Received: 28 February 2021 | Revised: 21 July 2021 | Accepted: 22 July 2021

DOI: 10.1002/hbm.25615

RESEARCH ARTICLE

WILEY

Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics

Norihide Maikusa[1,2] | Yinghan Zhu[1] | Akiko Uematsu[1] | Ayumu Yamashita[3] | Kousaku Saotome[1] | Naohiro Okada[4,5] | Kiyoto Kasai[4,5,6,7] | Kazuo Okanoya[1,4,6,7] | Okito Yamashita[3,8] | Saori C. Tanaka[3] | Shinsuke Koike[1,4,6,7]

[1]Center for Evolutionary Cognitive Sciences, Graduate School of Art and Sciences, The University of Tokyo, Tokyo, Japan

[2]Department of Radiology, National Center Hospital, National Center of Neurology and Psychiatry, Tokyo, Japan

[3]Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institutes International, Kyoto, Japan

[4]The International Research Center for Neurointelligence (WPI-IRCN), Institutes for Advanced Study (UTIAS), The University of Tokyo, Tokyo, Japan

[5]Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[6]The University of Tokyo Center for Integrative Science of Human Behavior (CiSHuB), Tokyo, Japan

[7]The University of Tokyo Institute for Diversity Adaptation of Human Mind (UTIDAHM), Tokyo, Japan

[8]Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan

Correspondence
Norihide Maikusa and Shinsuke Koike, Center for Evolutionary Cognitive Sciences, Graduate School of Art and Sciences, The University of Tokyo 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan.
Email: maikusa@g.ecc.u-tokyo.ac.jp (N. M.) and skoike-tky@umin.ac.jp (S. K.)

Funding information
Japan Agency for Medical Research and Development, Grant/Award Numbers: JP20dm0307001, JP20dm0307004, JP20dm0307008, JP20dm0307009; JSPS KAKENHI, Grant/Award Numbers: JP16H06280, JP20H03596; World Premier International Research Center for Neurointelligence (WPI-IRCN); UTokyo Center for Integrative Science of Human Behavior (CiSHuB)

Abstract

Multisite magnetic resonance imaging (MRI) is increasingly used in clinical research and development. Measurement biases—caused by site differences in scanner/image-acquisition protocols—negatively influence the reliability and reproducibility of image-analysis methods. Harmonization can reduce bias and improve the reproducibility of multisite datasets. Herein, a traveling-subject (TS) dataset including 56 T1-weighted MRI scans of 20 healthy participants in three different MRI procedures—20, 19, and 17 subjects in Procedures 1, 2, and 3, respectively—was considered to compare the reproducibility of TS-GLM, ComBat, and TS-ComBat harmonization methods. The minimum participant count required for harmonization was determined, and the Cohen's *d* between different MRI procedures was evaluated as a measurement-bias indicator. The measurement-bias reduction realized with different methods was evaluated by comparing test–retest scans for 20 healthy participants. Moreover, the minimum subject count for harmonization was determined by comparing test–retest datasets. The results revealed that TS-GLM and TS-ComBat reduced measurement bias by up to 85 and 81.3%, respectively. Meanwhile, ComBat showed a reduction of only 59.0%. At least 6 TSs were required to harmonize data obtained from different MRI scanners, complying with the imaging protocol predetermined for multisite investigations and operated with similar scan parameters. The results indicate that TS-based harmonization outperforms ComBat for measurement-bias reduction and is optimal for MRI data in well-prepared multisite investigations. One

drawback is the small sample size used, potentially limiting the applicability of ComBat. Investigation on the number of subjects needed for a large-scale study is an interesting future problem.

KEYWORDS
ComBat, FreeSurfer, harmonization, MRI, multisite, traveling subject

1 | INTRODUCTION

The archiving and sharing of large-scale clinical data from multisite brain magnetic resonance imaging (MRI) studies have gained considerable interest due to their potential to elucidate disease characteristics in the field of neuropsychiatric disorders. For example, the Global Alzheimer's Association Interactive Network project (Toga, Neu, Bhatt, Crawford, & Ashish, 2016), which aggregated data archives from 53 neuroimaging studies, such as the Alzheimer's Disease Neuroimaging Initiative (Jack et al., 2008), Open Access Series of Imaging Study (Marcus et al., 2007; Marcus, Fotenos, Csernansky, Morris, & Buckner, 2010), and Australian Imaging, Biomarker, & Lifestyle Flagship Study of Aging (Ellis et al., 2009), can be used to explore the data of more than 500,000 subjects. Other large-scale multisite imaging cohorts include the Human Connectome Project (HCP; Glasser et al., 2013), Human Brain Project (Rose, 2014), and UK Biobank (Sudlow et al., 2015). Multisite clinical research on MRI data has been expanding worldwide in recent years. Multisite MRI data can cause nonbiological measurement biases resulting from the differences in the properties of MRI scanners and procedures. These differences include scanner manufacturer, image-acquisition protocol, scanner coil, and field strength differences. Each of these can cause unwanted measurement biases that can negatively influence the reproducibility of the results and the ability to detect disease-related changes (He, Byge, & Kennedy, 2020; Koike et al., 2021; Ma et al., 2019).

In their initial multisite MRI research, Jack et al. (2008) standardized imaging protocols across sites in the ADNI project to reduce the effects of different imaging protocols on the MRI quality. Moreover, previous researchers have attempted to correct measurement biases using image pre-processing methods to treat raw MRI data. For instance, Sled, Zijdenbos, and Evans (1998) proposed a method to rectify the MRI-intensity inhomogeneity (also referred to as the "bias field") using the nonparametric nonuniformity intensity normalization (N3) approach. Further, Tustison et al. (2010) proposed an improved version of N3 referred to as Nick's N3 (N4). The N3 approach involves built-in prepossessing for the FreeSurfer pipeline. Janke, Zhao, Cowin, Galloway, and Doddrell (2004) developed a gradient nonlinearity correction software package called Gradunwarp to reduce the spatial-distortion bias based on the spherical-harmonic expansion, and finally, Maikusa et al. (2013) proposed a phantom-based distortion–correction method. Gradunwarp has been adopted by the ADNI project and HCP.

