


# Harmonizing T1-Weighted Images to Improve Consistency of Brain Morphology Among Different Scanner Manufacturers in Alzheimer's disease

Shilun Zhao, PhD,<sup>1,2</sup> Tianhao Zhang, PhD,<sup>1</sup> Wei Zhang, PhD,<sup>1,2</sup> Tingting Pan, PhD,<sup>3</sup> Ge Zhang, PhD,<sup>1,4</sup> Shuang Feng, MSc,<sup>3</sup> Xiwan Zhang, MSc,<sup>3</sup> Binbin Nie, PhD,<sup>1</sup> Hua Liu, PhD,<sup>1</sup> and Baoci Shan, PhD,<sup>1,2\*</sup>  For the Alzheimer's Disease Neuroimaging Initiative

**Background:** Brain MRI scanner variability can introduce bias in measurements. Harmonizing scanner variability is crucial.

**Purpose:** To develop a harmonization method aimed at removing scanner variability, and to evaluate the consistency of results in multicenter studies.

**Study Type:** Retrospective.

**Population:** Multicenter data from 170 healthy participants (males/females = 98/72; age =  $73.8 \pm 7.3$ ) and 170 Alzheimer's disease patients (males/females = 98/72; age =  $76.2 \pm 8.5$ ) were compared with reference data from another 340 participants.

**Field Strength/Sequence:** 3-T, magnetization prepared rapid gradient echo and turbo field echo; 1.5-T, inversion recovery prepared fast spoiled gradient echo T1-weighted sequences.

**Assessment:** Gray matter (GM) brain images, obtained through segmentation of T1-weighted images, were utilized to evaluate the performance of the harmonization method using common orthogonal basis extraction (HCOBE) and four other methods (removal of artificial voxel effect by linear regression, RAVEL; Z\_score; general linear model, GLM; ComBat). Linear discriminant analysis (LDA) was used to access the effectiveness of different methods in reducing scanner variability. The performance of harmonization methods in preserving GM volumes heterogeneity was evaluated by the similarity of the relationship between GM proportion and age in the reference and multicenter data. Furthermore, the consistency of the harmonized multicenter data with the reference data were evaluated based on classification results (train/test = 7/3) and brain atrophy.

**Statistical Tests:** Two-sample t-tests, area under the curve (AUC), and Dice coefficients were used to analyze the consistency of results from the reference and harmonized multicenter data. A *P*-value <0.01 was considered statistically significant.

**Results:** HCOBE reduced the scanner variability from 0.09 before harmonization to 0.003 (ideal: 0, RAVEL/Z\_score/GLM/ComBat = 0.087/0.003/0.006/0.013). GM volumes showed no significant difference (*P* = 0.52) between the reference and HCOBE-harmonized multicenter data. Consistency evaluation showed that AUC values of 0.95 for both reference and HCOBE-harmonized multicenter data (RAVEL/Z\_score/GLM/ComBat = 0.86/0.86/0.84/0.89), and the Dice coefficient increased from 0.73 before harmonization to 0.82 (ideal: 1, RAVEL/Z\_score/GLM/ComBat = 0.39/0.64/0.59/0.74).

**Data Conclusion:** HCOBE may help to remove scanner variability and could improve the consistency of results in multicenter studies.

**Level of Evidence:** 2

**Technical Efficacy Stage:** 1

J. MAGN. RESON. IMAGING 2024;59:1327–1340.

View this article online at [wileyonlinelibrary.com](http://wileyonlinelibrary.com). DOI: 10.1002/jmri.28887

Received Jan 4, 2023, Accepted for publication Jun 16, 2023.

\*Address reprint requests to: B.S., Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China. E-mail: [shanbc@ihep.ac.cn](mailto:shanbc@ihep.ac.cn)

Contract grant sponsor: National Natural Science Foundation of China; Contract grant numbers: 12205329, 12175268, and 11975249; Contract grant sponsor: NIH; Contract grant numbers: P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, and R01 EB009352.

From the <sup>1</sup>Beijing Engineering Research Center of Radiographic Techniques and Equipment, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China; <sup>2</sup>School of Nuclear Science and Technology, University of Chinese Academy of Sciences, Beijing, China; <sup>3</sup>School of Physics and Microelectronics, Zhengzhou University, Zhengzhou, China; and <sup>4</sup>School of Chemical Sciences, University of Chinese Academy of Sciences, Beijing, China

In brain research, MRI is widely used for the study of disorders and development.<sup>1,2</sup> Pooling MRI data from different centers brings several advantages, such as increasing the sample size, increasing the representation of pathological diversity, and promoting the development of robust and generalizable models.<sup>3</sup> However, it has been indicated that differences in MRI scanner characteristics (e.g., manufacturer, field strength, and software version) and variations in acquisition protocol parameters (e.g., echo time [TE], gradient orientation, and voxel size) can result in nonbiological measurement biases for voxel-based morphometry, regional cortical thickness, and brain volume assessments.<sup>4–7</sup> These biases are collectively known as the “center effect,” which describes the impact of the variability in scanner and acquisition protocol parameters.<sup>8,9</sup> Neglecting harmonization of center effects may lead to faulty and unreliable results, ultimately affecting the consistency of the results.<sup>10,11</sup> Therefore, it is crucial to remove center effects when pooling MRI data from multiple centers.

Several harmonization methods have been proposed to address the issue of center effects.<sup>12–15</sup> These methods can be broadly categorized into two groups: histogram-based methods and statistical covariate methods.<sup>10,16</sup> The former uses histograms to normalize MRI data from different centers, providing the advantage of simplicity and ease of implementation.<sup>10</sup> For instance, Fortin et al. proposed the removal of artificial voxel effect by linear regression (RAVEL) method, which utilizes cerebrospinal fluid (CSF) regions to correct gray matter (GM) brain regions affected by both scanner and disease characteristics.<sup>13</sup> Furthermore, Wachinger et al. used the Z-score normalization method to standardize MRI histograms from different centers, thereby removing intercenter differences.<sup>13,15</sup> The latter category utilized covariates to regress the effect of center on the data and could more accurately remove center differences. For example, Kostro et al. used a general linear model (GLM) to remove center differences by treating centers as covariates.<sup>14</sup> Fortin et al. used the ComBat method to harmonize fractional anisotropy in diffusion tensor imaging and cortical thickness in T1-weighted MRI data.<sup>12</sup> However, these existing methods fail to distinguish between center effects and biological heterogeneity caused by factors such as age, sex, and pathology, by indiscriminately removing both types of differences.<sup>17</sup>

These existing methods used the elimination of differences between images acquired at different centers as a reference standard to evaluate performance.<sup>12–15</sup> However, a harmonization method should ideally preserve biological heterogeneity (e.g., age- and pathology-related differences) while removing center effects to be effective. Such an approach can strengthen the consistency and reliability of research outcomes by removing extraneous nonbiological variance in downstream analyses.<sup>18</sup> Yet, there is no harmonization method available to address these requirements.

In this study, we aimed to develop a harmonization method using common orthogonal basis extraction (HCOBE). Furthermore, we used the single-center data as a reference. We evaluated the performance of HCOBE and four existing harmonization methods to correct the scanner and acquisition parameters, and also evaluated the consistency of the multicenter data harmonized by HCOBE and four existing methods with the reference data.

## Materials and Methods

This study obtained institutional review board (IRB) approval for each dataset. Informed consent was waived by the IRB because this study was a retrospective analysis of existing data.

### Data Selection

This study utilized T1-weighted images from two public datasets: OASIS (Open Access Series of Imaging Studies; <https://www.oasis-brains.org/#data>) and ADNI (Alzheimer’s Disease Neuroimaging Initiative; <https://adni.loni.usc.edu/data-samples/access-data/>).<sup>19,20</sup> Inclusion criteria for healthy control (HC) participants were as follows: age 18 years or older and absence of cognitive impairment as indicated by a clinical dementia rating (CDR) score of 0 and a mini-mental state examination (MMSE) score  $\geq 27$ . Patients with Alzheimer’s disease (AD) were required to meet NINCDS-ADRDA criteria and had to have cognitive scores of  $\text{CDR} \geq 0.5$  and  $\text{MMSE} < 27$ . We removed individuals with a history of neurological or psychiatric diseases, left-handed participants, pregnant or nursing women, and patients with implanted medical devices like pacemakers and drug pumps.

