Evaluating MRI Harmonization via Multiple Clinical Prediction Tasks[1*]

Tamoghna Chattopadhyay[1], Gautam Mehendale[1], Rahul H. Ankarath[1], Sophia I. Thomopoulos[1], Himanshu Joshi[2], Ganesan Venkatasubramanian[3], John P. John[2], Jose Luis Ambite[4], Greg Ver Steeg[5] and Paul M. Thompson[1]

[1]Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine,University of Southern California, Marina del Rey, CA, United States, [2]Multimodal Brain Image Analysis Laboratory, [3]Translational Psychiatry Laboratory, National Institute of Mental Health and Neuro Sciences (NIMHANS), Bengaluru, Karnataka, India, [4]Information Sciences Institute, University of Southern California, Marina del Rey, CA, United States, [5]University of California, Riverside, CA, United States
Email: tchattop@usc.edu, pthomp@usc.edu

Harmonization has become a cornerstone of modern neuroimaging studies as multi-site MRI datasets are increasingly combined to improve statistical power and generalizability. Even so, it is crucial to assess whether image harmonization boosts the performance of deep learning methods trained on brain MRI, yet most existing evaluations rely primarily on image-level similarity rather than benchmarking performance on biomedical tasks. Here, we compare two widely used harmonization approaches, deep learning-based CycleGAN and statistical ComBat from the dPABI toolbox, using four large neuroimaging datasets (ADNI, NACC, OASIS-3, and NIMHANS). To study how disease information affects harmonization, we trained two variants of each method: one using only cognitively normal (CN) subjects and another using multiple diagnostic groups (CN, mild cognitive impairment, and Alzheimer's disease). Beyond standard similarity metrics, we evaluated harmonization through a battery of downstream biomedical prediction tasks using convolutional neural networks, including sex classification, brain age estimation, Alzheimer's disease classification, and sobCDR prediction, in both zero-shot and pooled-training settings. This multi-task experimental design allows us to disentangle how harmonization strategies, training populations, and dataset composition jointly influence not only visual similarity but also performance on key medical imaging benchmark tasks. We find that harmonization is neither uniformly beneficial nor harmful: its effect is task- and training-dependent. Harmonization-aware training enhances performance for disease-related tasks, particularly when harmonization models are trained on all diagnostic groups. In contrast, applying harmonization only at test time can degrade performance due to representation shifts between the training and inference distributions. In contrast, tasks driven by global anatomical features, such as sex classification, are relatively robust to harmonization.

Keywords: Harmonization, Deep Learning, Cycle GAN, COMBAT, dPABI Toolbox

[1*] This work was supported by NIH grants U01AG068057, R01AG081571, S10OD032285 and U24AG074855.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Large-scale neuroimaging studies increasingly aggregate MRI data across scanners, sites, and protocols [1] to improve statistical power, enable machine learning, and support translational neuroscience; however, pooling multisite data comes with the cost of substantial non-biological variability introduced by hardware, acquisition, and preprocessing differences. This scanner- and site-related variability can dominate or confound disease-related effects, leading to biased models, reduced reproducibility, and poor generalization across cohorts [2,3,4]. Harmonization aims to model or mitigate these unwanted sources of variation while preserving biologically meaningful signals by mapping images or derived features into a common representational space [5]. In recent years, harmonization has taken center stage in neuroimaging research as more datasets become publicly available. In dementia research, for example, widely-used public datasets include ADNI [6], NACC [7], and OASIS [8], yet deep learning models trained on these datasets can be highly sensitive to domain shifts. Broadly, harmonization can be performed either at the feature level using statistical adjustment models or at the image level using generative models, with the shared goal of reducing site effects that are irrelevant to the task, while retaining disease-relevant structure. A wide range of harmonization strategies has been proposed. Statistical approaches such as the ComBat family of methods [9] use empirical Bayes modeling to remove scanner effects from imaging features while preserving effects of biological covariates, and have been applied in hundreds of studies of structural MRI, diffusion MRI, and cortical thickness measures [10,11]. Extensions of ComBat, including longitudinal and generalized additive model variants (ComBat-GAM), further improve flexibility in modeling nonlinear covariate effects [12,13]. The NeuroHarmonize toolbox [12] has made these methods accessible and is widely applied to derived numeric features in multi-site neuroimaging studies. In a separate line of work, deep learning-based image-level harmonization has gained traction, with CycleGANs proposed to learn unpaired mappings between scanner domains while preserving anatomical structure [14-16]. Subsequent work has adapted CycleGANs and related adversarial frameworks for structural MRI, diffusion MRI, and multi-contrast harmonization, demonstrating improved visual consistency across sites [17]. Variational autoencoders and disentangled representation learning have also been developed to separate site-related and biological factors, enabling harmonization through latent-space manipulation [18,19]. Domain adaptation and contrastive learning–based approaches further attempt to learn scanner-invariant representations for downstream prediction tasks [20,21]. More recent approaches have explored diffusion-based translation models [22], latent-space disentanglement strategies [23], and pan-contrast [24] or multi-domain harmonization frameworks that aim to more explicitly separate scanner-related and biological variation. Despite this growing

body of work, most harmonization models are trained exclusively on data from healthy (cognitively normal) subjects to avoid disease leakage (introduction of disease-related features into scans), and evaluation often prioritizes image similarity metrics or narrow downstream tasks such as brain age prediction. Ideally, we would test the effect of different harmonization approaches on multiple biomedical tasks to assess when and where the benefits are most evident, as it is unclear whether using harmonized images necessarily leads to better task performance.

