ComBat Harmonization: Empirical Bayes versus Fully Bayes Approaches

Maxwell Reynolds[a], Tigmanshu Chaudhary[a], Mahbaneh Eshaghzadeh Torbati[b], Dana L. Tudorascu[c,d], Kayhan Batmanghelich[a,1], for the Alzheimer's Disease Neuroimaging Initiative[2]

[a] Department of Biomedical Informatics, University of Pittsburgh School of Medicine, 5607 Baum Blvd. Suite 500, Pittsburgh, PA, 15206, USA
[b] Intelligent System Program, University of Pittsburgh School of Computing and Information, 210 South Bouquet Street, Pittsburgh, PA 15260, USA
[c] Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, PA 15213, USA
[d] Department of Biostatistics, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA 15213, USA

mar398@pitt.edu, tic48@pitt.edu, mae82@pitt.edu, dlt30@pitt.edu, kayhan@pitt.edu

Corresponding Author: Maxwell Reynolds (mar398@pitt.edu)

[1] Present Address: Department of Electrical and Computer Engineering, Boston University, 44 Cummington Mall, Boston, MA 02215, USA
[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

Studying small effects or subtle neuroanatomical variation requires large-scale sample size data. As a result, combining neuroimaging data from multiple datasets is necessary. Variation in acquisition protocols, magnetic field strength, scanner build, and many other non-biologically related factors can introduce undesirable bias into studies. Hence, harmonization is required to remove the bias-inducing factors from the data. ComBat is one of the most common methods applied to features from structural images. ComBat models the data using a hierarchical Bayesian model and uses the empirical Bayes approach to infer the distribution of the unknown factors. The empirical Bayes harmonization method is computationally efficient and provides valid point estimates. However, it tends to underestimate uncertainty. This paper investigates a new approach, fully Bayesian ComBat, where Monte Carlo sampling is used for statistical inference. When comparing fully Bayesian and empirical Bayesian ComBat, we found Empirical Bayesian ComBat more effectively removed scanner strength information and was much more computationally efficient. Conversely, fully Bayesian ComBat better preserved biological disease and age-related information while performing more accurate harmonization on traveling subjects. The fully Bayesian approach generates a rich posterior distribution, which is useful for generating simulated imaging features for improving classifier performance in a limited data setting. We show the generative capacity of our model for augmenting and improving the detection of patients with Alzheimer's disease. Posterior distributions for harmonized imaging measures can also be used for brain-wide uncertainty comparison and more principled downstream statistical analysis. Code for our new fully Bayesian ComBat extension is available at https://github.com/batmanlab/BayesComBat.

Keywords

MRI, Harmonization, Alzheimer's, Bayesian, ComBat, ADNI

1. Introduction

Large-scale neuroimaging datasets have been created in recent years to identify disease biomarkers, study brain development, and standardize image acquisition (Mueller et al., 2005).

These datasets have enabled the identification of disease-relevant features (King et al., 2009), population-wide examination of neurological phenotypes (Cury et al., 2015), individual brain trajectory modeling (Koval et al., 2021), and data-driven disease subtyping (Young et al., 2018). The open-access nature of these datasets has allowed for external validation of new findings (Cury et al., 2020), and large longitudinal datasets have enabled subject-specific prediction with ground truth validation (Marinescu et al., 2020; Nebli et al., 2020).

Projects like the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) and the Adolescence Brain Cognitive Development (ABCD) Study (Casey et al., 2018) have led to insights into Alzheimer's Disease, aging, development, and other biological processes. Most of these datasets use images acquired from many different clinical sites and scanners. Differences in magnetic resonance imaging (MRI) scanner hardware and acquisition processes from these multiple sites can introduce additional unwanted variance into neuroanatomical feature measurements (Han et al., 2006). For example, a 3T scanner may produce a higher quality tissue contrast than a 1.5T scanner, leading to higher estimates of grey matter volumes. Quantifying and correcting for nonbiological scanner factors, while maintaining biological information, is necessary to facilitate more accurate analysis so that scanner effects are not attributed to subject or population differences (Fortin et al., 2017). Harmonization addresses this issue by modeling and correcting for scanner effects in imaging features.

Many methods have been proposed for harmonizing raw images and imaging-derived features (e.g. regional grey matter thickness and volume). For image-level harmonization, recent work has viewed harmonization as a style transfer procedure and used deep learning approaches including variational auto-encoders (Zuo et al., 2021), generative adversarial networks (GANs) (Liu et al., 2021), and U-Nets (Torbati et al., 2021b) to harmonize images to a specific reference scanner. Deep learning image-based approaches are increasingly adopted for various applications in the medical domain, including harmonization and subsequent downstream tasks such as disease classification. However, feature-based methods that measure regional and global brain characteristics such as thickness and volume remain relevant for studies of

neurological disease pathology and progression due to their interpretable nature. Such features are readily related to the biological understanding of disease models. We therefore focus on image-derived feature harmonization and build on an existing widely used harmonization method.

Feature-level harmonization is often treated as a regression problem. One approach is to model scanner effects for each imaging feature as a fixed effect and residualize the effect from the data (Venkatraman et al., 2015). Another method, ME-Mega, uses a similar model but views scanner effects as random intercepts (Radua et al., 2020). ComBat, a batch harmonization technique originally proposed for gene expression microarrays (Johnson et al., 2007), adds a multiplicative (variance scaling) scanner effect term. Additionally, ComBat assumes that site effects come from a common distribution across regions of interest by using a hierarchical Bayesian model. This causes pooling of scanner effects towards a mean, making ComBat more robust to smaller within-scanner sample sizes (Johnson et al., 2007).

ComBat has recently been proposed for structural MRI-derived feature harmonization (Fortin et al., 2018, 2017) and has since been used routinely for harmonization in neuroimaging studies (Bartlett et al., 2018; Dima et al., 2021; Habes et al., 2021). The original ComBat model has also been further extended to accommodate repeated scans on the same subjects over time in longitudinal datasets such as ADNI (Beer et al., 2020).

ComBat uses a type of Bayesian inference called empirical Bayes (EB) (Carlin and Louis, 2000) to infer the distribution of the latent variables. In EB, the observed data is used to learn a point estimation of latent variables at the highest level of the hierarchical model (hyperparameters), rather than learning a probability distribution. Empirical Bayes is often computationally less expensive, especially for large models such as ComBat, but the hyperparameter point estimation ignores uncertainty in part of the model. This can lead to inaccurate and underestimated posterior uncertainty for the latent variables (van de Wiel et al., 2019). Additionally, the empirical Bayes approach confines the posterior of a parameter of interest to

a specific distribution (e.g., Normal or Inverse-Gamma). For models with conjugate priors, this assumption is valid. However, this limits the choice of a prior distribution. Using a fully Bayesian approach generally produces more accurate uncertainty measurements (Gelman et al., 2021; van de Wiel et al., 2019), allowing for more accurate posterior distribution inference even when some model parameters are misspecified (Piecuch et al., 2017).

Using fully Bayesian approaches has typically relied on slower Markov Chain Monte Carlo (MCMC) inference methods such as Metropolis-Hastings MCMC (Hastings, 1970) for inference of the posterior distribution of the model's latent variables. The sampling approach allows more flexible choice of prior distributions at the expense of the computational cost of inference. Recently however, more efficient samplers and parallel GPU computation have enabled computationally feasible fully Bayesian estimation for large models (Phan et al., 2019). As the efficiency difference between empirical and fully Bayesian inference narrows, fully Bayesian inference may offer a more principled approach without prohibitive computational costs.

With this work, we contribute to the literature in multiple ways: 1) We introduce a new ComBat formulation which infers a joint posterior distribution for the entire model in a single inference stage; 2) we investigate the performance for harmonization of features from T1-weighted structural images against EB ComBat using metrics to quantify biological (e.g. age and disease) information while removing non-biological information (e.g. scanner strength and test-retest feature differences); and 3) we introduce several novel use cases for FB harmonization which utilize its rich posterior distribution for augmentation and uncertainty quantification.

2. Materials and Methods

2.1 Data

We use T1-weighted structural images from the ADNI dataset, acquired using MPRAGE on Philips, Siemens, and GE scanners (Jack et al., 2010). ADNI is a public-private partnership launched in 2003 with the primary goal of testing if combinations of MRI, positron emission

tomography (PET), other biomarkers, and clinical and neuropsychological assessments can measure MCI and AD progression. More information can be found at www.adni-info.org (ADNI, 2016). All subjects gave informed consent in accordance with local Institutional Review Board Regulations (Petersen et al., 2010).

We obtain 3894 initial images from 809 patients grouped as either cognitively normal (CN), mild cognitive impaired (MCI), or Alzheimer's disease (AD). At baseline, 28.1% (n=227) of subjects are labeled CN, 49.1% (n=397) are labeled MCI, and 22.9% (n=185) are labeled AD. The subjects' sexes are 57% (n=467) male and 43.2% (n=342) female. The subjects' ages at baseline are between 51 and 91 years with a mean age of 75.3 years. Images were acquired on 83 different scanners. 58 of the scanners have a 1.5T field strength; the remaining 25 scanners are 3T.

2.2 Preprocessing

We use the FreeSurfer version 7.1.1 longitudinal pipeline (Reuter et al., 2012) on a Linux CentOS version 8.2 machine to segment various brain structures and obtain global and local cortical thickness and subcortical volume measurements. The first step in the Freesurfer Longitudinal pipeline is the standard cross-sectional "recon-all" function. This includes motion correction, N3 non-uniformity correction, brain extraction, subcortical segmentation, and cortical parcellation of each image. Next, a mean template from each within-subject image set is created and used for an unbiased initialization for a second "recon-all" run on each image.

2.3 Quality Control

After FreeSurfer processing, 401 images are dropped due to duplicate scans (subject scanned on the same scanner on the same date) or failure during FreeSurfer registration, segmentation, or parcellation stages. Next, 70 images with outlier imaging features (at least 5 standard deviations away from the feature mean) are manually inspected and excluded if noticeable errors existed such as brain extraction failure leading to segmentations labeling the skull as cortical gray matter.

2.4 Empirical Bayes ComBat Model

The EB ComBat model (Beer et al., 2020) is given by:

$$y_{ijv}(t) \sim N(\alpha_v + X_j^T(t)\beta_v + \eta_{jv} + \gamma_{iv}, \delta_{iv}^2 * {\sigma_v}^2) \tag{1}$$

where *i* is the scanner index, *j* is the subject index, *v* is the imaging feature index, and *t* represents time. $y_{ijv}(t)$ is the measured (unharmonized) value for feature *v* of subject *i* on scanner *j*. γ_{iv} is the additive scanner factor for scanner *i* and feature *v*. δ_{iv}^2 is the scaling scanner factor from scanner *i* and feature *v*. X_j^T is a covariate term for biological effects. A description of all variables is given in Table 1. We include age, sex, diagnosis, and diagnosis x age covariate terms in both EB and FB ComBat models (Beer et al., 2020; Sun et al., 2021). Of note, the subject-specific random effect η_{jv} was added in longitudinal ComBat (Beer et al., 2020) and showed improved harmonization in longitudinal datasets. We therefore include this parameter in both our EB and FB models.

In EB ComBat, α_v, β_v, η_{jv}, σ_v, and priors for γ_{iv} and δ_{iv}^2 are estimated using restricted maximum likelihood (REML) and method of moments, then conditional posteriors for γ_{iv} and δ_{iv}^2 are identified using an expectation-maximization (EM) algorithm.

Harmonized adjusted feature values follow the equation:

$$y_{ijv}^{EB}(t) = \frac{y_{ijv}(t) - \hat{\alpha}_v - X_j^T(t)\hat{\beta}_v - \hat{\eta}_{jv} - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + \hat{\alpha}_v + X_j^T(t)\hat{\beta}_v + \hat{\eta}_{jv} \tag{2}$$

where $\hat{\alpha}_v$, $\hat{\beta}_v$, $\hat{\eta}_{jv}$, $\hat{\gamma}_{iv}$, $\hat{\delta}_{iv}$, are the parameter estimates.

2.5 Fully Bayes ComBat Model

In FB ComBat, all high-level parameters are given weakly informative hyper-priors. A plate diagram including hyperparameters is shown in Figure 1. We choose prior distributions to be

centered at 0 for additive factors and 1 for variance parameters. Fat-tailed Cauchy and Half-Cauchy distributions are used for several parameters when values close to 0 are expected with the possibility of outliers (e.g. a very biased scanner additive factor).

We add the following constraints to scanner additive and multiplicative factors to ensure identifiability:

$$\sum_i n_i \hat{\gamma}_{iv} = 0 \tag{3}$$

$$\sum_i \frac{n_i}{n} \hat{\delta}_{iv} = 1 \tag{4}$$

where n is the total number of images, and n_i is the number of images from scanner i. The identifiability constraint on additive scanner parameters *(3)* is explicitly made in EB ComBat. The constraint on multiplicative scanner parameters *(4)* ensures that the average overall error variance of imaging features is not changed by harmonization. In EB ComBat, this constraint is made implicitly by a multi-step inference procedure of first estimating σ_v, then learning δ_{iv} and rescaling the features accordingly (Beer et al., 2020). FB ComBat samples all parameters jointly, so in this approach, we perform a transformation on the σ_v and δ_{iv} samples which preserves overall error variance while ensuring *(4)* is met.