Images acquired using the same scanner and acquisition protocol were grouped into a center. The reference center (reference data = RD) was identified by choosing the center with the largest subset of data in the OASIS dataset, which utilized Siemens 3 T scanners for image acquisition. The subsets with the highest numbers of images from 3 T Philips and 1.5 T General Electric (GE) scanners in the ADNI dataset (multicenter data-Philips = MD-P; multicenter data-GE = MD-G) were selected as the multicenter data to account for scanner differences (Fig. 1). Following selection, only T1-weighted images of 298 cognitively normal (CN) participants and 271 patients with AD were retained for further analysis after matching for age [50,97], gender, and handedness. The receive coils of all scanners are 8-channel head coils. Table 1 lists the detailed acquisition parameters for all participants.

We created two categories for CN participants and AD patients: HC and AD groups. Both HC and AD groups contained data for four centers: RD, MD-P, MD-G, and multicenter data-Siemens (MD-S). The RD data were designated as the reference standard for evaluating the consistency of multicenter results. The multicenter data consisted of the remaining data from the other three centers. Furthermore, MD-S was a subset of RD and ensured equal statistical power in the multicenter and reference data. After inclusion of the MD-S, the multicenter data consisted of 170 CN participants and 170 patients with AD. Similarly, the reference data consisted of 170 CN participants and 170 patients. Demographic information about all participants is provided in Table 2.

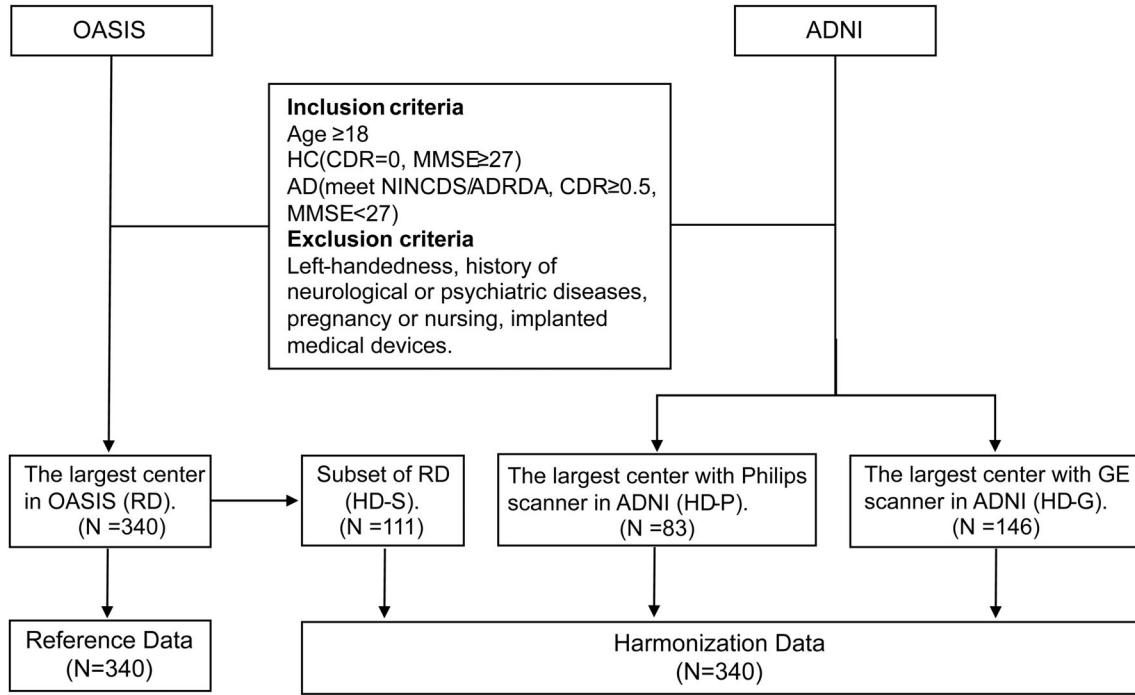


FIGURE 1: Flowchart of data acquisition. Acquisition criteria for reference data and multicenter data. Symbol N represents the total of CN participants and AD patients in this center. CDR = clinical dementia rating; MMSE = mini-mental state examination; RD = reference data; MD-S = multicenter data-Siemens; MD-P = multicenter data-Philips; MD-G = multicenter data-GE; ADNI = Alzheimer’s disease neuroimaging initiative; OASIS = open access series of imaging studies.

### Image Preprocessing

The T1-weighted images were initially segmented using the computational anatomy toolbox (CAT12; <https://neuro-jena.github.io/cat/>), which entailed brain images normalization to montreal neurological institute space with a voxel size of  $1.5 \times 1.5 \times 1.5$  mm and segmented the normalized brain image into GM, white matter, and CSF. Subsequently, all segmented GM images underwent smoothing with an  $8 \times 8 \times 8$  mm full width at half maximum (FWHM) Gaussian kernel to augment the signal-to-noise ratio. The preprocessing pipeline is presented schematically on the left of Fig. 2.

### A Harmonization Method Using Common Orthogonal Basis Extraction

**COMMON ORTHOGONAL BASIS EXTRACTION (COBE) METHOD.** The COBE method was initially introduced in multi-block data analysis to separate common and heterogeneity spaces.<sup>21,22</sup> In this study, we reformulated the COBE model for application on the segmented GM T1-weighted images. Let  $\mathcal{H}$  be a set of matrices encompassing  $\{Y_n \in \mathbb{R}^{D \times J_n}, n \in \mathcal{M}\}$ ,  $\mathcal{M} = \{1, 2, \dots, N\}$ , where  $Y_n$  represents the voxel matrix at center  $n$  ( $n = 1, 2, \dots, N$ ) and  $N$  is the total number of centers. The voxel matrix factorization of  $Y_n$  can be expressed as follows:

$$\min_{A_n, B_n} \|Y_n - A_n B_n^T\|^2, n \in \mathcal{M} \quad (1)$$

Let  $A_n = [\bar{A} \check{A}_n]$ ,  $n \in \mathcal{M}$  and  $B_n = [\bar{B} \check{B}_n]$ ,  $n \in \mathcal{M}$ , where  $\bar{A}$  contains the common feature shared by all voxel matrixes in  $\mathcal{M}$ , and  $\check{A}_n$

contains heterogeneous information in each  $Y_n$ . Furthermore,  $\bar{B}$  and  $\check{B}_n$  are the coefficients of  $\bar{A}$  and  $\check{A}_n$ . This allowed us to factorize voxel matrix in  $\mathcal{M}$  in a linked way so that

$$Y_n \approx A_n B_n^T = [\bar{A} \check{A}_n] \begin{bmatrix} \bar{B}_n^T \\ \check{B}_n^T \end{bmatrix} = \bar{A} \bar{B}_n^T + \check{A}_n \check{B}_n^T \quad (2)$$

The common feature matrix  $\bar{A}$  can be obtained using alternating least squares. Once  $\bar{A}$  is estimated,  $\bar{B}_n$  can be calculated by

$$\bar{B}_n = (Y_n^T - \check{B}_n \check{A}_n^T) \bar{A} (\bar{A}^T \bar{A})^{-1} = Y_n^T \bar{A}, n \in \mathcal{M} \quad (3)$$

Since  $Y_n$  contains common and heterogeneity information of T1-weighted images, it is necessary to map  $\bar{B}_n$  to have a common distribution, as in Eq. 4

$$\bar{B}_n^{\text{map}} = \text{std}(\bar{B}_{\text{ref}}) \frac{\bar{B}_n - \text{mean}(\bar{B}_n)}{\text{std}(\bar{B}_n)} + \text{mean}(\bar{B}_{\text{ref}}) \quad (4)$$

where  $\bar{B}_{\text{ref}}$  is the  $\bar{B}$  of reference center, and  $\text{mean}(\cdot)$  and  $\text{std}(\cdot)$  denote the mean and standard deviation. Once  $\bar{B}_n^{\text{map}}$  is estimated, we can find the common space  $\bar{A} \bar{B}_n^{\text{map}T}$  and heterogeneity space  $\check{A}_n \check{B}_n^T = Y_n - \bar{A} \bar{B}_n^{\text{map}T}$ .

Furthermore,  $\mathcal{F}_{\text{cobe}}(\cdot)$  denotes the common feature extracted by COBE, while  $\mathcal{F}_{\text{space}}^{\text{sitei}}(\cdot)$  denotes the heterogeneity space calculated by the common feature.