In this work, we address these limitations through a comprehensive evaluation of MRI harmonization that explicitly examines how harmonization strategy, training population, and dataset composition affect task performance on a range of common but challenging biomedical tasks in dementia research. We systematically compare a deep learning-based CycleGAN framework and the widely used ComBat method from the dPABI toolbox [25] in a study that pools multisite data from four large neuroimaging cohorts: ADNI, NACC, OASIS-3, and NIMHANS [26-28]. We investigate two training regimes for each method, one using only cognitively healthy subjects and another incorporating all diagnostic groups, to quantify the effects of disease-aware versus disease-agnostic harmonization. We move beyond visual similarity by evaluating the benefits of using harmonized data when training convolutional neural networks to perform four common benchmark tasks: sex classification, brain age estimation, Dementia classification, and prediction of dementia severity (using the widely-used 'sum-of-boxes' clinical dementia rating, or sobCDR [29], under both zero-shot and pooled-training settings. Together, these experiments provide a comprehensive assessment of MRI harmonization across a range of widely-used tasks. We demonstrate that improvements in image similarity do not necessarily translate to enhanced clinical inference, highlighting the need for task-aware evaluation frameworks in multi-site neuroimaging.

2. Data and Preprocessing

To evaluate MRI harmonization and its impact on downstream biomedical prediction tasks, we used data from four independent cohorts, including three large widely-used public datasets and one external clinical dataset. These included (1) the Alzheimer's Disease Neuroimaging Initiative (ADNI 2/GO/3; 1,754 participants: 74.37 ± 7.28 years, 795 F / 960 M; 588 cognitively normal (CN), 836 mild cognitive impairment (MCI), 330 dementia), (2) the Open Access Series of Imaging Studies (OASIS-3; 1,040 participants: 70.49 ± 9.46 years, 587 F / 467 M; 751 CN, 289 dementia), and (3) the National Alzheimer's Coordinating Center (NACC; 3,856 participants: 71.48 ± 10.61 years, 2,243 F / 1,613 M; 2,073 CN, 672 MCI, 1,111 dementia). In addition, we included an external dataset from NIMHANS, India (301 participants: 67.22 ± 7.86 years, 124 F / 177 M; 123 CN, 88 MCI, 90 dementia) to assess out-of-distribution generalization. ADNI was selected as the target reference domain for harmonization, and all other datasets were mapped to

match ADNI intensity and contrast characteristics. This choice was motivated by ADNI's central role in dementia research and its standardized acquisition protocol, which has informed and influenced subsequent large-scale national initiatives, including SCAN and CLARITI. All subjects underwent standardized preprocessing [30] of 3D T1-weighted brain MRI volumes prior to harmonization and model training. Images were first corrected for intensity non-uniformity using N4 bias field correction, followed by skull stripping to isolate brain tissue. Each scan was then rigidly registered to a common in-house anatomical template using a 6-degree-of-freedom transformation and resampled to isotropic 2 mm resolution, resulting in uniform image dimensions of 91 × 109 × 91 voxels. Finally, voxel intensities were normalized to the [0,1] range using min–max scaling to reduce inter-subject intensity variability and facilitate stable training of downstream deep learning models. This preprocessing pipeline ensured consistent spatial alignment and intensity characteristics across cohorts while preserving anatomical structure relevant to clinical prediction tasks.

For harmonization model development and downstream evaluation, we adopted two complementary training paradigms to study the impact of disease-aware versus disease-agnostic harmonization. In the CN-only setting, harmonization models were trained exclusively using cognitively normal subjects from ADNI and the corresponding source datasets (NACC and OASIS-3), while all non-ADNI datasets were held out for testing to enable zero-shot evaluation. In the ALL-subject setting, harmonization models were trained using subjects spanning all diagnostic categories (CN, MCI, and dementia) from ADNI and the source datasets, allowing us to assess how inclusion of pathological variation during training influences harmonization performance. For downstream prediction tasks, convolutional neural networks were trained either solely on ADNI data and evaluated on harmonized and unharmonized external datasets (zero-shot setting), or trained on pooled datasets combining ADNI with either harmonized or unharmonized external data (pooled-training setting). Test splits were strictly held out from both harmonization and predictive model training to prevent information leakage, ensuring a fair and reproducible assessment of harmonization effects across datasets and tasks.