The aforementioned attempts were made to standardize imaging protocols and image pre-processing, but the measurement bias could not be eliminated (Beer et al., 2020). Therefore, sites were utilized as covariates in a general linear model (GLM) using dummy variables in previous studies (Fortin et al., 2017, 2018). However, this method is limited in that the effects of individual subjects and sites cannot be separated. Thus, the existing GLM methods eliminate biologically meaningless and meaningful biases. Recently, several large-scale investigations have considered the application of the meta-analytic approach to multisite datasets (Koshiyama et al., 2020; Okada et al., 2016; Van Erp et al., 2018). However, this approach is limited in that (a) the publication bias, wherein negative results are less likely to be published, reveals positive results in the original data to be combined; (b) the quality of brain images and clinical assessments varies significantly; (c) individual-based statistics cannot be obtained; and (d) survey literature and/or records are missing.

ComBat is an empirical Bayesian-based harmonization method that was originally designed for genomic microarray data (Johnson, Li, & Rabinovic, 2007). Fortin et al. effectively harmonized fractional anisotropy and mean diffusivity data from diffusion tensor imaging using ComBat (Fortin et al., 2017) and estimated the cortical thickness (Fortin et al., 2018) using ComBat to improve the statistical and machine-learning classification power (Radua et al., 2020).

Nevertheless, site differences include measurement and sampling biases. A sampling bias is a difference in biological information (e.g., age, sex, and pathology) between sites and can affect MRI signals. Numerous subjects or sites are required to separate measurement bias from sampling bias; however, most multisite MRI studies have failed to perform this distinction because both bias types have been characterized based on their respective sites (Yamashita et al., 2019).

To eliminate the effects of measurement bias, Yamashita et al. (2019) extended GLM harmonization using a TS dataset. The machine and protocol availabilities for each site can be inferred from TS data, and therefore, TS measurements facilitate advanced preparation for multisite projects. Thus, because TS data are free from sampling bias, they can be considered to differentiate between measurement and sampling biases. In addition, the TS-GLM harmonization method can correct measurement bias and improve the signal-to-noise ratio of resting-state functional connectivity data (Yamashita et al., 2019).

Limited research has been conducted on the reproducibility of data obtained at different sites or by different scanners using the TS method. Furthermore, no researchers have compared the reproducibility of harmonized variables using the test–retest reproducibility, where only the sampling bias or number of subjects required for

harmonization could be considered. A test–retest dataset provides variables without measurement or sampling bias. Given that a TS dataset requires many subjects and incurs high scanning and travel costs, it would be useful to define the minimum number of subjects required for harmonization.

In this study, we evaluated the performance of three harmonization methods: TS-GLM, ComBat, and TS-ComBat (combination of TS and ComBat) harmonization. ComBat is a convenient method because it does not require additional scans and is adaptable to retrospective data. TS-GLM, on the other hand, has limitations in that it can only be applied to prospective data and has a high scanning cost (requires traveling), but it can be superior to separate biological bias and measurement bias (Yamashita et al., 2019). TS-ComBat is similar in principle. Therefore, a quantitative comparison of these methods in terms of their ability to improve reproducibility will provide useful insights to determine whether TS, which has a high imaging cost, should be implemented.

We assessed the abilities of the three methods to reduce the measurement bias and improve the reproducibility of T1-weighted MRI scans for the same subject. Furthermore, by evaluating the differences between the test and retest results, the minimum number of subjects required to achieve reproducibility was determined.

2 | MATERIALS AND METHODS

2.1 | TS and test–retest datasets

We evaluated the effects of scan-procedure differences on the structural characteristics of T1-weight brain images taken from 20 healthy control participants. We used two 3-T scanners and three procedures to acquire data (Table 1). In Procedure 1, a Philips Achieva with an 8-channel head coil was used; in Procedure 2, a Siemens Prisma with a 64-channel head coil was used; and in Procedure 3, the same Siemens Prisma with a 32-channel head coil was used. The protocols of Procedures 1 and 2 were determined according to previous Japanese multisite projects and had similar scan parameters (Koizumi et al., 2016; Taschereau-Dumouchel et al., 2018; Yamada et al., 2017; Yamashita et al., 2019; Yamashita, Hayasaka, Kawato, & Imamizu, 2017). The protocol of Procedure 3 was the same as that provided by the HCP (Glasser et al., 2013).

Because one subject was missing in Procedure 2 and three other subjects were missing in Procedure 3, 56 measurements were performed in total: 20, 19, and 17 measurements in Procedures 1, 2, and 3, respectively. The control participants included 7 women and 13 men, with a mean [*SD*] age of 24.3 [6.56] years, mean [*SD*] height of 168.1 [6.58] cm, and mean [*SD*] weight of 61.1 [10.4] kg at the first measurement. The median duration of the three scans was 22 days (range = 0–448 days). More detailed information, that is, age, sex, height, weight, and scan duration for each subject, can be found in Table S1.

To compare the harmonization methods in terms of their test–retest result reproducibility, we performed TS-independent scans of 40 images of 20 healthy participants (11 women and 9 men; mean [*SD*] age of 15.4 [0.42] years, height of 163.3 [7.03] cm, and weight of 52.1 [7.29] kg) with Procedure 3, considering a median interval of 2.5 days (range = 1–54 days) between successive scans. The preliminary analysis results reveal the Cohen's *d* between the test and retest datasets to be quite high for long intervals.

This study was approved by the Ethics Committee at the University of Tokyo (Approval No. 19-298), and all the participants provided informed consent to participate in this study prior to performing the initial measurement.

