Goal-specific brain MRI harmonization

Lijun An [a,b,c], Jianzhong Chen [a,b,c], Pansheng Chen [a,b,c], Chen Zhang [a,b,c], Tong He [a,b,c], Christopher Chen [d], Juan Helen Zhou [a,b,e], B.T. Thomas Yeo [a,b,c,f,*], for the Alzheimer's Disease Neuroimaging Initiative[1], and the Australian Imaging Biomarkers and Lifestyle Study of Aging[1]

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

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 10,085 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 re-gression 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; Ning et al., 2019; Blumberg 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 10,085 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 center (MACC) Harmonization cohort (Hilal et al., 2015; Chong et al., 2017; Hilal et al., 2020). Data collection was approved by the Institutional Review Board (IRB) at each corresponding institution. The analysis in the current study is approved by the National University of Singapore IRB. 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 to 3T scanners from different vendors (see Table S1 for more details). 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 (see Table S2 for more details). 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. Figs. 1 and 2 illustrate 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 (Fig. 1A). The distributions of age, sex, MMSE and clinical diagnosis of all participants and matched participants are shown in Fig. 3.

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. Care was taken to ensure that all scans of every participant were classified as either "matched" or "unmatched", and not split between the two categories. *P* values showing the quality of the matching procedure are found in Tables S3–S9.

The unmatched ADNI data was used to train goal-specific deep neural networks (DNN) for predicting MMSE and clinical diagnosis (Fig. 1B; details in Section 2.5). Here, clinical diagnosis categories were normal controls, mild cognitive impairment, and Alzheimer's disease dementia. The clinical diagnoses from all three datasets were determined by multiple criteria, including MRI and cognitive tests. The unmatched ADNI and AIBL participants were also used to fit ComBat and cVAE (Fig. 1C; details in Section 2.6). The unmatched AIBL participants and goal-specific DNN (from Fig. 1B) were utilized for training the gcVAE model (Fig. 1C). The same procedure was applied to ADNI and MACC.

The trained harmonization models were then applied to unharmonized brain volumes of all matched and unmatched participants (Fig. 2A). The harmonized data was evaluated with two criteria (Fig. 2B and C). The first criterion was dataset prediction performance, in which a machine learning algorithm was used to predict which dataset the harmonized data came from (Fig. 2B). Lower dataset prediction performance indicates better harmonization. More specifically, we trained a XGBoost classifier (Chen and Guestrin, 2016) using the harmonized ADNI and harmonized AIBL brain volumes from the unmatched participants (Fig. 2B). We then applied the classifier to the harmonized ADNI and AIBL brain volumes from the matched participants (details in Section 2.8). 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 (Fig. 2C). Here, we applied the goal-specific DNN (from Fig. 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 AIBL and unharmonized ADNI brain volumes from the matched participants (Fig. 2C). The same procedure was applied to ADNI and MACC.

We note that the goal-specific DNN (Fig. 1B), harmonization models (Fig. 1C) and dataset prediction classifier (Fig. 2B) were trained on unmatched data, while harmonization evaluation was performed on matched data (Fig. 2B and C). The matching procedure was important to ensure that prediction performance was comparable between matched
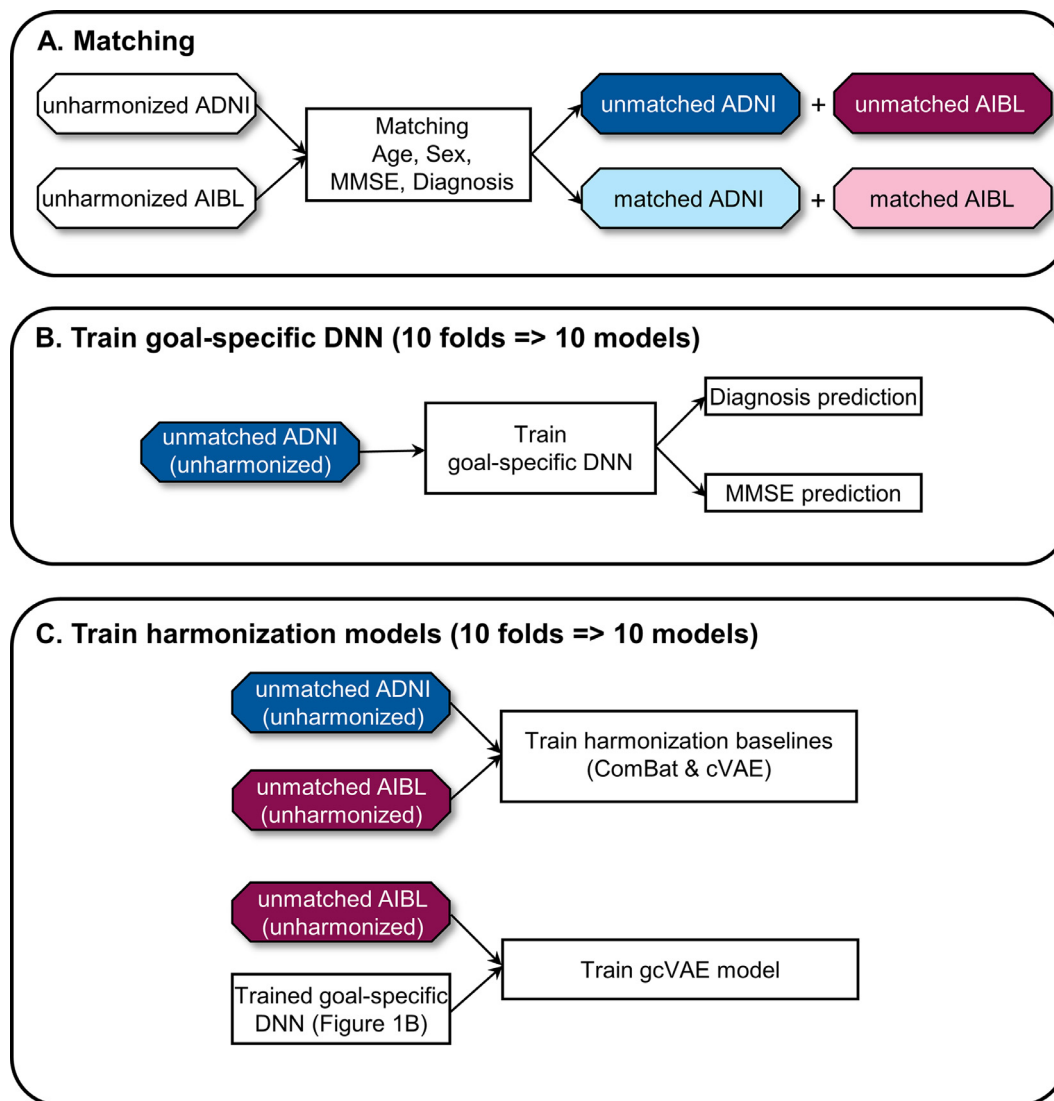
Fig. 1. *Workflow of current study for data matching and model training.* We illustrate the workflow using ADNI and AIBL. The same procedure was applied to ADNI and MACC. (A) Matching participants to derive test set for harmonization evaluation. (B) Train goal-specific deep neural network (DNN) using unmatched unharmonized ADNI data to predict clinical diagnosis and MMSE. (C) Train harmonization models using unmatched unharmonized data. We note that ComBat and cVAE were trained using unmatched unharmonized ADNI and AIBL data, while gcVAE was trained using unmatched unharmonized AIBL data and the goal-specific DNN (from Fig. 1B). Dark colors (e.g., dark red and dark blue) are used to indicate unmatched participants, while light colors (e.g., pink and light blue) are used to indicate matched participants.

ADNI and matched AIBI participants. Suppose we did the opposite: trained a clinical diagnosis prediction model on matched ADNI participants and then tested the model on unmatched ADNI and unmatched AIBL participants. In this scenario, the clinical diagnosis prediction performance would not be comparable between unmatched ADNI and unmatched AIBL participants. More specifically, suppose unmatched ADNI comprised mostly participants with AD and healthy participants, as well as few participants with mild cognitive impairment (MCI). On the other hand, suppose AIBL contained equal proportions of healthy participants, participants with MCI and participants with AD. In this scenario, because it is easier to distinguish between healthy controls and participants with AD, compared with distinguishing participants with MCI from the other two classes (participants with AD and healthy participants), the prediction performance would likely be better in unmatched ADNI compared with unmatched AIBL, even if there was no scanner difference between the two sites. By testing prediction performance on matched AIBL and matched ADNI participants, we ensure that any drop in prediction performance was due to unavoidable site differences, such as scanner differences.