Features are standardized to have a mean of 0 and a standard deviation of 1 before inference. Hamiltonian Monte Carlo (HMC) using a No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014) is performed for 40,000 samples on the entire model, yielding posterior distribution samples for all model parameters. Briefly, HMC sampling models the negative log model posterior space as a *p*-dimensional physical space where *p* is the number of parameters in a model. Sampling trajectories based on the gradients of the posterior space can produce accurate posterior distributions using far fewer samples than more traditional MCMC methods. The No-U-Turn Sampler (NUTS) adaptively tunes the HMC path for more efficient overall

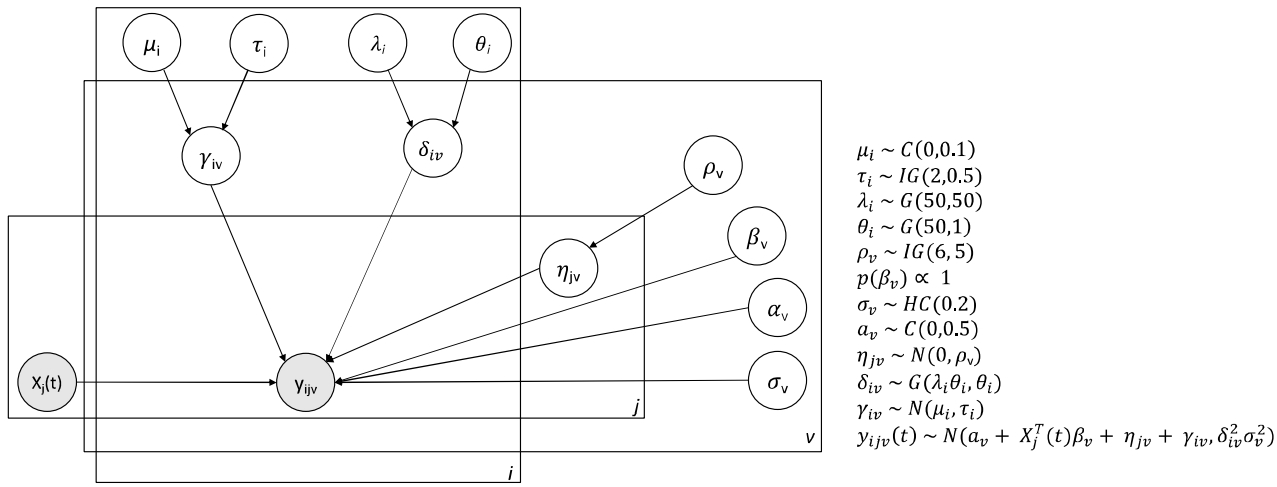


Figure 1: FB ComBat Plate Diagram

Plate diagram for FB ComBat model. Shaded circles represent observed measurements (covariates and imaging feature values). Unshaded circles represent latent parameters. Distributions C , IG, G, HC, N are Cauchy, Inverse Gamma, Gamma, Half-Cauchy, and Normal respectively.

Harmonized adjusted feature values are obtained by the equation:

$$y_{ijv}^{FB}(t) = \frac{y_{ijv} - \hat{\alpha}_v - X_j^T(t)\hat{\beta}_v - \hat{\eta}_{jv} - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + \hat{\alpha}_v + X_j^T(t)\hat{\beta}_v + \hat{\eta}_{jv} \tag{5}$$

where $\hat{\alpha}_v$, $\hat{\beta}_v$, $\hat{\eta}_{jv}$, $\hat{\gamma}_{iv}$, $\hat{\delta}_{iv}$, are the posterior parameter estimates. We perform this transformation on the joint posterior parameter distribution to obtain the posterior harmonized data distribution.

Variable	**Definition**
i	Scanner index

j	Subject index
v	Feature index
y_{ijv}	Unharmonized feature value
y_{ijv}^{EB}	Harmonized feature value using EB ComBat
y_{ijv}^{FB}	Harmonized feature value using FB ComBat
α_v	Feature mean
X_j	Covariate terms (age, sex, diagnosis, age x diagnosis interaction)
β_v	Covariate Coefficients
η_{jv}	Subject-specific intercept
γ_{iv}	Additive scanner factor
δ_{iv}	Multiplicative scanner factor
${\sigma_v}^2$	Feature-specific average variance
μ_i, τ_i	Hyperparameters for additive scanner factor
λ_i, θ_i	Hyperparameters for multiplicative scanner factor
ρ_v	Hyperparameter for variance of subject-specific intercepts

Table 1: ComBat equation variables

2.6 Implementation

NUTS inference is implemented in NumPyro Version 0.72 (Phan et al., 2019) using Python version 3.7.1. We run NUTS inference with 4 chains, 40000 samples, and 1000 warmup samples on a CentOS Version 8.2 Linux Machine using 4 Nvidia V-100 32GB GPUs. This inference takes about six hours to complete. Posterior distributions for all parameters are gathered from the 40000 MCMC samples. In both EB and FB ComBat, thickness and volume features are harmonized together.

3. Experiments and Results

3.1 Overview

We perform several experiments to evaluate the harmonization methods. First, we check HMC sample quality and convergence for the FB ComBat model. Next, we compare harmonization

performance in EB ComBat versus FB ComBat with respect to retaining biological (i.e. age and disease) information and removing scanner information. We also explore posterior uncertainty as a tool for dataset augmentation, overall regional measurement uncertainty, and uncertainty-aware association tests between brain regional measures and AD. Next, we perform a simulation study to test how harmonization affects disease effect identifiability. Finally, we perform sensitivity analysis to examine the effect of different priors on FB ComBat harmonization.

3.2 Sampling Validation

We check for the quality of our HMC sampling using two methods. We use the Effective Sample Size metric to ensure low auto-correlation of our sampling (Gelman et al., 2021). We also visually inspect the overall likelihood of the model and individual parameter chains to ensure that the inference algorithm converges to the stationary posterior distribution.

Overall model density is shown in Figure 2. After the warmup HMC parameter-tuning phase, all chains converge rapidly and explore the posterior distribution, indicating successful sampling. Effective sample sizes for various parameters are shown in Supplementary Table S1.

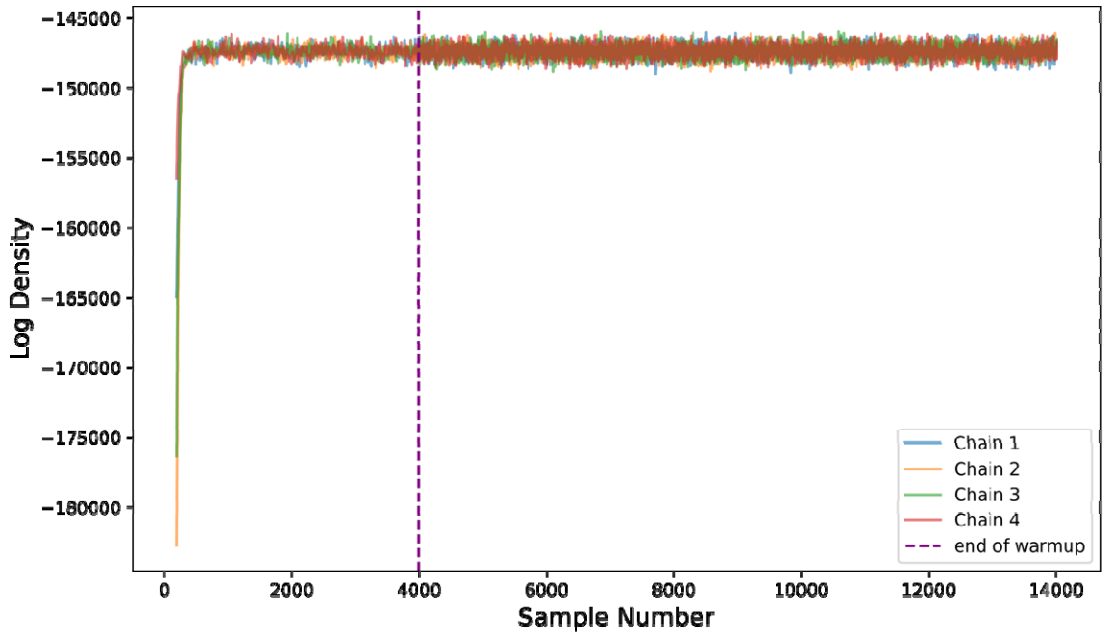


Figure 2: Sampling model density

Log of joint density for FB ComBat model given all parameters in each sample. All four chains are shown. The chains converge to the high probability region of the posterior distribution and exhibit good mixing (rapidly exploring the full region), and stationarity.

3.3 Posterior Distribution

Posterior distributions for harmonized imaging measurements are obtained for EB ComBat and FB ComBat harmonization. Posterior variances from FB ComBat are larger than those from EB ComBat, as they incorporate uncertainty from all parameters of the ComBat model. An example of a posterior distribution for a single measurement (left entorhinal cortex thickness) from one image is shown in Figure 3. Prior and posterior estimates for individual parameters are shown in Supplementary Figure S1.

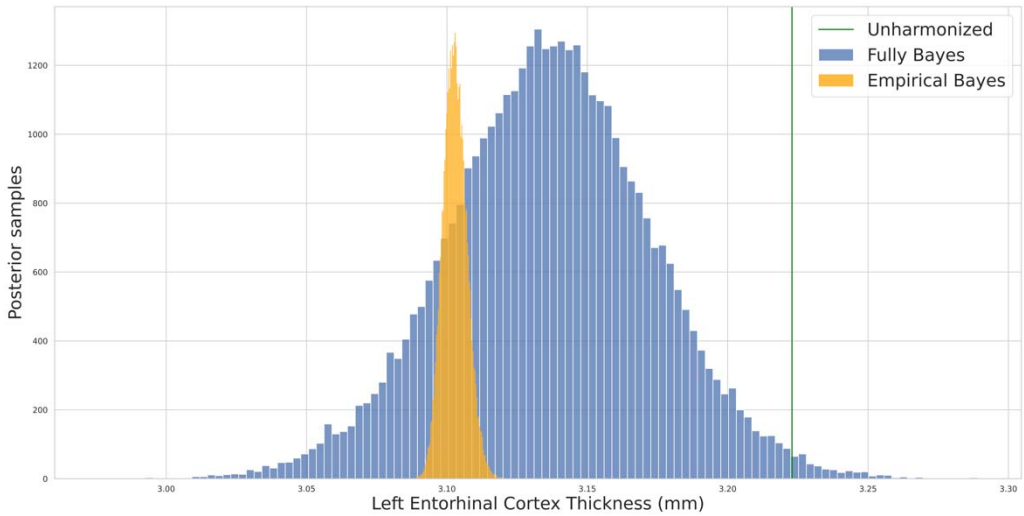


Figure 3: Single Measurement Posteriors

Harmonized posterior distribution for left entorhinal cortex thickness from a single image. The FB harmonized posterior is noticeably wider than the EB harmonized posterior. The unharmonized thickness value is shown in green.

3.4 Exploratory Analysis of Scanner Effects

We first explore and visualize the unharmonized and harmonized data for the presence of scanner effects. We check for significant additive and multiplicative scanner effects in all features using the Kenwald-Roger (Kenward and Roger, 1997) and Fligner-Killeen (Conover et al., 1981) tests using the longitudinal ComBat package (Beer et al., 2020). Of the 122 imaging features, significant ($p<0.05$) additive effects are present in 121 (of 122) features in unharmonized data and zero features in both EB ComBat-harmonized data and FB ComBat-harmonized data. Multiplicative effects are present in all 122 features in unharmonized data, 1 feature in EB ComBat-harmonized data, and 6 features in FB ComBat-harmonized data.

A visualization of scanner effects (after regressing out covariates and z-score normalization) across field strengths and manufacturers is shown in Figure 4. We plot the distribution of feature values for each image and summarize the distributions across each of the six field strength/manufacturer combinations. The field strength and manufacturer differences from

unharmonized data are partially (and to a similar degree) removed to a similar degree by EB and FB ComBat.

We also performed linear discriminant analysis (LDA) on the imaging features using the field strength and manufacturer together as the target variable. The first three components in unharmonized, EB ComBat-harmonized, and FB ComBat-harmonized data are shown in Figure 5. Both EB ComBat and FB ComBat remove most variation associated with scanner field strength and manufacturer, although some scanner-related variation is still evident in FB ComBat data (i.e. between 1.5T GE and 1.5T Siemens scanners in the first two LDA components).