3. Harmonization Methods

Image-level harmonization was performed using a 3D CycleGAN framework [16] designed to learn scanner-invariant representations without requiring paired images across sites. The model consists of two generator networks that learn bidirectional mappings between a source domain and a target domain, along with two discriminators that enforce domain realism. Each generator follows an encoder-residual-decoder structure, where residual blocks facilitate preservation of anatomical content while enabling contrast and intensity transformations. Instance normalization is employed throughout the network to reduce sensitivity to scanner-specific intensity

distributions. Adversarial learning encourages generated images to be indistinguishable from real images in the target domain, while a cycle-consistency constraint ensures that mapping an image from one domain to the other and back recovers the original structure. An identity constraint discourages unnecessary modifications when an image is already in the target domain. This architecture allows the model to selectively reduce scanner-related variability while preserving subject-specific anatomical features critical for downstream clinical tasks.

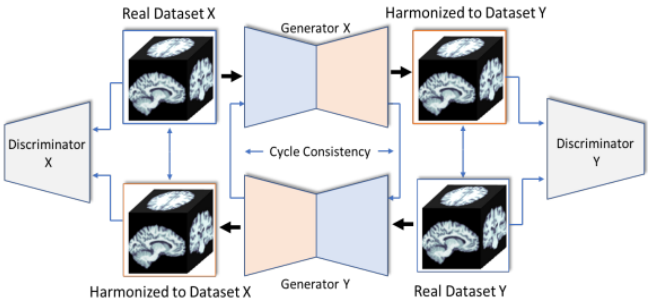


Fig. 1. **CycleGAN Architecture.** Two generators and discriminators are trained to harmonize scans from dataset X to match the protocol of dataset Y and vice versa. As with other non-paired methods, CycleGANs can be trained without requiring training data from the same subjects scanned with different scanners or protocols.

In addition to deep learning–based harmonization, we applied a statistical harmonization strategy based on the ComBat framework using the implementation provided in the DPABI toolbox. ComBat models voxel-wise intensities using a location–scale formulation that explicitly accounts for site-related batch effects while preserving biological covariate effects. For each voxel v, the observed intensity y_{iv} for subject i is modeled as

$$y_{iv} = \mu_v + x_i^T \beta_v + \gamma_{s(i)v} + \delta_{s(i)v}\, \varepsilon_{iv} \quad (1)$$

where μ_v is the global mean, x_i denotes biological covariates, and $\gamma_{s(i)v}$ and $\delta_{s(i)v}$ represent site-specific additive and multiplicative effects. Harmonized intensities are then obtained by removing these estimated site effects and recentering the data,

$$\overline{y}_{iv} = \frac{y_{iv} - \mu_v - x_i^T \beta_v - \gamma_{s(i)v}}{\delta_{s(i)v}} + \mu_v + x_i^T \beta_v \quad (2)$$

This empirical Bayes adjustment was applied voxel-wise within a predefined brain mask (BrainMask_05mm.nii at 0.5 mm resolution). The workflow was integrated with SPM12 to ensure consistent handling of NIfTI volumes. Both harmonization methods were applied to map external datasets into a common ADNI reference space. For CycleGAN, separate models were trained for each source–target pair, using ADNI as the target domain and either cognitively normal subjects or all diagnostic groups, depending on the experimental setting. Harmonized images were generated

by applying the trained generator to held-out test data. For ComBat, harmonization parameters were estimated using all available subjects from each dataset prior to downstream model training. This requirement is operationally important: ComBat needs a batch of scans from each deployment site to estimate site-specific location and scale parameters and therefore cannot harmonize a single incoming scan from a previously uncharacterized site. Once a CycleGAN mapping has been trained for a source acquisition domain, its generator can process one incoming scan at a time. However, neither method has been validated here for arbitrary unseen scanners or protocols; a CycleGAN applied outside its trained source domain may itself undergo domain shift and should be prospectively validated before clinical use.

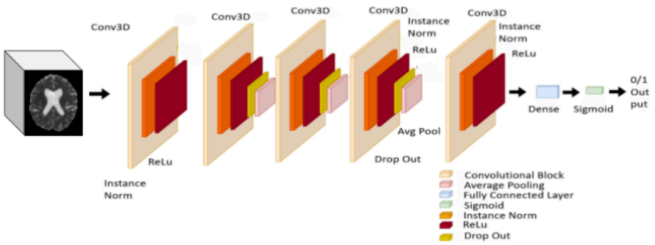


Fig. 2. 3D CNN Architecture that we trained for downstream tasks.While the architecture shown is used for classification tasks, the final binary output layer is replaced with a single linear regression output for continuous prediction tasks such as brain age and sobCDR, with all preceding layers kept identical..