2.4. Training, validation and test procedure

As mentioned in the previous section, the matched participants were used as the test set for evaluation (Fig. 2C). The unmatched participants were used for training the goal-specific DNN (Fig. 1B), harmonization (Fig. 1C) and dataset prediction (Fig. 2B) models. More specifically, we divided the unmatched participants into 10 groups. Recall that a participant might be scanned at multiple timepoints. Care was taken to ensure that all timepoints of any participant were assigned to be in a single group, and 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 (Fig. 2A), 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 for evaluation (Fig. 2B and C).
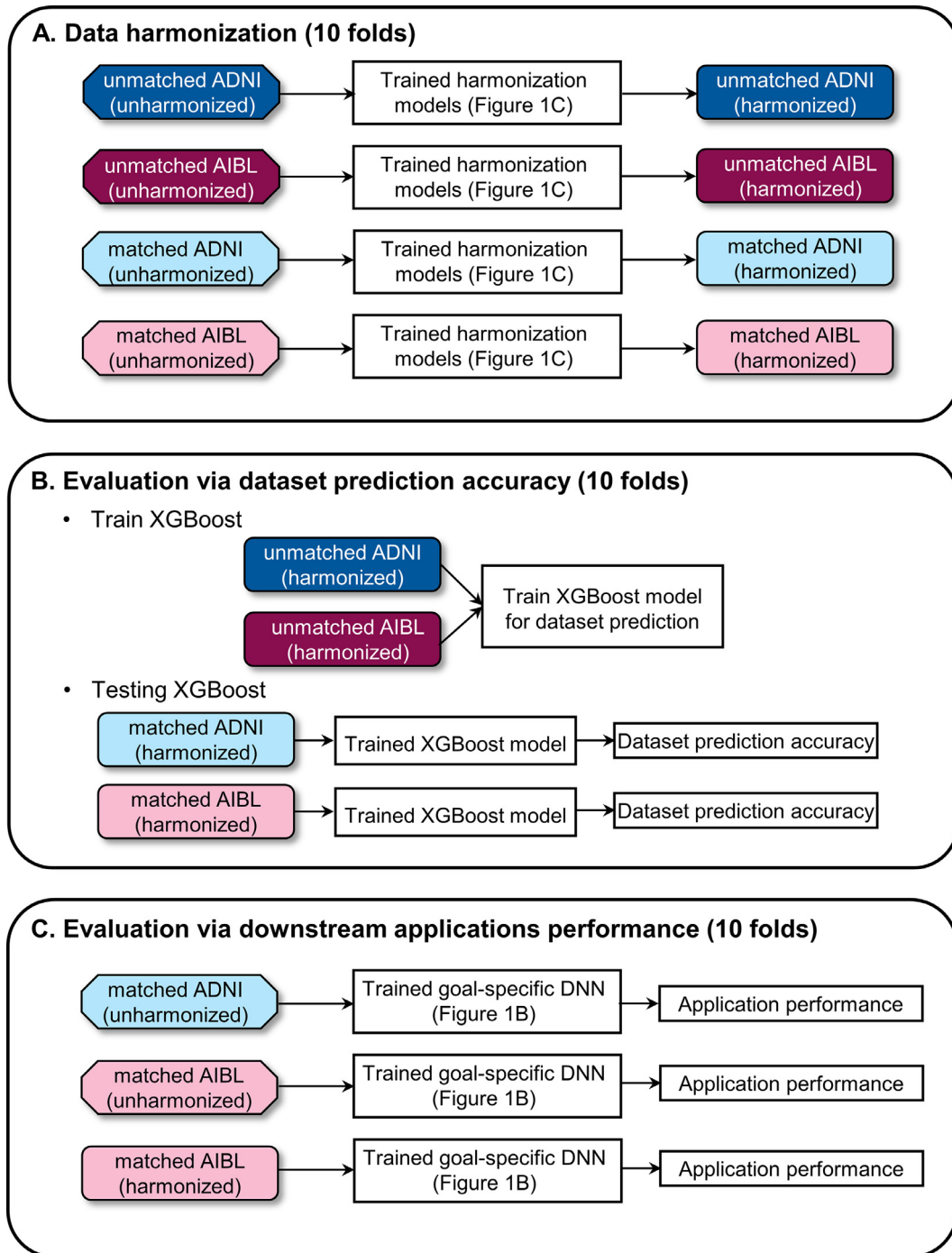
Fig. 2. *Workflow of current study for data harmonization and performance evaluation.* We illustrate the workflow using ADNI and AIBL. The same procedure was applied to ADNI and MACC. (A) Harmonize data using trained harmonization models from Fig. 1C. (B) Evaluate harmonization performance using XGBoost dataset prediction model. (C) Evaluate harmonization performance using goal-specific DNN (Fig. 1B) to predict MMSE and clinical diagnosis. We note that dark colors (e.g., dark red and dark blue) are used to indicate unmatched participants, while light colors (e.g., pink and light blue) are used to indicate matched participants. On the other hand, octagons are used to indicate unharmonized data, while rectangles (with rounded corners) are used to indicate harmonized data.

2.5. 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 train the gcVAE model (Fig. 1C) and evaluate the harmonization approaches (Fig. 2C). The inputs to the goal-specific DNNs were the brain ROI volumes. 10 DNNs were trained with a 10-fold cross-validation procedure (Section 2.4) using the *unmatched unharmo-*

nized ADNI MRI volumes (Fig. 1B). The training procedure utilized the 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.
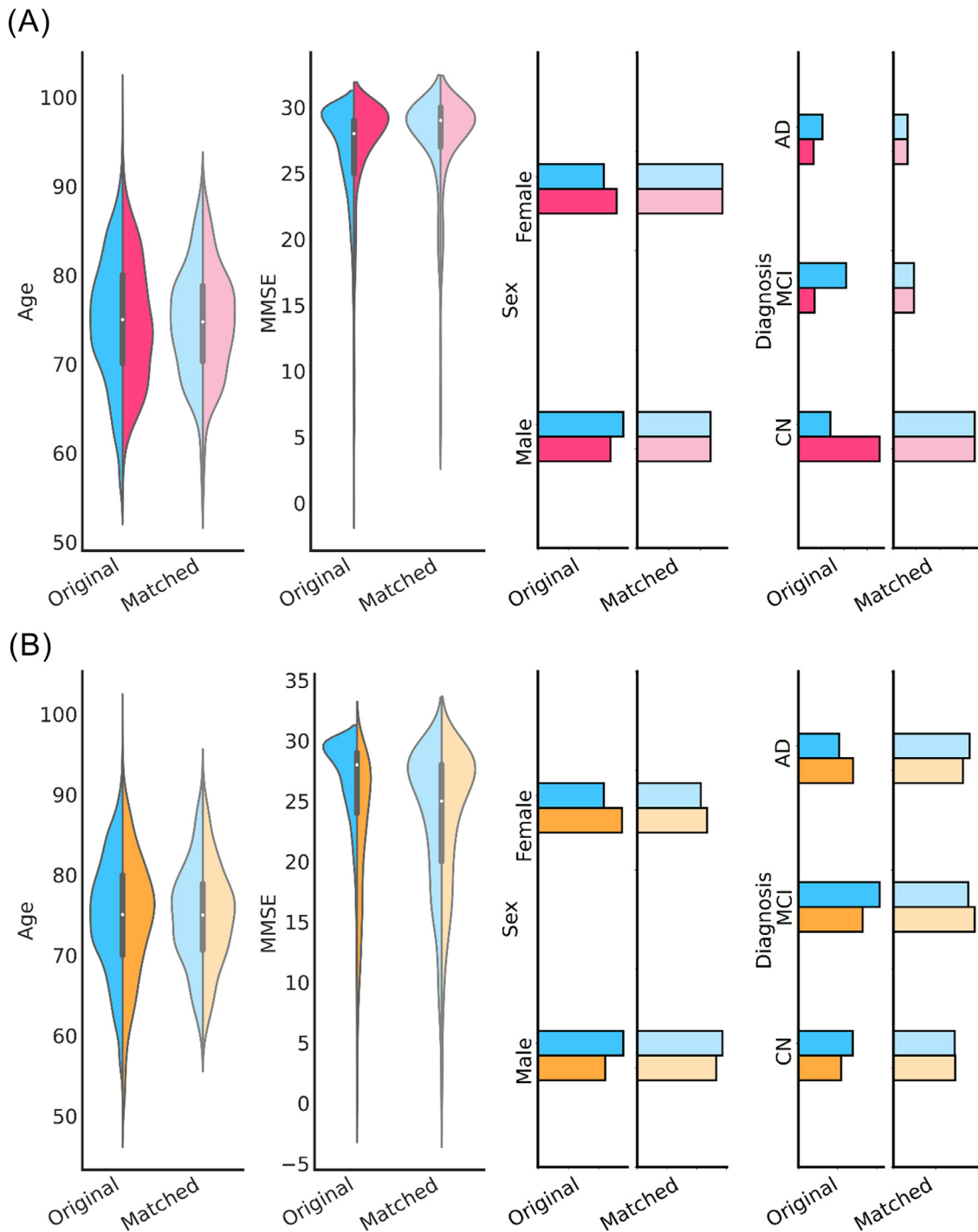
Fig. 3. *Age, MMSE, sex and clinical diagnosis distributions before and after matching.* (A) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and AIBL (red). Differences in the attributes between ADNI and AIBL were not significant after matching. (B) Distributions of age, sex, MMSE 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 S3–S9.