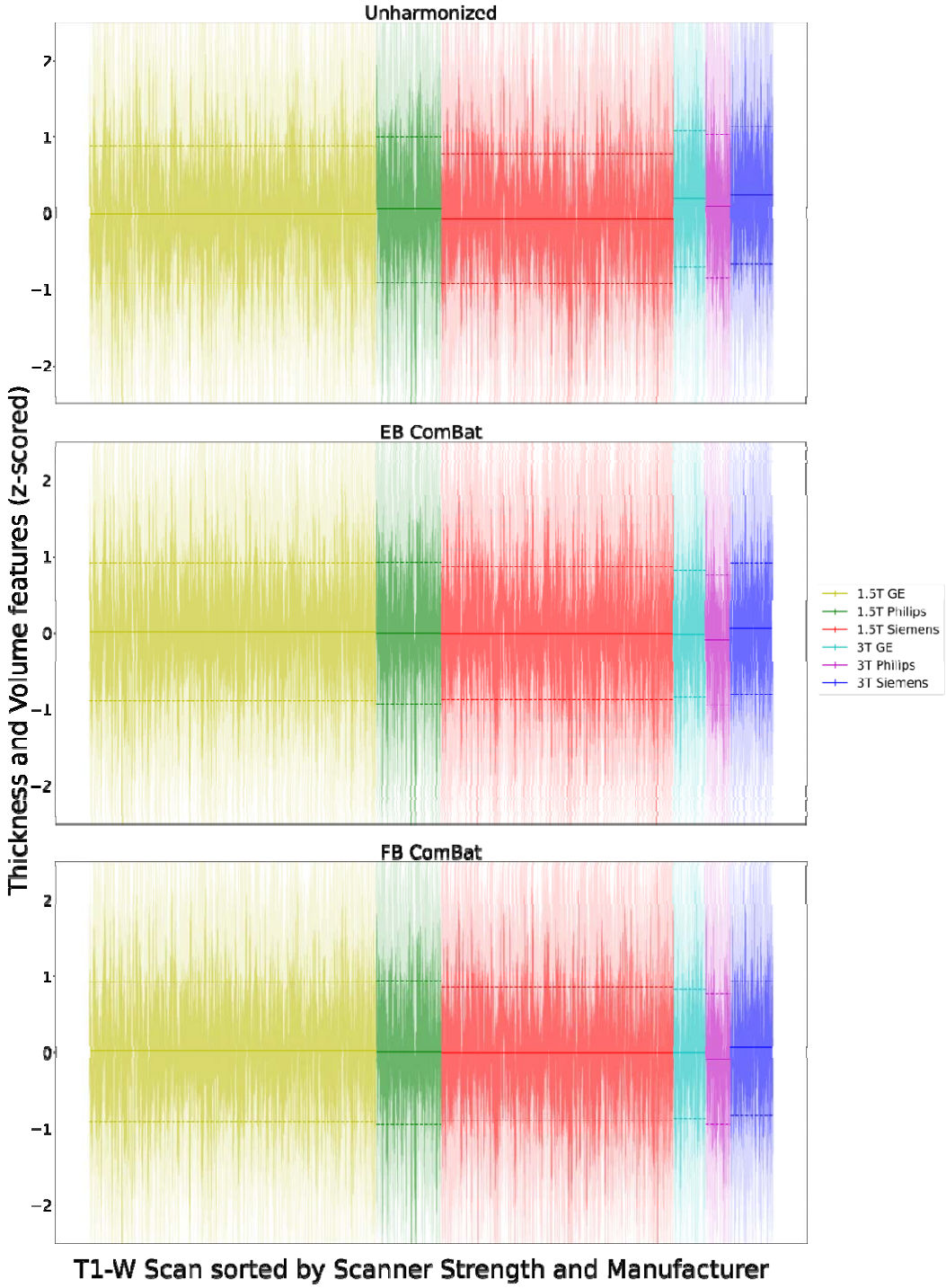


Figure 4: Scanner Effects across Field and Manufacturer

Each vertical line represents the distribution of normalized thickness and volume features for a single image; the region within one standard deviation of the image-specific mean is shaded darker. Horizontal lines show the means and 1 standard deviation intervals of normalized

feature means for each scanner strength/manufacturer combination. Covariate effects are regressed out before plotting. Distribution distances across field strength and manufacturer are partially but not completely removed by both EB ComBat and FB ComBat.

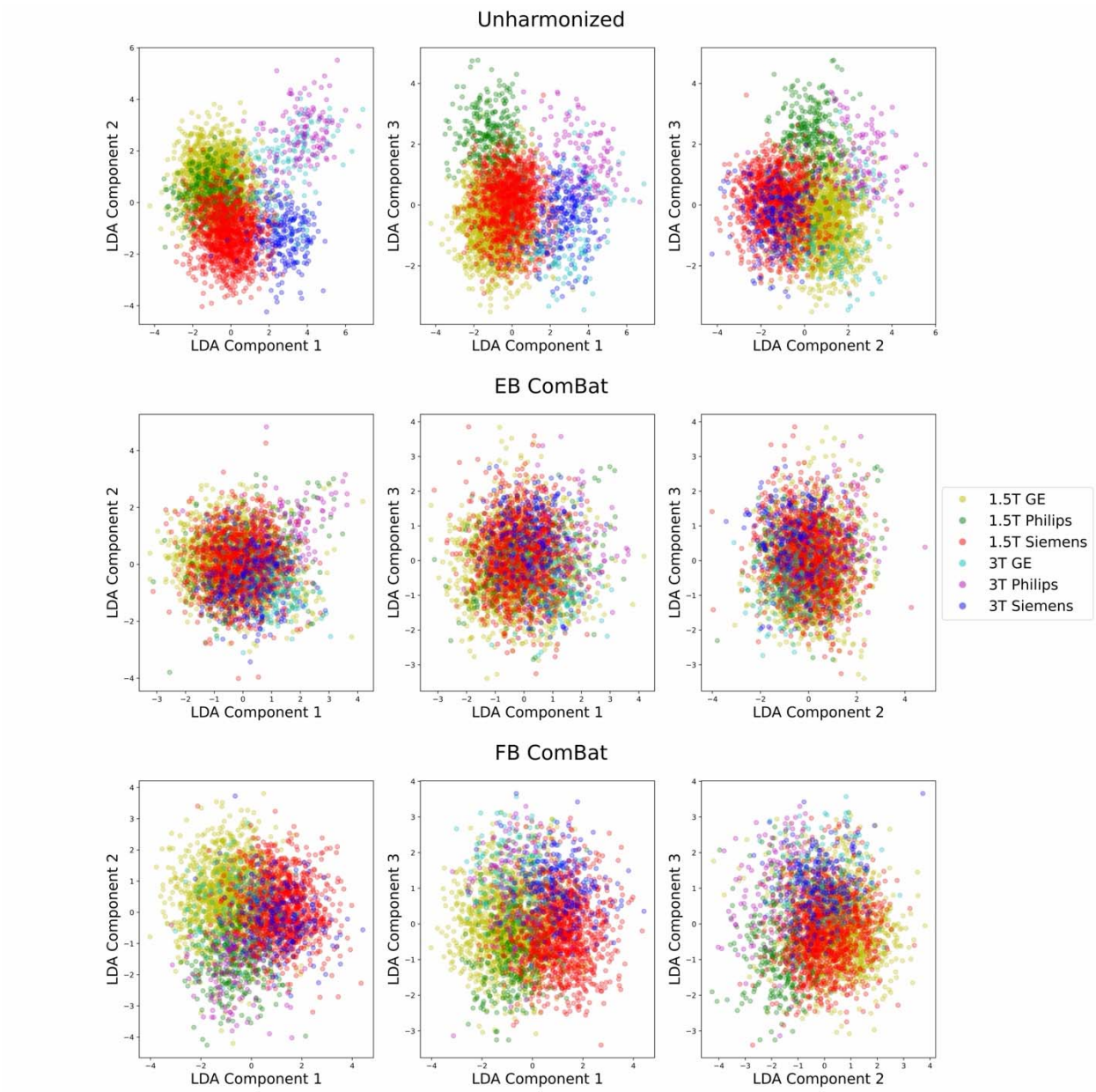


Figure 5: Linear Discriminant Analysis (LDA) with Respect to Field Strength and Manufacturer The first three LDA components of unharmonized and harmonized datasets (using the combination of scanner strength and manufacturer as the target variable) are shown. EB ComBat and FB ComBat remove most scanner-related variation in LDA components, although

some scanner-related variation is still visible in FB ComBat data. Covariates are regressed out from features used for LDA, and these features are z-score normalized.

3.5 Age Prediction

One objective of harmonization is to retain biological information in the adjustment. To evaluate whether the age information is preserved after harmonization, we train three separate random forest regression models (Breiman, 2001) from all 122 brain measures from 1) unharmonized data, 2) EB-harmonized data, and 3) FB-harmonized data. Age prediction performance of harmonized brain features has been used previously to validate the retention of biological information after harmonization (Fortin et al., 2018; Wachinger et al., 2021). We use three separate sets of predictors: unharmonized measurements, EB ComBat harmonized measurements, and FB ComBat posterior mean harmonized measurements, and compare mean absolute error (MAE) and R^2 for all three models using repeated k-fold validation with three repeats and 10 folds. We also include a dummy classifier that outputs the mean age value in the training set as a baseline that ignores input. We evaluate performance by using Wilcoxon signed-rank test for cross-validation MAE scores. In the data used for the age prediction task, the mean age is 76.3 years (min = 55 years, max = 93 years) and the data are approximately symmetric (skew = -0.38).

Age prediction results for unharmonized data, EB ComBat, and FB ComBat are shown in Table 2. FB ComBat results in a lower test MAE for age prediction than EB ComBat *($p<10^{-6}$)*, indicating that less age-related biological information is removed in FB ComBat. The test MAE using unharmonized data is not significantly lower than using FB ComBat *($p=0.39$)*.

	MAE (years)		R^2	
Method	*Train*	*Test*	*Train*	*Test*
Dummy Classifier	5.38 (0.19)	5.38 (0.19)	0.00 (0.01)	0.00 (0.01)
Unharmonized	1.05 (0.01)	2.79 (0.14)	0.96 (0.00)	0.70 (0.02)

EB ComBat	1.09 (0.01)	2.92 (0.15)	0.95 (0.00)	0.67 (0.02)
FB ComBat	1.05 (0.01)	2.80 (0.13)	0.96 (0.00)	0.70 (0.02)

Table 2: Age Prediction Results

Mean absolute error (MAE) and R^2 are shown for the age prediction task evaluating retention of biological (age-relevant) information after harmonization. Cross validation standard deviation is shown in parentheses. FB ComBat has a lower test MAE and higher test R^2 compared with EB ComBat indicating that FB ComBat performs slightly better than EB ComBat.

Differences in population characteristics across scanners may cause identifiability issues for scanner and covariate effects. We therefore check for mean differences in age across scanners using one-way analysis of variance (ANOVA). We check for age differences at two levels: individual scanners (83 groups) and field strength/manufacturer combinations (6 groups). We found significant age mean differences ($p=0.001$) at the individual scanner level but no significant age mean difference ($p=0.501$) at the field strength/manufacturer level.

3.6 Scanner Strength Prediction

The harmonization process should remove the effect of non-biological covariates that introduce bias in data. An example of such covariate is variation in the scanner strength. We train random forest binary classifier models using the three datasets (unharmonized, EB harmonized, and FB harmonized) to predict scanner strength (3.0T vs 1.5T). For this task, achieving high accuracy indicates that scanner information remains in the data. In other words, low classifier accuracy is an indicator of better harmonization. We report the area under the receiver operating characteristic curve (AUROC) for evaluation. We use Wilcoxon signed-rank test for cross-validation AUROC scores to assess the statistical significance between the performance of two models.

Area under the receiver operating characteristic curve (AUROC) curves and values of scanner strength prediction from the random forest classification model trained on unharmonized, EB ComBat, and FB ComBat data are shown in Figure 6. Training AUROC values are listed in Supplementary Table S2. Lower accuracy on this task suggests that scanner strength information was more effectively removed using EB ComBat than FB ComBat *($p<10^{-4}$)*, although both methods show improvement over unharmonized data.

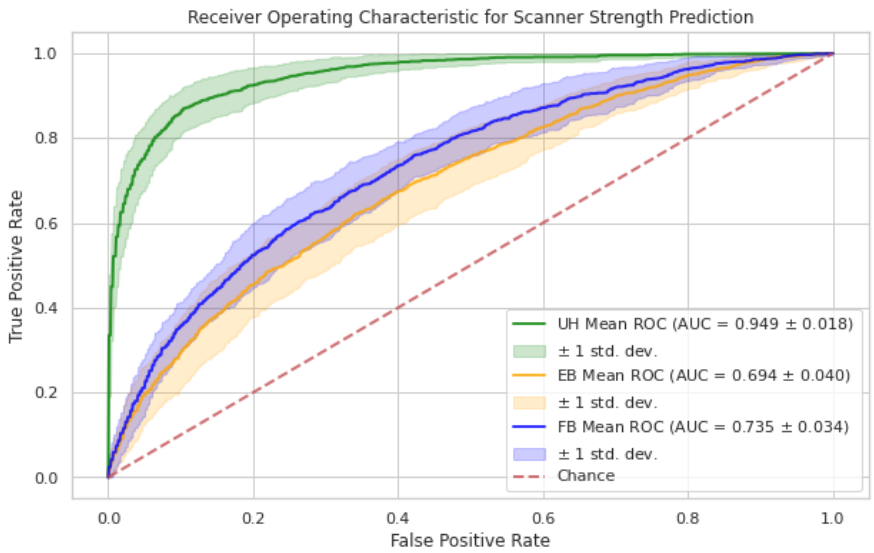


Figure 6: Scanner Strength Prediction AUROC

AUROC curves for scanner strength prediction in unharmonized (UH), EB harmonized, and FB harmonized imaging features. Means of cross-validation AUROC are shown by green, orange, and blue lines; coverage envelope of one standard deviation of cross-validation AUROC is shown by colored shaded regions. Lower AUROC indicates better performance for this task. Both EB and FB ComBat similarly reduce the AUROC closer to random chance, but there is still some scanner signal left in the harmonized data.

3.7 Test-retest using Paired Scan Evaluation

For the next phase of evaluation, we identified 184 imaging pairs where a subject was scanned on two different scanners on the same day. In all cases, the subject was scanned on both a 1.5T and a 3T scanner. The ground truth difference in brain thickness and volume should be

negligible between same-day measurements. Perfect harmonization therefore should remove any difference between these imaging pairs.

Following previous work (Torbati et al., 2021b, 2021a), we compare differences between the paired images on unharmonized, EB harmonized, and FB harmonized datasets for 22 imaging features that have previously been selected as regions of interest for studying AD (Pölsterl and Wachinger, 2020). The mean difference between paired 3T and 1.5T features (bias) and root mean squared deviation (RMSD), a measurement of variance, are computed in the three datasets. We use paired T-tests across the paired images to identify significant bias with respect to scanner strength in any of the three datasets. We also use paired T-tests on the mean absolute differences in image pairs to compare the harmonization performances of EB ComBat and FB ComBat. Significance thresholds for bias and absolute error are adjusted for multiple tests using Bonferroni correction.