2.2 | Image pre-processing

To extract the cortical and subcortical volumes and cortical thickness, we used FreeSurfer software (version 6.0) (Dale, Fischl, & Sereno, 1999; Fischl, 2012; Fischl et al., 2002, 2004; Fischl &

TABLE 1 Scanner information and demographics of participants

	Procedure 1	Procedure 2	Procedure 3
Manufacturer	PHILIPS	SIEMENS	SIEMENS
Scanner model	Achieva	Prisma	Prisma
Head coil (ch)	8	64	32
Repetition time (ms)	7	1900	2,400
Echo time (ms)	3.17	2.53	2.22
In-plate resolution (mm^2)	1.0 × 1.0	1.0 × 1.0	0.8 × 0.8
Matrix size	256 × 256	256 × 256	256 × 240
Slice thickness (mm)	1.2	1.2	0.8
Slice direction	AP	AP	AP
Slice orientation	Sagittal	Sagittal	Sagittal
Pulse sequence	MPRAGE	MPRAGE	MPRAGE
Flip angle (°)	9	9	8
Number of participants	20	19	17

Dale, 2000; Fischl, Sereno, & Dale, 1999) with a CentOS PC and the "recon-all" pipeline with the default parameters. The FreeSurfer pipeline performs N3 (Sled et al., 1998) as part of the pre-processing to minimize the effect of intensity inhomogeneity. We obtained the cortical volume and thickness from the rh.aparc.a2009s.stats and lh.aparc.a2009s.stats files, derived from "Destrieux Atlas," and included 74 anatomical cortical regions in each hemisphere. For the subcortical volume, we used the aseg.stats file, which included 41 subcortical anatomical regions (the left-WM-hypointensities, right-WM-hypointensities, left-non-WM-hypointensities, and right-non-WM-hypointensities were excluded because their values were zero). The cortical volume, cortical thickness, and subcortical volume were used to assess the reduction in measurement bias using the three kinds of harmonization.

2.3 | Harmonization methods

In this study, $y(i, j, v)$ was the vth FreeSufer variable, that is, cortical thickness, volume, and subcortical volume within the arbitrary anatomical label for imaging procedure i for the jth subject; k was the number of procedures; and n was the total number of traveling subjects. The harmonization methods considered in this study are described as follows.

2.3.1 | ComBat harmonization

ComBat is a tool that was initially developed to correct the batch effect in genomics (Johnson et al., 2007) and has more recently been applied to MRI datasets (Fortin et al., 2017, 2018). ComBat corrects a type of multivariate dataset using an empirical Bayesian estimation approach and can be used to analyze datasets obtained through different scanning procedures. The ComBat methodology can be described as

$$y(i,j,v) = a(v) + \boldsymbol{X}^T(i,j)\boldsymbol{\beta}(v) + \gamma(i,v) + \delta(i,v)\varepsilon(i,j,v), \quad (1)$$

where, $\alpha(v)$ is the average anatomical volume at the reference site within the vth anatomical variable, $\boldsymbol{\beta}(v)$ is the $p \times 1$ vector of coefficients associated with the design matrix of biological covariates of interest (age, sex, weight, and height in this study), $\boldsymbol{X}(i, j)$ is the design matrix of the vth anatomical variable, and p is the number of biological covariates. $\varepsilon(i, j, v)$ is the error term, following a normal distribution with a mean of zero and a variance of $\sigma^2(v)$. The terms $\gamma(i, v)$ and $\delta(i, v)$ represent the additive and multiplicative site effects of procedure i on the vth anatomical volume or thickness, respectively. In ComBat harmonization, an empirical Bayesian framework is used to estimate $\gamma^*(i, v)$ and $\delta^*(i, v)$. The final ComBat harmonized values can be expressed as

$$y^{\text{combat}}_{(i,j,v)} = \frac{y_{(i,j,v)} - \widehat{a}_{(v)} - \boldsymbol{X}_{(i,j)}\widehat{\boldsymbol{\beta}}_{(v)} - \gamma^*_{(i,v)}}{\delta^*_{iv}} + \widehat{a}_{(v)} + \boldsymbol{X}_{(i,j)}\widehat{\boldsymbol{\beta}}_{(v)}, \quad (2)$$

where, $\widehat{\beta}(v)$ and $\widehat{a}(v)$ represent estimated coefficients associated with the biological covariates of interest and estimated population mean of the vth anatomical variable.

2.3.2 | TS-GLM harmonization

The use of GLM is the most basic approach to remove the site effects. We followed the TS-GLM harmonization method reported by Yamashita et al. (2019), which extends the GLM harmonization model using a TS dataset. The TS-GLM harmonization model can be described as follows:

$$y(i,j,v) = \boldsymbol{X}_s^T(i,j)\boldsymbol{\beta}_s(v) + \boldsymbol{X}_p^T(i,j)\boldsymbol{\beta}_p(v) + \boldsymbol{\varepsilon}(i,j,v), \quad (3)$$

where, $\boldsymbol{\beta_p}(v)$ represents the participant factor and $\boldsymbol{X_p}(i, j)$ is the $n \times 1$ vector of the participant indicator. $\boldsymbol{\beta_s}(v)$ represents the coefficient of the site factor, namely, the measurement bias, and $\boldsymbol{X_s}(i, j)$ is the $k \times 1$ vector of the site indicator. To estimate the respective parameters, we calculated the inverse matrix for $\boldsymbol{X_p}(i, j)$ and $\boldsymbol{X_s}(i, j)$. In this study, all the subjects were healthy and identical at each site; thus, the sampling bias was not considered. However, the design matrix of the GLM was rank-deficient; thus, we used the Moore–Penrose pseudo inverse matrix as the "pinv" function in MATLAB (R2016b) to estimate $\widehat{\beta}_s(v)$ and $\widehat{\beta}_p(v)$. After estimating $\widehat{\beta}_s(v)$, the TS-GLM harmonized anatomical volumes and thicknesses were set as follows:

$$y^{TS\,glm}(i,j,v) = y(i,j,v) - \boldsymbol{X}_s^T(i,j)\widehat{\boldsymbol{\beta}}_s(v)\,. \quad (4)$$

2.3.3 | TS-ComBat harmonization

We also extended the ComBat harmonization model to a TS dataset. Conventional ComBat estimates covariates $\boldsymbol{\beta}$, such as age and gender, to exclude individual effects from the measurement values, but in this model, individual effects are estimated by the traveling subjects, as in TS-GLM. This approach enables TS-ComBat to have full-control sampling bias like TS-GLM.

