Goal-specific brain MRI harmonization

Lijun An[1,2,3], Jianzhong Chen[1,2,3], Pansheng Chen[1,2,3], Tong He[1,2,3], Christopher Chen[4], Juan Helen Zhou[1,2,5], B.T. Thomas Yeo[1,2,3,5,6]

for the Alzheimer's Disease Neuroimaging Initiative* and the Australian Imaging Biomarkers and Lifestyle Study of Aging*

[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore [2] Centre for Sleep and Cognition (CSC) & Centre for Translational Magnetic Resonance Research (TMR), National University of Singapore, Singapore [3] N.1 Institute for Health & Institute for Digital Medicine (WisDM), National University of Singapore, Singapore [4] Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore [5] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore [6] Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA

Address correspondence to:

B.T. Thomas Yeo

ECE, CSC, TMR, N.1 & WISDM

National University of Singapore

Email: thomas.yeo@nus.edu.sg

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL) database (www.aibl.csiro.au). As such, the investigators within the ADNI and AIBL contributed to the design and implementation of ADNI and AIBL 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: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

There is significant interest in pooling magnetic resonance image (MRI) data from multiple datasets to enable mega-analysis. Harmonization is typically performed to reduce heterogeneity when pooling MRI data across datasets. Most MRI harmonization algorithms do not explicitly consider downstream application performance during harmonization. However, the choice of downstream application might influence what might be considered as study-specific confounds. Therefore, ignoring downstream applications during harmonization might potentially limit downstream performance. Here we propose a goal-specific harmonization framework that utilizes downstream application performance to regularize the harmonization procedure. Our framework can be integrated with a wide variety of harmonization models based on deep neural networks, such as the recently proposed conditional variational autoencoder (cVAE) harmonization model. Three datasets from three different continents with a total of 2787 participants and 10085 anatomical T1 scans were used for evaluation. We found that cVAE removed more dataset differences than the widely used ComBat model, but at the expense of removing desirable biological information as measured by downstream prediction of mini mental state examination (MMSE) scores and clinical diagnoses. On the other hand, our goal-specific cVAE (gcVAE) was able to remove as much dataset differences as cVAE, while improving downstream cross-sectional prediction of MMSE scores and clinical diagnoses.

1 Introduction

Large scale MRI datasets from multiple sites have boosted the study of human brain structure and function (Yeo et al., 2011; Van Essen et al., 2013; Miller et al., 2016; Volkow et al., 2018). Combining datasets from multiple sites can potentially boost statistical power, so there is significant interest in pooling data across multiple sites (Thompson et al., 2017; Whelan et al., 2018; Tang et al., 2020; Lu et al., 2020). However, MRI data is sensitive to variation of scanners across different sites (Jovicich et al., 2006; Magnotta et al., 2012; Chen et al., 2014; Hawco et al., 2018), so post-acquisition harmonization is necessary for removing unwanted variabilities in pooling data across multiple studies.

A popular harmonization approach is the ComBat framework (Fortin et al., 2017, 2018; Yu et al., 2018) that utilizes a mixed effects regression model to remove additive and multiplicative site effects. Other ComBat variants have since been proposed (Garcia-Dias et al., 2020; Pomponio et al., 2020; Wachinger et al., 2021). However, most ComBat variants consider each brain region separately (but see Chen et al., 2019), so might not be able to remove nonlinear site differences that are distributed across brain regions.

These nonlinear distributed site differences might be more readily removed by harmonization approaches based on deep neural networks (DNNs; (Tanno et al., 2017; Blumberg et al., 2018; Ning et al., 2019). One popular approach is the use of the variational autoencoder (VAE) framework (Moyer et al., 2020; Russkikh et al., 2020; Zuo et al., 2021), which typically uses an encoder to generate site-invariant latent representations. Site information can then be added to the latent representations to "reconstruct" the MRI data. Another popular approach is the use of generative adversarial networks and cycle consistency constraints (Zhu et al., 2017; Zhao et al., 2019; Dewey et al., 2019; Modanwal et al., 2020; Bashyam et al., 2021).

However, most previously proposed harmonization approaches do not consider downstream applications in the harmonization procedure. It is important to note that the goal of MRI harmonization is to remove 'unwanted' dataset differences, while preserving relevant biological information. However, unwanted dataset differences depend on the application. For example, if our goal is to develop an Alzheimer's disease (AD) dementia prediction model that is generalizable across different racial groups, then 'race' might be considered an undesirable study difference. On the other hand, if we are interested in studying AD progression across different racial groups, then racial information needs to be preserved in the

harmonization process. Therefore, ignoring downstream applications in the harmonization procedure might potentially limit downstream performance.

In this study, we propose a goal-specific harmonization framework that utilizes downstream applications to regularize the harmonization model. Our approach can be integrated with most DNN-based harmonization approaches, such as the conditional VAE (cVAE) harmonization model (Moyer et al., 2020), which was previously applied to diffusion MRI data. We then compared the resulting goal-specific cVAE (gcVAE) model with cVAE and ComBat using three datasets comprising 2787 participants and 10085 anatomical MRI scans. The evaluation procedure tested the ability of different harmonization models to remove dataset differences while retaining biological information as measured by downstream cross-sectional prediction of mini mental state examination (MMSE) scores and clinical diagnoses.

2 Methods

2.1 Datasets

In this study, we considered T1 structural MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008, 2010), the Australian Imaging, Biomarkers and Lifestyle (AIBL) study (Ellis et al., 2009, 2010) and the Singapore Memory Ageing and Cognition Centre (MACC) Harmonization cohort (Hilal et al., 2015; Chong et al., 2017; Hilal et al., 2020). Across all three datasets, MRI data was collected at multiple timepoints.

In the case of ADNI (Jack et al., 2008, 2010), we considered data from ADNI1 and ADNI2/Go. For ADNI1, the MRI scans were collected from 1.5 and 3T scanners from different vendors. For ADNI2/Go, the MRI scans were collected from 3T scanners. There were 1735 participants with at least one T1 MRI scan. There was a total of 7955 MRI scans across the different timepoints of the 1735 participants.

In the case of AIBL (Ellis et al., 2009, 2010), the MRI scans were collected from 1.5T and 3T Siemens (Avanto, Tim Trio and Verio) scanners. There were 495 participants with at least one T1 MRI scan. There was a total of 933 MRI scans across the different timepoints of the 495 participants.

In the case of MACC (Hilal et al., 2015; Chong et al., 2017; Hilal et al., 2020), the MRI scans were collected from a Siemens 3T Tim Trio scanner. There were 557 participants with at least one T1 MRI scan. There was a total of 1197 MRI scans across the different timepoints of the 557 participants.

2.2 Data Preprocessing

Our goal is to harmonize volumes of regions of interest (ROIs) across datasets. Here, 108 cortical and subcortical ROIs were defined based on the FreeSurfer software (Fischl et al., 2002; Desikan et al., 2006). In the case of ADNI, we utilized the ROI volumes provided by ADNI. These ROIs were generated by ADNI after several preprocessing steps (http://adni.loni.usc.edu/methods/mri-tool/mri-pre-processing/) followed by the FreeSurfer version 4.3 (ADNI1) and 5.1 (ADNI2/GO) recon-all pipeline. In the case of AIBL and MACC, FreeSurfer version 6.0 recon-all pipeline was utilized. Therefore, differences between the datasets arose from both scanner and preprocessing differences.

2.3 Workflow overview

In this study, we sought to harmonize brain ROI volumes between ADNI and AIBL, as well as ADNI and MACC. Figure 1 illustrates the workflow in this study using AIBL as an illustration. The procedure is exactly the same for MACC. In the case of AIBL, we used the Hungarian matching algorithm (Kuhn, 1955) to first select pairs of ADNI and AIBI participants with matched number of timepoints, age, sex, MMSE and clinical diagnosis (Figure 1A). The distributions of age, sex, MMSE and clinical diagnosis of all participants and matched participants are shown in Figure 2.

There were 247 pairs of matched AIBI and ADNI participants with an average of 1.1 scans per participant. The same approach was applied to ADNI and MACC, yielding 277 pairs of matched MACC and ADNI participants with an average of 1.5 scans per participant. We note that not all timepoints have corresponding MMSE and clinical diagnosis information. Therefore, care was taken to ensure that all timepoints in the matched participants had both MMSE and clinical diagnosis. P values showing the quality of the matching procedure are found in Tables S1 to S7. The matched participants served as a test set to evaluate the harmonization procedures.

The unmatched ADNI data was used to train goal-specific deep neural networks (DNN) for predicting MMSE and clinical diagnosis (Figure 1B; details in Section 2.6). The unmatched ADNI and AIBL participants were also used to fit the various harmonization models (Figure 1B; details in Section 2.7 and Section 2.8). We note that the goal-specific DNN was also utilized for training the gcVAE model. The same procedure was applied to ADNI and MACC. The trained harmonization models were then applied to the unharmonized brain volumes (Figure 1C).

The harmonized data was then evaluated with two criteria (Figure 1D). The first criterion was dataset prediction performance when using a machine learning algorithm to predict which dataset the harmonized data came from. Lower dataset prediction performance indicates better harmonization. More specifically, we trained a XGBoost classifier (Chen & Guestrin, 2016) using the harmonized unmatched ADNI and AIBL brain volumes and then applied the classifier to the matched ADNI and AIBL brain volumes (details in Section 2.5). The same procedure was applied to ADNI and MACC.