Before harmonization, significant biases are found in 16 of 22 regional measurements, shown in Figure 7. Both EB and FB ComBat harmonization remove any significant bias in all regions. RMSD values are shown in Figure 8. All regions had the lowest RMSD after FB ComBat harmonization. Both FB ComBat and EB ComBat improve error variance in all regions compared to unharmonized data. For all thickness and volume measurements, both harmonization methods (EB and FB ComBat) decrease the mean absolute difference between paired scans across scanners, as shown in Figure 9. Additionally, FB ComBat paired scan values have significantly smaller absolute differences ($p<10^{-14}$) consistently for all 22 volume and thickness measurements compared to EB Combat.

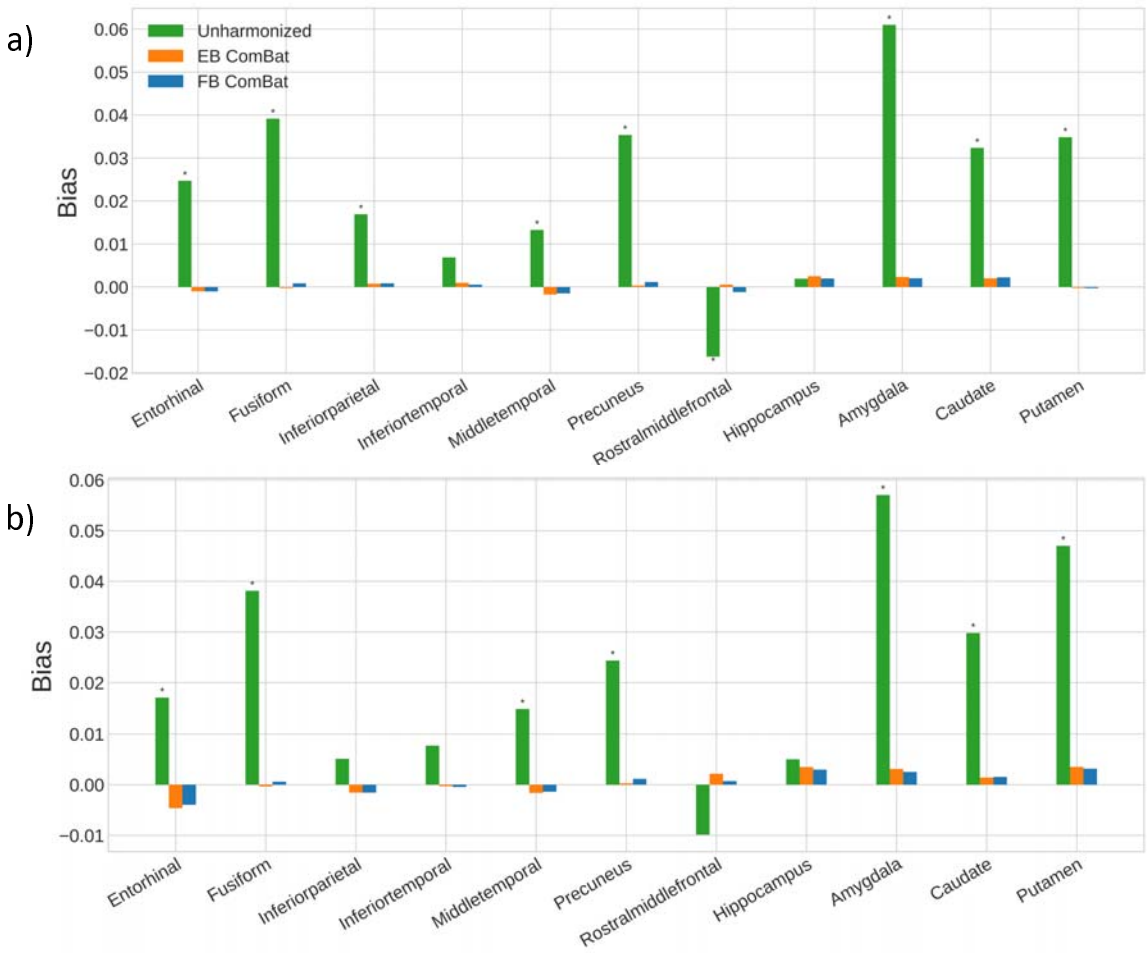


Figure 7: Test-retest Scanner Strength Bias

Bias (mean difference) between 3T and 1.5T test-retest scans in AD-relevant brain regions for left hemisphere (a) and right hemisphere (b) measures. Significant biases (p<0.05) are denoted with (*). Significant bias is present in all but six regions for unharmonized data and is removed in all regions with both EB ComBat and FB ComBat. Biases are normalized with respect to the mean feature value.

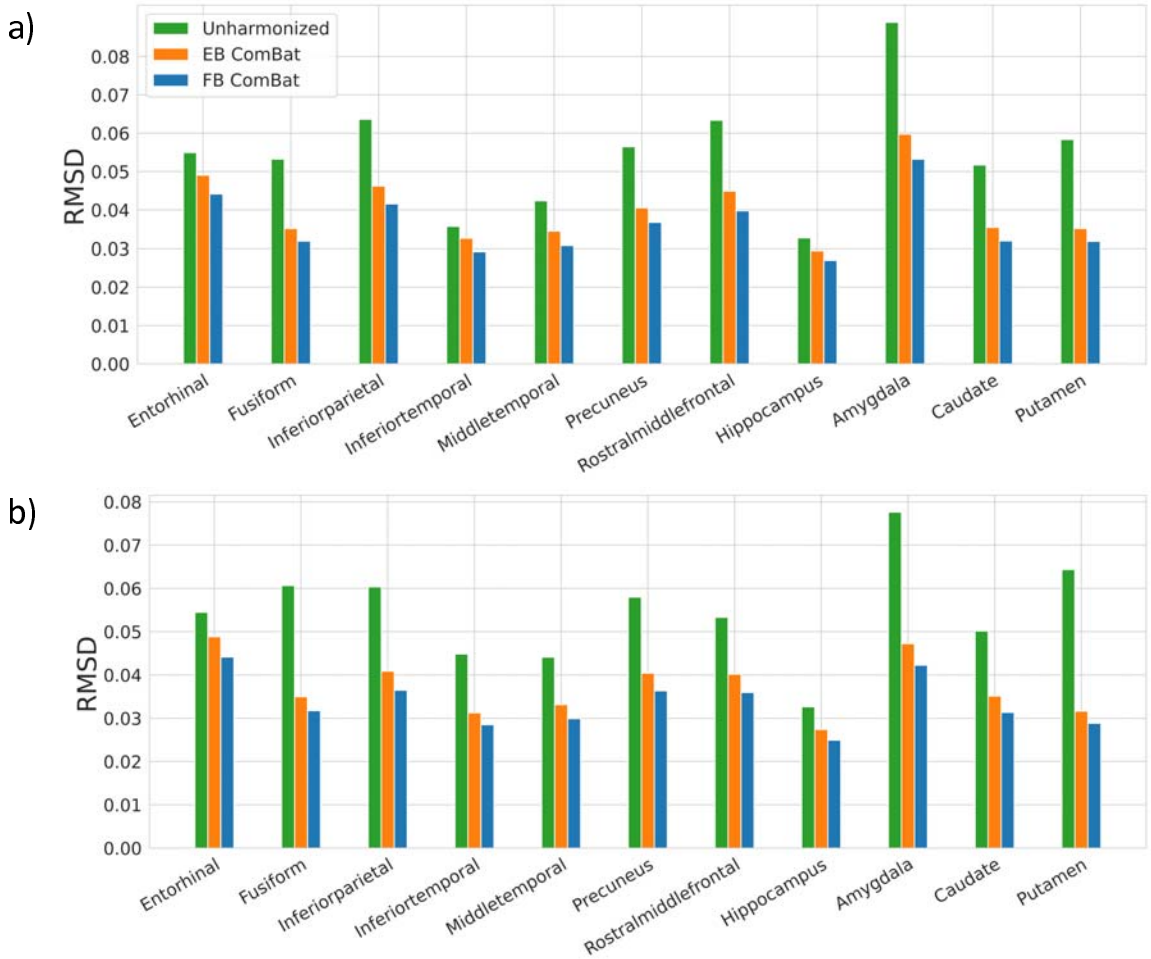


Figure 8: Test-retest Scanner Strength Variance (RMSD)

Variance (Root mean squared deviation) between 3T and 1.5T test-retest scans in left hemisphere (a) and right hemisphere (b) AD-relevant brain regions. Variances are normalized with respect to the mean feature value. All regions have the lowest RMSD after FB ComBat harmonization.

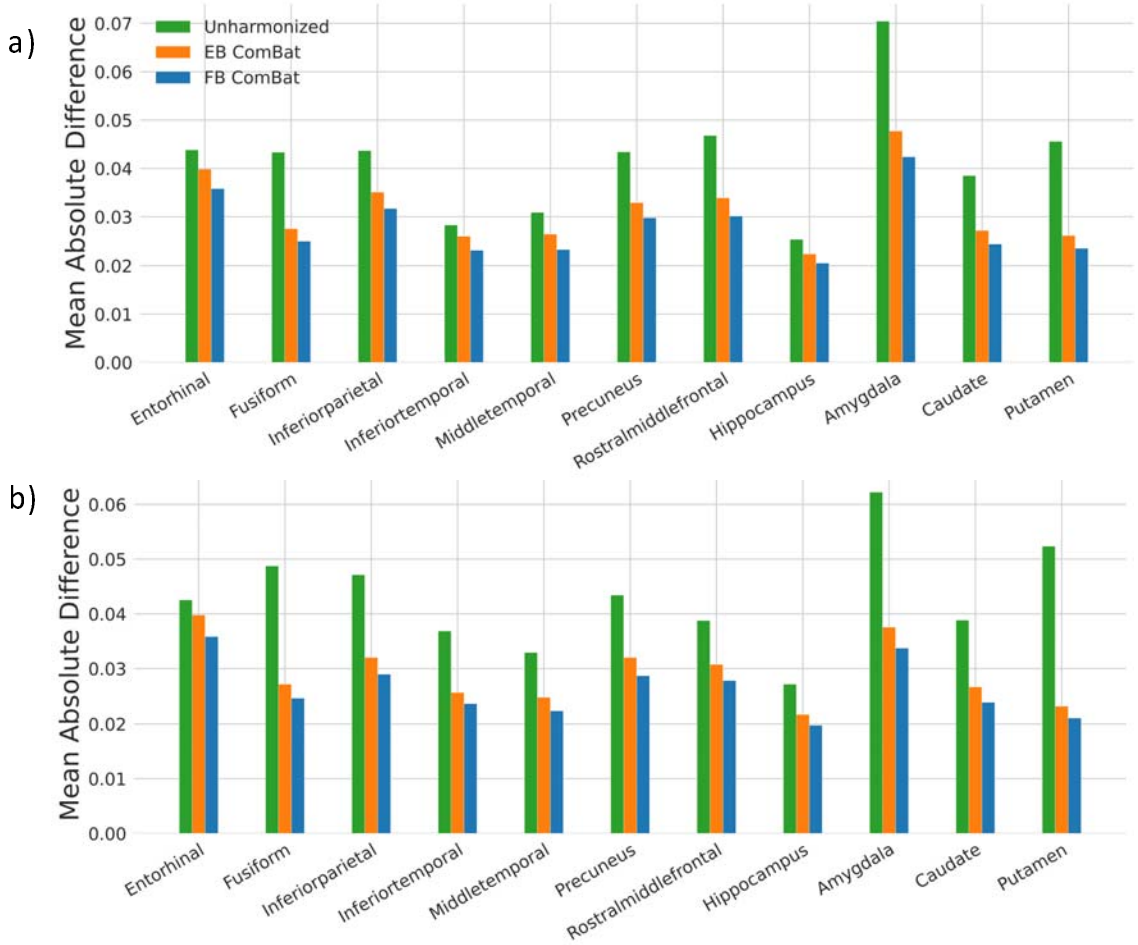


Figure 9: Test-retest Harmonization Error

Absolute mean difference in test-retest scans regions for AD-relevant left hemisphere (a) and right hemisphere (b) measures. FB ComBat paired scan values have significantly smaller absolute differences ($p<10^{-13}$) for all 22 volume and thickness measurements compared to EB ComBat, indicating better harmonization performance on the test-retest scans. Errors are normalized with respect to mean feature value.

3.8 Dataset Augmentation

Dataset augmentation involves artificially increasing the size of a dataset by modifying the existing data or creating synthetic data. Augmentation is often used to increase classifier performance (Wong et al., 2016). In the imaging domain, augmentation is performed by applying transformations to an image that do not change its content or class label. For tabular

data such as FreeSurfer regional thickness and volume features, data augmentation is not as straightforward. We hypothesize that the posterior probability distribution of our harmonized features can be sampled to augment tabular imaging feature datasets. Specifically, each harmonized image can be seen as a *V*-dimensional (corresponding to the number of imaging features) probability distribution. In addition to using the posterior mean of each image, as in typical ComBat harmonization, we also randomly sample from each distribution a number of times. This increases the size of our training set.

We evaluate EB ComBat and FB ComBat as tools for data augmentation by generating additional data samples for a prediction task, namely classifying a patient as having MCI or AD based on imaging features. We perform scanner-stratified train/test splits and repeat 20 times via different random seed numbers. The stratification ensures different scanner groups are used for train and test splits. Images from 75% of scanners are used for training, and images from the remaining 25% of scanners are used for validation. Of the 2408 MCI and AD images from 596 subjects, training sets include between 1609 (66.8%) and 1963 (81.5%) of the images. Five datasets are used in this evaluation. First, the three datasets—unharmonized, EB ComBat, and FB ComBat—are used. Augmented EB ComBat and FB ComBat datasets are then created. For every image in the training split in these two datasets, we draw 100 additional random samples from the posterior distribution of the image's harmonized thickness and volume features. An augmentation that improves classification performance is desirable. We use the AUROC of AD classification to evaluate the samples generated from EB ComBat and FB ComBat.