**TABLE 1. Description of the Scanning Parameters of all Participant**

Group	Center	Manufacturer	Platform	Sequence	Matrix	Field (T)	TR (msec)	TE (msec)	Angle (°)	TI (msec)	Database
HC	RD	Siemens	Trio Tim	MPRAGE	256 × 256 × 176	3	2400	3.16	8	1000	OASIS
	HD-S	Siemens	Trio Tim	MPRAGE	256 × 256 × 176	3	2400	3.16	8	1000	OASIS
	HD-P	Philips	Achieva	TFE	256 × 256 × 170	3	6.8	3.1	9	–	ADNI
	HD-G	GE	Signa Excite	IR-FSPGR	256 × 256 × 166	1.5	8.6	3.8	8	1000	ADNI
AD	RD-S	Siemens	Trio Tim	MPRAGE	256 × 256 × 176	3	2400	3.16	8	1000	OASIS
	HD-S	Siemens	Trio Tim	MPRAGE	256 × 256 × 176	3	2400	3.16	8	1000	OASIS
	HD-P	Philips	Achieva	TFE	256 × 256 × 170	3	6.8	3.1	9	–	ADNI
	HD-G	GE	Signa Excite	IR-FSPGR	256 × 256 × 166	1.5	8.6	3.8	8	1000	ADNI

RD = reference data; HD-S = harmonization data-Siemens; HD-P = harmonization data-Philips; HD-G = harmonization data-GE; HC = healthy control; AD = Alzheimer's disease; ADNI = Alzheimer's disease neuroimaging initiative; OASIS = open access series of imaging studies; MPRAGE = magnetization prepared rapid gradient echo; TFE = turbo field echo; IR-FSPGR = inversion recovery prepared fast spoiled gradient echo; TR = repetition time; TE = echo time; TI = Inversion time.

TABLE 2. Demographics of All Participant

Group	Center	<i>N</i>	Age range	Age	Gender (M/F)	Dataset	Function
HC (CDR = 0, MMSE ≥27)	RD	170	[50,95]	72.8 ± 6.7	98/72	OASIS	Reference data
	–	170	[50,95]	73.8 ± 7.3	98/72	OASIS + ADNI	Total
	MD-S	42	[50,95]	67.5 ± 7.8	34/8	OASIS	Multicenter data
	MD-P	47	[58,90]	74.4 ± 7.0	22/25	ADNI	
	MD-G	81	[70,90]	76.7 ± 4.7	42/39	ADNI	
AD (CDR ≥ 0.5, MMSE <27)	RD	170	[50,97]	76.6 ± 8.2	98/72	OASIS	Reference data
	–	170	[55,97]	76.2 ± 8.5	98/72	OASIS + ADNI	Total
	MD-S	69	[62,97]	78.2 ± 8.3	49/20	OASIS	Multicenter data
	MD-P	36	[56,91]	73.1 ± 8.8	13/23	ADNI	
	MD-G	65	[55,89]	75.7 ± 8.1	36/29	ADNI	

–Unless otherwise indicated, data are from a single-center.

CDR = clinical dementia rating; MMSE = mini-mental state examination; RD = reference data; MD-S = multicenter data-Siemens; MD-P = multicenter data-Philips; MD-G = multicenter data-GE; HC = healthy control; AD = Alzheimer's disease; ADNI = Alzheimer's disease neuroimaging initiative; OASIS = open access series of imaging studies.

**HCOBE METHOD.** We proposed a harmonization method using COBE (HCOBE) to harmonize scanner and acquisition protocol variability. Figure 2 illustrates the HCOBE framework for two centers ( $Y_1, Y_2$ ). For harmonizing multicenter data using HCOBE, the COBE method was initially employed to decompose intercenter data. The decomposed heterogeneity space  $S_{heter1}^{inter}$  and  $S_{heter2}^{inter}$  can be computed from the common biological feature as follows:

$$\begin{aligned} S_{heter1}^{inter} &= \mathcal{F}_{space}^{site1}(\mathcal{F}_{cobe}(Y_1, Y_2)) \\ S_{heter2}^{inter} &= \mathcal{F}_{space}^{site2}(\mathcal{F}_{cobe}(Y_1, Y_2)) \end{aligned} \quad (5)$$

Since center 1 and center 2 utilized different scanners and protocols, both biological heterogeneity space and center effects were present in  $S_{heter1}^{inter}$  and  $S_{heter2}^{inter}$ . To remove the center effects while preserving biological heterogeneity, we separated the center effect and biological heterogeneity further. We decomposed the intracenter data using COBE. Specifically, participants at center 1 and center 2 were randomly divided into two groups. Heterogeneity space of each center ( $S_{heter1}^{intra}, S_{heter2}^{intra}$ ) was then extracted using the following equation:

$$\begin{aligned} S_{heter1}^{intra} &= \mathcal{F}_{space}^{site1}(\mathcal{F}_{cobe}(Y_1)) \\ S_{heter2}^{intra} &= \mathcal{F}_{space}^{site2}(\mathcal{F}_{cobe}(Y_2)) \end{aligned} \quad (6)$$

Since intracenter data were acquired on the same scanner and with the same acquisition protocol, the decomposed heterogeneity space  $S_{heter1}^{intra}$  and  $S_{heter2}^{intra}$  only contained biological heterogeneity. Finally, we accurately obtained the center effects ( $Z$ ) by calculating the differences between intercenter and intracenter heterogeneity space:

$$Z_i = S_{heteri}^{inter} - S_{heteri}^{intra} \quad (7)$$

Once center effect  $Z_i$  has been estimated, we can accurately correct the data for center  $i$ :

$$Y_i^{HCOBE} = Y_i - Z_i \quad (8)$$

During the process of harmonizing the multicenter data using HCOBE, the data from the three centers were divided into two pairs (MD-S and MD-P, MD-S and MD-G) and successively forwarded to the HCOBE pipeline. In both harmonization processes, MD-S was used as the reference center.

A MATLAB (software version matlab 2020a, the MathWorks, Inc., Beijing, China) toolbox for this method is freely available at <https://github.com/zhaoslucas/MBIH>.

**EXISTING METHODS.** In this work, we compared the proposed HCOBE method with four existing methods: 1) RAVEL,<sup>13</sup> 2) Z-score normalization ( $Z_{score}$ ),<sup>15,23</sup> 3) GLM,<sup>14</sup> and 4) ComBat.<sup>12,24</sup> We briefly introduce these competing methods as follows.

RAVEL is a method that harmonizes GM brain regions affected by both scanner and disease using CSF regions that are primarily affected by scanner differences.<sup>13</sup>

The Z-score normalization method is utilized to standardize data across different centers, such that the histograms of the data from each center follow the same distribution.<sup>15,23</sup>

The GLM uses the center as a covariate to determine the relationship between biological variables, such as age and sex and brain image intensity. Center differences were removed by regressing the effects of center.<sup>14</sup>

ComBat was initially used to harmonize batch effects in genomics. The model assumes that centers have additive and