However, a simple way to achieve perfect dataset prediction results was to map all brain volumes to zero, thus losing all biological information. Therefore, the second criterion was downstream application performance. Here, we applied the goal-specific DNN (Figure 1B) to the harmonized AIBL brain volumes from the matched participants. To demonstrate

the effects of no harmonization, the goal-specific DNN was also applied to the unharmonized ADNI brain volumes from the matched participants. The same procedure was applied to ADNI and MACC.

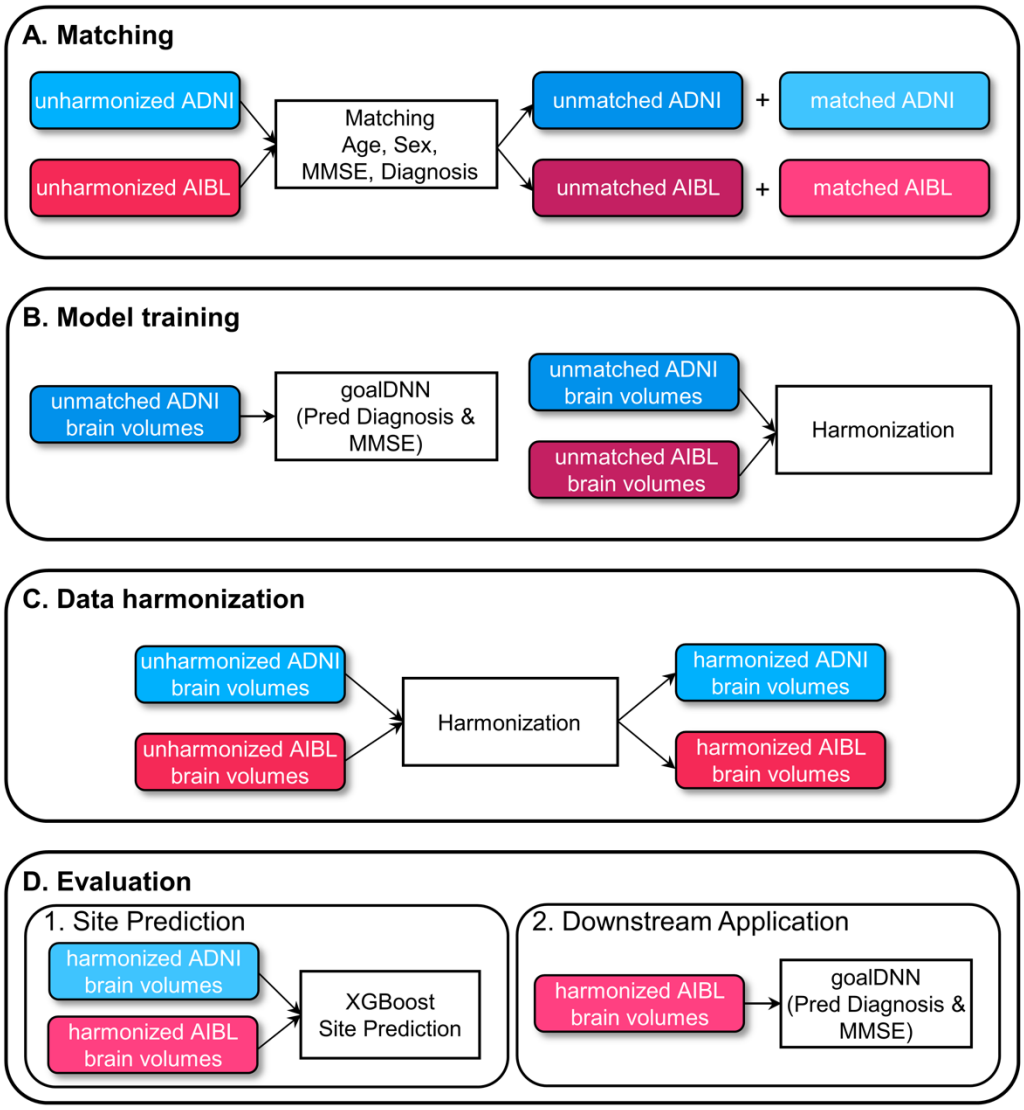


Figure 1. Workflow of current study. We illustrate the workflow using ADNI and AIBL. The same procedure was applied to ADNI and MACC. (A) Matching participants to derive test set for harmonization evaluation (B left) Train goal-specific deep neural network (DNN) using unmatched unharmonized ADNI data to predict clinical diagnosis and MMSE. (B right) Train harmonization models (ComBat, cVAE & gcVAE) using unmatched unharmonized ADNI and AIBL data (C) Harmonize ADNI and AIBL brain volumes using trained harmonized models from step B (D) Evaluate harmonization performance using XGBoost site prediction model and goal-specific DNN using harmonized ADNI and AIBL brain volumes from matched participants.

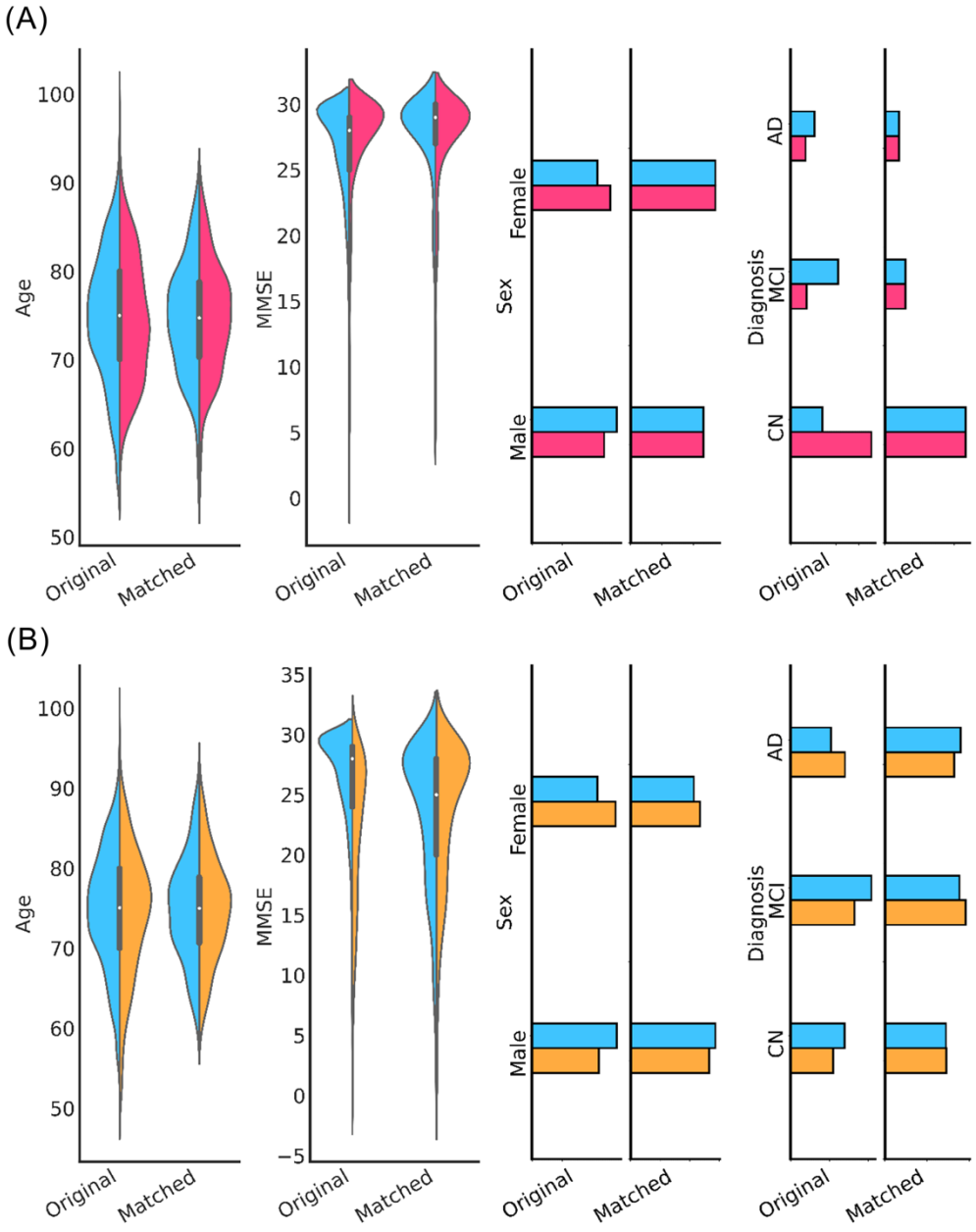


Figure 2. Age, MMSE, sex and clinical diagnosis distributions before and after matching. (A) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and AIBL (red). Differences in the attributes between ADNI and AIBL were not significant after matching. (B) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and MACC (yellow). Differences in the attributes between ADNI and MACC were not significant after matching. P values showing the quality of the matching procedure are found in Tables S1 to S7.