Predictive performance for classifying MCI versus AD patients with and without augmentation is shown in Table 3. Without any augmentation, FB ComBat outperformed EB ComBat ($p=0.002$). Using FB ComBat harmonization with posterior distribution resampling augmentation results in the highest AUROC. This is significantly higher than FB harmonization without augmentation ($p<0.0001$) and higher than EB harmonization with augmentation ($p<0.0001$).

	AUROC
Unharmonized	0.772 (0.028)
EB ComBat	0.777 (0.026)
FB ComBat	0.796 (0.030)
EB Combat with posterior resampling	0.799 (0.027)
FB ComBat with posterior resampling	0.821 (0.022)

Table 3: Disease Prediction Results

Evaluation of MCI versus AD classification with and without augmentation (sampling from the posterior distribution of harmonized imaging features). Area under the receiver operating characteristic curve (AUROC) is greatest in FB ComBat with augmentation.

As with age (see Section 3.5), understanding scanner distributions of disease diagnosis can reveal a potential source of confounding for scanner and covariate effects. We use chi-squared test of independence to check for an association between scanner and diagnosis. We find no significant association between scanner and diagnosis both at the individual scanner level ($p=0.91$) and field strength/manufacturer level ($p=0.38$).

3.9 Region-Level Uncertainty

To identify brain measurements most and least prone to measurement uncertainty, we use the posterior variance of the FB-harmonized imaging features. To compare uncertainty across regions of interest in the brain, it is necessary to normalize posterior variance, due to the difference in measurement scales (thickness vs. volume). Additionally, different regions also have different variances across the population, even after adjusting for mean thickness or volume. We devise a normalized uncertainty value as the ratio of average variance within all *individual* measurements (MSW- Mean Squared Within) to the variance of individual measurement posterior means among the population (MSA- Mean Squared Among) for each regional imaging feature (v):

$$Uncertainty_v = \frac{MSW_v}{MSA_v} \tag{6}$$

We compute relative uncertainty values for thickness and volume measurements of regions on the Desikan-Killiany cortical atlas (Desikan et al., 2006) and Freesurfer Volumetric Segmentation Atlas (Fischl et al., 2002).

Region-level uncertainty results are shown in Figure 10. Subcortical volumes generally have lower uncertainty values compared to cortical thickness. Among cortical thickness features, regions in the temporal lobe have lower overall uncertainty. Left and right pericalcarine thickness have the highest overall uncertainty among regional imaging features. Among subcortical volume features, mid-anterior and posterior regions of the corpus callosum have the highest posterior uncertainty.

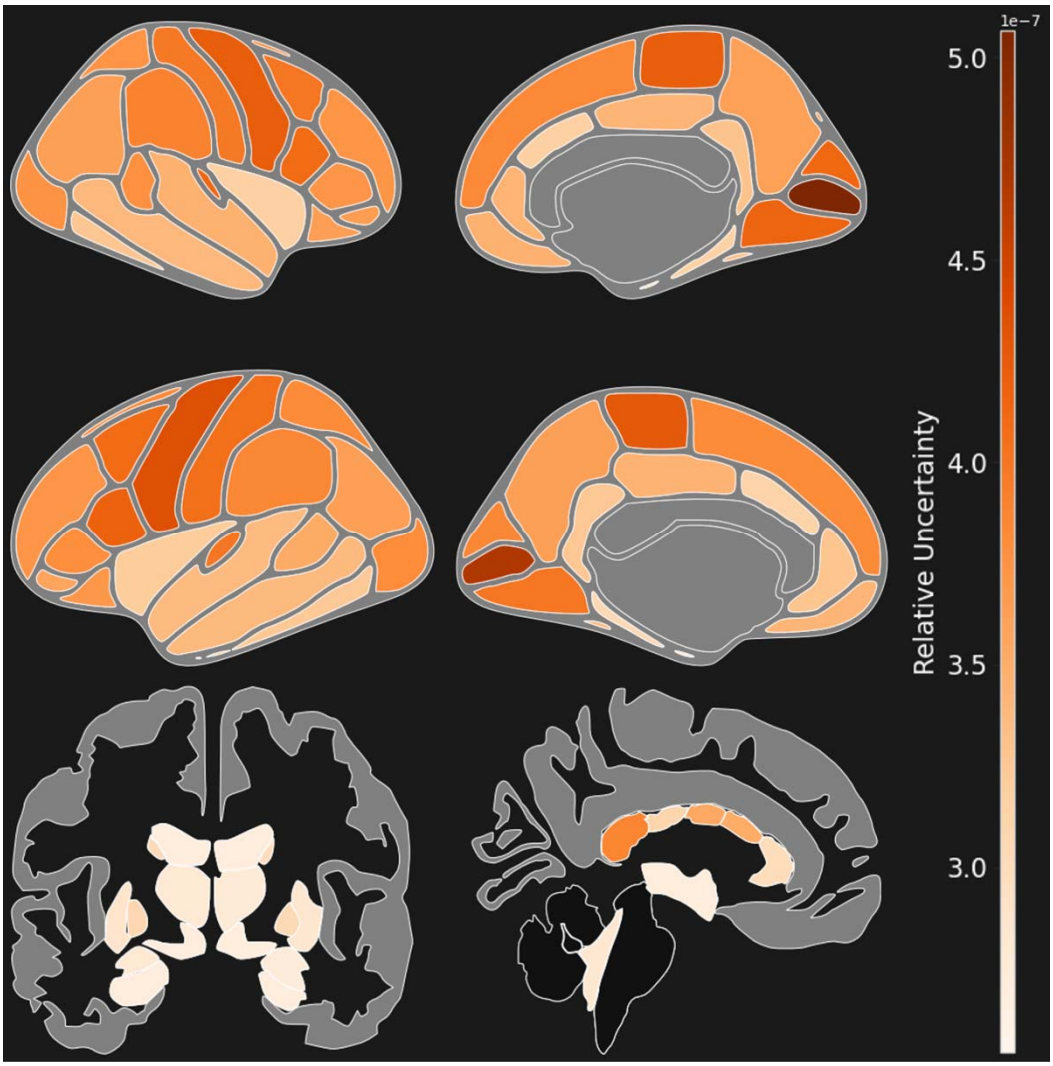


Figure 10: Brain-Wide Uncertainty

Regional uncertainty, defined as the ratio of variance within the posterior distribution of individual measurements (MSW- Mean Squared Within) to the variance among the population (MSA- Mean Squared Among). Uncertainty is shown in cortical thicknesses and subcortical volumes. Uncertainty is higher in cortical thickness values compared to subcortical volumes. Left and right pericalcarine thickness have the highest overall uncertainty of any region.

3.10 Uncertainty in Statistical Analysis

Statistical association models like least squares regression, which can be used to model brain regional associations with disease (Wang et al., 2011), consider all data points as equally reliable when minimizing the error term. In the brain imaging domain, this assumption may not

hold when scanner reliability introduces uncertainty in a multi-site study. We propose using the posterior variance of harmonized measurements as a measure of uncertainty. Data for a regression model can then be weighted by the inverse of this variance, resulting in a model fit that accounts for measurement uncertainty (Aitken, 1936).

We demonstrate the difference between ordinary least squares (OLS) and uncertainty-weighted least squares (WLS) by testing for association between different brain region measurements with Alzheimer's disease. We use the model:

$$y_{ijv}^{FB}(t) = X_j^{AD}(t)\beta_v^{AD} + X_j^{covar}(t)\beta_v^{covar} + \varepsilon \tag{7}$$

with null and alternative hypotheses:

$$H_0: \beta_v^{AD} = 0$$
$$H_A: \beta_v^{AD} \neq 0$$

where $X_j^{AD}(t)$ is an indicator variable for whether a patient has an Alzheimer's disease diagnosis at time t, β_v^{AD} is the association of Alzheimer's disease with feature v, $X_j^{covar}(t)\beta_v^{covar}$ is the covariate term, and ε is the random error term. Imaging features $y_{ijv}^{FB}(t)$ are adjusted by mean thickness or intracranial volume for regional thickness and regional volume measurements, respectively. A significant β_v^{AD} indicates an association between the feature v and Alzheimer's disease, adjusting for covariates (age and sex).

We test for association between 104 brain regions (from Desikan-Killiany and Freesurfer Volumetric Segmentation atlases) and Alzheimer's disease using OLS (baseline) and WLS weighted by the reciprocal of measurement posterior variance and examine differences in findings between the two methods.

Significant associations (after Bonferroni correction) between brain regions and Alzheimer's disease are shown in Figure 11. Several regions show differing significance when using WLS versus OLS regression including right lateral occipital thickness, left rostral anterior cingulate thickness, and left caudate volume. We also show results from using EB ComBat posterior variance in Supplementary Figures S2,S3.

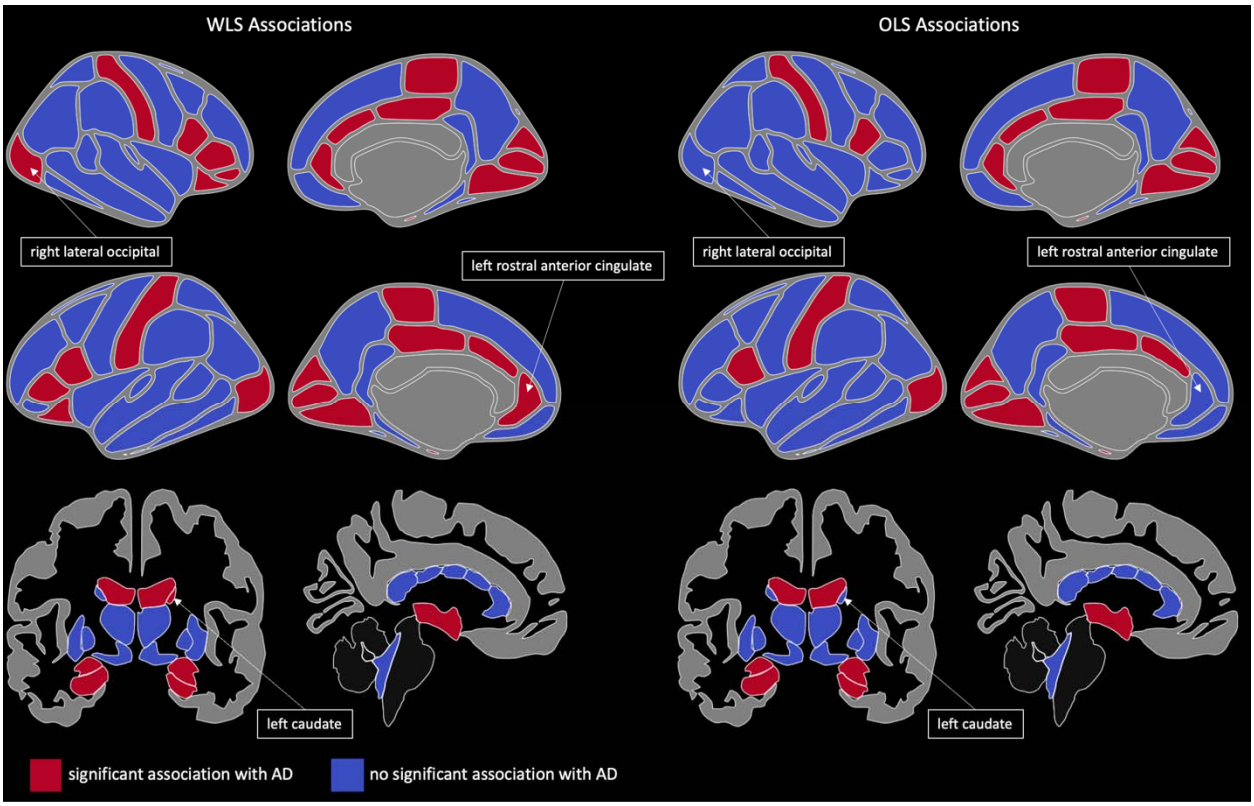


Figure 11: Cortical Thickness and Subcortical Volume Associations with AD

Cortical thickness and subcortical volume regions associated with Alzheimer's disease using Ordinary Least Squares (OLS) and Weighted Least Squares (WLS) weighted by the reciprocal of posterior measurement variance. Right lateral occipital, left rostral anterior cingulate thickness, and left caudate volume are highlighted as regions with different associations using OLS and WLS models.

3.11 Simulation Study

Following prior work (Beer et al., 2020), we conducted a simulation analysis to study to impact of harmonization on the ability to identify disease effects on neuroimaging features. First, we segregate the AD and CN patients from ADNI and fit the following linear mixed-effect model to each imaging feature separately:

$$y_j = \beta_{Age}\,(Age) + \beta_{AD}\,I(Diagnosis = AD) + \beta_{AD \times Age}\,(Age)\,I(Diagnosis = AD) + \beta_{Sex}\,I(Sex = M) + \eta_j + \gamma_i + \alpha + \varepsilon \tag{8}$$

where y_j is the feature value for subject *j*, β are coefficients for covariates (age, AD, age x AD interaction, and sex), η_j is a random subject-specific intercept, γ_i is a scanner-specific intercept, α is a feature intercept, and ε is the residual. We select six features (right hippocampus volume, left parahippocampal thickness, right accumbens area volume, left medial orbitofrontal thickness, right amygdala volume, and left thalamus volume) with significant β_{AD} and $\beta_{AD \times Age}$ coefficient values and choose these as "alternative hypothesis features". The remaining features are considered null features.

We then isolate the CN ADNI patients, randomly assign half to the AD group, manually add the identified AD and AD x age effects to the AD group patients, and harmonize the semi-synthetic dataset using EB ComBat and FB ComBat. We repeat the process of random group assignment and harmonization 500 times. Finally, we re-fit the linear mixed effect model (without the scanner additive term) on all 500 sets of unharmonized, EB ComBat-harmonized, and FB ComBat-harmonized data.