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 $+ \lambda_{DX}$ 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: $\frac{1}{2}$ 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 and Shoemaker, 2013; Ilievski et al., 2017; Eriksson, 2019) to find the best set of hyperparameters using the validation set (Table 1). The trained DNN after 100 epochs was utilized for subsequent analyses.

At the evaluation phase (Fig. 2C), the 10 goal-specific DNNs were applied to the harmonized brain volumes from the matched AIBL partic-

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.

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

ipants, as well as unharmonized brain volumes from the matched AIBL and ADNI participants. The prediction performance was averaged across all time points of each participant and the 10 goal-specific DNNs before averaging across participants. The same procedure was applied to ADNI and MACC participants.

2.6. Baseline harmonization models

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

2.6.1. 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:

$$x_{ijv} = \alpha_v + Y_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv}\epsilon_{ijv} \tag{1}$$

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 vth 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 (Fig. 1C). The estimated parameters can then be applied to a new participant i from site j with brain volume i from site x_{ijv} and covariates Y_{ij}

$$x_{ijv}^{ComBat} = \frac{x_{ijv} - \hat{a}_v - Y_{ij}^T \hat{\beta}_v - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + \hat{a}_v + Y_{ij}^T \hat{\beta}_v \tag{2}$$

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, this implied that MMSE and clinical diagnosis information were not available in the matched participants. Therefore, we could not utilize MMSE and clinical diagnosis as covariates in the ComBat model. Therefore, in the main results, we only utilized age and sex as covariates. However, as a control analysis (Section 2.9.3), we also considered a version of ComBat where age, sex, MMSE and clinical diagnosis were used as covariates.

Furthermore, since the goal-specific DNNs were trained with unmatched unharmonized ADNI data without distinguishing among the sites (Section 2.5), 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.

Note that Eq. (2) mapped both ADNI and AIBL data to an "intermediate" space, which is not an issue for the purpose of dataset prediction because the XGBoost classifier was trained from scratch (Fig. 2B;

(A) cVAE model structure
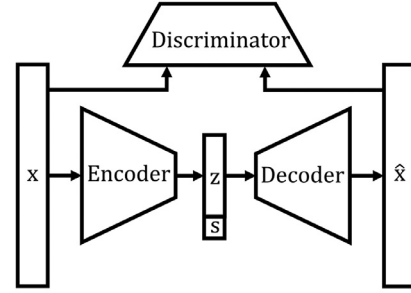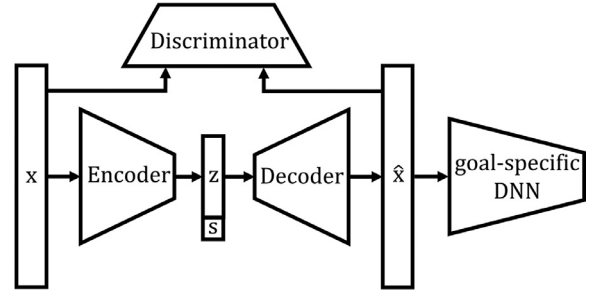

(B) gcVAE model structure

Fig. 4. *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 from Section 2.5 (Fig. 1B) was used to guide the cVAE harmonization process. During training of gcVAE, the weights of the goal-specific DNN were fixed.

Section 2.8). However, for the purpose of predicting MMSE and clinical diagnosis, since the goal-specific DNN was trained with unharmonized ADNI data, we used the ref.batch option in the ComBat package to map AIBL data to "ADNI-space" after harmonization. The same procedure was applied to ADNI and MACC.

2.6.2. 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 Fig. 4A. 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:

$$L_{cVAE} = L_{recon} + \alpha L_{prior} - \gamma L_{adv} + \lambda I(z, s) \tag{3}$$

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).

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

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.

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

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: $\frac{1}{2}$ MSE + Dataset Accuracy. The MSE 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 and Shoemaker, 2013; Ilievski et al., 2017; Eriksson, 2019) to find the best set of hyperparameters using the validation set (Table 2). The trained DNN after 1000 epochs was utilized for subsequent analyses.

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.

Similar to ComBat, for the purpose of dataset prediction, data were mapped to intermediate space by setting the site s to 0 during harmonization. On the other hand, for the purpose of predicting MMSE and clinical diagnosis, data from AIBL (and MACC) was mapped to ADNI space by setting the site s to correspond to ADNI.

2.7. Goal-specific cVAE (gcVAE)

To incorporate downstream application performance in the harmonization procedure, the outputs of the cVAE (Fig. 4A) were fed into the goal-specific DNN (Section 2.5). The resulting goal-specific cVAE (gcVAE) is illustrated in Fig. 4B. 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:

$$L_{gcVAE} = \alpha_{MMSE}MAE + \alpha_{DX}CrossEntropy \tag{4}$$

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 2.6.2) 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: $\frac{1}{2}$ MAE – Diagnosis Accuracy (same as Section 2.5). 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.

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

Similar to ComBat, for the purpose of dataset prediction, data were mapped to intermediate space by setting the site s to 0 during harmonization. On the other hand, for the purpose of predicting MMSE and clinical diagnosis, data from AIBL (and MACC) was mapped to ADNI space by setting the site s to correspond to ADNI.

2.8. Dataset prediction model

As one evaluation criterion, we utilized XGBoost to predict which dataset the harmonized brain volumes came from (Fig. 2B). 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. Recall that there were 10 groups of harmonized data because of our 10-fold cross-validation procedure (Section 2.4). Therefore, 10 XGBoost classifiers were trained using harmonized MRI volumes from unmatched ADNI and AIBL participants (Fig. 2B). For each XGBoost classifier, we used a grid search using the validation group to find the optimal set of hyperparameters.

For evaluation, the 10 XGBoost classifiers were applied to harmonized MRI volumes of matched ADNI and AIBL participants (Fig. 2B). The prediction accuracy was averaged across all time points of each participant and the 10 classifiers before averaging across participants. The same procedure was applied to ADNI and MACC participants.

Here, we chose to use XGBoost because it is a powerful classifier for unstructured or tabular data (Grinsztajn and Oyallon, 2022; Shwartz-Ziv and Armon, 2022). Using a DNN instead of XGBoost is unlikely to yield very different dataset prediction performance. On the other hand, XGBoost is less sensitive to the choice of hyperparameters compared with DNN, so hyperparameter tuning (and thus training) was a lot faster for XGBoost. Therefore, we chose to use XGBoost for dataset prediction. By contrast, a DNN was utilized for predicting MMSE and clinical diagnosis (i.e., goal-specific DNN), so that the gradients of the goal-specific DNN can be backpropagated to guide the training of the gcVAE model (Section 2.7).

2.9. Further analyses

We performed four additional analyses to study the effectiveness of the proposed gcVAE approach.

2.9.1. Effects of training set size
To investigate the effect of training set size on harmonization quality, we repeated the previous analyses (Figs. 1 and 2), except that when training harmonization models (Fig. 1C), the training set size was varied by sampling 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80 or 90% from the unmatched participants. We repeated this procedure 10 times.

2.9.2. Association analyses
We further investigated the association of the harmonized brain volumes with age, sex, MMSE and clinical diagnosis. We considered all 87 cortical and subcortical gray matter ROIs. For each continuous measure (age or MMSE) and for each ROI, we computed the Pearson's correlation between the harmonized ROI volume and the continuous measure across matched ADNI and matched AIBL participants. In the case of age, we expected a negative correlation between age and harmonized ROI volumes, so a stronger negative correlation indicates better harmonization. In the case of MMSE, we expect a positive correlation between MMSE and harmonized ROI volumes because lower MMSE indicates greater cognitive decline. Therefore, a greater positive correlation indicates better harmonization. For each discrete variable (clinical diagnosis or sex), we computed η^2 from running ANOVA on the matched ADNI and matched AIBL participants. Greater η^2 indicates greater differences across the groups (e.g., male versus female), suggesting better harmonization. The same procedure was applied to ADNI and MACC.

2.9.3. ComBat with additional covariates

As discussed previously, in our main analyses, we only used age and sex as covariates for the ComBat baseline (Section 2.6.1). Here, we also considered a ComBat variant, where age, sex, MMSE and clinical diagnosis were used as covariates. We note that this version of ComBat assumed that MMSE and clinical diagnosis information were known in the test set (matched participants). Therefore, the prediction performance of Com-Bat (with the additional covariates) was corrupted by test set leakage and was not valid.