2.4 Training, validation and test procedure

As mentioned in the previous section, the *matched* participants were used as the test set for evaluation. The *unmatched* participants were used for training the goal-specific DNN, harmonization and dataset prediction models. More specifically, we divided the unmatched participants into 10 groups. Care was taken to ensure that the timepoints of any participant were not split across multiple groups. To train the goal-specific DNN, harmonization and dataset prediction models, 9 groups were used for training, while the remaining group was used as a validation set to tune the hyperparameters. This procedure was repeated 10 times with a different group being the validation set. Therefore, we ended up with 10 sets of trained models. The 10 sets of harmonization models were applied to the unharmonized data (Figure 1C), yielding 10 sets of harmonized data. The 10 sets of XGBoost classifiers and goal-specific DNNs were applied to the 10 corresponding sets of harmonized data (Figures 1D). The performance was evaluated across the 10 sets of models.

2.5 Dataset prediction model

To evaluate the harmonization approaches, we utilized XGBoost to predict which dataset the brain volumes came from. The inputs to the XGBoost model were the brain volumes divided by the total intracranial volume (ICV) of each participant. We used logistic regression as the objective function and ensemble of trees as the model structure. 10 XGBoost classifiers were trained using the *unmatched harmonized* MRI volumes (see Section 2.4). For each XGBoost classifier, we used a grid search using the validation group to find the optimal set of hyperparameters. The prediction accuracy was averaged across all time points of each participant and the 10 classifiers before averaging across participants.

2.6 Goal-specific DNNs

Here we utilized DNNs to predict MMSE and clinical diagnosis (normal controls, mild cognitive impairment or Alzheimer's disease dementia) jointly. The goal-specific DNNs were used to evaluate the harmonization approaches and also helped to finetune gcVAE. The inputs to the goal-specific DNNs were the brain ROI volumes. 10 DNNs were trained using the *unmatched unharmonized* ADNI MRI volumes (see Section 2.4). Previous studies have suggested that training with large number of participants from multiple sites can improve generalization to new sites (Liem et al., 2017; Orban et al., 2018; Mårtensson et al., 2020). Indeed, with sufficient training data, there was no difference in performance between intra-site and inter-site prediction even without any harmonization (Abraham et al., 2017).

Therefore, in the current study, the training procedure utilized unharmonized ADNI data without differentiation among ADNI sites.

Recall that not all unmatched timepoints had MMSE and clinical diagnosis information. Therefore, we used the previous timepoint with available information to fill in the missing data (Lipton et al., 2016; Che et al., 2018; Nguyen et al., 2020). Note that this filling in procedure was only performed during training procedure for the unmatched participants.

The architecture of the goal-specific DNN was a generic feedforward neural network, where every layer was fully connected with the next layer. The nonlinear activation function ReLU (Maas et al., 2013) was utilized. The DNN loss function corresponded to the weighted sum of the mean absolute error (MAE) for MMSE prediction and cross entropy loss for clinical diagnosis prediction: $L_{goalDNN} = \lambda_{MMSE}$ MAE + λ_{DX} CrossEntropy. λ_{MMSE} and λ_{DX} were two hyperparameters that were tuned on the validation set.

The metric for tuning hyperparameters in the validation set was the weighted sum of MMSE MAE and clinical diagnosis accuracy: ½ MAE – Diagnosis Accuracy. The MAE term was divided by two so the two terms had similar ranges of values. We utilized the HORD algorithm (Regis & Shoemaker, 2013; Ilievski et al., 2017; Eriksson et al., 2020) to find the best set of hyperparameters using the validation set (Table 1). The trained DNN after 100 epochs was utilized for subsequent analyses.

Hyperparameter	Search range
Initial learning rate	1e-4 – 1e-3
Learning rate step	10 – 99
Dropout rate	0 – 0.5
λ_{MMSE}	0 – 1
λ_{DX}	0 – 1
Nodes for each layer	32 – 512
Number of layers	2 – 5

Table 1. Hyperparameters estimated from the validation set. We note that a learning rate decay strategy was utilized. After K training epochs (where K = learning rate step), the learning rate was reduced by a factor of 10.

At the evaluation phase (Figure 1D), the 10 goal-specific DNNs were applied to the harmonized brain volumes from the matched participants. The prediction performance was averaged across all time points of each participant and the 10 goal-specific DNNs before averaging across participants.

2.7 Baseline harmonization models

Here, we considered ComBat (Johnson et al., 2007) and cVAE (Moyer et al., 2020) as baseline models.

2.7.1 ComBat

ComBat is a linear mixed effects model that controls for additive and multiplicative site effects (Johnson et al., 2007). Here we utilized the R implementation of the algorithm (https://github.com/Jfortin1/ComBatHarmonization). The ComBat model is as follows:

$$x_{ijv} = \alpha_v + Y_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv}\epsilon_{ijv}\,,$$

where i is the site index, j is the participant index and v is the brain ROI index. x_{ijv} is the volume of the v-th brain ROI of subject j from site i. γ_{iv} is the addictive site effect. δ_{iv} is the multiplicative site effect. ϵ_{ijv} is the residual error term following a normal distribution with zero mean and variance δ_v^2. Y_{ij} are the covariates of subject j from site i.

The ComBat parameters α_v, β_v, γ_{iv} and δ_{iv} were estimated for each brain ROI using the unmatched unharmonized ROI volumes (Figure 1B). The estimated parameters can then be applied to a new participant i from site j with brain volume x_{ijv} and covariates Y_{ij}

$$x_{ijv}^{ComBAT} = \frac{x_{ijv} - \hat{\alpha}_v - Y_{ij}^T\hat{\beta}_v - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + \hat{\alpha}_v + Y_{ij}^T\hat{\beta}_v,$$

where ˆ indicates that the parameter was estimated from the *unmatched unharmonized* ROI volumes from ADNI and AIBL. A separate ComBat model was fitted for ADNI and MACC brain volumes. Observe that the equation required the covariates of the new participant. Given that we would like to predict MMSE and clinical diagnosis in the matched participants, we did not utilize MMSE and clinical diagnosis as covariates in the ComBat model. Therefore, in this study, we only utilized age and sex as covariates.

Furthermore, since the goal-specific DNNs were trained with unmatched unharmonized ADNI data without distinguishing among the sites (Section 2.6), for consistency, the ComBat procedure also treated ADNI as a single site despite the data coming from multiple sites and scanners. This was also the case for AIBL.

2.7.2 cVAE

The conditional variational autoencoder (cVAE) model was proposed by Moyer and colleagues to harmonize diffusion MRI data (Moyer et al., 2020). Here, we applied cVAE to harmonize brain ROI volumes. The cVAE model is illustrated in Figure 3A. Input brain volumes were passed through an encoder DNN yielding representation z. Site index s was concatenated with the latent representation z before feeding into the decoder DNN, resulting in the reconstructed brain volumes $\hat{x}$. By incorporating the mutual information $I(z, s)$ in the cost function, this encouraged the learned representation z to be independent of the site s. The resulting lost function is as follows:

$$L_{cVAE} = L_{recon} + \alpha L_{prior} - \gamma L_{adv} + \lambda I(z, s),$$

where L_{recon} is the mean square error (MSE) between x and $\hat{x}$, so this encouraged the harmonized volumes to be similar to the unharmonized volumes. To further encourage x and $\hat{x}$ to be similar, Moyer and colleagues added an additional term L_{adv}, which is the soft-max cross-entropy loss of an adversarial discriminator seeking to distinguish between x and $\hat{x}$. Finally, L_{prior} is the standard KL divergence between representation z and the multivariate Gaussian distribution with zero mean and identity covariance matrix (Sohn et al., 2015).