We are particularly interested in the ability to recover the AD effect and AD × Age interaction effect from alternative hypothesis features while minimizing Type I Error from the null features. Thus β_{AD} and $\beta_{AD \times Age}$ estimates from harmonized data should be near their true values; coefficient p-values for null features should be insignificant, and coefficient p-values for alternative hypothesis features should be significant.

FB ComBat generally reduces type I error and improves coefficient estimates for null features compared to EB ComBat; both methods still underperform in most null feature evaluations compared to unharmonized data. Estimated coefficient values for β_{AD} and $\beta_{AD \times Age}$ in null features are shown in Figure 12a. Unharmonized data yields the lowest β_{AD} mean absolute error for null features, while EB ComBat results in greater β_{AD} error compared to FB ComBat (*p=0.008*). No difference between $\beta_{AD \times Age}$ error in EB ComBat and FB ComBat for null features is found (*p=0.28*), while EB and FB ComBat both have lower $\beta_{AD \times Age}$ error compared to unharmonized data (*p<0.004*). β_{AD} and $\beta_{AD \times Age}$ Type I error rates for null features are shown in Figure 12b. Unharmonized null features have a significantly lower Type I error rate for β_{AD} and $\beta_{AD \times Age}$ compared to EB and FB ComBat ($p<10^{-5}$). β_{AD} Type I error rate is greater in EB ComBat than FB ComBat ($p<10^{-20}$); no significant difference for $\beta_{AD \times Age}$ error rate is found between EB ComBat and FB ComBat (*p=0.24)*.

EB ComBat leads to the highest sensitivity to β_{AD} for alternative hypothesis features, while FB ComBat data has the highest sensitivity to $\beta_{AD \times Age}$ interaction effects. Error in coefficient value estimates for the six alternative hypothesis features do not significantly differ across harmonization measures. Coefficient estimates are shown in Supplementary Figure S1. P-value distributions for these features are shown in Figure 13. The β_{AD} p-values for all six features are more significant in EB ComBat than FB ComBat ($p<10^{-4}$) and are more significant in both EB and FB ComBat compared to unharmonized data ($p<10^{-5}$). The $\beta_{AD \times Age}$ p-values for five of six features are more significant in FB ComBat than EB ComBat (*p<0.01*). All $\beta_{AD \times Age}$ p-values are more significant in FB ComBat than in unharmonized data, and four of six p-values are more significant in EB ComBat than in unharmonized data.

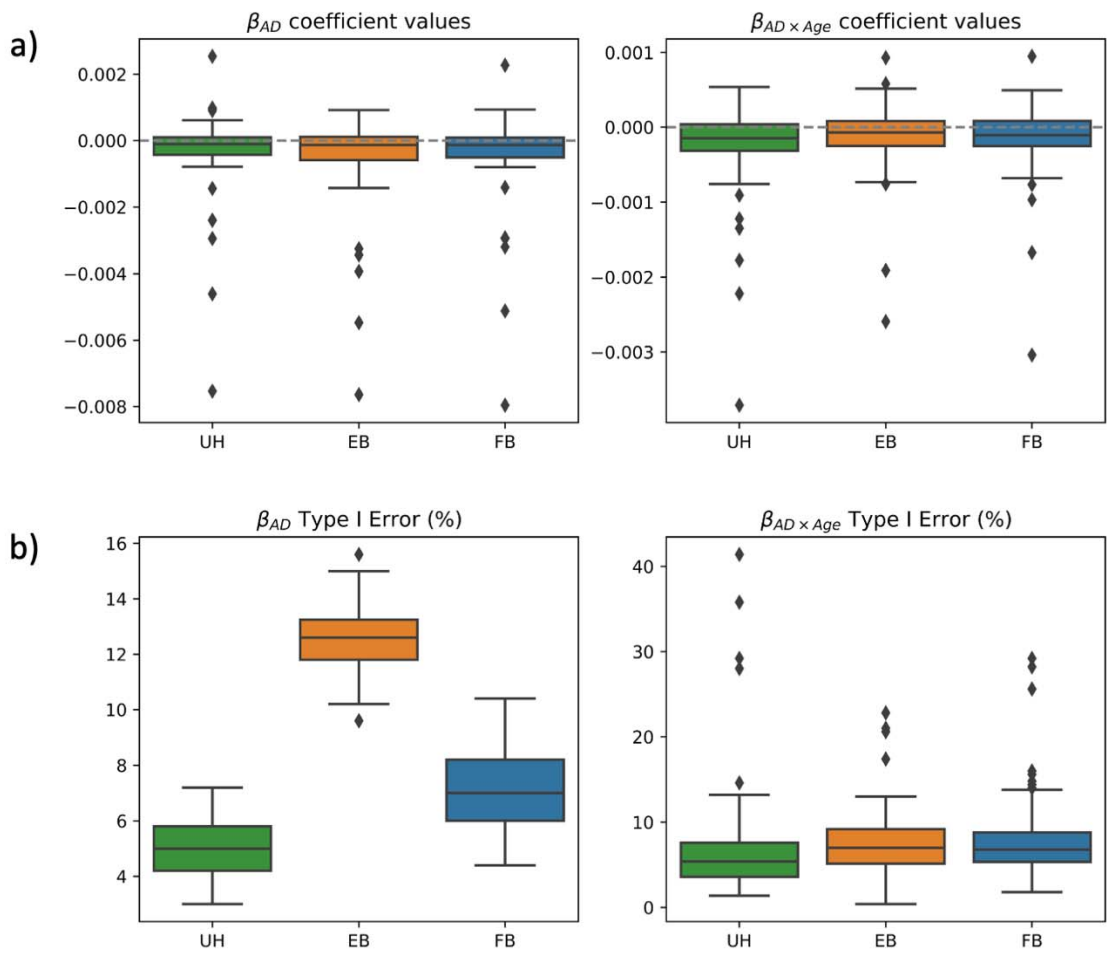
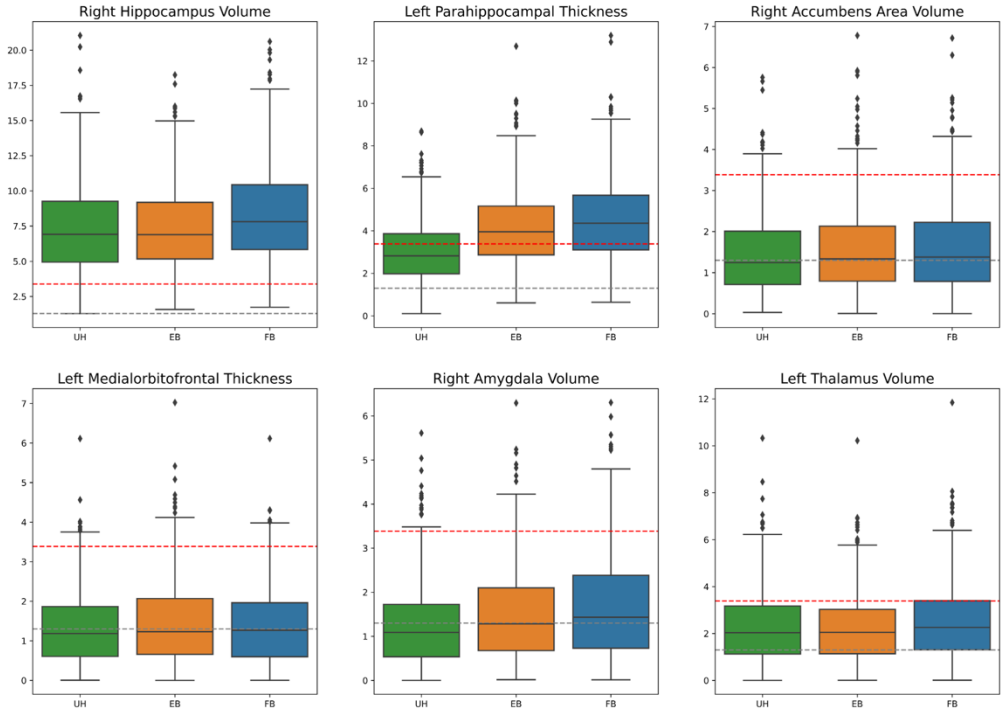


Figure 12: Simulation Study Null Feature Coefficients

Simulation data a) Coefficient values and b) Type I error rate for null features (zero AD and AD x Age effect sizes) from unharmonized, EB ComBat, and FB ComBat data. Coefficient values in a) are shown across all null features and simulations and are normalized with respect to feature means. Type I Error rate is calculated as the percentage of Type I error in 500 simulations for a given feature.

a)


Right Hippocampus Volume
Left Parahippocampal Thickness
Right Accumbens Area Volume
Left Medialorbitofrontal Thickness
Right Amygdala Volume
Left Thalamus Volume
UH
EB
FB

b)

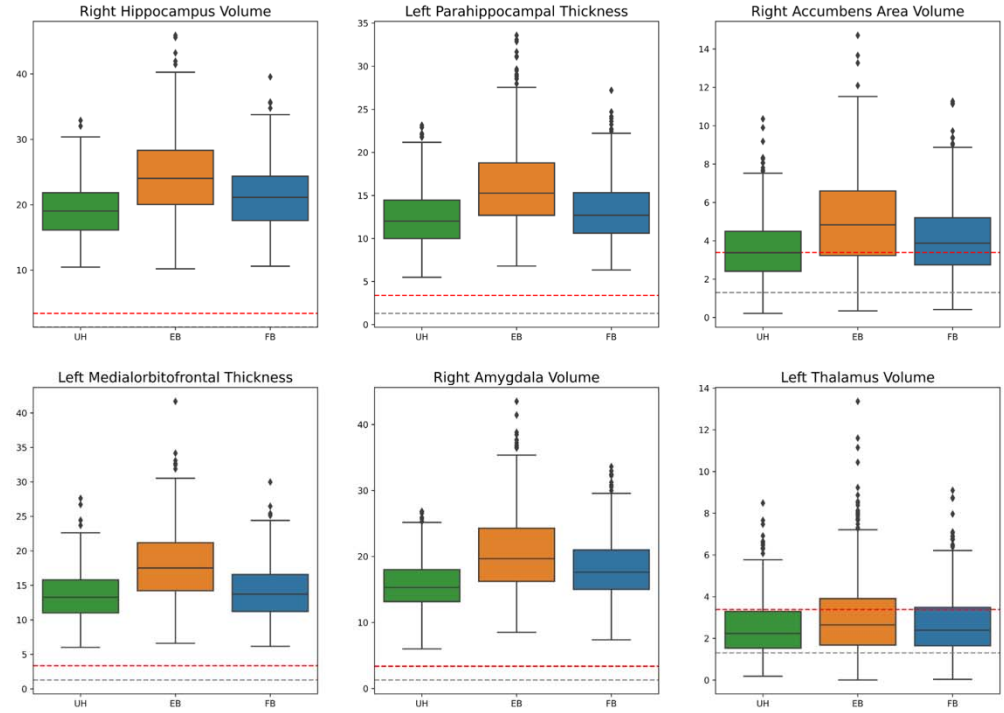

Right Hippocampus Volume
Left Parahippocampal Thickness
Right Accumbens Area Volume
Left Medialorbitofrontal Thickness
Right Amygdala Volume
Left Thalamus Volume
UH
EB
FB

Figure 13: Simulation p-value Disease Effect Distribution

Distribution of -$\log_{10}$ p-values in the true (alternative hypothesis) disease effect features for a) AD coefficient β_{AD} and b) AD x Age coefficient $\beta_{AD \times Age}$. The dotted grey line is at p=0.05, and the dotted red line is at the Bonferroni-corrected significant p-value (p=0.0004).

3.12 Sensitivity Analysis

We perform sensitivity analysis to explore the impact of prior distributions on harmonization. In one set of experiments, we vary the prior distributions τ_i parameter to study whether a strong prior on the presence or absence of scanner effect pooling affects the harmonization performance. We also check whether making the baseline FB ComBat priors stronger changes harmonization. A more in-depth description of our sensitivity analysis procedure can be found in Supplementary Section S4. Overall, we find that harmonization performance was fairly robust to various priors over the hyperparameters. Of note, a strong prior against pooling (high τ_i) leads to modest increases in significant additive scanner effects after harmonization. Stronger priors on all parameters, compared with weakly informative priors, lead to increases in significant multiplicative scanner effects after harmonization.

4. Discussion

Scanner harmonization, especially for image-derived thickness and volume features, is an important step of brain MRI pre-processing to reduce noise and potential biases. We built on ComBat, a popular statistical harmonization approach, by proposing a new Fully Bayes method for inference which leads to improved harmonization and additional ComBat use cases. While the FB method maintained the robustness of the EB approach, we also found that FB ComBat yielded harmonized features with greater retention of biologically relevant information and smaller differences in test-retest subjects. EB ComBat, on the other hand, removed scanner strength information more effectively and takes only a fraction of the runtime compared to FB ComBat. FB ComBat produced more realistic posterior distributions and uncertainty quantification, which is important for individualized disease diagnosis (Liu et al., 2020) and group-level statistical analysis (Aitken, 1936). Fully Bayesian inference allows us to draw

samples from a rich posterior distribution, which we used to augment a dataset to improve Alzheimer's Disease classifier performance. We also used the variance of the posterior distributions to perform a more principled analysis of brain regional associations with Alzheimer's disease (AD) and identify features that are more prone to measurement uncertainty.