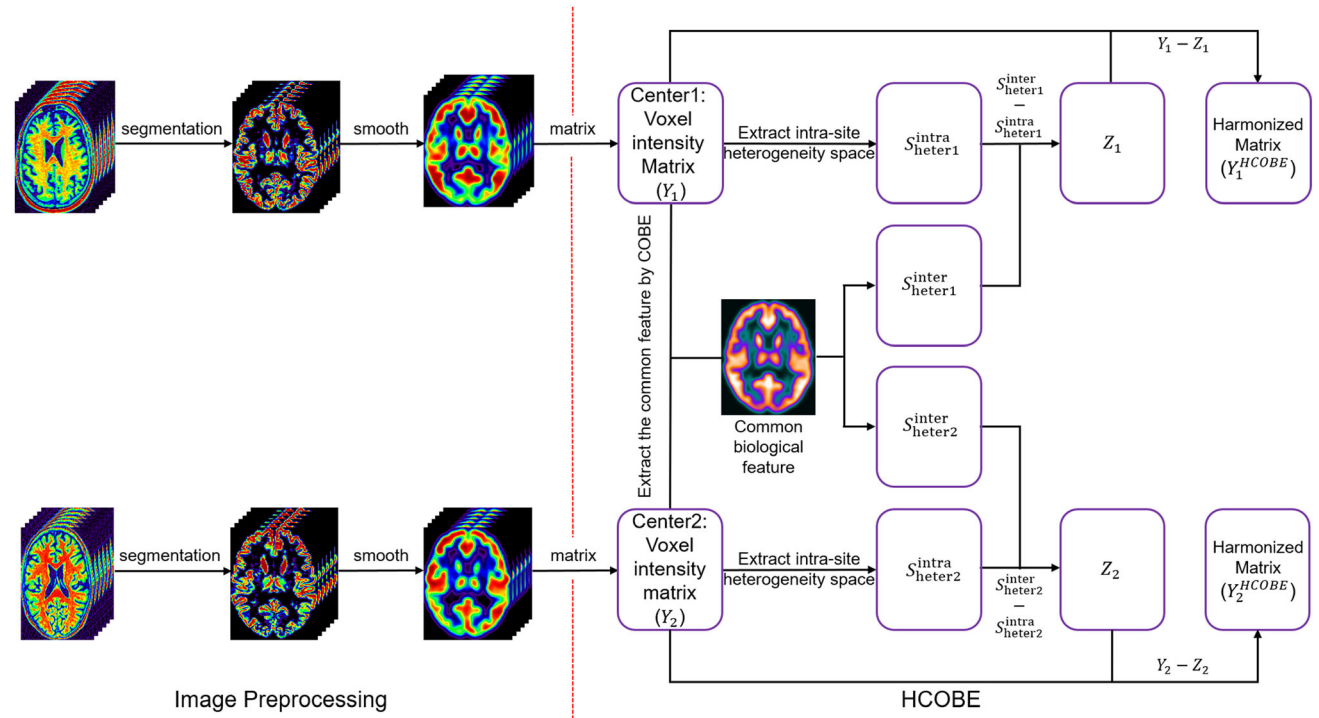


FIGURE 2: The pipeline of preprocessing and HCOBE. The left of the red dotted line is the preprocessing pipeline. The right of the red dotted line is the straightforward generalization of HCOBE. HCOBE = harmonization method using common orthogonal basis extraction.

multiplicative effects on the data. The parameters associated with the center were estimated using Empirical Bayes. The final ComBat-harmonized voxel intensity is calculated by subtracting the center effects from the original image.<sup>12,24</sup>

**Evaluation of Harmonization Methods**

**PERFORMANCE EVALUATION TO REMOVE THE CENTER EFFECTS.** The disease can influence the heterogeneity in brain structure and morphology of AD patients.<sup>25</sup> In contrast, multicenter CN participants, who do not have brain disease, show heterogeneity in brain structure and morphology primarily influenced by center effects. Therefore, we mainly assessed the performance of harmonization methods in the HC group. Here, we used linear discriminant analysis (LDA) to reduce the dimensionality of the multicenter data before and after harmonization. The primary objective of LDA was to identify a projection direction that maximized the distinction between participants with data from different centers, while minimizing the distance between sample points from the same center. The average voxel intensity of 90 GM brain regions defined by the Automated Anatomical Labeling (AAL) template in multicenter data were projected into two coordinates.<sup>26</sup> The center effect was represented by the difference in the coordinate distribution of the three centers. To further quantitatively analyze the performance of different harmonization methods in removing center effects, we introduced the objective function  $J$  of LDA. The objective function  $J$  is defined as

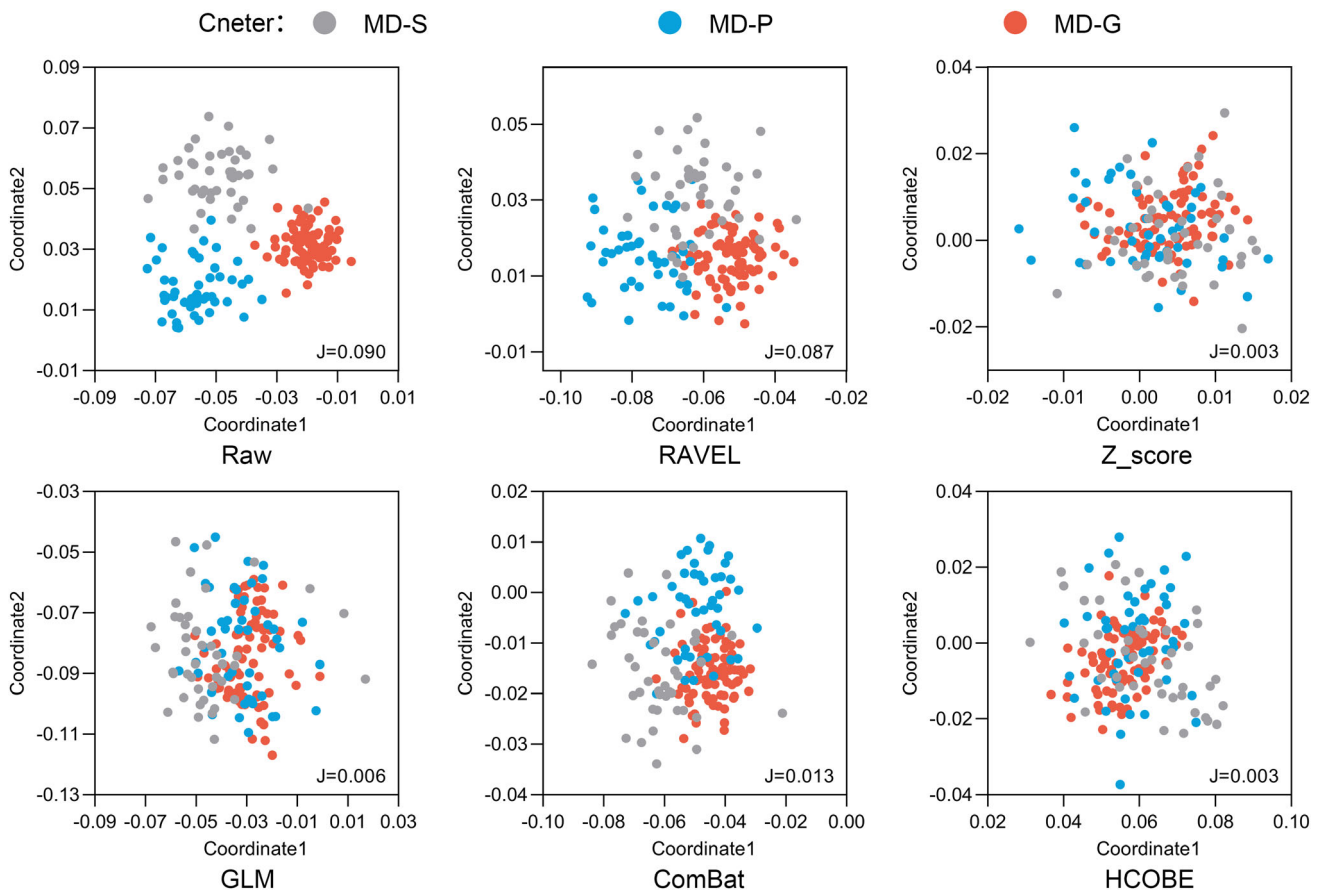
$$J = \frac{w^T S_b w}{w^T S_w w} \tag{9}$$

where  $w$  represents the projection direction vector solved by LDA,  $S_b$  is the between-class scatter matrix measuring the distance between individual center data means, and  $S_w$  is the within-class scatter matrix measuring the distance between each participant’s data and its central mean. When the center effect is removed, the mean values of each central data are identical, resulting in the objective function  $J$  being 0. Hence, the closer the value of the objective function  $J$  is to 0, the more effectively the center effect has been removed.

**PERFORMANCE EVALUATION TO PRESERVE BIOLOGICAL HETEROGENEITY.** To achieve successful harmonization, it is necessary to preserve biological heterogeneity while removing center effects. As a measure for the latter, we used the observation that GM volume was lower with age.<sup>27</sup> The GM volume was obtained by multiplying the CAT12-calculated modulation factor with the sum of voxels. To standardize the GM volume, we used the GM proportion that measures GM in relation to the whole brain. We fitted linear regression lines for GM proportion and age for reference and multicenter data separately. We refer to the regression coefficient as the “atrophy rate.” The disparity between the harmonization line and the reference line demonstrated changes in GM volume after harmonization. The closer the harmonization line to the reference line, the better the preservation of biological heterogeneity in the harmonized multicenter data.

**Consistency Evaluation of the Results of Harmonized Multicenter Data**

**CONSISTENCY EVALUATION OF CLASSIFICATION RESULTS.** Differences in brain structure and morphology have



**FIGURE 3:** Bivariate scatter plots of multicenter data (MD-S, MD-P, and MD-G) for HC group before and after harmonization, colored by center. The smaller the optimal value of the LDA objective function  $J$ , the cleaner the removal of the center effect. MD-S = multicenter data-Siemens; MD-P = multicenter data-Philips; MD-G = multicenter data-GE; RAVEL = removal of artificial voxel effect by linear regression; GLM = general linear model; HCOBE = harmonization method using common orthogonal basis extraction.

been observed between CN and AD participants, and these findings may impact the accuracy of clinical diagnosis for AD disease.<sup>28</sup> To determine the impact of center effects and biological heterogeneity on the consistency of multicenter study results, we utilized a support vector machine (SVM) model with a radial basis function kernel to compare HCOBE and four existing methods before and after harmonization for CN and AD classification. We tested three variants, which were: only intracenter data (HC: HD-G; AD: HD-G), only intercenter data (HC: HD-P; AD: HD-G), and both (HC: multicenter data; AD: multicenter data). The model input included age, sex, and the mean GM volume of 90 brain regions in the AAL template. Our model training used a 70/30 data split, with 70% used for the training set and 30% for testing. The model was assessed in the training set using 3-fold cross-validation. We used the classification outcomes of CN and AD in the single-center reference data as a standard for assessment. The similarity between the harmonized classification outcomes and the reference standard was evaluated to determine the consistency of the multicenter study.