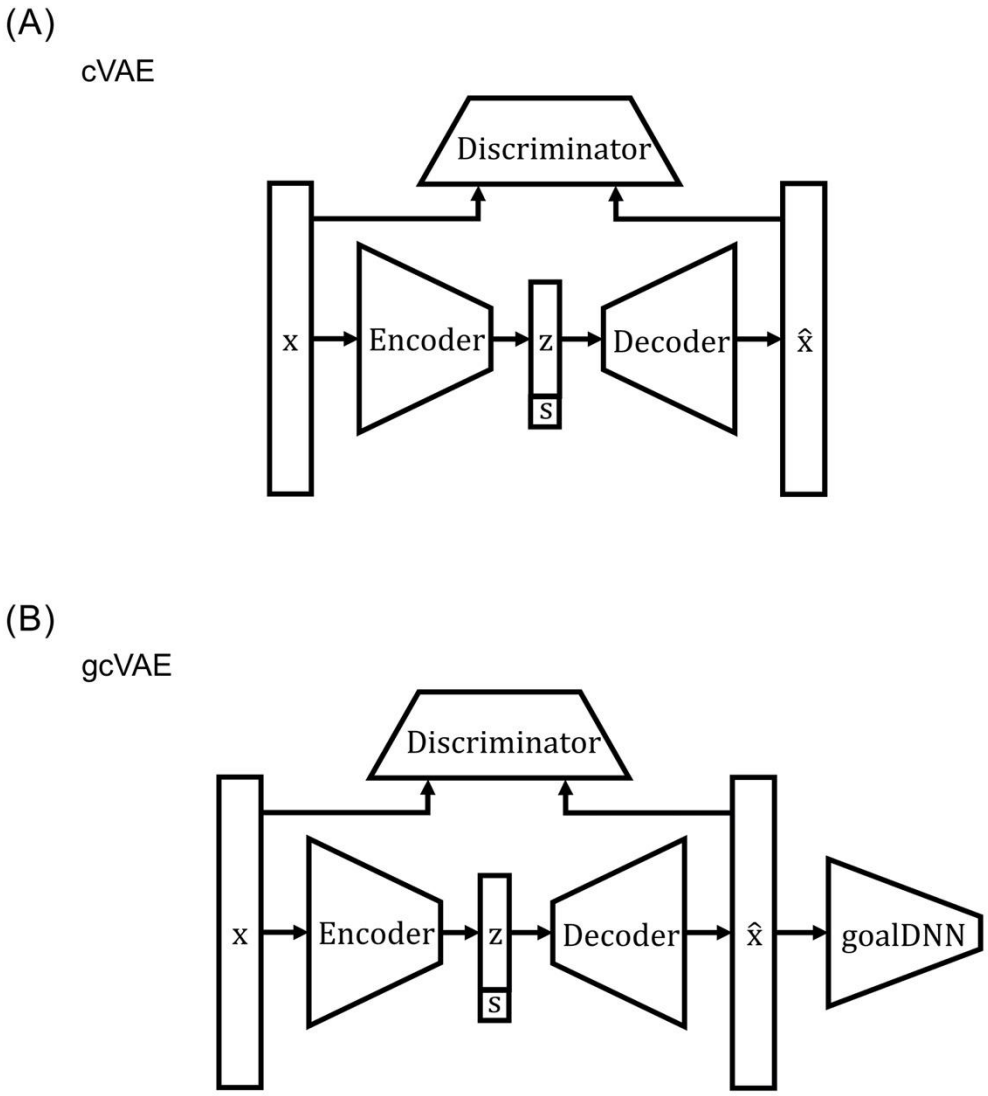


Figure 3. cVAE and gcVAE model structures. (A) Model structure for the cVAE model. Encoder, decoder, and discriminator were all fully connected feedforward DNNs. s was the site we wanted to map the brain volumes to. (B) Model structure for the gcVAE model. The goal-specific DNN (goalDNN) from Section 2.6 was used to guide the cVAE harmonization process. During training of gcVAE, the weights of the goal-specific DNN were fixed.

Both the decoder and encoder were instantiated as generic feedforward neural networks, where every layer was fully connected with the next layer. Following Moyer and colleagues, the nonlinear activation function tanh (Maas et al., 2013) was utilized. During the training process, s is the true site information for input brain volumes x. After training, we could map input x to any site by changing s. The metric for tuning hyperparameters in the validation set was the weighted sum of the reconstruction loss (MSE between x and $\hat{x}$) and the subject-level dataset prediction accuracy: ½ MAE + Dataset Accuracy. The MAE reconstruction loss was divided by two so the two terms had similar ranges of values. Dataset prediction accuracy was obtained by training a XGBoost classifier on the training set and

applying to the validation set. We utilized the HORD algorithm (Regis & Shoemaker, 2013; Ilievski et al., 2017; Eriksson et al., 2020) to find the best set of hyperparameters using the validation set (Table 2). The trained DNN after 1000 epochs was utilized for subsequent analyses (Figure 1C).

Similar to ComBat, the cVAE model was trained using *unmatched unharmonized* brain volumes from ADNI and AIBL. A separate model was trained using ADNI and MACC. For consistency, the cVAE model also treated ADNI and AIBL as single sites.

Hyperparameter	Search range
Initial learning rate	1e-2 – 1e-1
Learning rate step	10 - 999
Dropout rate	0 – 0.5
α	0.01 - 1
γ	0.01 - 10
λ	0.01 - 1
Nodes for each layer	32 - 512
Number of layers	2 - 4
Node for z	32 - 512

Table 2. Hyperparameters estimated from the validation set. We note that a learning rate decay strategy was utilized. After K training epochs (where K = learning rate step), the learning rate was reduced by a factor of 10.

2.8 Goal-specific cVAE (gcVAE)

To incorporate downstream application performance in the harmonization procedure, the outputs of the cVAE (Figure 3A) were fed into the goal-specific DNN (Section 2.6). The resulting goal-specific cVAE (gcVAE) is illustrated in Figure 3B. The loss function of the gcVAE was given by corresponded to the weighted sum of the mean absolute error (MAE) for MMSE prediction and cross entropy loss for clinical diagnosis prediction:

$$\mathrm{L}_{\mathrm{gcVAE}} = \alpha_{MMSE}\mathrm{MAE} + \alpha_{DX}\mathrm{CrossEntropy},$$

where α_{MMSE} and α_{DX} were two hyperparameters to be tuned with the validation set. The loss function was used to finetune the trained cVAE model (Section 2.7.2) using the training set with a relatively small learning rate. We note that the weights of the goal-specific DNN model were frozen during this finetuning procedure.

The metric for tuning hyperparameters in the validation set was the weighted sum of MMSE MAE and clinical diagnosis accuracy: ½ MAE – Diagnosis Accuracy (same as Section 2.6). Since there were only three hyperparameters (learning rate, α_{MMSE} and α_{DX}), a grid search was performed using the validation set to find the best set of hyperparameters.

Similar to ComBat and cVAE, the gcVAE model was trained using *unmatched unharmonized* brain volumes from ADNI and AIBL. A separate model was trained using ADNI and MACC. For consistency, the gcVAE model also treated ADNI and AIBL as single sites.

2.9 Deep neural network implementation

All DNNs were implemented using PyTorch (Paszke et al., 2017) and computed on NVIDIA RTX 3090 GPUs with CUDA 11.0. To optimize the DNNs, we used the Adam optimizer (Kingma & Ba, 2017) with default PyTorch settings.

2.10 Statistical tests

Two-sided two-sample t-tests were utilized to test for differences in age and MMSE between matched participants of AIBI and ADNI (as well as MACC and ADNI). In the case of sex and clinical diagnoses, we utilized chi-squared tests.

As discussed in Sections 2.5 and 2.6, prediction performance was averaged across all time points of each participant and across the 10 sets of models, yielding a single prediction performance for each participant. When comparing dataset prediction performance and goal-specific prediction performance between two harmonization approaches (Figure 1D), we utilized the permutation test with 10,000 permutations.

Multiple comparisons were corrected with a false discovery rate (FDR) of $q < 0.05$.

2.11 Data and code availability

Code for the various harmonization algorithms can be found here (GITHUB_LINK). One of the co-authors (PC) reviewed the code before merging it into the GitHub repository to reduce the chance of coding errors.

The ADNI and the AIBL datasets can be accessed via the Image & Data Archive (https://ida.loni.usc.edu/). The MACC dataset can be obtained via a data-transfer agreement with the MACC (http://www.macc.sg/).

3 Results

3.1 Matched participants were more similar after VAE harmonization

Figure 4A illustrates the Pearson's correlation of each brain ROI volume between matched ADNI and AIBL participants before and after harmonization. Before harmonization, the average correlation (across ROIs) was 0.16. After applying ComBat, the correlation remained low with an average correlation of 0.15. After applying cVAE and gcVAE, the correlations increased to an average of 0.30.

Similar results were obtained with ADNI and MACC (Figure 4B). Before harmonization, the average correlation (across ROIs) was 0.20. After applying ComBat, the correlation remained low with an average correlation of 0.19. After applying cVAE and gcVAE, the correlations increased to an average of 0.44.

Given that matched participants had similar age, sex, MMSE and clinical diagnosis, the results suggest that cVAE and gcVAE appeared to remove more dataset-specific differences than ComBat.

3.2 cVAE & gcVAE removed more dataset differences than ComBat

Figure 5A shows the dataset prediction performance for the matched ADNI and AIBL participants. Before harmonization, the XGBoost classifier was able to predict which dataset a participant came from with 100% accuracy. After applying ComBat, the prediction accuracy dropped to 0.626 ± 0.410 (mean ± std), suggesting significant removal of dataset differences. After applying cVAE and gcVAE, dataset prediction performance dropped to 0.595 ± 0.381 and 0.603 ± 0.382 respectively, which were significantly lower than ComBat (Table 3). There was no statistical difference between cVAE and gcVAE. However, dataset prediction accuracies for cVAE and gcVAE were still better than chance ($p = 1e\text{-}4$), suggesting residual dataset differences.

Similar results were obtained for matched ADNI and MACC participants (Figure 5B). Before harmonization, the XGBoost classifier was able to predict which dataset a participant came from with 100% accuracy. Dataset prediction accuracies after ComBat, cVAE and gcVAE were 0.721 ± 0.392, 0.603 ± 0.391 and 0.598 ± 0.398 respectively. There was no statistical difference between cVAE and gcVAE. Both cVAE and gcVAE had statistically lower dataset prediction performance than ComBat (Table 4).

Overall, cVAE and gcVAE appeared to remove more dataset differences than ComBat. However, dataset prediction accuracies for cVAE and gcVAE were still better than chance (p = 1e-4), suggesting residual dataset differences.