To assess harmonization quality, we used multiple complementary image similarity and distributional metrics widely used in computer vision and medical image harmonization. Peak Signal-to-Noise Ratio (PSNR [31]) measures voxel-wise fidelity between original and harmonized images; higher values indicate lower distortion. The Structural Similarity Index Measure (SSIM [32]) evaluates perceptual similarity by jointly considering luminance, contrast, and structural information: values closer to one indicate greater similarity. Mutual Information (MI [33]) captures shared information between images and is particularly useful for assessing multimodal or intensity-shifted data. Fréchet Inception Distance (FID [34]) compares feature distributions between real and harmonized images to quantify global realism, while Maximum Mean Discrepancy (MMD [35]) measures distributional alignment between datasets in a reproducing kernel Hilbert space. Together, these metrics comprehensively assess both local image fidelity and global distributional consistency. To evaluate the impact of harmonization on clinically meaningful inference, we trained a 3D convolutional neural network for several downstream prediction tasks. The network architecture consisted of four 3D convolutional layers with 3×3 kernels, followed by a 1×1 convolution and a final fully connected layer. Rectified linear unit activations and instance normalization were applied throughout to promote stable optimization. To reduce overfitting, we applied dropout with a rate of 0.5 after the second, third, and fourth convolutional layers, along with a 3D average pooling operation using a 2×2 kernel. The model

was optimized using the Adam optimizer with an initial learning rate of 10^{-4}, decayed exponentially during training. Training was performed for 100 epochs with a batch size of 8, using task-appropriate loss functions and early stopping to prevent overfitting. Model performance was first benchmarked on sex classification, where ground-truth labels are unambiguous, using balanced accuracy and F1 score, and subsequently extended to brain age prediction, Alzheimer's disease classification, and sobCDR regression to assess the clinical relevance of harmonization.

4. Experimental Results and Discussion

Table 1 summarizes quantitative image similarity and distributional alignment metrics for all harmonization experiments, spanning both harmonization methods (CycleGAN and ComBat), training regimes (CN-only and ALL-subject), and evaluation datasets (NACC, OASIS-3, and NIMHANS), with ADNI serving as the reference domain. Higher PSNR, SSIM, and Mutual Information (MI) indicate better voxel-level and structural fidelity, while lower Fréchet Inception Distance (FID) and Maximum Mean Discrepancy (MMD) indicate improved alignment between image distributions.

Table 1. Image similarity and distributional metrics for MRI harmonization into ADNI space. An upward arrow (↑) denotes higher values are better.

Har. Model	Training Set DX	Dataset	PSNR↑	SSIM↑	MI↑	FID↓	MMD↓
CycleGAN	CN-Only	NACC	32.46	0.99	1.86	3.74	0.01
CycleGAN	CN-Only	OASIS3	34.86	0.99	1.88	3.22	0.03
CycleGAN	CN-Only	NIMHANS	31.40	0.99	1.97	5.06	0.16
CycleGAN	CN, MCI, Dem	NACC	23.36	0.93	1.33	29.22	0.06
CycleGAN	CN, MCI, Dem	OASIS3	33.75	0.99	1.84	3.46	0.05
CycleGAN	CN, MCI, Dem	NIMHANS	31.40	0.99	1.97	5.06	0.16
ComBat	CN-Only	NACC	8.80	0.32	0.75	115.80	0.11
ComBat	CN-Only	OASIS3	13.30	0.50	0.78	88.15	0.80
ComBat	CN, MCI, Dem	NACC	9.30	0.30	0.80	96.30	0.60
ComBat	CN, MCI, Dem	OASIS3	14.00	0.60	0.80	84.96	0.40

Across all datasets and training regimes, CycleGAN-based harmonization consistently achieved substantially higher PSNR, SSIM, and mutual information compared to ComBat, indicating superior preservation of voxel-level intensity relationships and structural similarity relative to the ADNI reference domain. This trend was particularly pronounced for SSIM, where CycleGAN models approached near-perfect structural similarity, showing that they can perform nonlinear intensity and contrast transformations while maintaining anatomical consistency. In contrast, ComBat produced markedly lower PSNR and SSIM values, consistent with its feature-level statistical adjustment that does not explicitly model spatial structure at the voxel level. Distributional metrics also highlight differences between the two approaches. CycleGAN harmonization resulted in substantially lower FID and MMD scores, indicating a closer alignment between the global image distributions of the harmonized and reference datasets. Notably,

CycleGAN models trained on cognitively normal subjects achieved the lowest FID values across datasets, whereas models trained on all diagnostic groups exhibited increased FID, particularly for NACC, indicating that inclusion of pathological variability during training introduces additional distributional complexity. ComBat reduced scanner-related batch effects, but consistently exhibited higher FID values, reflecting limited capacity to align higher-order image statistics. To evaluate the effect of harmonization on clinical prediction, we report results across four experimental configurations: Task 1, models trained on ADNI and evaluated on unharmonized external datasets; Task 2, models trained on ADNI and evaluated on harmonized external datasets (zero-shot harmonization); Task 3, models trained on pooled ADNI and unharmonized external datasets and evaluated on unharmonized test data; and Task 4, models trained on pooled ADNI and harmonized external datasets and evaluated on harmonized test data. This design enables us to disentangle the effects of harmonization at test time from those of harmonization-aware training. In Tables 2–5, boldface identifies the best value within the corresponding dataset and training-population comparison (including ties). The ADNI/UH and ADNI+UH/UH columns are unharmonized baselines; because no harmonization model is applied in those configurations, the same baseline is intentionally repeated across the CycleGAN and ComBat row groups to support direct within-row comparison. The repeated entries are therefore not duplicate experiments.