TS-ComBat is defined as follows:

$$y(i,j,v) = \alpha(v) + \boldsymbol{X_p}(i,j)\boldsymbol{\beta_p}(v) + \gamma(i,v) + \delta(i,v)\varepsilon(i,j,v). \quad (5)$$

Thus, the TS-ComBat harmonized volumes can be set as follows:

$$y^{\text{combat}}_{(i,j,v)} = \frac{y_{(i,j,v)} - \widehat{a}_{(v)} - \boldsymbol{X_p}_{(i,j)}\widehat{\boldsymbol{\beta}}_{p_{(v)}} - \gamma^*_{(i,v)}}{\delta^*_{iv}} + \widehat{a}_{(v)} + \boldsymbol{X}_{(i,j)}\widehat{\boldsymbol{\beta}}_{(v)}. \quad (6)$$

ComBat, TS-GLM, and TS-ComBat all assume a normal distribution. We used the Kolmogorov–Smirnov test (KS-test) for all input data to check the guarantee of normality.

2.4 | Evaluation metrics

To investigate and compare the reproducibility of the different procedures before and after the implementation of these harmonization methods, we computed the Cohen's *d* effect size of different MRI procedures as the metric of the measurement bias or reproducibility of the anatomical variables—cortical volume/thickness and subcortical volume—between the different procedures.

$$sc = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}},$$

$$\text{Cohen's } d = \frac{|\bar{x}_1 - \bar{x}_2|}{sc}. \quad (7)$$

In the above expression, n_1 and n_2 denote the numbers of subjects in groups 1 and 2, respectively, and $\bar{x}_1$ and $\bar{x}_2$ and s_1 and s_2 denote the average and *SD* of each variable in groups 1 and 2, respectively. In this study, Cohen's *d* was calculated between Procedures 1 and 2, 1 and 3, and 2 and 3 for each FreeSurfer variable, that is, cortical thickness, cortical volume, and subcortical volume within the brain region. If there is no difference between the procedures, Cohen's *d* must equal zero.

2.5 | Statistical analyses

2.5.1 | Comparison of three harmonization methods

To explore the effects of the harmonization methods on reproducibility, we employed a general linear mixed model (GLMM) to estimate Cohen's *d* as a dependent variable, with the procedure and harmonization method as independent variables and the anatomical structures as within-subject variables. In this manner, we investigate whether each harmonization method improves the Cohen's *d* between Procedures 1 and 2, 1 and 3, and 2 and 3 with respect to the corresponding raw values. Furthermore, by comparison with the Cohen's *d* between test and retest, we investigate whether each harmonization method achieves the same reproducibility as the test–retest dataset. To assess the potential differences in the associations between the dependent and independent variables among the anatomical structures, we set a random effect for the intercept of anatomical structures. The GLMMs were estimated using the "lmer" function in the "lmerTest" package for R, version 3.1.2. A *p* value <.05 was considered significant. The Bonferroni correction for multiple comparisons to control the familywise error (*FWE*) was used for post-hoc analyses (*FWE*-corrected $p = .05/3 = .0166$).

Next, we tested the differences in Cohen's *d* between the TS dataset (no harmonization, TS-GLM, ComBat, and TS-ComBat) and the test–retest dataset using a two-sample *t*-test. The Bonferroni correction was also applied (*FWE*-corrected $p = .05/4 = .0125$).

To test the effect of scan duration on the harmonization, we obtained the effect size, that is, Cohen's *d*, for cortical thickness and cortical/subcortical volume across MRI procedures, and then tested Cohen's *d* (dependent variable) across MRI procedure difference, as scan duration as an independent variable and subject as a within-subject factor, using a repeated-measures analysis of variance (*ANOVA*).

2.5.2 | Minimum number of participants for TS harmonization

We re-sampled *s* subjects from the all *S* TSs corresponding to all combinations (${}_{S-1}C_s$) and calculated Cohen's *d* as a function of *s*, that is, *d*(*s*) after ComBat, TS-GLM, and TS-ComBat harmonization. Subsequently, we performed a two-sample *t*-test to compare the values of *d*(*s*) obtained for the test and retest scans. We defined the minimum number of subjects required for TS as the minimum *s* for which the null hypothesis of a difference from test–retest was rejected.

We performed a preliminary assessment to ensure that the all sampled data followed a normal distribution in the KS-test. We applied the Benjamini–Hochberg procedure to control the false discovery rate (*FDR*) of a family of hypotheses ($q < 0.05$) because Bonferroni correction is conservative and, therefore, could overestimate the required number of TSs.

3 | RESULTS

3.1 | Evaluation of three harmonization methods

The GLMM showed that all harmonization methods significantly reduced Cohen's *d* across all the procedures ($p < .001$, Figure 1). ComBat harmonization reduced the averaged Cohen's *d* for each FreeSurfer variable, namely, the cortical thickness, cortical volume, and subcortical volume, in the corresponding brain region by 59.0, 29.1, and 40.1% when comparing Procedures 1 and 2, 2 and 3, and 1 and 3, respectively. Similarly, TS-GLM and TS-ComBat reduced the averaged Cohen's *d* by 85.0%, 50.0%, and 68.5% and 81.3%, 48.1%, and 65.6% in the three above-mentioned comparisons, respectively.

Cohen's *d* before harmonization was significantly greater than the test–retest difference. Meanwhile, Cohen's *d* after ComBat harmonization was significantly greater than the test–retest difference when comparing Procedures 2 and 3 (0.187 [0.0191] vs. the test–retest effect size, *FWE*-corrected $p < .001$) and Procedures 1 and 3, but no significant difference was found when comparing Procedures 1 and 2. The TS-GLM and TS-ComBat harmonization methods did not differ significantly from the test–retest reproducibility when comparing Procedures 2 and 3; the values of Cohen's *d* after TS-GLM and TS-ComBat were significantly smaller than the test–retest value when comparing Procedures 1 and 2 (*FWE*-corrected *p*s < .0001) and Procedures 1 and 3 (*FWE*-corrected *p*s < .005).

