of fMRI in translational medicine and clinical practice. *Nat Rev Neurosci* 7, 732-744.

McTeague, L.M., et al., 2017. Identification of Common Neural Circuit Disruptions in Cognitive Control Across Psychiatric Disorders. *Am J Psychiatry* 174, 676-685.

Megumi, F., et al., 2015. Functional MRI neurofeedback training on connectivity between two regions induces long-lasting changes in intrinsic functional network. *Front Hum Neurosci* 9, 160.

Menon, V., 2015. Salience Network. 597-611.

Milham, M.P., 2012. Open neuroscience solutions for the connectome-wide association era. *Neuron* 73, 214-218.

Mill, R.D., et al., 2017. From connectome to cognition: The search for mechanism in human functional brain networks. *Neuroimage* 160, 124-139.

Minzenberg, M.J., et al., 2009. Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Arch Gen Psychiatry* 66, 811-822.

Mulders, P.C., et al., 2015. Resting-state functional connectivity in major depressive disorder: A review. *Neurosci Biobehav Rev* 56, 330-344.

Munafò, M.R., et al., 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 0021.

Ng, B., et al., 2014. Transport on Riemannian manifold for functional

connectivity-based classification. *Med Image Comput Comput Assist Interv* 17, 405-412.

Nieuwenhuis, M., et al., 2017. Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. *Neuroimage* 145, 246-253.

Niv, S., 2013. Clinical efficacy and potential mechanisms of neurofeedback. *Personality and Individual Differences* 54, 676-686.

Noble, S., et al., 2017. Multisite reliability of MR-based functional connectivity. *Neuroimage* 146, 959-970.

Norman, K.A., et al., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10, 424-430.

Nosek, B.A., Errington, T.M., 2017. Making sense of replications. *Elife* 6.

Ogawa, S., et al., 1990a. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A* 87, 9868-9872.

Ogawa, S., et al., 1990b. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn Reson Med* 14, 68-78.

Ogawa, S., et al., 1992. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc Natl Acad Sci U S A* 89, 5951-5955.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97-113.

Orban, P., et al., 2018. Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. *Schizophr Res* 192, 167-171.

Pearlson, G., 2009. Multisite collaborations and large databases in psychiatric neuroimaging: advantages, problems, and challenges. *Schizophr Bull* 35, 1-2.

Peng, X., et al., 2018. Insular subdivisions functional connectivity dysfunction within major depressive disorder. *J Affect Disord* 227, 280-288.

Poldrack, R.A., et al., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci* 18, 115-126.

Poldrack, R.A., Farah, M.J., 2015. Progress and challenges in probing the human brain. *Nature* 526, 371-379.

Power, J.D., et al., 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* 84, 320-341.

Raichle, M.E., 2010. The brain's dark energy. *Sci Am* 302, 44-49.

Raichle, M.E., 2015a. The brain's default mode network. *Annu Rev Neurosci* 38, 433-447.

Raichle, M.E., 2015b. The restless brain: how intrinsic activity organizes brain function. *Philos Trans R Soc Lond B Biol Sci* 370.

Raichle, M.E., et al., 2001. A default mode of brain function. *Proc Natl Acad Sci U S A* 98, 676-682.

Rao, A., et al., 2017. Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage* 150, 23-49.

Reggente, N., et al., 2018. Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive-compulsive disorder. *Proc Natl Acad Sci U S A* 115, 2222-2227.

Rosenberg, M.D., et al., 2016. A neuromarker of sustained attention from whole-brain functional connectivity. *Nat Neurosci* 19, 165-171.

Roy, C.S., Sherrington, C.S., 1890. On the Regulation of the Blood-supply of the Brain. *The Journal of physiology* 11, 85-158.117.

Salomons, T.V., et al., 2014. Resting-state cortico-thalamic-striatal connectivity predicts response to dorsomedial prefrontal rTMS in major depressive disorder. *Neuropsychopharmacology* 39, 488-498.

Scheinost, D., et al., 2013. Orbitofrontal cortex neurofeedback produces lasting changes in contamination anxiety and resting-state connectivity. *Transl Psychiatry* 3, e250.

Scholvinck, M.L., et al., 2010. Neural basis of global resting-state fMRI activity. *Proc Natl Acad Sci U S A* 107, 10238-10243.

Sheline, Y.I., et al., 2010. Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proc Natl Acad Sci U S A* 107, 11020-11025.

Shen, X., et al., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* 82, 403-415.

Shibata, K., et al., 2016. Differential Activation Patterns in the Same Brain Region Led to Opposite Emotional States. *PLOS Biology* 14, e1002546.

Shibata, K., et al., 2011. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* 334, 1413-1415.

Shmuel, A., Leopold, D.A., 2008. Neuronal correlates of spontaneous fluctuations in fMRI signals in monkey visual cortex: Implications for functional connectivity at rest. *Hum Brain Mapp* 29, 751-761.