2.9.4. Reversing the roles of the matched and unmatched participants

In the original analyses (Figs. 1 and 2), the harmonization models, goal-specific DNNs and dataset prediction models were trained on un-matched participants. The evaluations were then performed on matched participants (Fig. 2B and C).

In this analysis, we reversed the roles of the matched and unmatched participants (Figs. S1 and S2) with two exceptions. First, the prediction performance of unmatched unharmonized ADNI and unmatched un-harmonized AIBL participants was not comparable, so the downstream application performance was only evaluated on unmatched unharmo-nized AIBL and unmatched harmonized AIBL data (compare Fig. S2C and Fig. 2C).

Second, given the number of matched participants were so small, the training of the goal-specific DNN would not be effective. Therefore, the goal-specific DNN was trained with all (both matched and unmatched) ADNI participants (compare Fig. S1B and Fig. 1B). We note that this is not an issue since the downstream application performance no longer utilized any ADNI data (Fig. S2C).

2.10. 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 et al., 2014) with default PyTorch settings.

2.11. 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 2.8, 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 par-ticipant. Therefore, for each dataset, harmonization approach and eval-uation metric, we obtained a performance vector where each element represented one participant. When comparing dataset prediction per-formance (or goal-specific prediction performance) between two har-monization approaches, a permutation test with 10,000 permutations. Each permutation involves randomly swapping the entries between the performance vectors of the two approaches. Fig. S3 illustrates this per-mutation procedure in more details.

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

2.12. Data and code availability

Code for the various harmonization algorithms can be found here (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/predict_phenotypes/An2022_gcVAE). Two co-authors (PC and CZ) 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. cVAE & gcVAE removed more dataset differences than ComBat

Fig. 5A shows the dataset prediction performance for the matched ADNI and AIBL participants. Before harmonization, the XGBoost clas-sifier 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 predic-tion performance dropped to 0.595 ± 0.381 and 0.603 ± 0.382 respec-tively, 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-4), suggesting residual dataset differences.

Similar results were obtained for matched ADNI and MACC par-ticipants (Fig. 5B). Before harmonization, the XGBoost classifier was able to predict which dataset a participant came from with 100% ac-curacy. 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 differ-ences than ComBat. However, dataset prediction accuracies for cVAE and gcVAE were still better than chance (p = 1e-4), suggesting residual dataset differences.

3.2. gcVAE outperformed cVAE for clinical diagnosis prediction

Fig. 6A shows the clinical diagnosis prediction accuracies for matched ADNI and AIBL participants. Because the matched partici-pants had similar age, sex, MMSE and clinical diagnosis (Fig. 3), com-parison between unharmonized ADNI and unharmonized AIBL partic-ipants would indicate whether there was a drop in prediction perfor-mance due to dataset differences (e.g., scanner differences). Unexpect-edly, there was no statistical difference in clinical diagnosis prediction performance between unharmonized ADNI and unharmonized AIBL par-ticipants (Table 5).

Applying ComBat resulted in a statistically significant drop in predic-tion performance (p = 7e-4) compared with no harmonization. This sug-gests that ComBat removed biological information in addition to dataset differences (Fig. 5A). cVAE exhibited an even bigger drop in predic-tion performance compared with ComBat (p = 1e-4), suggesting that the better removal of dataset differences (Fig. 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).

Fig. 6B shows the clinical diagnosis prediction accuracies for matched ADNI and MACC participants. As expected, there was a sig-nificant drop in clinical diagnosis prediction performance between un-harmonized ADNI and unharmonized MACC participants (p = 1e-4). The decrease in clinical diagnosis performance was worsened by Com-Bat and cVAE, once again suggesting that the removal of dataset dif-ferences (Fig. 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 signifi-cantly worse than unharmonized ADNI, suggesting potential room for improvement.

3.3. gcVAE outperformed cVAE in MMSE prediction

Fig. 7A shows the MMSE prediction mean absolute error (MAE) for matched ADNI and AIBL participants. Because the matched participants
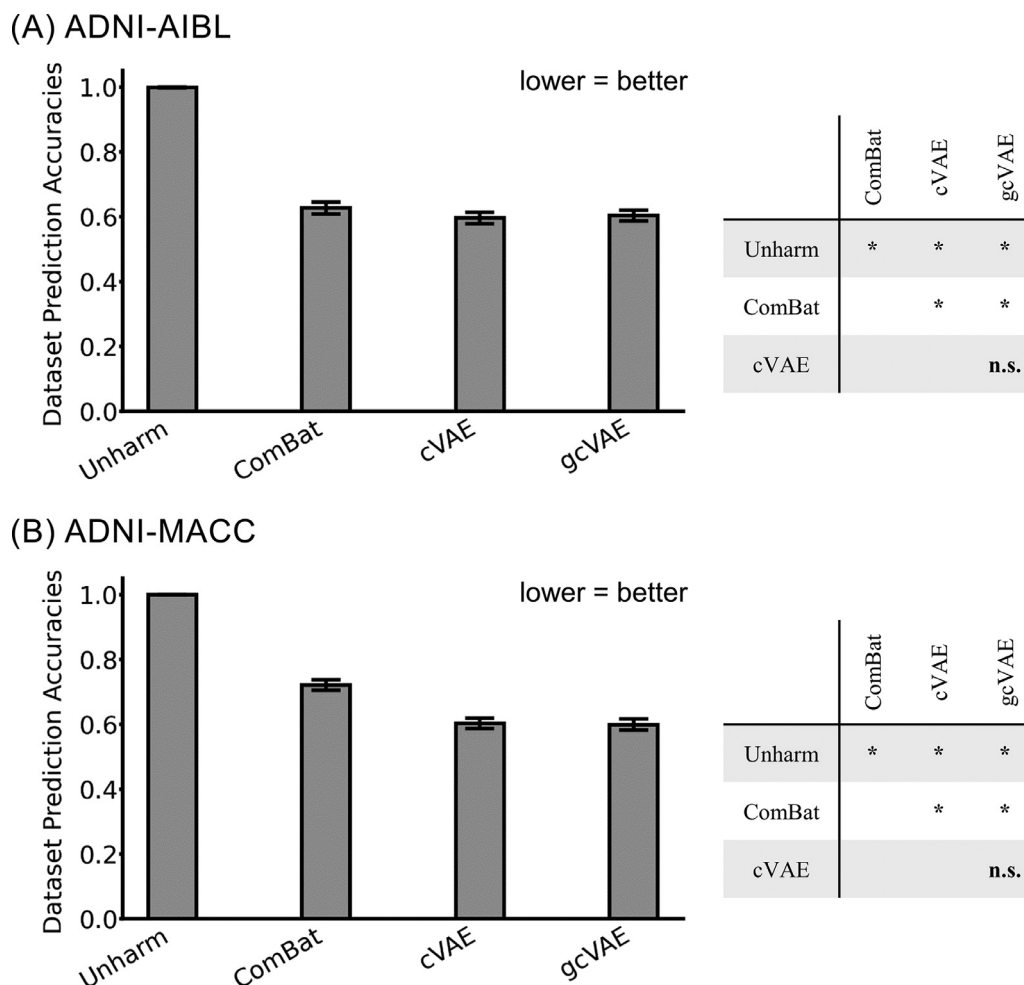
(A) ADNI-AIBL

(B) ADNI-MACC

Fig. 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.

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 (1.000 ± 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)				

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 (e.g., scanner differences). As expected, there was a drop in MMSE prediction performance (increased MAE) for unharmonized AIBL participants compared with unharmonized ADNI participants (p = 1e-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.

Fig. 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 statistically better than unharmonized MACC and not statistically different from unharmonized ADNI (Table 8).

3.4. Further analyses

3.4.1. Effects of training set size

We investigated the effects of varying the training set size used for fitting the harmonization models (Fig. 1C). Across all sample sizes, cVAE

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.