Repeated measure ANOVA did not show significant main effect of scan duration in Cohen's *d* for cortical thickness ($F[1, 45] = 0.508$, $p = .480$) and subcortical/ cortical volume ($F[1, 45] = 1.130$,

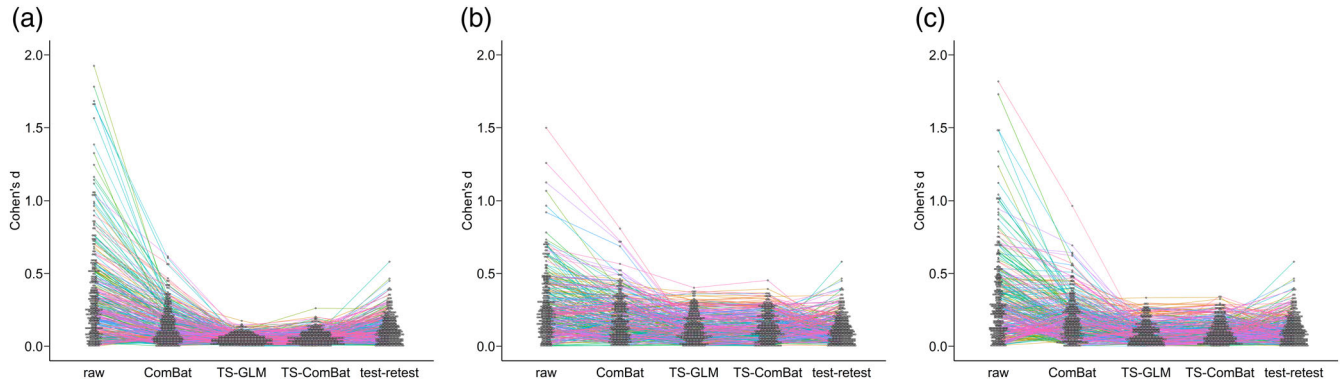

FIGURE 1 Bee-swarm plots for Cohen's *d* values before and after harmonization. Cohen's *d* values were derived from comparison of (a) Procedures 1 and 2, (b) Procedures 2 and 3, and (c) Procedures 1 and 3. The test–retest results have been plotted in all the subplots for comparison. The colored line indicates Cohen's *d* of an arbitrary FreeSurfer's anatomical label between procedures

$p = .293$).The mean Cohen's *d* for the maximum scan duration (448 days) was 0.01.

3.2 | Spatial distribution of the different procedures

There was a trend toward a higher averaged Cohen's *d* in the medial prefrontal cortex and inferior occipital cortex for both volume and thickness; specifically, the right medial orbital sulcus has the largest Cohen's *d* before harmonization (Figure 2). After ComBat harmonization, the Cohen's *d* values corresponding to the volume and thickness of the medial prefrontal cortex are reduced. However, they nonetheless exceed those observed in other regions. After applying the TS-GLM and TS-ComBat harmonization methods, Cohen's *d* values were lower in all the cortical and subcortical regions.

The spatial distributions of Cohen's *d* between Procedures 1 and 2, 1 and 3, and 2 and 3 are shown in Figures S1, S2, and S3, respectively. Irrespective of the procedures, a high Cohen's *d* was consistently observed around the medial prefrontal cortex before harmonization, which was well corrected by TS-GLM and TS-Combat. ComBat had a moderate effect on harmonization; in other words, ComBat could not remove the strong site effect around the medial frontal cortex when comparing Procedures 1 and 3 (Figure S2).

3.3 | Minimum number of participants required for harmonization

After TS-GLM harmonization, Cohen's *d* between Procedures 1 and 2 was not significantly different from the test–retest difference when the number of TSs was at least 6 (*FDR*-corrected $p > .05$). Thus, the minimum number of subjects required was 6 in Procedure 1, which involved different MRI scanners with a similar MRI protocol (SRPB) (Figure 3a). Similarly, with TS-ComBat harmonization, the minimum number of TSs was 13 in Procedure 1. Furthermore, the minimum number of TSs was 12 for TS-GLM and 14 for TS-ComBat in Procedure 2, which involved the same MRI scanner but different MRI protocols, that is, SRPB and CRHD (Figure 3b). In addition, the minimum number of TSs for TS-GLM was 19; however, the Cohen's *d* value after TS-ComBat harmonization remained significantly higher than the test–retest difference in Procedure 3, which involved different MRI scanners and protocols. (Figure 3(c)). In contrast, ComBat harmonization consistently showed significantly higher Cohen's *d* values than the test–retest differences (*FDR*-corrected $p < .05$), regardless of the scanning procedure.

4 | DISCUSSION

We compared the three harmonization methods with three measurement procedures using the TS dataset, as well as a test–retest dataset. Although considerable measurement bias was confirmed prior to harmonization, the TS-based harmonization results obtained by applying the TS-GLM and TS-ComBat approaches to the test and retest results were observed to be comparable and, hence, reproducible. Because the test–retest dataset did not have measurement bias, Cohen's *d* was expected to be zero, but the actual results were different. We expected additional factors to be present that could have affected the reproducibility such as image analysis error, individual errors, and measurement bias. The advantage of the TS-GLM and TS-ComBat methods are that they can harmonize these factors without modeling; therefore, these TS-based harmonization methods showed better reproducibility in the test–retest case (Figure 1).

In contrast, ComBat harmonization yielded a Cohen's *d* higher than the test–retest difference. These facts indicate that the biological covariates used in this study were not sufficient to estimate individual effects and that the measurement bias and individual effects could not be separated.

When we focused on the spatial distribution of Cohen's *d*, a greater measurement bias was found in the medial prefrontal cortex. The results indicate that the measurement bias between the procedures has a moderate effect size, irrespective of procedure differences and structural characteristics. Larger effect sizes were observed