Table 2. Sex Classification Results across four experimental configurations. B.A. stands for Balanced Accuracy and F1 stands for F1 Score.

Harmonization Model	Training Set DX	Dataset	ADNI/UH		ADNI/H		ADNI+UH/UH		ADNI+H/H	
			B.A.	F1	B.A.	F1	B.A.	F1	B.A.	F1
CycleGAN	CN-Only	NACC	**0.82**	**0.81**	0.77	0.76	0.79	0.78	0.78	0.78
CycleGAN	CN-Only	OASIS3	0.87	0.87	**0.88**	**0.88**	0.80	0.78	0.83	0.82
CycleGAN	CN-Only	NIMHANS	0.79	0.82	0.79	0.81	0.80	0.81	**0.83**	**0.87**
CycleGAN	CN, MCI, Dem	NACC	**0.82**	**0.81**	0.77	0.76	0.79	0.78	0.81	0.80
CycleGAN	CN, MCI, Dem	OASIS3	0.87	0.87	**0.88**	**0.88**	0.80	0.78	0.82	0.82
CycleGAN	CN, MCI, Dem	NIMHANS	0.79	0.82	0.79	0.81	0.80	0.81	**0.84**	**0.88**
ComBat	CN-Only	NACC	**0.82**	**0.81**	0.64	0.60	0.79	0.78	0.68	0.69
ComBat	CN-Only	OASIS3	0.87	0.87	0.55	0.58	0.80	0.78	0.62	0.61
ComBat	CN, MCI, Dem	NACC	**0.82**	**0.81**	0.70	0.65	0.79	0.78	0.72	0.71
ComBat	CN, MCI, Dem	OASIS3	0.87	0.87	0.60	0.61	0.80	0.78	0.63	0.64

Sex classification results across the four tasks are summarized in Table 2. In Task ADNI/UH, models trained solely on ADNI achieved strong generalization to unharmonized external datasets, indicating that sex-related anatomical differences are robust to scanner variability. In Task ADNI/H, CycleGAN-based harmonization largely preserved performance, with minor fluctuations across datasets, whereas ComBat harmonization worsened performance, suggesting the removal of sex-discriminative signals. Tasks ADNI+UH/UH and ADNI+H/H demonstrate that pooled training mitigates some of the performance loss observed in zero-shot settings, particularly for CycleGAN, indicating that exposure to multi-site variability during training improves robustness.

However, even under pooled training, ComBat-harmonized data continued to underperform, highlighting the risk of overcorrection in feature-level harmonization.

Brain age prediction results are shown in Table 3 using mean absolute error (MAE) and R-squared scores. Brain age prediction is a commonly used task in neuroimaging, where a machine learning model is trained to predict a person's chronological age from their scan. Although the gap between this measure and the person's true age is often used as a brain health assessment or heuristic, here we just use the accuracy of the brain age estimate in healthy controls as a benchmarking task. In Task 1, unharmonized pooled data already yielded reasonable performance, particularly for OASIS-3, reflecting partial scanner robustness of age-related features. Task 2 demonstrates that CycleGAN harmonization at test time improves age prediction for OASIS-3 but has minimal effect for NACC and NIMHANS, suggesting dataset-dependent benefits. In Task 3, pooled training on unharmonized data yields marginal gains, while Task 4 consistently achieves the best or comparable performance for CycleGAN, indicating that harmonization-aware training helps align age-related representations across sites. In contrast, ComBat harmonization substantially increased MAE across all configurations, implying loss of continuous age-related structural information.

Table 3. Brain Age Prediction Results across four experimental configurations. MAE stands for Mean Absolute Error.

Harmonization Model	Training Set DX	Dataset	ADNI/UH		ADNI/H		ADNI+UH/UH		ADNI+H/H	
			MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
CycleGAN	CN-Only	NACC	7.05	0.22	7.07	0.23	7.19	0.20	7.18	0.20
CycleGAN	CN-Only	OASIS3	5.25	0.42	4.86	0.53	5.50	0.37	**4.82**	**0.53**
CycleGAN	CN-Only	NIMHANS	**5.30**	**0.32**	**5.30**	**0.32**	5.40	0.30	5.38	0.35
CycleGAN	CN, MCI, Dem	NACC	7.05	0.22	7.00	0.22	7.19	0.20	**6.76**	**0.32**
CycleGAN	CN, MCI, Dem	OASIS3	5.25	0.42	5.30	0.43	5.50	0.37	4.87	0.53
CycleGAN	CN, MCI, Dem	NIMHANS	**5.30**	**0.32**	6.51	0.31	5.40	0.30	5.38	0.30
ComBat	CN-Only	NACC	7.05	0.22	10.20	0.04	7.19	0.20	9.60	0.05
ComBat	CN-Only	OASIS3	5.25	0.42	9.50	0.08	5.50	0.37	8.10	0.11
ComBat	CN, MCI, Dem	NACC	7.05	0.22	9.90	0.10	7.19	0.20	8.30	0.17
ComBat	CN, MCI, Dem	OASIS3	5.25	0.42	9.10	0.13	5.50	0.37	7.70	0.23