Two important metrics of harmonization are that, first, biological information is maintained after harmonization and, second, that scanner-related information and other confounding nuisance are successfully removed. To evaluate EB and FB ComBat's ability to retain biological information, we trained models to predict age and AD status, and we evaluated their performance on datasets with unharmonized versus harmonized data. Harmonizing brain imaging feature data using FB ComBat resulted in stronger age and AD classification performance than EB ComBat, suggesting that FB ComBat was able to more effectively retain biologically-relevant brain structural information. Our age regression findings align with a recent study finding that ComBat harmonization methods tend to either modestly obscure or not affect biological effects compared to unharmonized data (Richter et al., 2022). Conversely, our AD classification study showed that FB ComBat improved disease identifiability. Richter et al. (2022) designed a more prospective study where subjects were scanned at an initial time point, then several years later at both the initial scanner and a separate scanner, allowing direct comparison of imaging feature change over time with separate scanners compared to a ground truth change. FB ComBat should be evaluated in a similar design against other ComBat approaches to further study its impact on biological and non-biological effect identifiability.

We also trained a model to predict scanner strength from harmonized imaging features. In this task, lower classification accuracy was ideal, indicating that information about scanner strength was removed during harmonization. EB ComBat performed best on this task, indicating that it removed more non-biological information related to scanner strength compared to FB ComBat. Both EB ComBat and FB ComBat also similarly removed detectable effects from scanner

strength and manufacturer. Furthermore, we showed that the harmonization performance of FB ComBat was very stable when varying the prior distributions.

One concern with covariate and scanner prediction on imaging features is the potential confounding of covariate and scanner effects. This complicates the identifiability of covariate and scanner model parameters, which can lead to the inadvertent removal of covariate effects or preservation of scanner effects. Additionally, when predicting covariates to evaluate harmonization, models could be learning scanner effects. Models predicting scanner strength may also be capturing covariate effects. Our ANOVA and chi-squared tests found that, among scanner strength and manufacturer, two possible contributors to scanner effect differences (Han et al., 2006), no significant disease or age differences exist. However, some age difference was found across individual scanners. This indicates the possibility that scanner effects related to some factor besides magnetic field strength or manufacturer could be learned by an age prediction model. Due to a large number of scanners, pairwise follow-up group comparisons to examine the extent of age differences were infeasible. Future large-scale imaging studies should ensure that covariates (including diagnostic groups) are balanced across scanners so that scanner and covariate effect confounding is eliminated. Furthermore, in studies with fewer sites, traveling subject-based study design and inferring the ComBat model on only traveling subjects can remove the impact of site-based sampling bias and improve harmonization performance (Maikusa et al., 2021; Yamashita et al., 2019). Additionally, researchers using multiple datasets should carefully consider population differences when pooling and harmonizing the datasets together.

We checked whether the harmonization methods would bring same-day, same-subject imaging features from different scanners closer together. This metric should be seen as a “ground-truth” check, as brain anatomy should not change in such a short time. Our FB ComBat harmonization resulted in the largest reduction in imaging feature measurement difference between repeat scans for all tested regions, indicating that our model specification removed scanner artifact most effectively for this repeat-scan subset of patients.

We further performed a simulation study to determine how both harmonization methods affect the identifiability of disease signals in brain imaging features. Both methods modestly improved identifiability of true effects. EB ComBat led to highest sensitivity in AD effect while FB ComBat led to highest sensitivity in AD × Age interaction effect. We also found that, while both methods increased type I error rate compared to unharmonized data, EB ComBat led to a type I error rate inflation in AD effect estimation compared to FB ComBat. The AD × Age interaction term did not have as pronounced type I error rate inflation, similar to prior simulation studies of EB ComBat (Beer et al., 2020). Further research is still needed to determine strategies for mitigating type I error in harmonization. One potential approach may be to only apply the harmonization transformation when a chosen credible interval (e.g. 95%) of a scanner error term lies above or below 0 (for additive error) or 1 (for multiplicative error). We leave this approach and additional exploration of type I error mitigation for future research.

We also presented the novel use of generative harmonization models to make a downstream classification model more robust with respect to limited training data. Large-scale imaging datasets have grown but are still relatively small in the medical field due to acquisition costs and privacy concerns. Datasets may not contain sufficient variation to train robust classifiers without augmentation. Several studies have examined the effect of image-level augmentations such as random affine transformations, elastic deformation, intensity shifts, or GAN-based image generation on voxel-level tasks such as tumor segmentation or disease classification (Dufumier et al., 2021; Li et al., 2020; Nalepa et al., 2019). However, as far as we are aware, augmentation of regional thickness and volume measurements has not been studied. We presented a new augmentation method that uses the EB and FB harmonization models' posterior distributions for a rich data generation tool that can improve classifier performance. Our augmentation method draws from our post-harmonization uncertainty regarding measurement error of an image. While EB harmonization generates a posterior distribution, it underestimates posterior uncertainty. Sampling from the EB ComBat posterior for

augmentation produces data with less variation, which may explain why FB ComBat performed better than EB ComBat in the augmentation task.

We explored posterior distributions to determine which imaging regions have the most uncertainty, compared to overall population variance. Subcortical volume measurements and temporal thickness were generally less prone to uncertainty than other cortical thickness measurements. The difference in uncertainty between cortical thickness versus structural volumes may be due to the difficulty of surface parcellation compared to segmentation. Gyral-based parcellation, used in FreeSurfer, is inherently difficult because gyri are connected without a clear visible boundary between connected regions (Meng et al., 2015). Our results suggest that subcortical volume measurements may be more reliable than cortical thickness, a finding verified by recent test-retest analysis of Freesurfer measures (Hedges et al., 2022).

Finally, we propose the use of uncertainty-based measurement weighting in association tests. Commonly used models such as ordinary least squares assume that regressors have no error and response variables have uniform uncertainty. We demonstrated that tests involving regional association with Alzheimer's Disease can vary depending on whether uncertainty is considered (using weighted least squares regression), or it is ignored (using ordinary least squares regression). FB ComBat's uncertainty measurement should be used for principled downstream statistical analysis. EB ComBat underestimates uncertainty, so we do not suggest using it for uncertainty-aware downstream tasks. Further work might investigate uncertainty-aware models for predictive tasks such as Alzheimer's disease conversion. Causal discovery that incorporates measurement error (Zhang et al., 2017) is another potential area for further exploration that may benefit from using FB ComBat measurement uncertainty. Further work might also investigate the impact of different experimental settings, such as the number of features included in FB ComBat harmonization, on posterior measurement uncertainty.

We suggest two important practical guidelines for the use of FB ComBat. First, imaging features should be standardized (i.e. z-scored) before inference time. This prevents misspecification in

the hierarchical model parameters which assume that scanner effects derive from a common distribution. Conversely, EB ComBat performs standardization based on fixed effect estimates from the first stage of inference so standardization before harmonization is not necessary. Next, we encourage the use of weakly informative priors, especially in scanner-specific parameters $\mu_i, \tau_i, \lambda_i, \theta_i$. For example, the parameter τ_i might use a distribution like *InverseGamma(2,0.5)*. This constrains 91% of the prior density to within 0 and 1. In other words, we allow for a 9% prior probability that the standard deviation of additive effects within a scanner varies by more than one feature-standard deviation. The use of weakly informative prior distributions makes the MCMC sampler more efficient while having a minimal impact on the posterior distribution.

Several limitations exist for ComBat harmonization, and for large-scale Bayesian inference. Harmonization evaluation is inherently limited because ground truth imaging feature measurements are unknown. Developing a quantitative metric to evaluate harmonization algorithms is challenging. Test-retest experiments where traveling subjects are scanned on two scanners in a short time, as well as simulation studies where known effects are added to patient subsets, are our best ground-truth tools for evaluating harmonization. Metrics for explicitly studying the retention of biological information and removal of scanner information are also informative and include tasks like age prediction and scanner strength prediction. However, confounding between biological and scanner factors may limit the usefulness of these metrics. Additionally, these tasks are only approximations and do not include all possible biological and non-biological variables of interest, many of which are unobserved. Including additional covariates in the ComBat model, such as total intracranial volume, may improve harmonization performance (Pomponio et al., 2020). Additionally, the linear nature of our model may leave out important non-linear covariate and scanner effects. Extensions of EB ComBat have shown improved performance by modeling scanner covariance effects (Chen et al., 2021) and non-linear covariate effects (Pomponio et al., 2020). While our work compares EB and FB approaches to Longitudinal ComBat (Beer et al., 2020), more complicated FB models should be compared against their corresponding EB ComBat extensions. These extensions are

straightforward to implement in the FB setting; the probabilistic programming approach used for FB ComBat just requires specification of the model, then sampling is done automatically. FB ComBat should also be studied for other imaging features that have benefitted from ComBat harmonization such as PET standardized uptake values (Orlhac et al., 2018), functional MRI (Yu et al., 2018), and white matter imaging features such as fractional anisotropy, mean diffusivity and regional volumes (Fortin et al., 2017; Richter et al., 2022). Finally, FB inference is inherently slower than an EB approach. EB ComBat uses expectation-maximization optimization while FB ComBat relies on much slower MCMC sampling. With large models like ComBat, the difference in inference speed is significant. If speed or computational resources are a concern (e.g. if datasets are constantly updated and re-harmonized), EB ComBat may be preferred. However, harmonization is generally performed just once before imaging features are analyzed, so we expect that in many cases the uncertainty quantification and improved harmonization performance for most metrics outweigh the efficiency drawback.

5. Conclusion

We have compared EB and FB approaches to ComBat brain MRI feature harmonization. FB harmonization performed slightly better in most harmonization tasks. We also demonstrated that the posterior distributions of FB harmonized data should be used for any study where the accurate estimation of uncertainty is important. We provided three examples, namely data augmentation, association tests, and brain-wide feature uncertainty quantification, which utilize the posterior distribution given by FB ComBat. The code for FB ComBat is available at *https://github.com/batmanlab/BayesComBat*.

Acknowledgements

This work was supported by the Pennsylvania Department of Health [grant number 4100087331], National Institutes of Health [grant numbers R01HL141813, 1R01-AG063752, 5T15LM007059], the National Science Foundation [grant number 1839332], Tripod+X, and the SAP SE. This work used the Bridges-2 system, which is supported by NSF award number OAC-1928147 at the Pittsburgh Supercomputing Center (PSC). Data collection and sharing for this

project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Declaration of Interest: none

References

ADNI, 2016. Alzheimer's Disease Neuroimaging Initiative (ADNI) DATA USE AGREEMENT.

Aitken, A.C., 1936. IV.—On Least Squares and Linear Combination of Observations. Proc. R. Soc. Edinburgh 55, 42–48. https://doi.org/10.1017/S0370164600014346

Bartlett, E.A., DeLorenzo, C., Sharma, P., Yang, J., Zhang, M., Petkova, E., Weissman, M., McGrath, P.J., Fava, M., Ogden, R.T., Kurian, B.T., Malchow, A., Cooper, C.M., Trombello, J.M., McInnis, M., Adams, P., Oquendo, M.A., Pizzagalli, D.A., Trivedi, M., Parsey, R. V., 2018. Pretreatment and early-treatment cortical thickness is associated with SSRI

treatment response in major depressive disorder. Neuropsychopharmacology 43, 2221–2230. https://doi.org/10.1038/s41386-018-0122-9

Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., Linn, K.A., 2020. Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. Neuroimage 220, 117129. https://doi.org/10.1016/j.neuroimage.2020.117129

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Carlin, B.P., Louis, T.A., 2000. Empirical Bayes: Past, Present and Future. J. Am. Stat. Assoc. 95, 1286–1289. https://doi.org/10.1080/01621459.2000.10474331

Casey, B.J., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C.A., Wager, T.D., Banich, M.T., Speer, N.K., Sutherland, M.T., Riedel, M.C., Dick, A.S., Bjork, J.M., Thomas, K.M., Chaarani, B., Mejia, M.H., Hagler, D.J., Daniela Cornejo, M., Sicat, C.S., Harms, M.P., Dosenbach, N.U.F., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J.R., Kuperman, J.M., Fair, D.A., Dale, A.M., 2018. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 32, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001

Chen, A.A., Beer, J.C., Tustison, N.J., Cook, P.A., Shinohara, R.T., Shou, H., 2021. Mitigating site effects in covariance for machine learning in neuroimaging data. Hum. Brain Mapp. 1179–1195. https://doi.org/10.1002/hbm.25688

Conover, W.J., Johnson, M.E., Johnson, M.M., 1981. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23, 351–361. https://doi.org/10.1080/00401706.1981.10487680

Cury, C., Scelsi, M.A., Toro, R., Frouin, V., Artiges, E., Grigis, A., Heinz, A., Lemaitre, H., Martinot, J.L., Poline, J.B., Smolka, M.N., Walter, H., Schumann, G., Altmann, A., Colliot, O., 2020. Genome wide association study of incomplete hippocampal inversion in adolescents. PLoS One 15, 1–18. https://doi.org/10.1371/journal.pone.0227355

Cury, C., Toro, R., Cohen, F., Fischer, C., Mhaya, A., Samper-González, J., Hasboun, D., Mangin, J.F., Banaschewski, T., Bokde, A.L.W., Bromberg, U., Buechel, C., Cattrell, A., Conrod, P.,

Flor, H., Gallinat, J., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Lemaitre, H., Martinot, J.L., Nees, F., Paillére Martinot, M.L., Orfanos, D.P., Paus, T., Poustka, L., Smolka, M.N., Walter, H., Whelan, R., Frouin, V., Schumann, G., Glaunés, J.A., Colliot, O., 2015. Incomplete hippocampal inversion: A comprehensive MRI study of over 2000 subjects. Front. Neuroanat. 9, 1–12. https://doi.org/10.3389/fnana.2015.00160