Singh, I., Rose, N., 2009. Biomarkers in psychiatry. *Nature* 460, 202-207.

Sitaram, R., et al., 2017. Closed-loop brain training: the science of neurofeedback. *Nat Rev Neurosci* 18, 86-100.

Smith, S.M., et al., 2013. Functional connectomics from resting-state fMRI. *Trends Cogn Sci* 17, 666-682.

Stam, C.J., 2014. Modern network science of neurological disorders. *Nat Rev Neurosci* 15, 683-695.

Stoeckel, L.E., et al., 2014. Optimizing real time fMRI neurofeedback for therapeutic discovery and development. *Neuroimage Clin* 5, 245-255.

Strikwerda-Brown, C., et al., 2015. Mapping the relationship between subgenual cingulate cortex functional connectivity and depressive symptoms across adolescence. *Soc Cogn Affect Neurosci* 10, 961-968.

Sulzer, J., et al., 2013. Real-time fMRI neurofeedback: progress and challenges. *Neuroimage* 76, 386-399.

Takagi, Y., et al., 2017. A Neural Marker of Obsessive-Compulsive Disorder from Whole-Brain Functional Connectivity. *Sci Rep* 7, 7538.

Takizawa, R., et al., 2014. Neuroimaging-aided differential diagnosis of the depressive state. *Neuroimage* 85 Pt 1, 498-507.

Taschereau-Dumouchel, V., et al., 2018. Towards an unconscious neural reinforcement intervention for common fears. *Proc Natl Acad Sci U S A* 115, 3470-3475.

Thompson, G.J., et al., 2013. Short-time windows of correlation between large-scale functional brain networks predict vigilance intraindividually and interindividually. *Hum Brain Mapp* 34, 3280-3298.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 267-288.

Tompson, S., et al., 2018. Network Approaches to Understand Individual Differences in Brain Connectivity: Opportunities for Personality Neuroscience. *Personal Neurosci* 1.

Tzourio-Mazoyer, N., et al., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273-289.

Vincent, J.L., et al., 2007. Intrinsic functional architecture in the anaesthetized monkey brain. *Nature* 447, 83-86.

Vos, T., et al., 2015. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* 386, 743-800.

Wallace, B.C., et al., 2011. Class Imbalance, Redux. 754-763.

Wang, L., et al., 2012. A systematic review of resting-state functional-MRI studies in major depression. *J Affect Disord* 142, 6-12.

Watanabe, T., et al., 2017. Advances in fMRI Real-Time Neurofeedback. *Trends Cogn Sci* 21, 997-1010.

Weiskopf, N., et al., 2006. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. *Neuroimage* 33, 493-504.

Weiskopf, N., et al., 2004. Self-regulation of local brain activity using real-time functional magnetic resonance imaging (fMRI). *J Physiol Paris* 98, 357-373.

Whelan, R., Garavan, H., 2014. When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol Psychiatry* 75, 746-748.

Winder, A.T., et al., 2017. Weak correlations between hemodynamic signals and ongoing neural activity during the resting state. *Nat Neurosci* 20, 1761-1769.

Wise, J., 2008. Consortium hopes to sequence genome of 1000 volunteers. *BMJ (Clinical research ed.)* 336, 237-237.

Woo, C.W., et al., 2017. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci* 20, 365-377.

Xia, C.H., et al., 2018. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat Commun* 9, 3003.

Xia, M., He, Y., 2017. Functional connectomics from a "big data" perspective. *Neuroimage* 160, 152-167.

Yahata, N., et al., 2017. Computational neuroscience approach to biomarkers and treatments for mental disorders. *Psychiatry Clin Neurosci* 71, 215-237.

Yahata, N., et al., 2016. A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nat Commun* 7, 11254.

Yamada, T., et al., 2017. Resting-State Functional Connectivity-Based Biomarkers and Functional MRI-Based Neurofeedback for Psychiatric Disorders: A Challenge for Developing Theranostic Biomarkers. *Int J Neuropsychopharmacol* 20, 769-781.

Yamashita, A., et al., 2017. Connectivity Neurofeedback Training Can Differentially Change Functional Connectivity and Cognitive Performance. *Cereb Cortex* 27, 4960-4970.

Young, K.D., et al., 2017. Randomized Clinical Trial of Real-Time fMRI Amygdala Neurofeedback for Major Depressive Disorder: Effects on Symptoms and Autobiographical Memory Recall. *Am J Psychiatry* 174, 748-755.

Young, K.D., et al., 2018. Amygdala real-time functional magnetic resonance imaging neurofeedback for major depressive disorder: A review. *Psychiatry Clin Neurosci* 72, 466-481.

Young, K.D., et al., 2014. Real-time FMRI neurofeedback training of amygdala activity in patients with major depressive disorder. *PLoS One* 9, e88785.

Yu, M., et al., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum Brain Mapp*.