Dementia classification results across configurations are summarized in Table 4. In Task 1, models trained on ADNI generalized well to unharmonized external datasets. Task 2 reveals that zero-shot harmonization can slightly reduce classification performance for CycleGAN, particularly when trained on CN-only data, likely due to attenuation of disease-specific contrast. Task 3 demonstrates that pooled training on unharmonized data improves robustness, while Task 4 recovers or modestly improves performance for CycleGAN, especially when trained on all subjects. These findings suggest that disease-aware harmonization is most effective when the classifier is trained on harmonized data. ComBat harmonization consistently degraded AD classification across all tasks, indicating substantial loss of disease-discriminative signal.

Table 4. Dementia Classification Results across four experimental configurations. B.A. stands for Balanced Accuracy and F1 stands for F1 Score.

Harmonization Model	Training Set DX	Dataset	ADNI/UH		ADNI/H		ADNI+UH/UH		ADNI+H/H	
			B.A.	F1	B.A.	F1	B.A.	F1	B.A.	F1
CycleGAN	CN-Only	NACC	0.84	0.83	0.75	0.77	**0.86**	**0.86**	0.75	0.76
CycleGAN	CN-Only	OASIS3	0.76	0.70	0.74	0.70	**0.78**	**0.72**	0.75	0.70
CycleGAN	CN-Only	NIMHANS	0.84	0.82	0.81	0.80	**0.85**	**0.83**	0.80	0.79
CycleGAN	CN, MCI, Dem	NACC	0.84	0.83	0.85	0.85	**0.86**	**0.86**	0.79	0.79
CycleGAN	CN, MCI, Dem	OASIS3	0.76	0.70	0.74	0.72	**0.78**	**0.72**	0.77	0.71
CycleGAN	CN, MCI, Dem	NIMHANS	0.84	0.82	0.83	0.81	**0.85**	**0.83**	0.82	0.82
ComBat	CN-Only	NACC	0.84	0.83	0.52	0.52	**0.86**	**0.86**	0.66	0.60
ComBat	CN-Only	OASIS3	0.76	0.70	0.61	0.53	**0.78**	**0.72**	0.61	0.59
ComBat	CN, MCI, Dem	NACC	0.84	0.83	0.71	0.65	**0.86**	**0.86**	0.80	0.75
ComBat	CN, MCI, Dem	OASIS3	0.76	0.70	0.60	0.55	**0.78**	**0.72**	0.68	0.65

sobCDR prediction results are presented in Table 5. Task 1 shows strong baseline performance on unharmonized data, particularly for OASIS-3. In Task 2, CycleGAN harmonization maintains or slightly improves performance when trained on all subjects, whereas CN-only models show minimal benefit, suggesting that severity-related information benefits from disease-inclusive harmonization. Task 3 yields stable performance under pooled unharmonized training, while Task 4 consistently produces the best results for CycleGAN, with lower MAE and higher correlation. ComBat harmonization severely degrades performance across all configurations, often resulting in near-zero or negative explained variance, underscoring its unsuitability for severity-sensitive regression tasks.

Table 5. sobCDR Prediction Results across four experimental configurations. MAE stands for Mean Absolute Error.

Harmonization Model	Training Set DX	Dataset	ADNI/UH		ADNI/H		ADNI+UH/UH		ADNI+H/H	
			MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
CycleGAN	CN-Only	NACC	1.93	0.26	1.99	0.02	0.47	0.96	**0.37**	**0.97**
CycleGAN	CN-Only	OASIS3	1.58	0.03	1.61	0.01	0.40	0.94	**0.25**	**0.96**
CycleGAN	CN, MCI, Dem	NACC	1.93	0.26	1.90	0.27	0.47	0.96	0.40	0.96
CycleGAN	CN, MCI, Dem	OASIS3	1.58	0.03	1.59	0.01	0.40	0.94	0.28	0.94
ComBat	CN-Only	NACC	1.93	0.26	3.51	-0.30	1.93	0.26	3.01	-0.10
ComBat	CN-Only	OASIS3	1.58	0.03	3.11	-0.21	1.58	0.03	2.88	-0.06
ComBat	CN, MCI, Dem	NACC	0.47	0.96	3.01	-0.15	0.47	0.96	2.71	0.10
ComBat	CN, MCI, Dem	OASIS3	0.40	0.94	2.83	0.01	0.40	0.94	2.55	0.18