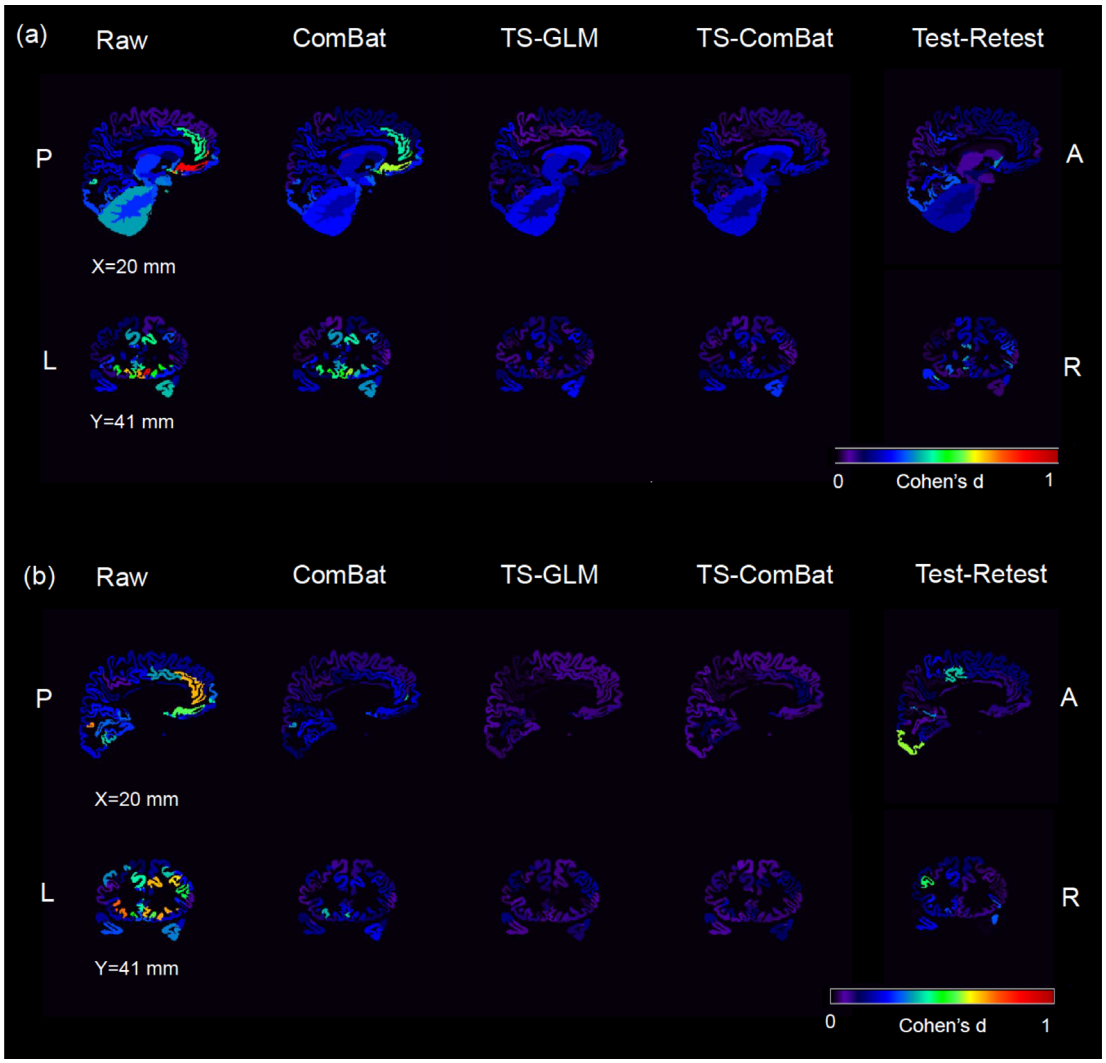

FIGURE 2 Averaged Cohen's *d* maps overlaid on aparc.a2009s + aseg.mgz file. The upper and lower rows show sagittal and coronal images, respectively. The columns indicate raw, ComBat, TS-GLM, TS-ComBat, and test–retest results obtained using each harmonization method. Cohen's *d* values were calculated from (a) the cortical and subcortical volumes and (b) the cortical thickness

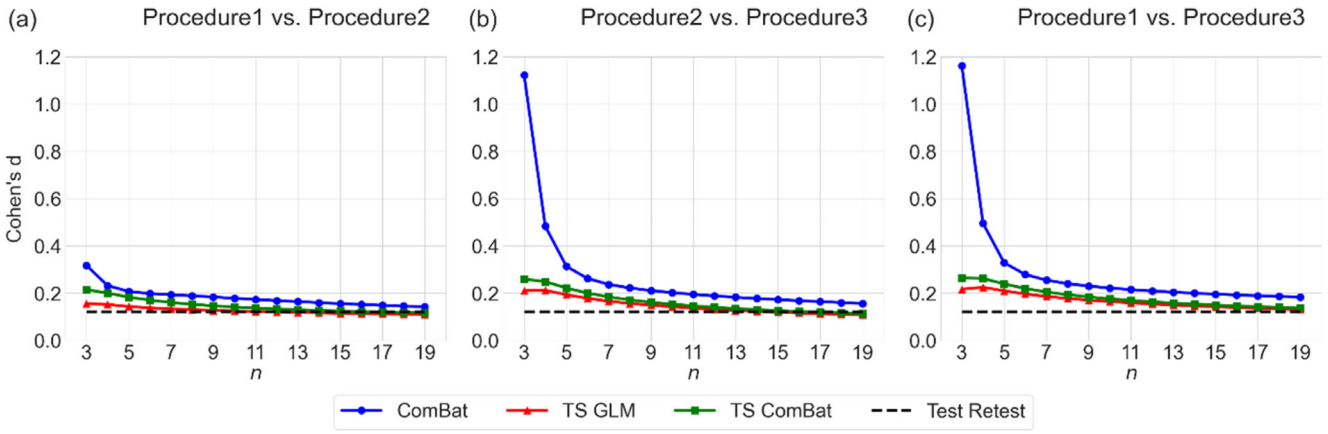

FIGURE 3 Average Cohen's *d* according to number of re-sampling subjects. Cohen's *d* as a function of *s*, the number of subjects resampled, from the comparison of (a) Procedure 1 and 2, (b) Procedures 2 and 3, and (c) Procedures 1 and 3

in the ventral and medial parts of the frontal cortex (i.e., medial prefrontal cortex). The results agree with those of previous studies; the specific bias in this region coincides with the location of high geometric distortion (Li, Williams, Frisk, Arnold, & Smith, 1995; Maikusa et al., 2013). According to Li et al. (1995), the boundary of the nasal cavity below the medial prefrontal cortex induces strong geomatical distortion (i.e., measurement bias) in these areas. Although harmonization methods cannot provide information on these characteristics, they can harmonize and statistically correct differences between scanners without the need to know the details of these characteristics. This is the advantage of TS-based harmonization methods.