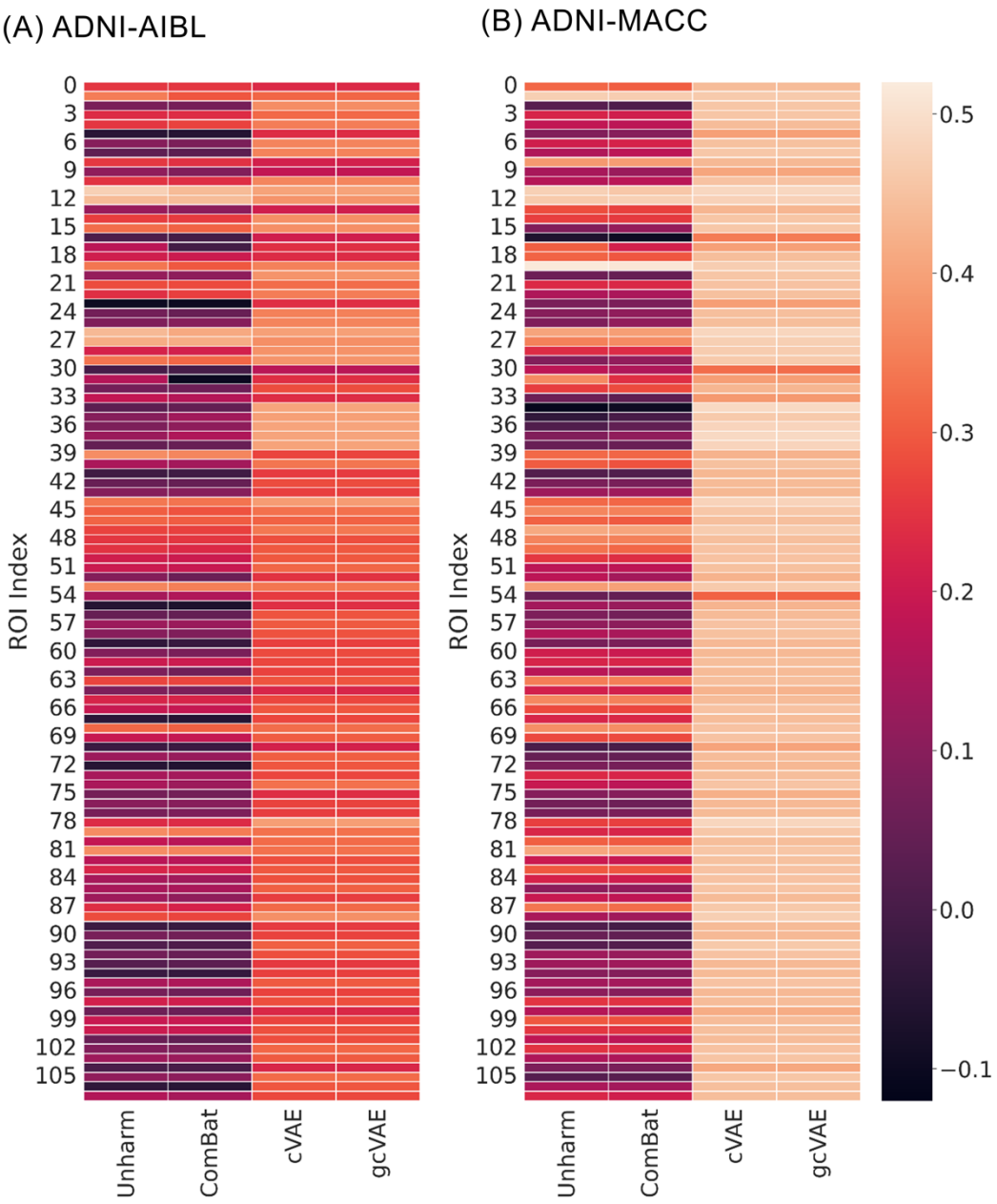


Figure 4. Correlation of each brain volume across matched participants before and after harmonization. (A) Correlation between ADNI and AIBL matched participants. (B) Correlation between ADNI and MACC matched participants. The higher correlations for cVAE and gcVAE suggest better removal of dataset-specific differences.

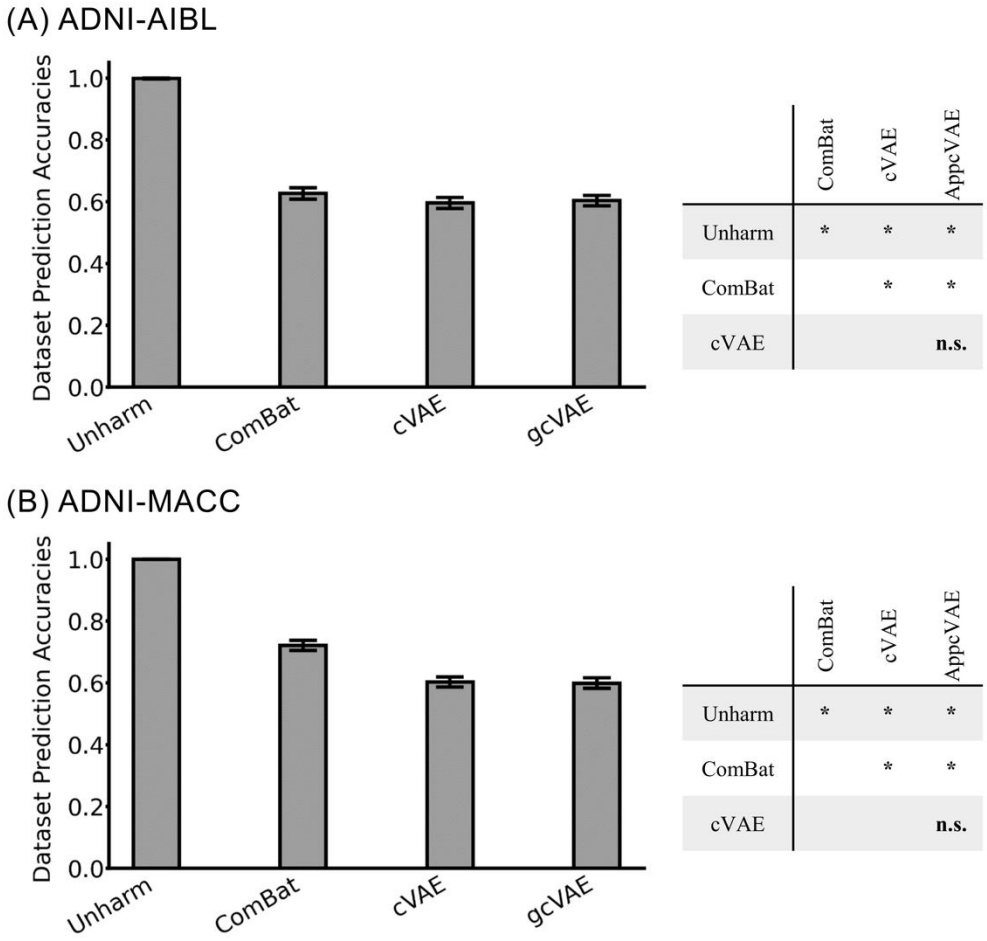


Figure 5. Dataset prediction accuracies. (A) Left: Dataset prediction accuracies for matched ADNI and AIBL participants. Right: p values of differences between different approaches. "*" indicates statistical significance after surviving FDR correction ($q < 0.05$). "n.s." indicates not significant. (B) Same as (A) but for matched ADNI and MACC participants. All p values are reported in Tables 3 and 4.

Dataset Prediction Accuracies (mean ± std)	p values			
	Unharm	ComBat	cVAE	gcVAE
Unharmonized (1.000 ± 0.027)		**1e-4**	**1e-4**	**1e-4**
ComBat (0.626 ± 0.410)			**0.0055**	**0.0410**
cVAE (0.595 ± 0.381)				0.1754
gcVAE (0.603 ± 0.382)				

Table 3. Dataset prediction accuracies with p values of differences between different approaches for matched ADNI and AIBL participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

Dataset Prediction Accuracies (mean ± std)	p values			
	Unharm	ComBat	cVAE	gcVAE
Unharmonized (1.00 ± 1e-16)		**1e-4**	**1e-4**	**1e-4**
ComBat (0.721 ± 0.392)			**1e-4**	**1e-4**
cVAE (0.603 ± 0.391)				0.3584
gcVAE (0.598 ± 0.398)				

Table 4. Dataset prediction accuracies with p values of differences between different approaches for matched ADNI and MACC participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

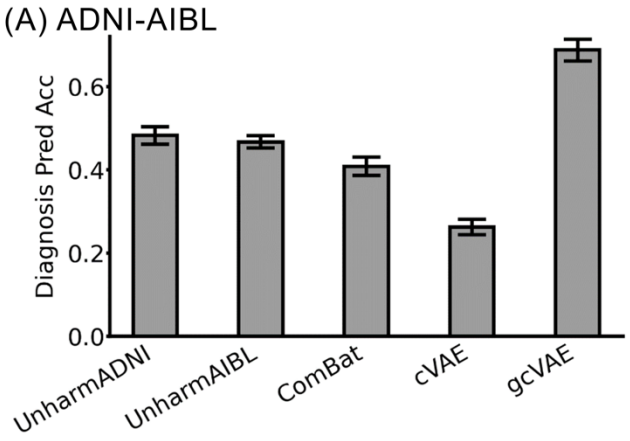
3.3 gcVAE outperformed cVAE for clinical diagnosis prediction

Figure 6A shows the clinical diagnosis prediction accuracies for matched ADNI and AIBL participants. Because the matched participants had similar age, sex, MMSE and clinical diagnosis, comparison between unharmonized ADNI and unharmonized AIBL participants would indicate whether there was a drop in prediction performance due to dataset differences. Unexpectedly, there was no statistical difference in clinical diagnosis prediction performance between unharmonized ADNI and unharmonized AIBL participants (Table 5).