**CONSISTENCY EVALUATION OF BRAIN ATROPHY RESULTS.** A major pathological feature of AD is the atrophy of brain tissues.<sup>27</sup> Since the multicenter data matched the single-center reference data for age and sex, the brain atrophy of multicenter data

should be consistent with that of the reference data. The Dice coefficient was used to evaluate the consistency of brain atrophy from the multicenter data and the reference data. Dice coefficient is given by van Rijsbergen<sup>29</sup>:

$$\text{Dice}_i = \frac{\text{Vol}(V_{\text{reference}} \cap V_i)}{(\text{Vol}(V_{\text{reference}}) + \text{Vol}(V_i))/2} \quad (10)$$

where  $\text{Vol}(\ast)$  indicates the count of voxel values that satisfies the condition of being equal to 1 in the mask;  $V_{\text{reference}}$  is a mask of brain atrophy in the reference center, and  $V_i$  is the mask of brain atrophy in the multicenter data after harmonization using method  $i$ . The ideal measure of Dice coefficient is equivalent to 1, which indicates a complete overlap between the two outcomes.

### Statistical Analysis

MATLAB 2020a (the MathWorks, Inc., Beijing, China) was used for statistical analysis. To evaluate the impact of preserving biological heterogeneity following the harmonization of data in the HC group using different methods, we performed a two-sample  $t$ -test on the reference data and the multicenter data before and after harmonization, specifically examining the GM proportion.

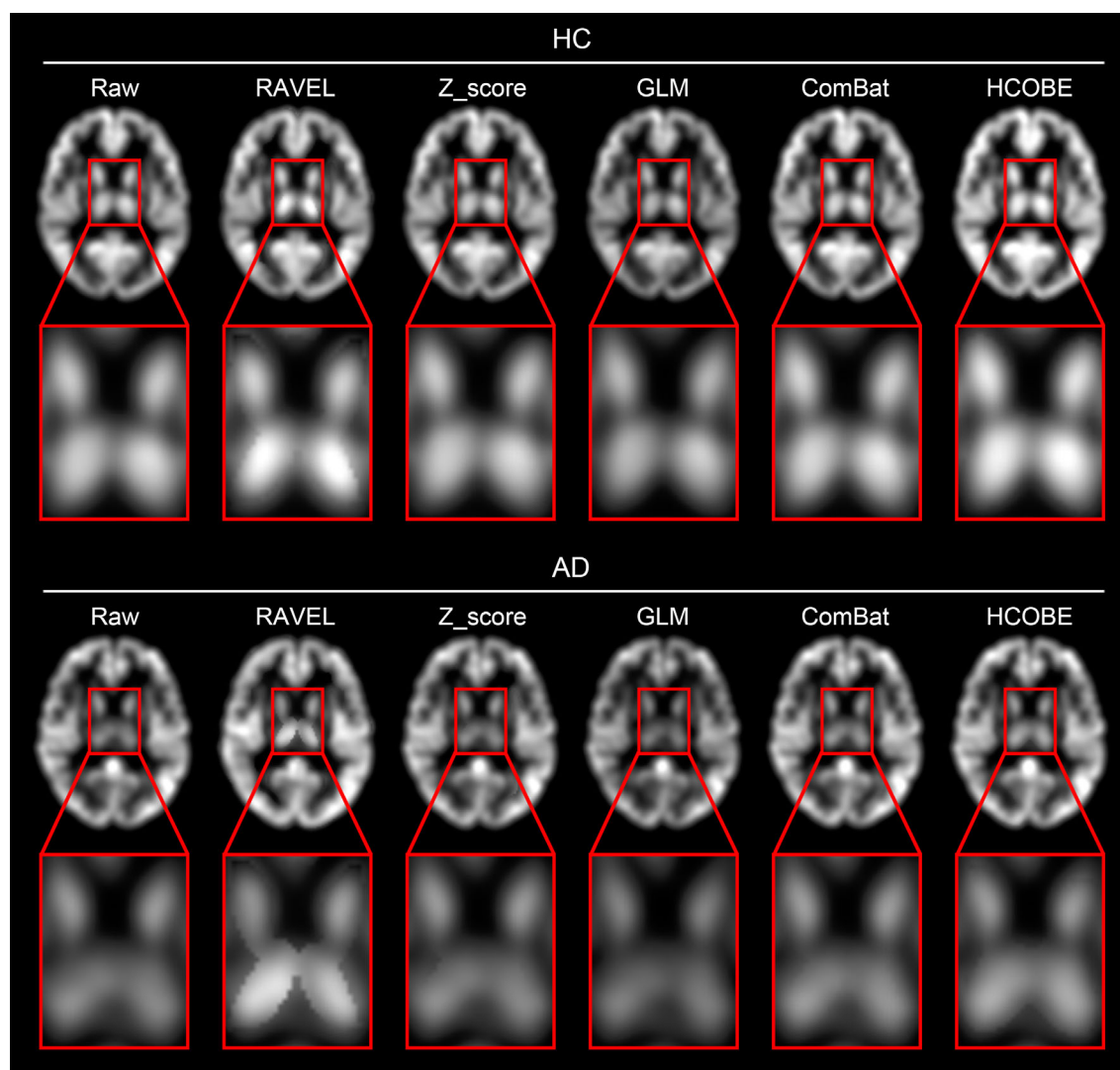


FIGURE 4: T1-weighted MRI before and after harmonization. HC = healthy control; AD = Alzheimer's disease; RAVEL = removal of artificial voxel effect by linear regression; GLM = general linear model; HCOBE = harmonization method using common orthogonal basis extraction.

The classification results were evaluated for accuracy, precision, recall, F1 score, receiver operating characteristics (ROC), and area under the curve (AUC) based on the test set.

Brain atrophy can be assessed by performing a two-sample  $t$ -test on the T1-weighted images of AD and CN. This procedure included age, sex, and total intracranial volume (TIV) as covariates. To control the false discovery rate, we implemented the Benjamini-Hochberg procedure, with a significance threshold set at  $P < 0.01$ .