We defined the minimum TS sample size for nonsignificance between the harmonization and test–retest reproducibility. For TS-GLM, it was 6 with different MRI scanners but similar protocols (Procedures 1 and 2). TS-GLM required a minimum TS sample size of 12 and 19 when comparing Procedures 1 and 3 and Procedures 2 and 3, respectively. To the best of our knowledge, this study was the first to compare the effectiveness of harmonization methods and to suggest a sample-size requirement for TS-based harmonization. Furthermore, the procedure comparisons, in the decreasing order of minimum TS sample sizes, were Procedures 1 versus 2, Procedures 2 versus 3, and Procedures 1 versus 3, which coincides with the result presented in Section 3.1, that is, the procedure comparisons, in the increasing order of reduction rates of Cohen's *d* after harmonization, were Procedures 1 versus 2, Procedures 1 versus 3, and Procedures 2 versus 3.

The measurement bias could not be fully corrected when ComBat harmonization was used, perhaps because ComBat harmonization does not exhibit the ability to harmonize the test–retest dataset with 20 subjects or less. In contrast, both TS-GLM and TS-ComBat successfully corrected the measurement bias, irrespective of procedural and brain region differences. A previous multisite fMRI study revealed the well-harmonized factors from a functional connectivity matrix using TS-GLM harmonization (Yamashita et al., 2019), suggesting the applicability of this method to the structural characteristics of the brain.

Although TS-based harmonization methods have caused measurement bias to decrease, considerable effort should be devoted toward TS recruitment and scan-schedule preparation within a short duration to obtain the TS dataset. Thus, determining the required number of TSs will help minimize the use of resources. As observed, although the highest number of subjects was required to compare Procedures 1 and 3, that is, different MRI scanners and protocols, only six TSs were required to compare Procedures 1 and 2, which involved different MRI scanners but similar scan parameters predetermined for a multisite investigation (Koizumi et al., 2016; Taschereau-Dumouchel et al., 2018; Yamada et al., 2017; Yamashita et al., 2017, 2019). The findings suggest that the required number of subjects varies depending on the procedure, and the attempt to unify the parameters highlights the importance of unifying imaging protocols when using MRI data obtained from different vendors and MRI scanners in multisite studies. It is considered ideal to scan up to 20 TSs; however, the operational costs increase with increasing site count, and it is difficult for participants to travel to multiple sites. It is practical to change the TS count depending on the differences between scanner configurations. We believe that TS should be implementation for all scanner vendors and MRI protocols to investigate the relation between the minimal sample size and measurement bias. However, this is not realistic, because scanning traveling subjects incurs high cost. Therefore, our study involved a minimal sample size and limited scenarios, that is, two different scanners with similar imaging protocols (the SRPB protocol), different protocols on the same scanner (SRPB and CRHD protocols), and different scanners and protocols. Our study provides guidance on the minimum number of TSs for limited scenarios; in particular, it provides guidance for the scenario of different scanners with similar protocols, which is similar to the scenario in a recent multi-site imaging study. In addition, we do not have sufficient longitudinal data to discuss whether the proposed harmonization methods are applicable to longitudinal data; we would like to examine this possibility with a new dataset in future work.

Van Erp et al. (2018) reported that, when compared with healthy subjects, schizophrenic subjects had lower thickness in the left and right cortices, with Cohen's $d = -0.530$ and -0.516, respectively; bilateral fusiform, temporal (inferior, middle, and superior), and left superior frontal gyri; right pars opercularis; and bilateral insula. Therefore, it is necessary to reduce the measurement bias so that it does not affect the size of the target disease. Our results showed that the averaged Cohen's *d* (measurement bias) for whole brain cortical thickness in all scenarios was 0.259 before harmonization, which is approximately half the effect size for the above diseases, and TS-GLM can reduce this value to 0.0710. We plan to consider the number of TSs required for the assumed effect sizes between different groups, such as disease and healthy control groups.

Our study has some limitations. First, the datasets were used for harmonization and validation. Ideally, independent TS validation datasets should be utilized. Moreover, TS-based harmonization is a method of estimating the measurement bias at the time of scanning, and the TS-scanning intervals may affect the harmonization accuracy; this tendency was not fully investigated in this study. In addition, we verified ComBat harmonization for only 20 subjects, which may not represent larger imaging studies, and the small sample size may have prevented better ComBat estimations. A large-scale TS project is currently underway, and we hope to have more detailed validation and analyses possible in future works.

Second, we only investigated the cortical thickness and volume obtained from the FreeSurfer analysis. Although the initial TS-GLM harmonization was confirmed using functional connectivity during a resting state (Yamashita et al., 2019), it is possible that other modalities—other MRI and positron-emission tomography imaging sequences—exhibit different trends. Third, the dropout of the TSs meant that there was an imbalance in number between the scanning procedures. For example, the result that the optimal number of TSs was 19 in the Section 3.3 was precisely the result of using 19 subjects in Procedure 1 and 17 subjects in Procedure 3. This imbalance in the TS count between procedures due to dropout has been identified and

considered in similar extant studies. Lastly, we did not investigate how much of the TS-scanning interval could be feasibly harmonized.

Next, our TS data have a wide range of scan durations, which might have led to sampling bias caused by brain changes with normal aging. In this study, the minimum and maximum age of TSs were 20 and 40 years, respectively; brain changes due to aging are minimal in this age group. Therefore, we believe that brain changes are negligible, even if these scan durations were quite wide (maximum of 448 days). Repeated measures ANOVA for cortical thickness and subcortical/cortical volume did not show significant difference between scan durations to intra-subject Cohen's *d*. In fact, intra-subject Cohen's *d* was 0.01 at the maximum scan duration of 448 days.

Finally, there is risk that a harmonization method eliminates not only measurement bias but also biological information; in other words, it could cause a sampling bias. A harmonization method requires the separation of sampling bias from measurement bias as well as verity. However, in this study, sampling bias did not occur because the TSs showed the same sampling bias across the sites. Therefore, we would like to verify this risk using another dataset in future work.

In conclusion, our study showed that TS-based harmonization methods, namely, TS GLM and TS-ComBat, outperform ComBat harmonization. Furthermore, we demonstrated that at least six subjects are required when the dataset is scanned using different scanners with a similar scanning procedure. As a future endeavor, we intend to undertake a large-scale TS project to explain and resolve such problems associated with TS harmonization as the TS-scanning interval and validation of an independent test set.