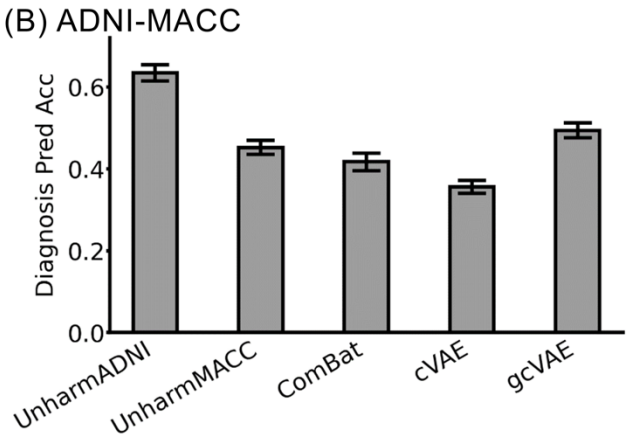
Applying ComBat resulted in a statistical significant drop in prediction performance (p = 7e-4) compared with no harmonization. This suggests that ComBat removed biological information in addition to dataset differences (Figure 5A). cVAE exhibited an even bigger drop in prediction performance compared with ComBat (p = 1e-4), suggesting that the better removal of dataset differences (Figure 5A) came at the expense of removing even more biological information. gcVAE yielded the best prediction performance with statistically significant improvements over all other approaches, including unharmonized ADNI (see p values in Table 5).

Figure 6B shows the clinical diagnosis prediction accuracies for matched ADNI and MACC participants. As expected, there was a significant drop in clinical diagnosis prediction performance between unharmonized ADNI and unharmonized MACC participants (p = 1e-4). The decrease in clinical diagnosis performance was worsened by ComBat and cVAE, once again suggesting that the removal of dataset differences (Figure 5B) came at the expense of also removing biological information. gcVAE recovered a significant portion of the decrease in prediction performance, such that it was not statistically different from

unharmonized MACC (Table 6). However, it was still significantly worse than unharmonized ADNI, suggesting potential room for improvement.



	Unharm AIBL	ComBat	cVAE	AppcVAE
Unharm ADNI	**n.s.**	*	*	*
Unharm AIBL		*	*	*
ComBat			*	*
cVAE				*



	Unharm MACC	ComBat	cVAE	AppcVAE
Unharm ADNI	*	*	*	*
Unharm MACC		*	*	**n.s.**
ComBat			*	*
cVAE				*

Figure 6. Clinical diagnosis prediction accuracies. (A) Left: Clinical diagnosis prediction accuracies for matched ADNI and AIBL participants. Right: p values of differences between different approaches. "*" indicates statistical significance after surviving FDR correction ($q < 0.05$). "n.s." indicates not significant. (B) Same as (A) but for matched ADNI and MACC participants. All p values are reported in Tables 5 and 6.

Clinical Diagnosis Prediction Accuracies (mean ± std)	p values				
	Unharm ADNI	Unharm AIBL	ComBat	cVAE	gcVAE
Unharm ADNI (0.48 ± 0.33)		0.5171	**0.0077**	**1e-4**	**2e-4**
Unharm AIBL (0.47 ± 0.23)			**7e-4**	**1e-4**	**1e-4**
ComBat (0.41 ± 0.34)				**1e-4**	**1e-4**
cVAE (0.26 ± 0.29)					**1e-4**
gcVAE (0.69 ± 0.41)					

Table 5. Clinical diagnosis prediction accuracies with p values of differences between different approaches for matched ADNI and AIBL participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

Clinical Diagnosis Prediction Accuracies (mean ± std)	p values				
	Unharm ADNI	Unharm MACC	ComBat	cVAE	gcVAE
Unharm ADNI (0.63 ± 0.33)		**1e-4**	**1e-4**	**1e-4**	**1e-4**
Unharm MACC (0.45 ± 0.29)			**0.0124**	**1e-4**	0.0545
ComBat (0.42 ± 0.35)				**2e-4**	**0.0065**
cVAE (0.36 ± 0.26)					**1e-4**
gcVAE (0.49 ± 0.30)					

Table 6. Clinical diagnosis prediction accuracies with p values of differences between different approaches for matched ADNI and MACC participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

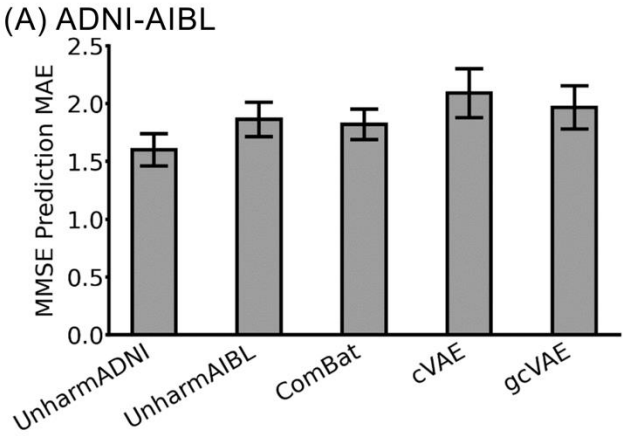
3.4 gcVAE outperformed cVAE in MMSE prediction

Figure 7A shows the MMSE prediction mean absolute error (MAE) for matched ADNI and AIBL participants. Because the matched participants had similar age, sex, MMSE and clinical diagnosis, comparison between unharmonized ADNI and unharmonized AIBL participants would indicate whether there was a drop in prediction performance due to dataset differences. As expected, there was a drop in MMSE prediction performance (increased MAE) for unharmonized AIBL participants compared with unharmonized ADNI participants ($p = 1e\text{-}4$).

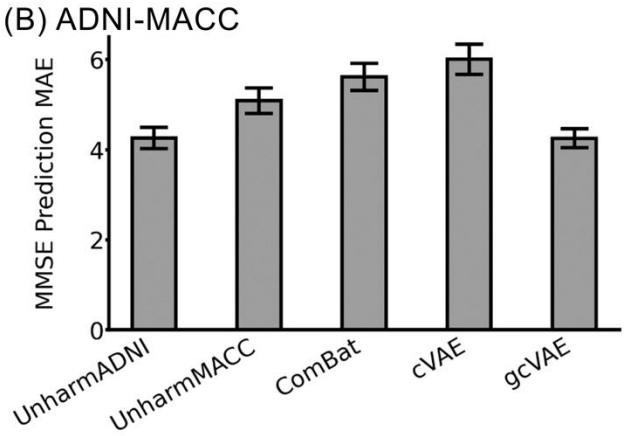
There was no statistical difference between ComBat and the unharmonized AIBL participants. cVAE had statistically worse MMSE prediction performance compared with all

other approaches (p values in Table 7). gcVAE recovered a significant portion of the decrease in prediction performance, such that prediction performance was not statistically different from ComBat and unharmonized AIBL (Table 7). However, it was still statistically worse than unharmonized ADNI, suggesting further room for improvement.

Figure 7B shows the MMSE prediction MAE for matched ADNI and MACC participants. As expected, there was a drop in MMSE prediction performance (increased MAE) for unharmonized MACC participants compared with unharmonized ADNI participants (p = 1e-4). Both ComBat and cVAE caused further drop in prediction performance (p values in Table 8). gcVAE had the best prediction performance (lowest MAE), such that prediction performance was not statistically different from unharmonized ADNI (Table 8).



	Unharm AIBL	ComBat	cVAE	AppcVAE
Unharm ADNI	*	*	*	*
Unharm AIBL		n.s.	*	n.s.
ComBat			*	n.s.
cVAE				*



	Unharm MACC	ComBat	cVAE	AppcVAE
Unharm ADNI	*	*	*	n.s.
Unharm MACC		*	*	*
ComBat			*	*
cVAE				*

Figure 7. MMSE prediction errors as measured by mean absolute error (MAE). (A) Left: MMSE prediction errors for matched ADNI and AIBL participants. Right: p values of differences between different approaches. "*" indicates statistical significance after surviving FDR correction ($q < 0.05$). "n.s." indicates not significant. (B) Same as (A) but for matched ADNI and MACC participants. All p values are reported in Tables 7 and 8.

MMSE Prediction MAE (mean ± std)	p values				
	Unharm ADNI	Unharm AIBL	ComBat	cVAE	gcVAE
Unharm ADNI (1.60 ± 2.17)		**1e-4**	**0.0061**	**1e-4**	**1e-4**
Unharm AIBL (1.86 ± 2.32)			**0.4339**	**0.0054**	0.0756
ComBat (1.82 ± 2.07)				**0.0322**	0.1473
cVAE (2.09 ± 3.29)					**0.0023**
gcVAE (1.97 ± 2.93)					

Table 7. MMSE prediction errors with p values of differences between different approaches for matched ADNI and AIBL participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

MMSE Pred MAE (mean ± std)	p values				
	Unharm ADNI	Unharm MACC	ComBat	cVAE	gcVAE
Unharm ADNI (4.26 ± 3.87)		**1e-4**	**1e-4**	**1e-4**	0.9570
Unharm MACC (5.09 ± 4.66)			**1e-4**	**1e-4**	**1e-4**
ComBat (5.61 ± 5.03)				**1e-4**	**1e-4**
cVAE (5.96 ± 5.50)					**1e-4**
gcVAE (4.61 ± 3.57)					

Table 8. MMSE prediction errors with p values of differences between different approaches for matched ADNI and MACC participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

4 Discussion

In this study, we proposed a flexible harmonization framework to utilize downstream application performance to regularize the harmonization model. Our proposed approach could be integrated with most harmonization approaches based on DNNs. Here, we integrated our approach with the cVAE model. Using three large-scale datasets, we demonstrated that gcVAE compared favorably with ComBat and cVAE.