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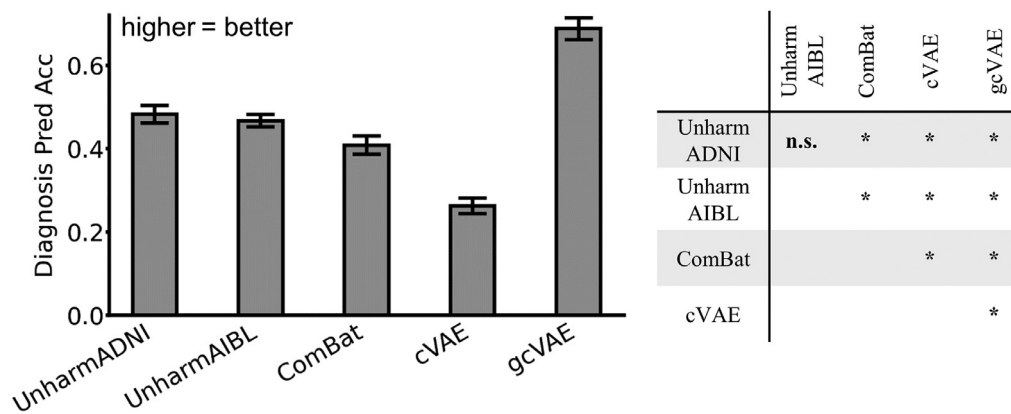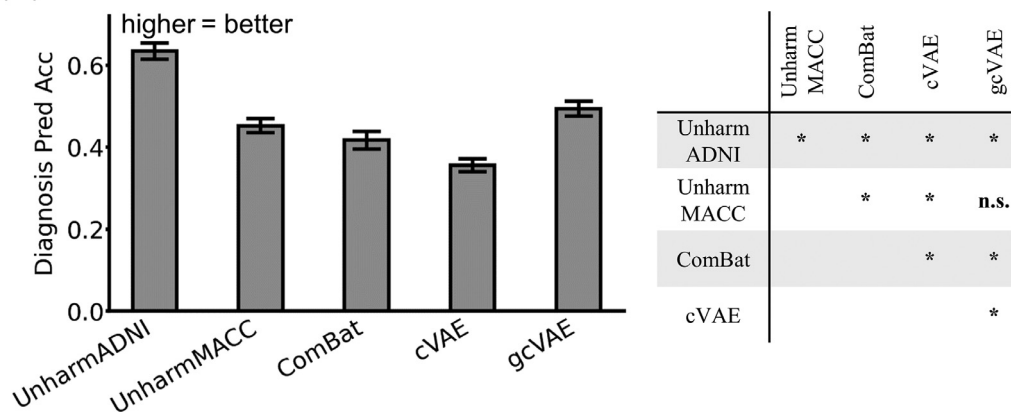
Fig. 6. *Clinical diagnosis prediction accuracies*. (A) Left: Clinical diagnosis prediction accuracies for matched ADNI and AIBL participants. Right: pvalues of differences between different approaches. "*" indicates statistical significance after surviving FDR correction (q < 0.05). "n.s." indicates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

and gcVAE generally achieved lower dataset prediction accuracies than ComBat (Fig. 8A). Across all sample sizes, gcVAE achieved better clinical diagnosis prediction than cVAE and ComBat (Fig. 8B). Across all sample sizes for MACC, gcVAE achieved better MMSE prediction than cVAE and ComBat (Fig. 8C2). Across all sample sizes for AIBL (Fig. 8C1), gcVAE achieved better MMSE prediction than cVAE; gcVAE achieved worse prediction than ComBat. Overall, across all sample sizes, gcVAE compared favorably with cVAE and ComBat.

In the case of ComBat, larger sample sizes led to worse dataset prediction accuracies (i.e., better harmonization). However, sample sizes have minimal effect on clinical diagnosis and MMSE prediction. In the case of cVAE, greater sample sizes led to better MMSE prediction for both AIBL and MACC participants, better clinical diagnosis prediction for MACC participants, worse clinical diagnosis prediction for AIBL participants, and better dataset prediction accuracies. In the case of gcVAE, greater sample sizes led to better MMSE and clinical diagnosis prediction for both AIBL and MACC participants, as well as better dataset predic-

tion accuracies. Overall, for both cVAE and gcVAE, larger sample sizes appeared to improve downstream application performance (i.e., MMSE and clinical diagnosis prediction), but at the expense of dataset prediction performance.

3.4.2. Association analyses

Fig. 9 shows the association analyses between gray matter ROI volumes and four variables (age, sex, MMSE and clinical diagnoseis) among matched ADNI and AIBL participants. Fig. 10 shows the same analyses for matched ADNI and MACC participants. For each scatter plot, more dots in the green region indicates better gcVAE performance compared with the baseline. gcVAE clearly outperformed no harmonization (Figs. 9A and 10A) and ComBat (Figs. 9B and 10B) in both datasets. On the other hand, cVAE and gcVAE exhibited comparable performance (Figs. 9C and 10C) in both datasets.

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 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 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.

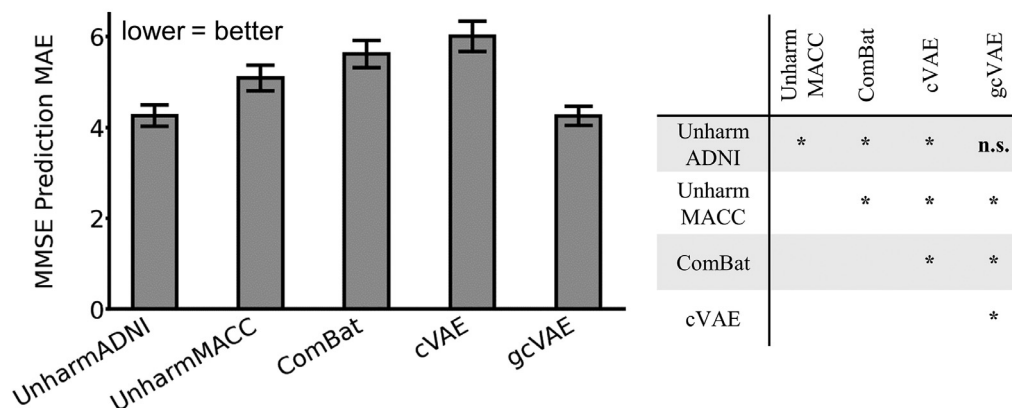| Clinical Diagnosis Prediction Accuracies (mean ± std)	p values				
	Unharm ADNI	Unharm MACC	ComBat	cVAE	gcVAE
Unharm ADNI (0.63 ± 0.33)		**1e-4**	**1e-4**	**1e-4**	**1e-4**
Unharm MACC (0.45 ± 0.29)			**0.0124**	**1e-4**	0.0545
ComBat (0.42 ± 0.35)				**2e-4**	**0.0065**
cVAE (0.36 ± 0.26)					**1e-4**
gcVAE (0.49 ± 0.30)					
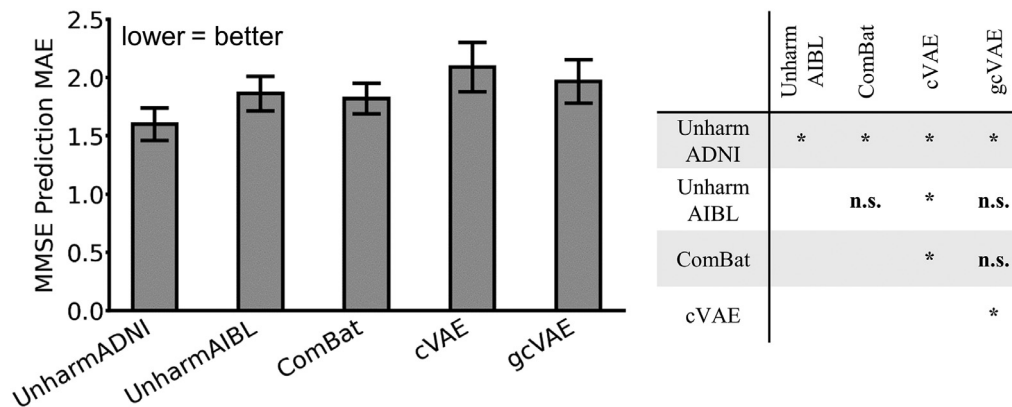
Fig. 7. *MMSE prediction errors as measured by mean absolute error (MAE).* (A) Left: MMSE prediction errors for matched ADNI and AIBL participants. Right: p values of differences between different approaches. "*" indicates statistical significance after surviving FDR correction (q < 0.05). "n.s." indicates not significant. (B) Same as (A) but for matched ADNI and MACC participants. All p values are reported in Tables 7 and 8.

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 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 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.

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 (4.25 ± 3.57)					

3.4.3. ComBat with additional covariates

Our main analysis utilized ComBat with age and sex as covariates. Here, we considered ComBat with age, sex, MMSE and clinical diagnosis as covariates. We note that this version of ComBat assumed that MMSE and clinical diagnosis information were known in the test set (matched participants). Therefore, the prediction performance of ComBat (with the additional covariates) was corrupted by test set leakage and was not valid.

The additional covariates led to better MMSE and clinical diagnosis prediction by ComBat (Tables S10 and S11). In the case of AIBL, clinical diagnosis prediction remained statistically worse than gcVAE, but MMSE prediction was now statistically better than gcVAE. In the case of MACC, clinical diagnosis prediction was now comparable with gcVAE, but MMSE prediction remained worse than gcVAE. Interestingly, the additional covariates led to greater dataset prediction accuracies for both ADNI-AIBL and ADNI-MACC, suggesting worse harmonization. Together, gcVAE remained better than ComBat.