## Results

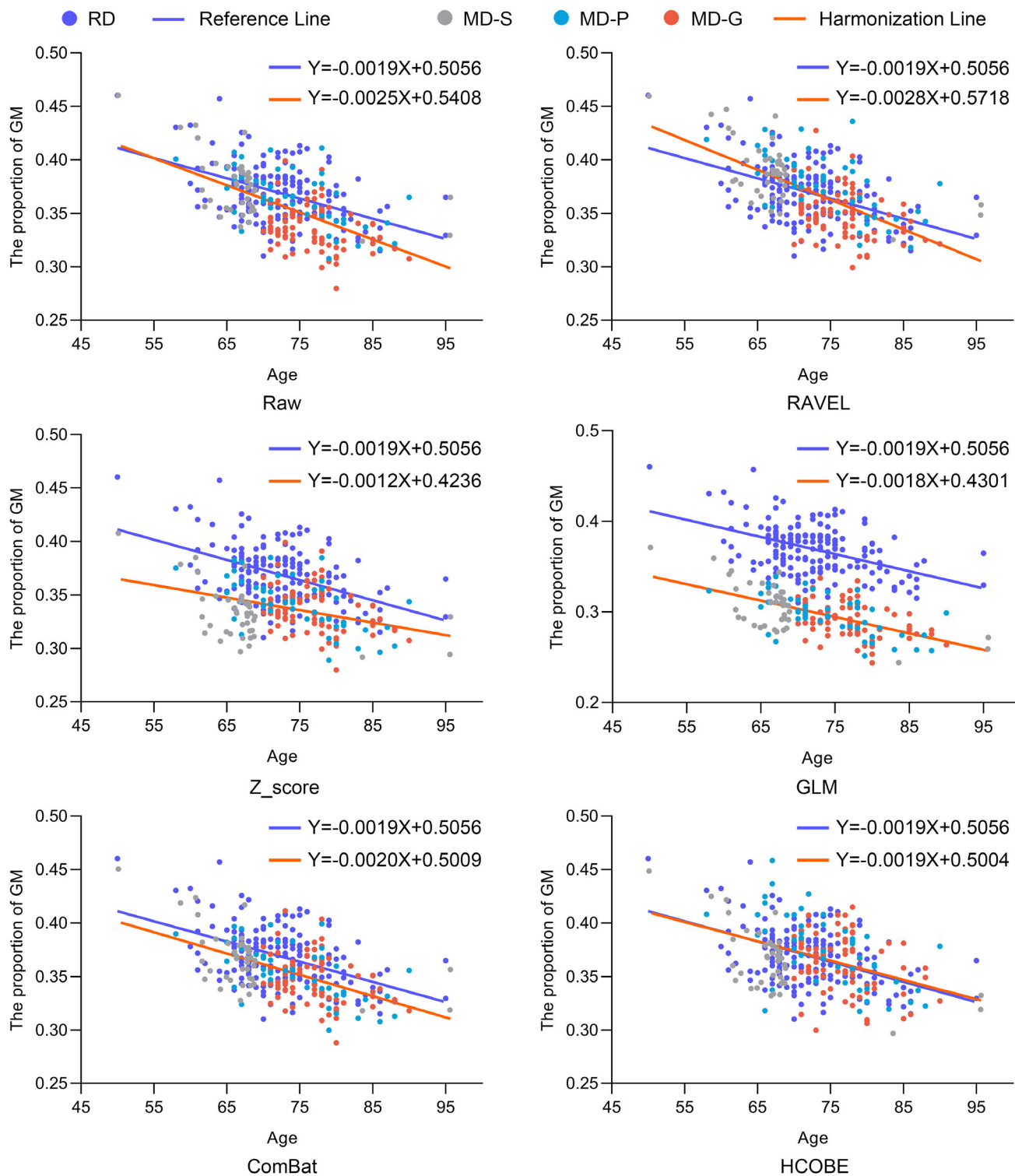
### Performance Evaluation to Remove the Center Effects

Figure 3 displays bivariate scatter plots before and after harmonization for HC multicenter data, with the differences in the coordinate distribution of the three centers represented as the center effect in each panel. The T1-weighted MRI data before and after harmonization were shown in Fig. 4. These

results indicated that all the harmonization methods were effective in reducing the center effects to varying degrees. Among these methods, HCOBE, GLM, and Z\_score minimized the center effects, while RAVEL did not appear to be as effective. Our results further indicated that HCOBE ( $J = 0.003$ ), GLM ( $J = 0.006$ ), and Z\_score ( $J = 0.003$ ) provided robust performance in removing the center effects.

### Performance Evaluation to Preserve Biological Heterogeneity

In Fig. 5, we noted that RAVEL showed an atrophy rate ( $r$ ) of  $r = -0.0028$ , indicating accelerated atrophy, while Z\_score showed an atrophy rate of  $r = -0.0012$ , indicating reduced atrophy. However, HCOBE ( $-0.0019$ ), ComBat ( $-0.0020$ ), and GLM ( $-0.0018$ ) demonstrated atrophy rates similar to that of the reference atrophy rate ( $-0.0019$ ). The distance between the harmonization line and the reference



**FIGURE 5:** The relationship between the proportion of GM and age in HC. For each harmonization method, we plot the relationship between the proportion of GM and age after harmonization. The reference data (RD) and its fitted line (reference line) were regarded as the reference standard. *P* values represented the results of two-sample *t*-test for the proportion of GM in the reference data and multicenter data. There was no significant difference in age between the multicenter and reference data ( $P = 0.062$ ). RD = reference data; MD-S = multicenter data-Siemens; MD-P = multicenter data-Philips; MD-G = multicenter data-GE; RAVEL = removal of artificial voxel effect by linear regression; GLM = general linear model; HCOBE = harmonization method using common orthogonal basis extraction.

line indicated the changes in GM volume after harmonization. We noticed that the GLM-harmonized multicenter data resulted in the largest change in GM volume, while the

ComBat method preserved the variability of age better. The HCOBE method performed remarkably well in preserving age variability individually at each center.

**TABLE 3. Classification Performance of SVM Models on Multicenter Data and Reference Data**

Methods	Range	Accuracy	Precision	Recall	F1 score	AUC
Raw	Intra	0.837	0.875	0.737	0.800	0.910
	Inter	0.909	0.944	0.895	0.919	0.977
	Intra + Inter	0.775	0.833	0.686	0.753	0.830
RAVEL	Intra	0.721	0.667	0.737	0.700	0.884
	Inter	0.788	0.875	0.737	0.800	0.910
	Intra + Inter	0.794	0.813	0.765	0.788	0.861
Z_score	Intra	0.767	0.846	0.579	0.688	0.868
	Inter	0.788	0.800	0.842	0.821	0.876
	Intra + Inter	0.794	0.800	0.784	0.792	0.856
GLM	Intra	0.791	0.750	0.790	0.769	0.855
	Inter	0.818	0.810	0.895	0.850	0.865
	Intra + Inter	0.716	0.824	0.549	0.659	0.837
ComBat	Intra	0.837	0.833	0.790	0.811	0.893
	Inter	0.818	0.810	0.895	0.850	0.910
	Intra + Inter	0.794	0.768	0.843	0.804	0.894
HCOBE	Intra	0.837	0.773	0.895	0.829	0.947
	Inter	0.849	0.850	0.895	0.872	0.944
	Intra + Inter	0.892	0.833	0.980	0.901	0.954
Reference	Intra	0.873	0.896	0.843	0.869	0.942

RAVEL = removal of artificial voxel effect by linear regression; GLM = general linear model; HCOBE = harmonization method using common orthogonal basis extraction; AUC = area under the curve.

In addition, the result of the two-sample  $t$ -test on the HC reference data and the multicenter data showed that only the HCOBE-harmonized multicenter data demonstrated similarity to the reference data with a  $P$ -value of 0.52. However, the multicenter data harmonized by the other four methods differed significantly from the reference data.

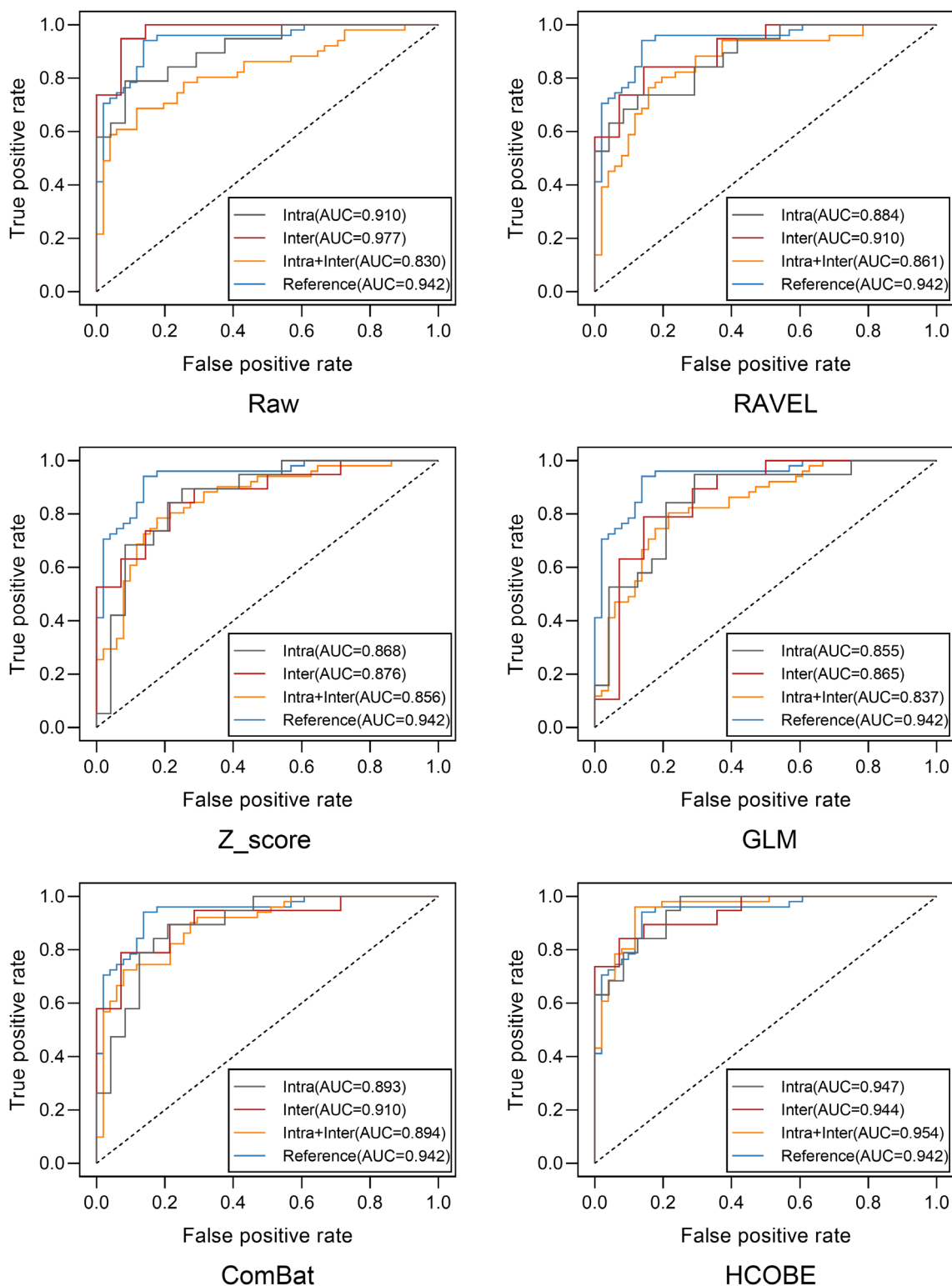
### Consistency Evaluation of Classification Results