We found that cVAE was able to significantly remove more dataset differences than ComBat (Figure 5). This makes intuitive sense given that cVAE considered all brain regions jointly, so should theoretically be able to remove multivariate site effects distributed across brain regions. However, the removal of more dataset differences came at the expense of also removing relevant biological information as measured by downstream application performance (Figures 6 and 7).

Indeed, the removal of relevant biological information was an issue not just for cVAE, but also for ComBat. In the case of predicting clinical diagnosis and MMSE, the use of ComBat led to similar or worse performance than not harmonizing at all. By constraining the harmonization with goal-specific DNNs, the cVAE models were able to yield better prediction of MMSE and clinical diagnosis (Figures 6 and 7), while removing as much dataset differences as cVAE (Figure 5). In the case of clinical diagnosis prediction, gcVAE was able to yield better prediction performance than no harmonization. In the case of MMSE prediction, gcVAE was able to yield better prediction performance than no harmonization in the MACC dataset, but was only able to yield comparable prediction performance than no harmonization in the AIBL dataset.

However, the strength of gcVAE is also its main limitation. The reliance of goal-specific DNNs led to better downstream performance, but the resulting improvements might not generalize to new downstream applications. Therefore, the training procedure might have to be repeated for each new downstream application. Future research is necessary to address this limitation.

5 Conclusion

In this study, we proposed a goal-specific brain MRI harmonization framework, which took into account downstream application performance in the harmonization process. Using three large-scale datasets, we demonstrated that our approach compared favorably with existing approaches in terms of preserving relevant biological information, while removing site differences.

Acknowledgment

Our research is currently supported by the Singapore National Research Foundation (NRF) Fellowship (Class of 2017), the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), NMRC STaR (STaR20nov-0003), and the USA NIH (R01MH120080). Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore NRF or the Singapore NMRC. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, *147*, 736–745. https://doi.org/10/f9xjqn

Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., Wolf, D. H., Gur, R. E., Gur, R. C., … The iSTAGING and PHENOM consortia. (2021). Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors. *Journal of Magnetic Resonance Imaging*, jmri.27908. https://doi.org/10/gmzt7m

Blumberg, S. B., Tanno, R., Kokkinos, I., & Alexander, D. C. (2018). Deeper Image Quality Transfer: Training Low-Memory Neural Networks for 3D Images. *ArXiv:1808.05577 [Cs, q-Bio]*. http://arxiv.org/abs/1808.05577

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, *8*(1), 6085. https://doi.org/10/gdfncx

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., Shou, H., & the Alzheimer's Disease Neuroimaging Initiative. (2019). *Removal of Scanner Effects in Covariance Improves Multivariate Pattern Analysis in Neuroimaging Data* [Preprint]. Neuroscience. https://doi.org/10.1101/858415

Chen, J., Liu, J., Calhoun, V. D., Arias-Vasquez, A., Zwiers, M. P., Gupta, C. N., Franke, B., & Turner, J. A. (2014). Exploration of scanning effects in multi-site structural MRI studies. *Journal of Neuroscience Methods*, *230*, 37–50. https://doi.org/10/f56f36

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10/gdp84q

Chong, J. S. X., Liu, S., Loke, Y. M., Hilal, S., Ikram, M. K., Xu, X., Tan, B. Y., Venketasubramanian, N., Chen, C. L.-H., & Zhou, J. (2017). Influence of cerebrovascular disease on brain networks in prodromal and clinical Alzheimer's disease. *Brain*, *140*(11), 3012–3022. https://doi.org/10/gcj7wj

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10/b2jf74

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., van Zijl, P. C. M., & Prince, J. L. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, *64*, 160–170. https://doi.org/10/ggbzsg

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., … the AIBL Research Group. (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/dg74qm

Ellis, K. A., Rowe, C. C., Villemagne, V. L., Martins, R. N., Masters, C. L., Salvado, O., Szoeke, C., Ames, D., & Group, A. research. (2010). Addressing population aging and Alzheimer's disease through the Australian Imaging Biomarkers and Lifestyle study: Collaboration with the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *6*(3), 291–296. https://doi.org/10/btmfdj

Eriksson, D., Pearce, M., Gardner, J. R., Turner, R., & Poloczek, M. (2020). Scalable Global Optimization via Local Bayesian Optimization. *ArXiv:1910.01739 [Cs, Stat]*. http://arxiv.org/abs/1910.01739

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation. *Neuron*, *33*(3), 341–355. https://doi.org/10/dbx4cf

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

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. https://doi.org/10/gcmg6f

Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W. H. L., Corvin, A., Redolfi, A., Nelson, B., Crespo-Facorro, B., McDonald, C., Tordesillas-Gutiérrez, D., Cannon, D., Mothersill, D., Hernaus, D., Morris, D., Setien-Suero, E., Donohoe, G., Frisoni, G., Tronchin, G., … Mechelli, A. (2020). Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *NeuroImage*, *220*, 117127. https://doi.org/10.1016/j.neuroimage.2020.117127

Hawco, C., Viviano, J. D., Chavez, S., Dickie, E. W., Calarco, N., Kochunov, P., Argyelan, M., Turner, J. A., Malhotra, A. K., Buchanan, R. W., & Voineskos, A. N. (2018). A longitudinal human phantom reliability study of multi-center T1-weighted, DTI, and resting state fMRI data. *Psychiatry Research: Neuroimaging*, *282*, 134–142. https://doi.org/10/gg48ch

Hilal, S., Chai, Y. L., Ikram, M. K., Elangovan, S., Yeow, T. B., Xin, X., Chong, J. Y., Venketasubramanian, N., Richards, A. M., Chong, J. P. C., Lai, M. K. P., & Chen, C. (2015). Markers of Cardiac Dysfunction in Cognitive Impairment and Dementia. *Medicine*, *94*(1), e297. https://doi.org/10/gpfkhg

Hilal, S., Tan, C. S., van Veluw, S. J., Xu, X., Vrooman, H., Tan, B. Y., Venketasubramanian, N., Biessels, G. J., & Chen, C. (2020). Cortical cerebral microinfarcts predict cognitive decline in memory clinic patients. *Journal of Cerebral Blood Flow & Metabolism*, *40*(1), 44–53. https://doi.org/10/gpfttm

Ilievski, I., Akhtar, T., Feng, J., & Shoemaker, C. A. (2017). Efficient Hyperparameter Optimization of Deep Learning Algorithms Using Deterministic RBF Surrogates. *ArXiv:1607.08316 [Cs, Stat]*. http://arxiv.org/abs/1607.08316

Jack, C. R., Bernstein, M. A., Borowski, B. J., Gunter, J. L., Fox, N. C., Thompson, P. M., Schuff, N., Krueger, G., Killiany, R. J., DeCarli, C. S., Dale, A. M., Carmichael, O. W., Tosun, D., & Weiner, M. W. (2010). Update on the Magnetic Resonance Imaging core of the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *6*(3), 212–220. https://doi.org/10/b6dx5v

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P.,

Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., … ADNI Study. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10/d6mhb9

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/dsf386

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., Fischl, B., & Dale, A. (2006). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, *30*(2), 436–443. https://doi.org/10/c2hrm5

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980 [Cs]*. http://arxiv.org/abs/1412.6980

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1–2), 83–97. https://doi.org/10/b2k5tg

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., Lampe, L., Rahim, M., Abraham, A., Craddock, R. C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M. L., Witte, A. V., Villringer, A., & Margulies, D. S. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, *148*, 179–188. https://doi.org/10/f9zpss

Lipton, Z. C., Kale, D. C., & Wetzel, R. (2016). *Modeling Missing Data in Clinical Time Series with RNNs*. 17.

Lu, B., Li, H.-X., Chang, Z.-K., Li, L., Chen, N.-X., Zhu, Z.-C., Zhou, H.-X., Li, X.-Y., Wang, Y.-W., Cui, S.-X., Deng, Z.-Y., Fan, Z., Yang, H., Chen, X., Thompson, P. M., Castellanos, F. X., Yan, C.-G., & for the Alzheimer's Disease Neuroimaging Initiative. (2020). *A Practical Alzheimer Disease Classifier via Brain Imaging-Based Deep Learning on 85,721 Samples* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.08.18.256594

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *In ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Magnotta, V. A., Matsui, J. T., Liu, D., Johnson, H. J., Long, J. D., Bolster, B. D., Mueller, B. A., Lim, K., Mori, S., Helmer, K. G., Turner, J. A., Reading, S., Lowe, M. J., Aylward, E., Flashman, L. A., Bonett, G., & Paulsen, J. S. (2012). MultiCenter Reliability of Diffusion Tensor Imaging. *Brain Connectivity*, *2*(6), 345–355. https://doi.org/10/gk82p6

Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M. G., Taylor, J.-P., Hort, J., Snædal, J., Kulisevsky, J., Blanc, F., Antonini, A., Mecocci, P., Vellas, B., Tsolaki, M., … Westman, E. (2020). The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. *Medical Image Analysis*, *66*, 101714. https://doi.org/10/ggx6gm

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10/f89khc

Modanwal, G., Vellal, A., Buda, M., & Mazurowski, M. A. (2020). MRI image harmonization using cycle-consistent generative adversarial network. In H. K. Hahn

& M. A. Mazurowski (Eds.), *Medical Imaging 2020: Computer-Aided Diagnosis* (p. 36). SPIE. https://doi.org/10/gmzt6h