3.4.4. Reversing the roles of the matched and unmatched participants

In this analysis, we reversed the roles of matched and unmatched participants (Section 2.9.4). Similar to the original main analyses, we found that gcVAE compared favorably with both ComBat and cVAE (Figs. S4–S6; Tables S12–S17).

More specifically, recall that there were six evaluation metrics (two for dataset prediction, two for diagnosis prediction and two for MMSE prediction). gcVAE was statistically better than ComBat for both dataset prediction metrics and two downstream application performance metrics, while being statistically worse than ComBat in one downstream application performance metric (Figs. S4–S6; Tables S12–S17). On the other hand, gcVAE was statistically worse than cVAE for the two dataset prediction metrics, but statistically better than cVAE for the four downstream application performance metrics (Figs. S4–S6; Tables S12–S17). Therefore, similar to the main results, cVAE removed more dataset differences at the expense of removing more biological information.

One interesting deviation from the main results was that in the current setup (where harmonization models were trained on matched participants), ComBat was statistically better than no harmonization across all six evaluation metrics. On the other hand, in the main analysis (Figs. 5–7; Tables 3–8), ComBat was statistically better than no harmonization for both dataset prediction metrics, but statistically worse than no har-

monization for all four downstream application performance metrics. On the other hand, for the main analysis, gcVAE was statistically better than no harmonization for both dataset prediction metrics and two downstream application performance metrics. In the current analysis, gcVAE was statistically better than no harmonization for both dataset prediction metrics and three downstream application performance metrics, but was statistically worse for one application performance metric. Therefore, gcVAE appeared more robust than ComBat to covariate differences during the harmonization procedure.

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 (Fig. 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 (Figs. 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 gcVAE models were able to yield better prediction of MMSE and clinical diagnosis (Figs. 6 and 7), while removing as much dataset differences as cVAE (Fig. 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.

Our main analyses (Figs. 6 and 7) showed that gcVAE facilitated the translation of goal-specific DNNs from ADNI to new datasets (AIBL and
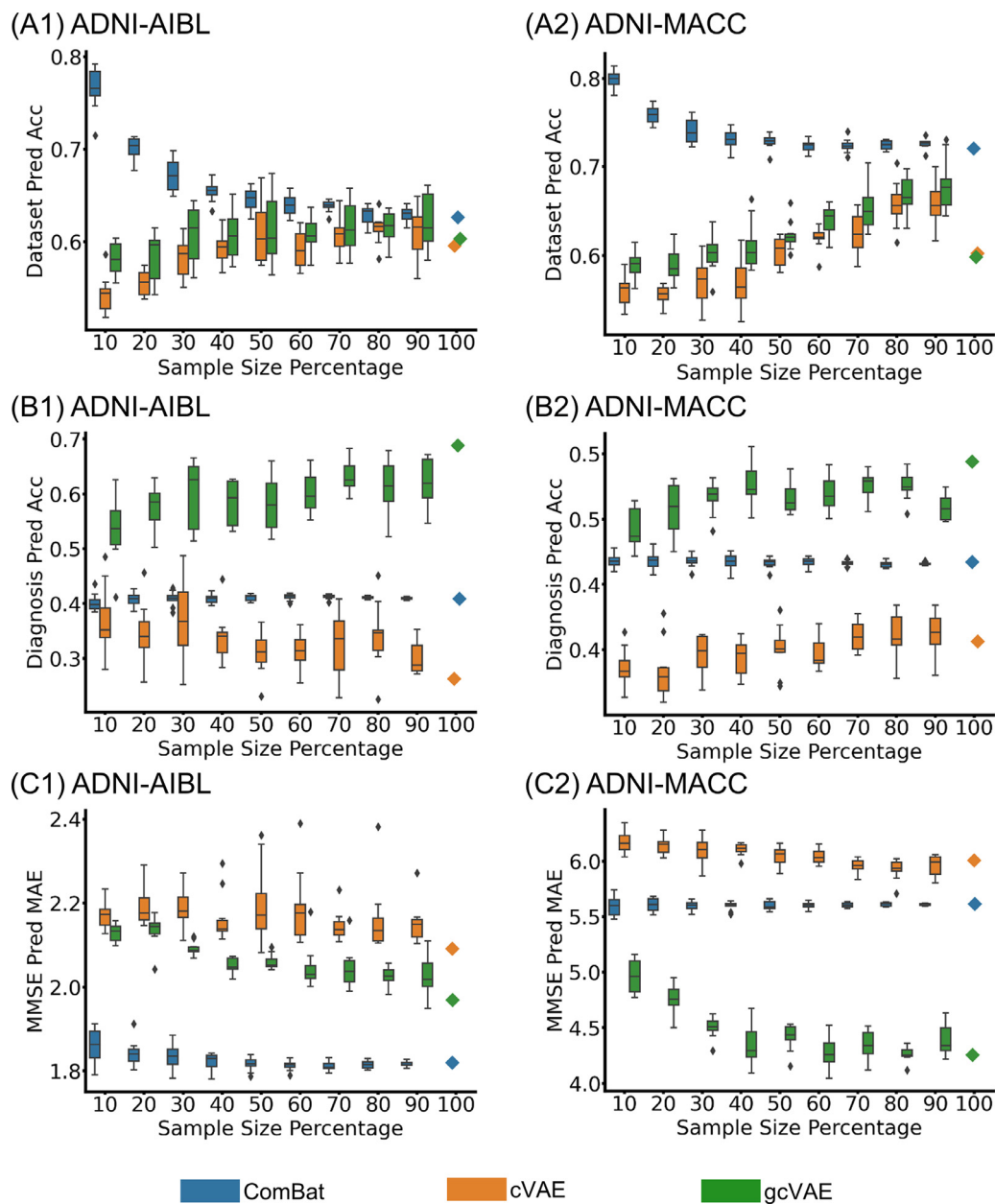
Fig. 8. *Performance of harmonization models trained with different sample sizes.* (A1) Dataset prediction accuracies for matched ADNI-AIBL participants; (B1) The clinical diagnosis prediction accuracies for matched AIBL participants; (C1) MMSE prediction errors for matched AIBL participants. (A2), (B2), and (C2) are the same as (A1), (B1), and (C1), but for matched ADNI and MACC participants.

MACC). Another common application of harmonization is to facilitate the pooling of datasets for some joint analysis. Here, we investigated the association of the brain volumes with multiple variables across the harmonized datasets. We found that gcVAE clearly outperformed no harmonization (Figs. 9A and 10A) and ComBat (Figs. 9B and 10B). On the other hand, cVAE and gcVAE exhibited comparable performance (Figs. 9C and 10C).

4.1. Matched versus unmatched participants

We note that our workflow utilized unmatched participants to train the harmonization models, dataset prediction models and goal-specific DNNs, while evaluation was performed on the matched participants (Figs. 1 and 2). The setup allowed us to compare downstream application performance between unharmonized data from matched ADNI participants and matched AIBL participants. Because age, sex, MMSE

and clinical diagnosis were similar between matched ADNI and AIBL participants, the drop in downstream application performance (clinical diagnosis or MMSE prediction) could be attributed to a lack of harmonization. Since the goal-specific DNNs were trained on ADNI (Fig. 2B), the prediction performance on matched unharmonized ADNI participants served as an upper bound on the prediction performance after harmonization.