The results across Tasks 1-4 reveal several consistent patterns. In Task 1, models trained solely on ADNI and evaluated on unharmonized external datasets suggest that certain neuroanatomical features, such as those associated with sex and age, are sufficiently robust to scanner-induced variability to support reasonable generalization. Deep CNNs can partially learn scanner-invariant representations when trained on a large, heterogeneous reference dataset, even without explicit harmonization. In Task 2, where harmonization is applied only at test time, performance frequently degrades for disease-related tasks. One plausible explanation is representation mismatch as classifiers trained on unharmonized ADNI images learn feature distributions that no

longer align with harmonized test data, even if the harmonized images appear visually similar. Harmonization alters intensity statistics, texture distributions, and higher-order spatial relationships, which can shift the learned feature embeddings away from the training distribution. This effect is particularly pronounced for Alzheimer's disease classification and sobCDR prediction, where subtle regional atrophy patterns and disease progression markers are critical. In contrast, sex classification is less sensitive to such shifts, as it relies on more global anatomical differences that are preserved under moderate harmonization.

Task 3 demonstrates that pooled training on unharmonized multi-site data improves robustness by exposing the classifier to a broader range of scanner-specific variations. From a learning theory perspective, this acts as a form of implicit domain augmentation, encouraging the network to focus on invariant features that generalize across sites. However, because scanner-induced biases remain present, this strategy does not fully resolve inter-site distribution mismatch and may still limit performance on more sensitive regression tasks. When harmonization is helpful, its largest improvements over the corresponding unharmonized condition generally occur in Task 4, where both training and test data are harmonized and pooled; Task 4 is not uniformly superior across every dataset, task, or metric. This configuration aligns the data distributions seen during training and inference and can reduce covariate shift, but its benefit remains task- and cohort-dependent. For CycleGAN-based harmonization, this effect is particularly pronounced because nonlinear image-level transformations explicitly reduce site-specific intensity and contrast differences while preserving spatial structure. As a result, harmonization-aware training enables improved performance for disease classification and severity prediction, especially when harmonization models are trained on all diagnostic groups.

The contrast between CN-only and ALL-subject harmonization further highlights the role of disease-aware modeling. Harmonization models trained exclusively on cognitively normal subjects may inadvertently suppress disease-related variability when applied to pathological brains, effectively projecting abnormal anatomy toward a normative template. This may explain the observed performance drops for AD classification and sobCDR prediction under CN-only harmonization. In contrast, models trained on data from all diagnostic groups learn transformations that accommodate pathological variation, preserving disease severity signals while still reducing scanner effects. The consistently poor downstream performance observed with ComBat-based harmonization may be due to its reliance on linear, feature-level adjustments that do not account for spatial structure or nonlinear scanner effects. While ComBat effectively reduces batch effects in summary features, voxel-wise application may attenuate biologically meaningful variance, particularly for tasks requiring fine-grained spatial information. This limitation becomes most evident in regression tasks such as brain age and sobCDR prediction, where continuous anatomical gradients are essential. These results underscore the need to evaluate harmonization

methods not only through image similarity metrics, but through their interaction with learning objectives and clinical endpoints.

5. Conclusion and Future Work

In this work, we presented a comprehensive evaluation of MRI harmonization strategies that extends beyond image similarity to benchmarking on a range of clinically meaningful prediction tasks. By comparing CycleGAN- and ComBat-based harmonization across four training and evaluation configurations, we found that harmonization applied only at test time can introduce harmful distribution shifts, while harmonization-aware training yields more stable and generalizable performance. Our results further show that harmonization models trained on all diagnostic groups better preserve disease-related information than those trained exclusively on cognitively normal subjects. While deep learning-based harmonization achieved superior image-level alignment, its benefits for downstream tasks were task-dependent, underscoring that visual similarity alone is insufficient for evaluating harmonization quality. The immediate value of these findings is methodological: they guide researchers who pool multi-site MRI and teams evaluating whether a model developed on one cohort can transfer to another. They do not establish clinical utility or deployment readiness. In a clinical workflow, adding a harmonizer only at inference could reproduce the representation-shift failure observed in Task 2; any harmonizer and predictor should instead be evaluated as a coupled pipeline, with site-specific calibration, subgroup error analysis, prospective validation, and clinically defined operating thresholds. Together, these findings motivate clinically grounded, task-aware evaluation frameworks. This study has several limitations that motivate future investigation. First, although we evaluated harmonization across multiple large cohorts, all datasets were mapped into a single reference domain (ADNI), and alternative reference choices or multi-reference harmonization strategies may yield different outcomes. Second, our downstream evaluation focused on T1-weighted structural MRI and a fixed 3D CNN architecture; extending this analysis to other modalities such as diffusion MRI, functional MRI, and multimodal fusion, as well as to alternative model architectures, may further elucidate the interaction between harmonization and predictive learning. Third, previously unseen sites and acquisition protocols were not evaluated prospectively; the NIMHANS cohort provides external-cohort evidence but does not establish generalization to arbitrary scanners. Fourth, while disease-inclusive harmonization improved performance for several common tasks, the optimal balance between preserving pathological variation and removing scanner effects remains an open problem, particularly for early or subtle disease stages. Future work will examine task-adaptive and representation-level harmonization approaches, as well as joint optimization of harmonization and prediction models, to align scanner invariance with clinical objectives better.