Table 3 shows the classification results of the SVM model for AD and CN based on 3-fold cross-validation. The results showed that the maximum values for accuracy, precision, recall, and AUC were obtained for the classification in raw intercenter data. Furthermore, the classification outcomes for CN and AD in the raw intracenter data were inferior to the results obtained intercenter. In three variations, the classification results of the HCOBE-harmonized multicenter data were comparable to those of the reference data. Figure 6 visualizes the ROC curves and AUC values of the SVM model for classifying AD and CN under 3-fold cross-

validation. The results showed that the SVM model had similar ROC curves and AUC values on the HCOBE-harmonized multicenter data and the single-center reference data. However, the classification results of the multicenter data harmonized by the other harmonization methods were not as good as the intracenter and intercenter classification results.

### Consistency Evaluation of Brain Atrophy Results

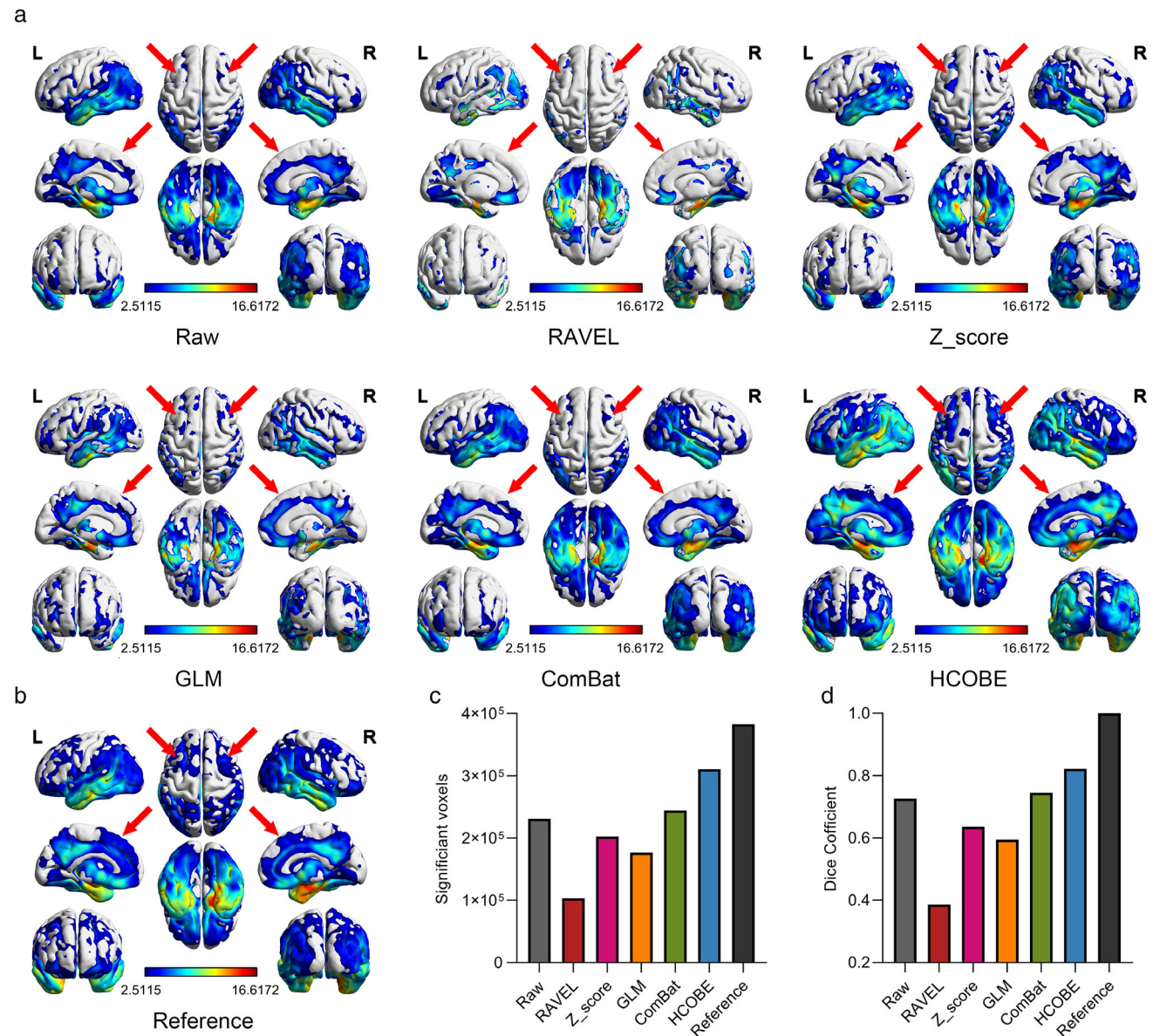
Figure 7a shows the results of the two-sample  $t$ -test before and after harmonization for the multicenter data. Figure 7b shows the result of the two-sample  $t$ -test in the single-center reference data, indicating the regions of brain atrophy in AD patients. Both the reference data and the multicenter data had the same statistical power. Hence, the two-sample  $t$ -test results of the harmonized multicenter data should be consistent with the two-sample  $t$ -test results of the reference data. It was noted that the harmonized results of existing methods had less brain atrophy in the



**FIGURE 6:** ROC curves and their associated AUC values for the SVM model. T-fold cross validation was used to test the classification of CN and AD before and after harmonization for the SVM model. The intra-center classification was between CN and AD in HD-G, the intercenter classification was between CN in HD-P and AD in HD-G, and the classification of both was between CN and AD for multicenter data. AUC = Area under the curve; RAVEL = removal of artificial voxel effect by linear regression; GLM = general linear model; HCOBE = harmonization method using common orthogonal basis extraction.

frontal lobe. However, the results of HCOBE-harmonized multicenter data were more consistent with the reference data for brain atrophy.

In Fig. 7c, we present the number of voxels with significant differences for the reference data and the multicenter data. The figure shows that the HCOBE harmonization



**FIGURE 7:** Two-sample *t*-tests of AD and CN on the reference data and the multicenter data. (a) The two-sample *t*-tests results of 170 AD and 170 CN in the multicenter data (MD-S, MD-P, and MD-G) before and after correction. (b) The two-sample *t*-tests results of 170 AD and 170 CN in the reference data (RD). The arrows pointed to the frontal lobe. (c) The number of significant voxels in the reference data and multicenter data. (d) The Dice coefficient of the harmonized data. RAVEL = removal of artificial voxel effect by linear regression; GLM = general linear model; HCOBE = harmonization method using common orthogonal basis extraction.

approach exhibited superior performance as compared to the other methods used in this study. The number of voxels generated by HCOBE (310,609) was found to be closest to the reference dataset (382,936 voxels). In Fig. 7d, we show the quantitative results of the Dice coefficients. The HCOBE method had the highest Dice coefficient value (Dice = 0.82). These results demonstrated that HCOBE significantly improved the consistency of results in multicenter studies.

## Discussion

In this study, we proposed the HCOBE methodology to harmonize multicenter data. Our evaluation of available

harmonization methods revealed that HCOBE outperforms other methods by removing center effects while preserving biological heterogeneity. In addition, we classified AD and CN in three variations and analyzed the brain atrophy regions in patients with AD. By comparing the similarity of the resulting harmonized multicenter data with the single-center reference data, we found that HCOBE may significantly improve the consistency of results in multicenter studies.