Moyer, D., Ver Steeg, G., Tax, C. M. W., & Thompson, P. M. (2020). Scanner invariant representations for diffusion MRI harmonization. *Magnetic Resonance in Medicine*, *84*(4), 2174–2189. https://doi.org/10.1002/mrm.28243

Nguyen, M., He, T., An, L., Alexander, D. C., Feng, J., & Yeo, B. T. T. (2020). Predicting Alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, *222*, 117203. https://doi.org/10.1016/j.neuroimage.2020.117203

Ning, L., Bonet-Carne, E., Grussu, F., Sepehrband, F., Kaden, E., Veraart, J., Blumberg, S. B., Khoo, C. S., Palombo, M., Coll-Font, J., Scherrer, B., Warfield, S. K., Karayumak, S. C., Rathi, Y., Koppers, S., Weninger, L., Ebert, J., Merhof, D., Moyer, D., … Tax, C. W. M. (2019). Muti-shell Diffusion MRI Harmonisation and Enhancement Challenge (MUSHAC): Progress and Results. In E. Bonet-Carne, F. Grussu, L. Ning, F. Sepehrband, & C. M. W. Tax (Eds.), *Computational Diffusion MRI* (pp. 217–224). Springer International Publishing. https://doi.org/10.1007/978-3-030-05831-9_18

Orban, P., Dansereau, C., Desbois, L., Mongeau-Pérusse, V., Giguère, C.-É., Nguyen, H., Mendrek, A., Stip, E., & Bellec, P. (2018). Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. *Schizophrenia Research*, *192*, 167–171. https://doi.org/10/gc64cv

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. *Undefined*. https://www.semanticscholar.org/paper/Automatic-differentiation-in-PyTorch-Paszke-Gross/b36a5bb1707bb9c70025294b3a310138aae8327a

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Wolf, D. H., … Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, *208*, 116450. https://doi.org/10.1016/j.neuroimage.2019.116450

Regis, R. G., & Shoemaker, C. A. (2013). Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. *Engineering Optimization*, *45*(5), 529–555. https://doi.org/10/gkg6w7

Russkikh, N., Antonets, D., Shtokalo, D., Makarov, A., Vyatkin, Y., Zakharov, A., & Terentyev, E. (2020). Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis. *Bioinformatics*, *36*(20), 5076–5085. https://doi.org/10.1093/bioinformatics/btaa624

Sohn, K., Lee, H., & Yan, X. (2015). Learning Structured Output Representation using Deep Conditional Generative Models. *Advances in Neural Information Processing Systems*, *28*. https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html

Tang, S., Sun, N., Floris, D. L., Zhang, X., Di Martino, A., & Yeo, B. T. T. (2020). Reconciling Dimensional and Categorical Models of Autism Heterogeneity: A Brain Connectomics and Behavioral Study. *Biological Psychiatry*, *87*(12), 1071–1082. https://doi.org/10/gpfkpx

Tanno, R., Worrall, D. E., Ghosh, A., Kaden, E., Sotiropoulos, S. N., Criminisi, A., & Alexander, D. C. (2017). Bayesian Image Quality Transfer with CNNs: Exploring Uncertainty in dMRI Super-Resolution. *ArXiv:1705.00664 [Cs]*. http://arxiv.org/abs/1705.00664

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10/b3t59j

Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., Buckner, R. L., Buitelaar, J. K., Bulayeva, K. B., Cannon, D. M., Cohen, R. A., Conrod, P. J., Dale, A. M., Deary, I. J., Dennis, E. L., de Reus, M. A., Desrivieres, S., Dima, D., Donohoe, G., … Ye, J. (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage*, *145*, 389–408. https://doi.org/10/f9jx4v

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10/f46ktq

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., Deeds, B. G., Dowling, G. J., Grant, S., Howlett, K. D., Matochik, J. A., Morgan, G. D., Murray, M. M., Noronha, A., Spong, C. Y., … Weiss, S. R. B. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7. https://doi.org/10/gd8ctx

Wachinger, C., Rieckmann, A., & Pölsterl, S. (2021). Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, *67*, 101879. https://doi.org/10/gh5vwj

Whelan, C. D., Altmann, A., Botía, J. A., Jahanshad, N., Hibar, D. P., Absil, J., Alhusaini, S., Alvim, M. K. M., Auvinen, P., Bartolini, E., Bergo, F. P. G., Bernardes, T., Blackmon, K., Braga, B., Caligiuri, M. E., Calvo, A., Carr, S. J., Chen, J., Chen, S., … Sisodiya, S. M. (2018). Structural brain abnormalities in the common epilepsies assessed in a worldwide ENIGMA study. *Brain*, *141*(2), 391–408. https://doi.org/10/gfz93s

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., & Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, *39*(11), 4213–4227. https://doi.org/10/gff7m4

Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., the UNC/UMN Baby Connectome Project Consortium, Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Shen, D., & Li, G. (2019). Harmonization of Infant Cortical Thickness Using Surface-to-Surface Cycle-Consistent Adversarial Networks. In S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Vol. 11767, pp. 475–483). Springer International Publishing. https://doi.org/10.1007/978-3-030-32251-9_52

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

Zuo, L., Dewey, B. E., Liu, Y., He, Y., Newsome, S. D., Mowry, E. M., Resnick, S. M., Prince, J. L., & Carass, A. (2021). Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, *243*, 118569. https://doi.org/10/gmzt6k

Supplementary

	Timepoint	ADNI value	AIBL value	P value
AGE	1	71.0±5.5	70.8±5.3	0.96
	2	72.5±5.5	72.6±5.5	0.98
	3	74.2±5.5	73.9±5.6	0.93
	4	75.7±5.5	75.6±5.5	0.99
MMSE	1	29.3±0.9	29.2±0.9	1.00
	2	29.5±0.5	29.5±0.5	1.00
	3	29.7±0.5	29.7±0.5	1.00
	4	29.5±0.8	29.5±0.8	1.00
AD diagnosis	1	100%-0%-0%	100%-0%-0%	1.00
	2	100%-0%-0%	100%-0%-0%	1.00
	3	100%-0%-0%	100%-0%-0%	1.00
	4	100%-0%-0%	100%-0%-0%	1.00
Sex	-	50%	50%	1.00

Table S1. ADNI-AIBL matching results for participants having 4 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	AIBL value	P value
AGE	1	73.3±3.3	73.1±3.3	0.96
	2	74.8±3.3	75.2±3.3	0.94
	3	76.3±3.3	76.1±3.3	0.97
MMSE	1	29.0±0.0	20.0±0.0	1.00
	2	30.0±0.0	30.0±0.0	1.00
	3	30.0±0.0	30.0±0.0	1.00
AD diagnosis	1	100%-0%-0%	100%-0%-0%	1.00
	2	100%-0%-0%	100%-0%-0%	1.00
	3	100%-0%-0%	100%-0%-0%	1.00
Sex	-	50%	50%	1.00

Table S2. ADNI-AIBL matching results for participants having 3 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated

from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	AIBL value	P value
AGE	1	74.4±9.8	74.5±9.8	0.99
	2	76.1±9.8	76.1±9.9	0.99
MMSE	1	27.9±2.8	27.9±2.8	1.00
	2	27.8±2.8	27.8±2.8	1.00
AD diagnosis	1	57%-43%-0%	57%-43%-0%	1.00
	2	57%-43%-0%	57%-43%-0%	1.00
Sex	-	88%	88%	1.00

Table S3. ADNI-AIBL matching results for participants having 2 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	AIBL value	P value
AGE	1	74.8±5.9	74.8±5.9	1.00
MMSE	1	27.3±3.9	27.3±3.9	0.98
AD diagnosis	1	68%-19%-13%	68%-19%-13%	1.00
Sex	-	43%	43%	1.00

Table S4. ADNI-AIBL matching results for participants having 1 time point (scan). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	MACC value	P value
AGE	1	71.5±6.8	72.3±6.7	0.67
	2	73.5±6.8	73.8±6.8	0.91
	3	75.9±6.9	75.5±6.6	0.81
MMSE	1	26.9±3.7	27.0±3.5	0.94
	2	26.1±4.5	26.1±4.5	0.98
	3	24.9±6.3	25.2±6.3	0.87
AD diagnosis	1	39%-46%-15%	36%-54%-10%	0.72
	2	43%-36%-21%	46%-36%-18%	0.88
	3	43%-36%-21%	46%-32%-22%	0.91
Sex	-	57%	57%	1.00

Table S5. ADNI-MACC matching results for participants having 3 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	MACC value	P value
AGE	1	73.6±5.7	73.9±5.6	0.78
	2	75.8±5.6	75.5±5.6	0.71
MMSE	1	24.7±4.9	24.8±4.6	0.86
	2	23.4±6.9	23.5±6.6	0.91
AD diagnosis	1	35%-38%-27%	35%-40%-25%	0.80
	2	37%-30%-33%	37%-35%-28%	0.49
Sex	-	51%	58%	0.20

Table S6. ADNI-MACC matching results for participants having 2 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	MACC value	P value
AGE	1	75.7±6.7	75.7±6.7	0.97
MMSE	1	21.0±5.9	21.0±5.9	0.94
AD diagnosis	1	14%-34%-52%	14%-38%-48%	0.64
Sex	-	52%	56%	0.34

Table S7. ADNI-MACC matching results for participants having 1 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.