6. Acknowledgments

This work was supported by NIH grants U01AG068057, R01AG081571, S10OD032285 and U24AG074855.

References

1. Thompson, P. M., et al. (2020). ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. Translational Psychiatry, 10(1), 100.
2. Takao, H., et al. (2011). Effect of scanner in longitudinal studies of brain volume changes. Journal of Magnetic Resonance Imaging, 34(2), 438-444.
3. Shinohara, R. T., et al. (2017). Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. American Journal of Neuroradiology, 38(8), 1501-1509.
4. Zhu, T., et al. (2011). Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study. NeuroImage, 56(3), 1398-1411.
5. Dennis, E.L., et al. (2023). Harmonization strategies and considerations in neuroimaging studies. Frontiers in Neurology, 14, 1165263.
6. Jack Jr, C. R., et al. (2008). 'The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods.' Journal of Magnetic Resonance Imaging, 27(4), 685-691.
7. Beekly, D.L., et. al. (2007). The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. Alzheimer Disease and Associated Disorders, 21(3), 249–258.
8. LaMontagne, P. J., et al. (2019). 'OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease.' medrxiv, 2019-12.
9. Orlhac, F., et al. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. Journal of Nuclear Medicine, 63(2), 172-179.
10. Fortin, J. P., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. Neuroimage, 167, 104-120.
11. Fortin, J. P., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. Neuroimage, 161, 149-170.
12. Pomponio, R., et al. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage, 208, 116450.
13. Beer, J. C., et al. (2020). Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. NeuroImage, 220, 117129.
14. Zhu, J. Y., et al. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

15. Dewey, B. E., et al. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magnetic resonance imaging, 64, 160-170.
16. Komandur, D., et al. (2023, November). Unsupervised harmonization of brain MRI using 3D CycleGANs and its effect on brain age prediction. In 2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM) (pp. 1-5). IEEE.
17. Modanwal, G., et al. (2020, March). MRI image harmonization using cycle-consistent generative adversarial network. In Medical Imaging 2020: Computer-Aided Diagnosis (Vol. 11314, pp. 259-264). SPIE.
18. Moyer, D., et al. (2020). Scanner invariant representations for diffusion MRI harmonization. Magnetic Resonance in Medicine, 84(4), 2174-2189.
19. Dinsdale, N. K., et al. (2021). Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. NeuroImage, 228, 117689.
20. Ganin, Y., et al. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1-35.
21. Abbasi, S., et al. (2024). Deep learning for the harmonization of structural MRI scans: a survey. BioMedical Engineering OnLine, 23(1), 90.
22. Zhang, S. et al. "Diffusion Bridge Models for 3D Medical Image Translation," 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 2025, pp. 1-7.
23. Zuo, L., et al. (2023, April). A latent space for unsupervised MR image quality control via artifact assessment. In Medical Imaging 2023: Image Processing (Vol. 12464, pp. 286-291). SPIE.
24. Adams R, et al. UltimateSynth: MRI Physics for Pan-Contrast AI. bioRxiv [Preprint]. 2025 Aug doi: 10.1101/2024.12.05.627056.
25. Yan, C. G., et al. (2016). DPABI: data processing & analysis for (resting-state) brain imaging. Neuroinformatics, 14(3), 339-351.
26. Bharath S., et al. (2017). A Multimodal Structural and Functional Neuroimaging Study of Amnestic Mild Cognitive Impairment. American Journal of Geriatric Psychiatry, 25(2), 158–169.
27. Joshi H., et al. (2019). Differentiation of Early Alzheimer's Disease, Mild Cognitive Impairment, and Cognitively Healthy Elderly Samples Using Multimodal Neuroimaging Indices. Brain Connectivity, 9(9), 730–741.
28. John J. P. , et al. (2023). India ENIGMA initiative for Global Aging & Mental Health–a globally coordinated study of brain aging and Alzheimer’s Disease. Alzheimer’s Dement., 19, e076394.

29. Cedarbaum, J. M., et al. (2013). ‘Rationale for use of the Clinical Dementia Rating Sum of Boxes as a primary outcome measure for Alzheimer’s disease clinical trials.’ Alzheimer's & Dementia, 9(1), S45-S55.
30. Lam P., et al., 3-D Grid-Attention Networks for Interpretable Age and Alzheimer’s Disease Prediction from Structural MRI, arXiv (2020).
31. Hore, A., et al., (2010, August). Image quality metrics: PSNR vs. SSIM. In 2010 20th International Conference on Pattern Recognition (pp. 2366-2369). IEEE.
32. Wang, Z., et al., (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612.
33. Studholme, C., et al., (1999). An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition, 32(1), 71-86.
34. Heusel, M., et al., (2017). Gans trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems, 30.
35. Gretton, A., et al., “A kernel two-sample test,” J. Mach. Learn. Res., vol. 13, pp. 723–773, 2012.