The RAVEL method has been found to perform poorly in removing center effects and preserving biological heterogeneity, which was not unexpected since the method relies on the control region.<sup>13</sup> Specifically, using CSF intensities as the

surrogate for center instead of GM intensities resulted in poor RAVEL performance. The Z\_score and GLM methods were similarly unsuccessful at preserving biological heterogeneity after the center effects was removed, which may explain why they mistakenly removed biological heterogeneity as the center effects. In addition, HCOBE was compared with ComBat. Although ComBat effectively preserved biological heterogeneity, it was less successful at removing the center effects due to its assumption that all voxels share the same common distribution, thereby rendering it incapable of accurately determining model parameters for voxels with different distributions.<sup>16</sup>

The results of the consistency assessment showed a superior classification performance of the raw data between centers. This can be attributed to the inclusion of center effects in addition to biological heterogeneity in the differences observed for CN and AD between centers. Furthermore, the classification results of the multicenter raw data showed inferior performance compared to the intracenter data. This observation suggests that center effects have an impact on the biological heterogeneity of CN and AD. ComBat exhibited the highest AUC value and Dice coefficient compared to the other available harmonization methods. In contrast to RAVEL, GLM, and Z\_score, ComBat significantly improved the consistency of results in multicenter studies. Similar findings were observed in the HCOBE-harmonized multicenter data and single-center reference data when using the identical analytical method. The HCOBE-harmonized multicenter data displayed higher consistency when contrasted with the raw multicenter data and the data that underwent harmonization using existing methods. This reflected that HCOBE accurately separated the center effects from biological heterogeneity and preserved the biological heterogeneity while removing the center effects.

Multicenter studies have become more common, highlighting the issue of scanner variability on experimental results.<sup>30–32</sup> In the particular context of T1-weighted images, we have demonstrated the strong performance of HCOBE. However, the HCOBE methodology is versatile and can be applied beyond the T1-weighted images. Studies have shown that software updates can affect the consistency of longitudinal data.<sup>33,34</sup> Furthermore, HCOBE can also be used to harmonize MRI scans over multiple time points, including functional MRI and longitudinal structural MRI of the same patient. In addition, the HCOBE method can be applied to measurements at the region of interest (ROI) level and radiomics.<sup>8</sup> The performance of the HCOBE method is not affected by dataset size or image category, and it holds great promise as a harmonization method for multicenter studies.

### Limitations

One limitation of this study is that participant heterogeneity was present despite the lack of significant differences in age,

sex, and handedness between the multicenter and single-center groups. As such, achieving agreement between the results from multicenter and single-center was challenging. In addition, this study focused on CN participants and AD patients, and further validation of the effect of HCOBE on other populations or diseases is needed. Additionally, the MRI data were acquired using older model scanners such as Tim Trio, Achieva, and Signa Excite, and the performance of HCOBE on the new model needs further investigation.

### Conclusion

The HCOBE method may facilitate the aggregation of brain images acquired from multiple centers. The HCOBE method could improve the consistency of results with multicenter studies after removing the center effects.

### Acknowledgments

Data were provided in part by OASIS-3 Principal Investigators: T. Benzinger, D. Marcus, J. Morris. Data were provided in part by the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](https://adni.loni.usc.edu)). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [https://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf)

### REFERENCES

1. Kaufmann T, van der Meer D, Doan NT, et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat Neurosci* 2019;22:1617-1623.
2. Machado-Rivas F, Gandhi J, Choi JJ, et al. Normal growth, sexual dimorphism, and lateral asymmetries at fetal brain MRI. *Radiology* 2022;303:162-170.
3. Parekh P, Bhalerao GV, John JP, Venkatasubramanian G, consortium A. Sample size requirement for achieving multisite harmonization using structural brain MRI features. *Neuroimage* 2022;264:119768.
4. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology* 2019;291:52-58.
5. Takao H, Hayashi N, Ohtomo K. Effect of scanner in longitudinal studies of brain volume changes. *J Magn Reson Imaging* 2011;34:438-444.
6. Takao H, Hayashi N, Ohtomo K. Effects of study design in multi-scanner voxel-based morphometry studies. *Neuroimage* 2014;84:133-140.
7. Tax CMW, Grussu F, Kaden E, et al. Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. *Neuroimage* 2019;195:285-299.
8. Orlhac F, Lecler A, Savatovski J, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol* 2021;31:2272-2280.
9. Pagani E, Storelli L, Pantano P, et al. Multicenter data harmonization for regional brain atrophy and application in multiple sclerosis. *J Neurol* 2023;270:446-459.

10. Tong QQ, Gong T, He HJ, et al. A deep learning-based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols. *Magn Reson Imaging* 2020;73:31-44.
11. Wengler K, Cassidy C, van der Pluijm M, et al. Cross-scanner harmonization of neuromelanin-sensitive MRI for multisite studies. *J Magn Reson Imaging* 2021;54:1189-1199.
12. Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 2018;167:104-120.
13. Fortin JP, Sweeney EM, Muschelli J, Crainiceanu CM, Shinohara RT, Alzheimer DN. Removing inter-subject technical variability in magnetic resonance imaging studies. *Neuroimage* 2016;132:198-212.
14. Kostro D, Abdulkadir A, Durr A, et al. Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *Neuroimage* 2014;98:405-415.
15. Wachinger C, Rieckmann A, Polsterl S, Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flag-ship study of ageing. Detect and correct bias in multi-site neuroimaging datasets. *Med Image Anal* 2021;67:101879.
16. Zhong J, Wang Y, Li J, et al. Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: Application to neonatal white matter development. *Biomed Eng Online* 2020;19:4.
17. Maikusa N, Zhu YH, Uematsu A, et al. Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. *Hum Brain Mapp* 2021;42:5278-5287.
18. Zuo XN, Xu T, Milham MP. Harnessing reliability for neuroscience research. *Nat Hum Behav* 2019;3:768-771.
19. Jack CR, Bernstein MA, Fox NC, et al. The Alzheimer's disease Neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging* 2008;27:685-691.
20. Lamontagne PJ, Benzinger TL, Morris JC, Keefe S, Marcus D. OASIS-3: Longitudinal Neuroimaging, clinical, and cognitive dataset for Normal aging and Alzheimer disease. *MedRxiv* 2019.
21. Zhou GX, Cichocki A, Zhang Y, Mandic DP. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems* 2016;27:2426-2439.
22. Zhou GX, Zhao QB, Zhang Y, Adali T, Xie SL, Cichocki A. Linked component analysis from matrices to high-order tensors: Applications to biomedical data. *Proc IEEE* 2016;104:310-331.
23. Yan CG, Craddock RC, Zuo XN, Zang YF, Milham MP. Standardizing the intrinsic brain: Towards robust measurement of inter-individual variation in 1000 functional connectomes. *Neuroimage* 2013;80:246-262.
24. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;8:118-127.
25. Raji CA, Lopez OL, Kuller LH, Carmichael OT, Becker JT. Age, Alzheimer disease, and brain structure. *Neurology* 2009;73:1899-1905.
26. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 2002;15:273-289.
27. Pini L, Pievani M, Bocchetta M, et al. Brain atrophy in Alzheimer's disease and aging. *Ageing Res Rev* 2016;30:25-48.
28. Niemeyer F, Galbusera F, Tao YP, Kienle A, Beer M, Wilke HJ. A deep learning model for the accurate and reliable classification of disc degeneration based on MRI data. *Invest Radiol* 2021;56:78-85.
29. Fahmy AS, Neisius U, Chan RH, et al. Three-dimensional deep convolutional neural networks for automated myocardial scar quantification in hypertrophic cardiomyopathy: A multicenter multivendor study. *Radiology* 2020;294:52-60.
30. Bashyam VM, Doshi J, Erus G, et al. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *J Magn Reson Imaging* 2022;55:908-916.
31. Schwartz DL, Tagge I, Powers K, et al. Multisite reliability and repeatability of an advanced brain MRI protocol. *J Magn Reson Imaging* 2019;50:878-888.
32. Tong QQ, He HJ, Gong T, et al. Reproducibility of multi-shell diffusion tractography on traveling subjects: A multicenter study prospective. *Magn Reson Imaging* 2019;59:1-9.
33. Jovicich J, Marizzoni M, Bosch B, et al. Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects. *Neuroimage* 2014;101:390-403.
34. Takao H, Hayashi N, Kabasawa H, Ohtomo K. Effect of scanner in longitudinal diffusion tensor imaging studies. *Hum Brain Mapp* 2012;33:466-477.