Surprisingly, in the case of clinical diagnosis prediction in the AIBL dataset, gcVAE was better than the upper bound (Fig. 6A). On the other hand, in the case of MMSE prediction in the AIBL dataset, gcVAE only achieved similar performance as no harmonization and was worse than the upper bound (Fig. 7A). One possible reason for this discrepancy is that when tuning the hyperparameters, the weights tradeoff the prediction of MMSE and clinical diagnosis were fixed, so in the case of AIBL, this might have inadvertently favored clinical diagnosis prediction more than MMSE prediction.
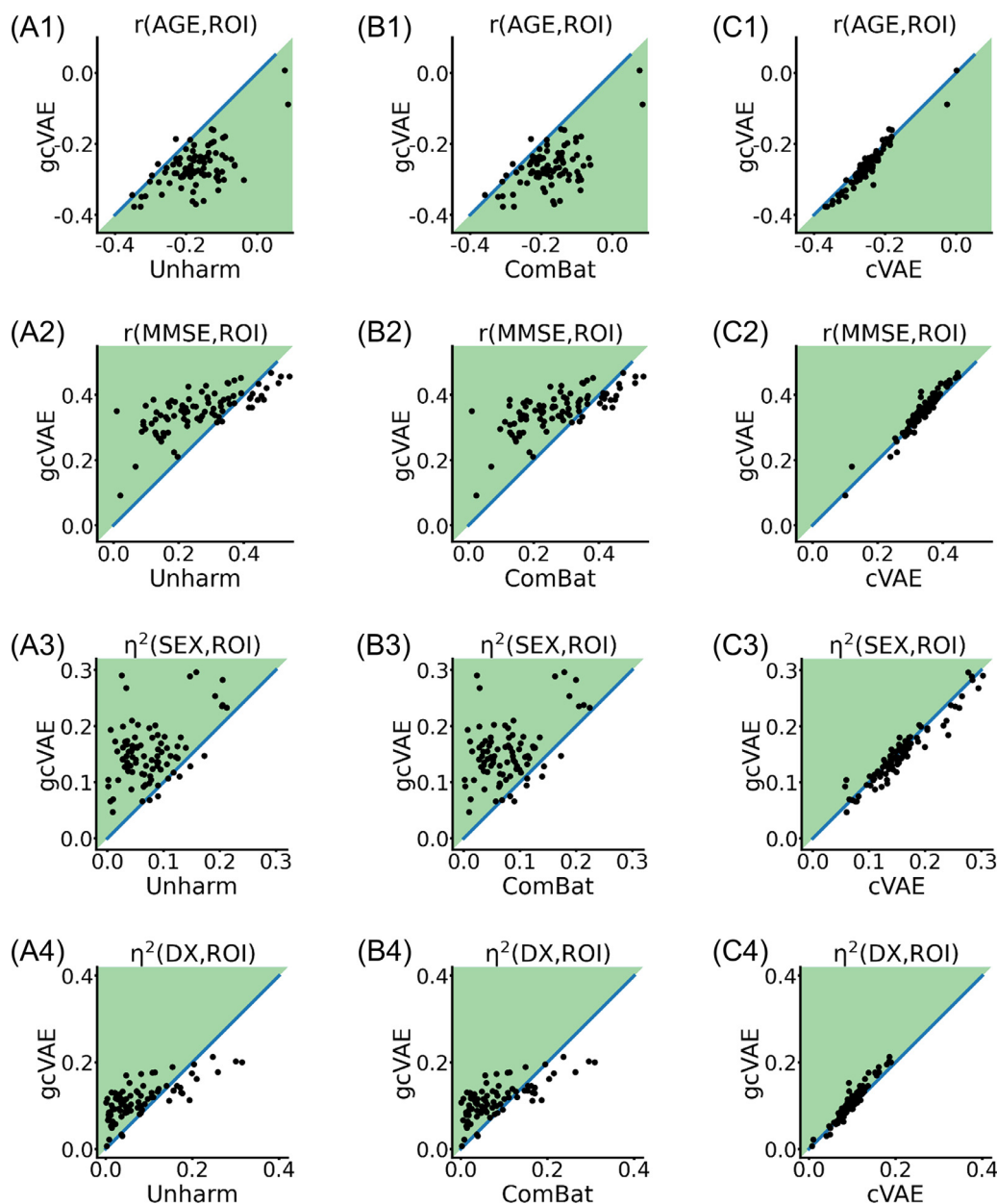
Fig. 9. Association analyses between gray matter ROI volumes and four variables (age, MMSE, sex and clinical diagnosis) for matched ADNI and AIBL participants. First row shows association with age. Second row shows association with MMSE. Third row shows association with sex. Fourth row shows association with clinical diagnosis. (A) Comparison between gcVAE and no harmonization. (B) Comparison between gcVAE and ComBat. (C) Comparison between gcVAE and cVAE. Each black dot represents one gray matter ROI. Dots in the green area indicates better gcVAE performance compared with baseline. gcVAE clearly outperforms no harmonization and ComBat. gcVAE and cVAE exhibited similar performance.

However, we note that the current workflow of training on unmatched participants can prove challenging for ComBat (Nygaard et al., 2016; Zindler et al., 2020) because of covariate differences between ADNI and AIBL (as well ADNI and MACC). Therefore, we considered a control analysis in which the roles of the matched and unmatched participants were swapped. Consistent with the main analyses, we found that gcVAE compared favorably with both ComBat and cVAE (Figs. S4–S6; Tables S12–S17). Furthermore, in the control analysis, ComBat was better than no harmonization for both dataset prediction and downstream application performance. On the other hand, in the control analysis, gcVAE was statistically better than no harmonization for both dataset prediction metrics and three downstream application performance metrics, but was statistically worse for one application performance metric. Overall, this suggests that gcVAE was more robust than ComBat to co-

variate differences between datasets used for the harmonization procedure.

4.2. Sample size

Deep neural networks are often thought to be data hungry. Across different sample sizes (Fig. 8), gcVAE was better than cVAE for all four downstream application performance. On the other hand, across all sample sizes, gcVAE was better than ComBat for three downstream prediction metrics. Interestingly, gcVAE was worse than ComBat for MMSE prediction in the AIBL dataset across all sample sizes but given the rapid improvement trajectory of gcVAE (Fig. 8C1), we might expect the gap to close rapidly with more data. Surprisingly, as the sample size increases, the downstream performance of gcVAE improved at the expense
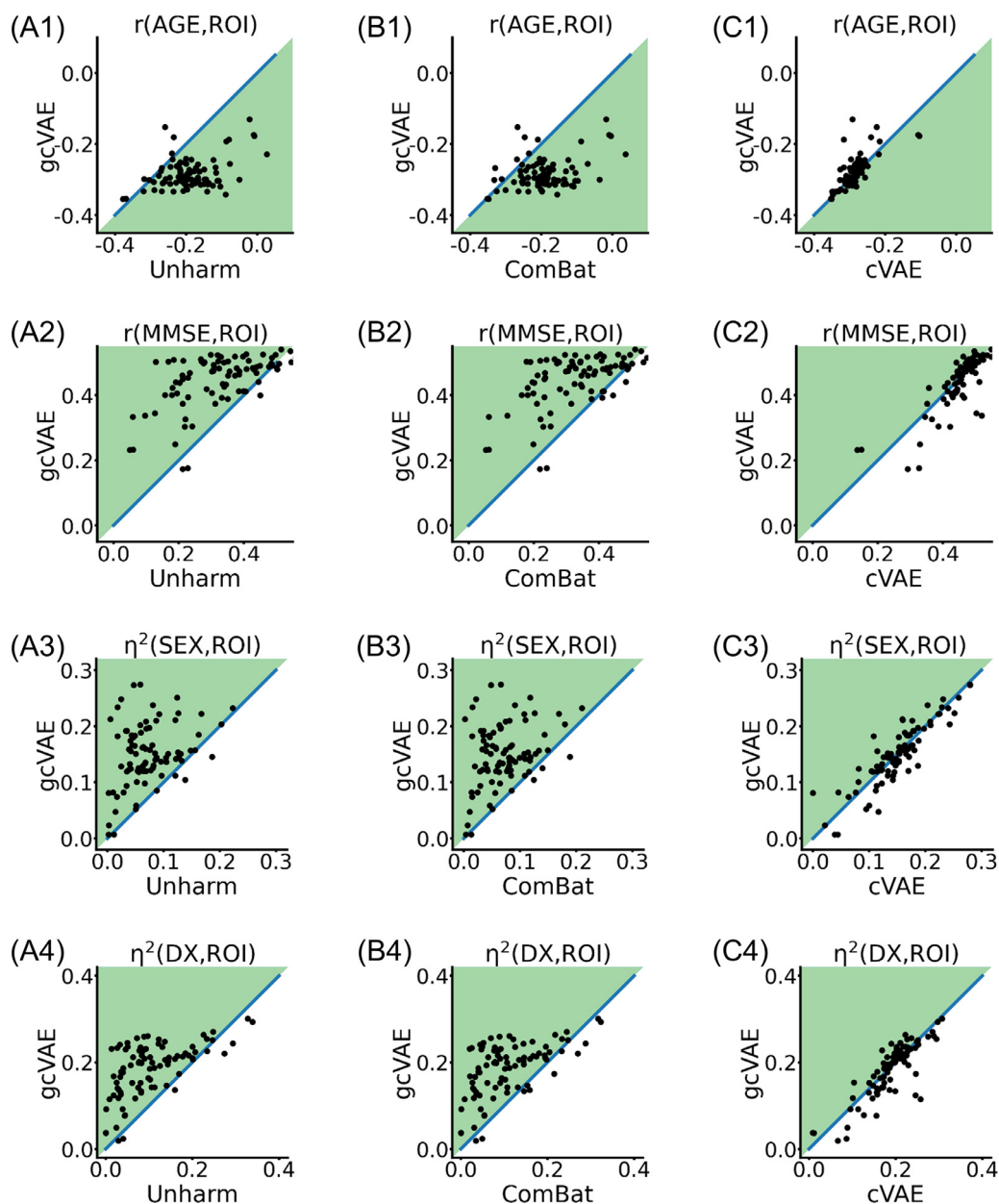
Fig. 10. Association analyses between gray matter ROI volumes and four variables (age, MMSE, sex and clinical diagnosis) for matched ADNI and MACC participants. First row shows association with age. Second row shows association with MMSE. Third row shows association with sex. Fourth row shows association with clinical diagnosis. (A) Comparison between gcVAE and no harmonization. (B) Comparison between gcVAE and ComBat. (C) Comparison between gcVAE and cVAE. Each black dot represents one gray matter ROI. Dots in the green area indicates better gcVAE performance compared with baseline. gcVAE clearly outperforms no harmonization and ComBat. gcVAE and cVAE exhibited similar performance.

of dataset prediction performance. However, the dataset prediction accuracies of gcVAE continued to be worse (i.e., better harmonization) than ComBat even with the full set of data (Fig. 8A).