ACKNOWLEDGMENTS

This study was supported by the Japan Agency for Medical Research and Development (AMED; JP20dm0307001, JP20dm0307004, JP20dm0307008 and JP20dm0307009), JSPS KAKENHI (JP16H06280 and JP20H03596), the UTokyo Center for Integrative Science of Human Behavior (CiSHuB), and the World Premier International Research Center for Neurointelligence (WPI-IRCN).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The datasets required the approval of an ethical review board. Please contact the corresponding author (S.K.).

ORCID

Norihide Maikusa https://orcid.org/0000-0003-0943-4684

REFERENCES

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., … Alzheimer's Disease Neuroimaging Initiative. (2020). Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, *220*, 117129. https://doi.org/10.1016/j.neuroimage.2020.117129

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., … Yastrubetskaya, O. (2009). The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672–687. https://doi.org/10.1017/S1041610209009405

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11050–11055. https://doi.org/10.1073/pnas.200033797

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

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

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*(2), 195–207. https://doi.org/10.1006/nimg.1998.0396

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., … Shinohara, R. T. (2018). NeuroImage harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

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

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., … Jenkinson, M. (2013). The minimal pre-processing pipelines for the human connectome project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

He, Y., Byge, L., & Kennedy, D. P. (2020). Nonreplication of functional connectivity differences in autism spectrum disorder across multiple sites and denoising strategies. *Human Brain Mapping*, *41*(5), 1334–1350. https://doi.org/10.1002/hbm.24879

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., … Weiner, M. W. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049

Janke, A., Zhao, H., Cowin, G. J., Galloway, G. J., & Doddrell, D. M. (2004). Use of spherical harmonic deconvolution methods to compensate for nonlinear gradient effects on MRI images. *Magnetic Resonance in Medicine*, *52*(1), 115–122. https://doi.org/10.1002/mrm.20122

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

Koike, S., Tanaka, S. C., Okada, T., Aso, T., Yamashita, A., Yamashita, O., … Brain/MINDS Beyond Human Brain MRI Group. (2021). Brain/MINDS beyond human brain MRI project: A protocol for multi-level harmonization across brain disorders throughout the lifespan. *NeuroImage: Clinical*, *30*, 102600. https://doi.org/10.1016/j.nicl.2021.102600.

Koizumi, A., Amano, K., Cortese, A., Shibata, K., Yoshida, W., Seymour, B., … Lau, H. (2016). Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour*, *1*, 6. https://doi.org/10.1038/s41562-016-0006

Koshiyama, D., Miura, K., Nemoto, K., Okada, N., Matsumoto, J., Fukunaga, M., & Hashimoto, R. (2020). Neuroimaging studies within

cognitive genetics collaborative research organization aiming to replicate and extend works of ENIGMA. *Human Brain Mapping*, 1–12. https://doi.org/10.1002/hbm.25040

Li, S., Williams, G., Frisk, T., Arnold, B., & Smith, M. (1995). A computer simulation of the static magnetic field distribution in the human head. *Magnetic Resonance in Medicine*, *34*(2), 268–275. http://doi.org/10.1002/mrm.1910340219

Ma, D., Popuri, K., Bhalla, M., Sangha, O., Lu, D., Cao, J., … Initiative, A.'s. D. N. (2019). Quantitative assessment of field strength, total intracranial volume, sex, and age effects on the goodness of harmonization for volumetric analysis on the ADNI database. *Human Brain Mapping*, *40*(5), 1507–1527. https://doi.org/10.1002/hbm.24463

Maikusa, N., Yamashita, F., Tanaka, K., Abe, O., Kawaguchi, A., Kabasawa, H., … Japanese Alzheimer's Disease Neuroimaging Initiative. (2013). Improved volumetric measurement of brain structure with a distortion correction procedure using an ADNI phantom. *Medical Physics*, *40*(6), 062303. https://doi.org/10.1118/1.4801913

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, *19*(9), 1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, *22*(12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407

Okada, N., Fukunaga, M., Yamashita, F., Koshiyama, D., Yamamori, H., Ohi, K., … Hashimoto, R. (2016). Abnormal asymmetries in subcortical brain volume in schizophrenia. *Molecular Psychiatry*, *21*(10), 1460–1466.

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., … Pineda-Zapata, J. (2020). Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage*, *218*, 116956. https://doi.org/10.1016/j.neuroimage.202.116956

Rose, R. (2014). The human brain project: Social and ethical challenges. *Neuron*, *82*(6), 1212–1215. https://doi.org/10.1016/j.neuron.2014.06.001

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, *17*(1), 87–97. https://doi.org/10.1109/42.668698

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UKbiobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J. D., Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(13), 3470–3475. https://doi.org/10.1073/pnas.1721572115

Toga, W., Neu, S. C., Bhatt, P., Crawford, K. L., & Ashish, N. (2016). The global Alzheimer's association interactive network. *Alzheimer's and Dementia*, *12*, 49–54. https://doi.org/10.1162/jocn.2009.21407

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Van Erp, T. G., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., … Turner, J. A. (2018). Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through MetaAnalysis (ENIGMA) consortium. *Biological Psychiatry*, *84*(9), 644–654. https://doi.org/10.1016/j.biopsych.2018.04.023

Yamada, T., Hashimoto, R. I., Yahata, N., Ichikawa, N., Yoshihara, Y., Okamoto, Y., … Kawato, M. (2017). Resting-state functional connectivity-based biomarkers and functional MRI-based neurofeedback for psychiatric disorders: A challenge for developing theranostic biomarkers. *International Journal of Neuropsychopharmacology*, *20*(10), 769–781. https://doi.org/10.1093/ijnp/pyx059

Yamashita, A., Hayasaka, S., Kawato, M., & Imamizu, H. (2017). Connectivity neurofeedback training can differentially change functional connectivity and cognitive performance. *Cerebral Cortex*, *27*(10), 4960–4970. https://doi.org/10.1093/cercor/bhx177

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., … Imamizu, H. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biology*, *17*(4), e3000042. https://doi.org/10.1371/journal.pbio.3000042

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S. C., & Koike, S. (2021). Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. *Human Brain Mapping*, *42*(16), 5278–5287. https://doi.org/10.1002/hbm.25615