4.3. Methodological considerations

To illustrate the use of gcVAE, when harmonizing ADNI and a new dataset, the researcher could validate gcVAE by repeating the same procedure as the current study (Figs. 2 and 3). Once the researcher is satisfied with the performance, the researcher could then train the model on 90% of the data and tune the hyperparameters on the remaining 10% of the data without the need of a 10-fold cross-validation procedure. The final model can then be applied to the full dataset.

An interesting methodological consideration is the handling of confound variables when using gcVAE. For example, age is likely related to clinical diagnosis. Therefore, when training gcVAE to harmonize ADNI and AIBL, the algorithm might seek to preserve age-related brain patterns related to clinical diagnosis. However, we note that this may or may not be an issue depending on the study. For example, if our goal is clinical diagnosis, then it would be counterproductive to exclude age in the diagnosis procedure. After all, demographics are often used for differential diagnosis in actual clinical practice.

There might indeed be situations, where the related variables are indeed confounds. For example, if a study is interested in dementia risks above and beyond aging, then age does become a confound. In that scenario, researchers could consider regressing age from the imaging features and/or target variables before training the goal-specific DNN.

Another approach is to include an adversarial cost when training the goal-specific DNN to ensure the intermediate layers could not be used to predict the confound variable (e.g., age).

The theoretical advantage of gcVAE over ComBat is its multivariate nature, which allowed cVAE to remove site differences distributed across brain regions. This advantage is clearly demonstrated in the dataset prediction experiments (Fig. 5). More recent ComBat variants, such as CovBat (Chen et al., 2019) allowed the harmonization of interregional covariance. Given their multivariate nature, cVAE and gcVAE should also in principle remove site variation in covariance.

Finally, our current study only demonstrated results from harmonizing pairs of datasets (ADNI and AIBL, as well as ADNI and MACC). However, the cVAE framework is highly flexible and the cVAE machinery can be easily extended to multiple datasets. Similarly, the goal-specific DNN could also be trained on multiple datasets. So overall, gcVAE could in principle be applied to harmonize multiple datasets jointly. However, this is not something we have demonstrated in this study, which we leave for future work.

4.4. Limitations

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.

Data Availability

The authors do not have permission to share data.

Credit authorship contribution statement

Lijun An: Conceptualization, Methodology, Software, Validation, Project administration, Writing – original draft. **Jianzhong Chen:** Methodology, Writing – review & editing. **Pansheng Chen:** Software, Validation. **Chen Zhang:** Software, Validation. **Tong He:** Methodology, Writing – review & editing. **Christopher Chen:** Conceptualization, Writing – review & editing. **Juan Helen Zhou:** Conceptualization, Writing – review & editing. **B.T. Thomas Yeo:** Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition.

Acknowledgments

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119570.

References

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. J. Magn. Reson. Imaging 27908 jmridoi: 10/gmzt7m.

Blumberg, S. B., Palombo, M., Khoo, C. S., Tax, C. M., Tanno, R., Alexander, D. C., 2019. October). Multi-stage prediction networks for data harmonization. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, pp. 411–419.

Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y., 2018. Recurrent neural networks for multivariate time series with missing values. Sci. Rep. 8 (1), 6085 doi: 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 doi: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. J. Neurosci. Methods 230, 37–50 doi: 10/f56f36.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 doi: 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 doi: 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 doi: 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. Magn. Reson. Imaging 64, 160–170 doi: 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. Int. Psychogeriatr. 21 (4), 672–687 doi: 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 Dement. 6 (3), 291–296 doi: 10/btmfdj.

Eriksson, David, et al., 2019. Scalable global optimization via local bayesian optimization. Advances in neural information processing systems 32.

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 doi: 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. doi: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 doi: 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. doi: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 Res. Neuroimaging 282, 134–142 doi: 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 (Baltimore)doi: 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. J. Cereb. Blood Flow Metab. 40 (1), 44–53 doi: 10/gpfttm.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on tabular data?." *arXiv preprint arXiv:2207.08815* (2022).

Ilievski, I., Akhtar, T., Feng, J., & Shoemaker, C. (2017, February). Efficient hyperparameter optimization for deep learning algorithms using deterministic rbf surrogates. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

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 Dement. 6 (3), 212–220 doi: 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. J. Magn. Reson. Imaging 27 (4), 685–691 doi: 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 doi: 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 doi: 10/c2hrm5.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Naval Res. Logist. Q. 2 (1–2), 83–97 doi: 10/b2k5tg.

Lipton, Z. C., Kale, D. C., Wetzel, R., 2016. Modeling missing data in clinical time series with rnns. Machine Learning for Healthcare 56, 253–270.

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 doi:10.1101/2020.08.18.256594.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the 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. Multi center reliability of diffusion tensor imaging. Brain Connect. 2 (6), 345–355 doi: 10/gk82p6.

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. Nat. Neurosci. 19 (11), 1523–1536 doi: 10/f89khc.

Modanwal, G., Vellal, A., Buda, M., Mazurowski, M.A, Hahn, H.K., Mazurowski, M.A., 2020. MRI image harmonization using cycle-consistent generative adversarial network. In: Medical Imaging 2020: Computer-Aided Diagnosis. SPIE, p. 36 doi: 10/gmzt6h.

Moyer, D., Ver Steeg, G., Tax, C.M.W., Thompson, P.M., 2020. Scanner invariant representations for diffusion MRI harmonization. Magn. Reson. Med. 84 (4), 2174–2189. doi: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. doi: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., Bonet-Carne, E., Grussu, F., Ning, L., Sepehrband, F., Tax, C.M.W., 2019. Muti-shell diffusion MRI harmonisation and enhancement challenge (MUSHAC): progress and results. In: Computational Diffusion MRI. Springer International Publishing, pp. 217–224. doi:10.1007/978-3-030-05831-9_18.

Nygaard, V., Rødland, E.A., Hovig, E., 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17 (1), 29–39. doi:10.1093/biostatistics/kxv027.

Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... & Lerer, A. (2017). Automatic differentiation in pytorch.

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. doi: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. Eng. Optim. 45 (5), 529–555 doi: 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. doi:10.1093/bioinformatics/btaa624.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: deep learning is not all you need. Inf. Fusion 81, 84–90. doi:10.1016/j.inffus.2021.11.011.

Sohn, K., Lee, H., Yan, X., 2015. Learning structured output representation using deep conditional generative models. Adv. Neural Inf. Process. Syst. 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. Biol. Psychiatry 87 (12), 1071–1082 doi: 10/gpfkpx.

Tanno, R., Worrall, D. E., Ghosh, A., Kaden, E., Sotiropoulos, S. N., Criminisi, A., Alexander, D. C., 2017, September. Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, pp. 611–619.

Yeo, B.T.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. J. Neurophysiol. 106 (3), 1125–1165 doi: 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 doi: 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 doi: 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. Dev. Cogn. Neurosci. 32, 4–7 doi: 10/gd8ctx.

Wachinger, C., Rieckmann, A., Pölsterl, S., 2021. Detect and correct bias in multi-site neuroimaging datasets. Med. Image Anal. 67, 101879 doi: 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 doi: 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. Hum. Brain Mapp. 39 (11), 4213–4227 doi: 10/gff7m4.

Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Shen, D., Li, G., Zhou, S., Yap, P.T., Khan, A.The UNC/UMN Baby Connectome Project Consortium, 2019. Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, 11767. Springer International Publishing, pp. 475–483. doi:10.1007/978-3-030-32251-9_52 Vol.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. doi:10.1109/ICCV.2017.244.

Zindler, T., Frieling, H., Neyazi, A., Bleich, S., Friedel, E., 2020. Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. BMC Bioinform. 21 (1), 271. doi:10.1186/s12859-020-03559-6.

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 doi: 10/gmzt6k.