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, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Dima, D., Modabbernia, A., Papachristou, E., Doucet, G.E., Agartz, I., Aghajani, M., Akudjedu, T.N., Albajes-Eizagirre, A., Alnæs, D., Alpert, K.I., Andersson, M., Andreasen, N.C., Andreassen, O.A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., Bonvino, A., Boomsma, D.I., Borgwardt, S., Bourque, J., Brandeis, D., Breier, A., Brodaty, H., Brouwer, R.M., Buitelaar, J.K., Busatto, G.F., Buckner, R.L., Calhoun, V., Canales-Rodríguez, E.J., Cannon, D.M., Caseras, X., Castellanos, F.X., Cervenka, S., Chaim-Avancini, T.M., Ching, C.R.K., Chubar, V., Clark, V.P., Conrod, P., Conzelmann, A., Crespo-Facorro, B., Crivello, F., Crone, E.A., Dale, A.M., Davey, C., de Geus, E.J.C., de Haan, L., de Zubicaray, G.I., den Braber, A., Dickie, E.W., Di Giorgio, A., Doan, N.T., Dørum, E.S., Ehrlich, S., Erk, S., Espeseth, T., Fatouros-Bergman, H., Fisher, S.E., Fouche, J.P., Franke, B., Frodl, T., Fuentes-Claramonte, P., Glahn, D.C., Gotlib, I.H., Grabe, H.J., Grimm, O., Groenewold, N.A., Grotegerd, D., Gruber, O., Gruner, P., Gur, R.E., Gur, R.C., Harrison, B.J., Hartman, C.A., Hatton, S.N., Heinz, A., Heslenfeld, D.J., Hibar, D.P., Hickie, I.B., Ho, B.C., Hoekstra, P.J., Hohmann, S., Holmes, A.J., Hoogman, M., Hosten, N., Howells, F.M., Hulshoff Pol, H.E., Huyser, C., Jahanshad, N., James, A., Jernigan, T.L., Jiang, J., Jönsson, E.G., Joska, J.A., Kahn, R., Kalnin, A., Kanai, R., Klein, M., Klyushnik, T.P., Koenders, L., Koops, S., Krämer, B., Kuntsi, J., Lagopoulos, J., Lázaro, L., Lebedeva, I., Lee, W.H., Lesch, K.P., Lochner, C., Machielsen, M.W.J., Maingault, S., Martin, N.G., Martínez-Zalacaín, I., Mataix-Cols, D., Mazoyer, B., McDonald, C., McDonald, B.C., McIntosh, A.M., McMahon, K.L., McPhilemy, G., Menchón, J.M., Medland, S.E., Meyer-Lindenberg, A., Naaijen, J., Najt,

P., Nakao, T., Nordvik, J.E., Nyberg, L., Oosterlaan, J., de la Foz, V.O.G., Paloyelis, Y., Pauli, P., Pergola, G., Pomarol-Clotet, E., Portella, M.J., Potkin, S.G., Radua, J., Reif, A., Rinker, D.A., Roffman, J.L., Rosa, P.G.P., Sacchet, M.D., Sachdev, P.S., Salvador, R., Sánchez-Juan, P., Sarró, S., Satterthwaite, T.D., Saykin, A.J., Serpa, M.H., Schmaal, L., Schnell, K., Schumann, G., Sim, K., Smoller, J.W., Sommer, I., Soriano-Mas, C., Stein, D.J., Strike, L.T., Swagerman, S.C., Tamnes, C.K., Temmingh, H.S., Thomopoulos, S.I., Tomyshev, A.S., Tordesillas-Gutiérrez, D., Trollor, J.N., Turner, J.A., Uhlmann, A., van den Heuvel, O.A., van den Meer, D., van der Wee, N.J.A., van Haren, N.E.M., van't Ent, D., van Erp, T.G.M., Veer, I.M., Veltman, D.J., Voineskos, A., Völzke, H., Walter, H., Walton, E., Wang, L., Wang, Y., Wassink, T.H., Weber, B., Wen, W., West, J.D., Westlye, L.T., Whalley, H., Wierenga, L.M., Williams, S.C.R., Wittfeld, K., Wolf, D.H., Worker, A., Wright, M.J., Yang, K., Yoncheva, Y., Zanetti, M. V., Ziegler, G.C., Thompson, P.M., Frangou, S., 2021. Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3–90 years. Hum. Brain Mapp. https://doi.org/10.1002/hbm.25320

Dufumier, B., Gori, P., Battaglia, I., Victor, J., Grigis, A., Duchesnay, E., 2021. Benchmarking CNN on 3D Anatomical Brain MRI: Architectures, Data Augmentation and Deep Ensemble Learning 1–26.

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, 341–355. https://doi.org/10.1016/s0896-6273(02)00569-x

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

Fortin, J.P., Parker, D., Tunç, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., Schultz, R.T., Verma, R., Shinohara, R.T., 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2021. Bayesian Data Analysis, 3rd ed.

Habes, M., Pomponio, R., Shou, H., Doshi, J., Mamourian, E., Erus, G., Nasrallah, I., Launer, L.J., Rashid, T., Bilgel, M., Fan, Y., Toledo, J.B., Yaffe, K., Sotiras, A., Srinivasan, D., Espeland, M., Masters, C., Maruff, P., Fripp, J., Völzk, H., Johnson, S.C., Morris, J.C., Albert, M.S., Miller, M.I., Bryan, R.N., Grabe, H.J., Resnick, S.M., Wolk, D.A., Davatzikos, C., 2021. The Brain Chart of Aging: Machine-learning analytics reveals links between brain aging, white matter disease, amyloid burden, and cognition in the iSTAGING consortium of 10,216 harmonized MR scans. Alzheimer's Dement. 17, 89–102. https://doi.org/10.1002/alz.12178

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., Fischl, B., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. Neuroimage 32, 180–194. https://doi.org/10.1016/j.neuroimage.2006.02.051

Hastings, W.K., 1970. Monte carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109. https://doi.org/10.1093/biomet/57.1.97

Hedges, E.P., Dimitrov, M., Zahid, U., Brito Vega, B., Si, S., Dickson, H., McGuire, P., Williams, S., Barker, G.J., Kempton, M.J., 2022. Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream. Neuroimage 246, 118751. https://doi.org/10.1016/j.neuroimage.2021.118751

Hoffman, M.D., Gelman, A., 2014. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. 15, 1593–1623.

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, 212–220. https://doi.org/10.1016/j.jalz.2010.03.004

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data

using empirical Bayes methods. Biostatistics 8, 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kenward, M.G., Roger, J.H., 1997. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood Author (s): Michael G . Kenward and James H . Roger Published by: International Biometric Society Stable URL: https://www.jstor.org/stable/2533558 REFERENCES Linked references. Biometrics 53, 983–997.

King, R.D., George, A.T., Jeon, T., Hynan, L.S., Youn, T.S., Kennedy, D.N., Dickerson, B., 2009. Characterization of atrophic changes in the cerebral cortex using fractal dimensional analysis. Brain Imaging Behav. 3, 154–166. https://doi.org/10.1007/s11682-008-9057-9

Koval, I., Bône, A., Louis, M., Lartigue, T., Bottani, S., Marcoux, A., Samper-González, J., Burgos, N., Charlier, B., Bertrand, A., Epelbaum, S., Colliot, O., Allassonnière, S., Durrleman, S., 2021. AD Course Map charts Alzheimer's disease progression. Sci. Rep. 11, 1–16. https://doi.org/10.1038/s41598-021-87434-1

Li, Q., Yu, Z., Wang, Y., Zheng, H., 2020. Tumorgan: A multi-modal data augmentation framework for brain tumor segmentation. Sensors (Switzerland) 20, 1–16. https://doi.org/10.3390/s20154203

Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N., 2021. Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 313–322. https://doi.org/10.1007/978-3-030-87199-4_30

Liu, Yuan, Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G.S., Peng, L.H., Webster, D.R., Ai, D., Huang, S.J., Liu, Yun, Dunn, R.C., Coz, D., 2020. A deep learning system for differential diagnosis of skin diseases. Nat. Med. 26, 900–908. https://doi.org/10.1038/s41591-020-0842-3

Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S.C., Koike, S., 2021. Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. Hum. Brain Mapp.

42, 5278–5287. https://doi.org/10.1002/hbm.25615

Marinescu, R. V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Eshaghi, A., Toni, T., Salaterski, M., Lunina, V., Ansart, M., Durrleman, S., Lu, P., Iddi, S., Li, D., Thompson, W.K., Donohue, M.C., Nahon, A., Levy, Y., Halbersberg, D., Cohen, M., Liao, H., Li, T., Yu, K., Zhu, H., Tamez-Pena, J.G., Ismail, A., Wood, T., Bravo, H.C., Nguyen, M., Sun, N., Feng, J., Yeo, B.T.T., Chen, G., Qi, K., Chen, S., Qiu, D., Buciuman, I., Kelner, A., Pop, R., Rimocea, D., Ghazi, M.M., Nielsen, M., Ourselin, S., Sorensen, L., Venkatraghavan, V., Liu, K., Rabe, C., Manser, P., Hill, S.M., Howlett, J., Huang, Z., Kiddle, S., Mukherjee, S., Rouanet, A., Taschler, B., Tom, B.D.M., White, S.R., Faux, N., Sedai, S., Oriol, J. de V., Clemente, E.E. V., Estrada, K., Aksman, L., Altmann, A., Stonnington, C.M., Wang, Y., Wu, J., Devadas, V., Fourrier, C., Raket, L.L., Sotiras, A., Erus, G., Doshi, J., Davatzikos, C., Vogel, J., Doyle, A., Tam, A., Diaz-Papkovich, A., Jammeh, E., Koval, I., Moore, P., Lyons, T.J., Gallacher, J., Tohka, J., Ciszek, R., Jedynak, B., Pandya, K., Bilgel, M., Engels, W., Cole, J., Golland, P., Klein, S., Alexander, D.C., 2020. The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up 1–60.

Meng, Y., Li, G., Gao, Y., Shen, D., 2015. Automatic parcellation of cortical surfaces using random forests. Proc. - Int. Symp. Biomed. Imaging 2015-July, 810–813. https://doi.org/10.1109/ISBI.2015.7163995

Modanwal, G., Vellal, A., Buda, M., Mazurowski, M.A., 2020. MRI image harmonization using cycle-consistent generative adversarial network 36. https://doi.org/10.1117/12.2551301

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15, 869–877. https://doi.org/10.1016/j.nic.2005.09.008

Nalepa, J., Marcinkiewicz, M., Kawulok, M., 2019. Data Augmentation for Brain-Tumor Segmentation: A Review. Front. Comput. Neurosci. 13, 1–18. https://doi.org/10.3389/fncom.2019.00083

Nebli, A., Kaplan, U.A., Rekik, I., 2020. Deep EvoGraphNet Architecture for Time-Dependent Brain Graph Data Synthesis from a Single Timepoint. Lect. Notes Comput. Sci. (including

Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 12329 LNCS, 144–155. https://doi.org/10.1007/978-3-030-59354-4_14

Orlhac, F., Boughdad, S., Philippe, C., Stalla-Bourdillon, H., Nioche, C., Champion, L., Soussan, M., Frouin, F., Frouin, V., Buvat, I., 2018. A postreconstruction harmonization method for multicenter radiomic studies in PET. J. Nucl. Med. 59, 1321–1328. https://doi.org/10.2967/jnumed.117.199935

Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., Trojanowski, J.Q., Weiner, M.W., 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI). Neurology 74, 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Phan, D., Pradhan, N., Jankowiak, M., 2019. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro 1–10.

Piecuch, C.G., Huybers, P., Tingley, M.P., 2017. Comparison of full and empirical Bayes approaches for inferring sea-level changes from tide-gauge data. J. Geophys. Res. Ocean. 122, 2243–2258. https://doi.org/10.1002/2016JC012506

Pölsterl, S., Wachinger, C., 2020. Estimation of Causal Effects in the Presence of Unobserved Confounding in the Alzheimer's Continuum.

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., Gur, Raquel, Gur, Ruben, Morris, J., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, R.N., Wolk, D.A., Shinohara, R.T., Shou, H., Davatzikos, C., 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage 208. https://doi.org/10.1016/j.neuroimage.2019.116450

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M.J., Weickert, C.S., Weickert, T., Bruggemann, J., Kircher, T., Nenadić, I., Cairns, M.J., Seal, M., Schall, U., Henskens, F., Fullerton, J.M., Mowry, B., Pantelis, C., Lenroot, R., Cropley, V., Loughland, C., Scott, R., Wolf, D., Satterthwaite, T.D., Tan, Y., Sim, K., Piras, Fabrizio, Spalletta, G., Banaj, N., Pomarol-Clotet, E., Solanes, A., Albajes-Eizagirre, A., Canales-Rodríguez, E.J., Sarro, S., Di

Giorgio, A., Bertolino, A., Stäblein, M., Oertel, V., Knöchel, C., Borgwardt, S., du Plessis, S., Yun, J.Y., Kwon, J.S., Dannlowski, U., Hahn, T., Grotegerd, D., Alloza, C., Arango, C., Janssen, J., Díaz-Caneja, C., Jiang, W., Calhoun, V., Ehrlich, S., Yang, K., Cascella, N.G., Takayanagi, Y., Sawa, A., Tomyshev, A., Lebedeva, I., Kaleda, V., Kirschner, M., Hoschl, C., Tomecek, D., Skoch, A., van Amelsvoort, T., Bakker, G., James, A., Preda, A., Weideman, A., Stein, D.J., Howells, F., Uhlmann, A., Temmingh, H., López-Jaramillo, C., Díaz-Zuluaga, A., Fortea, L., Martinez-Heras, E., Solana, E., Llufriu, S., Jahanshad, N., Thompson, P., Turner, J., van Erp, T., Glahn, D., Pearlson, G., Hong, E., Krug, A., Carr, V., Tooney, P., Cooper, G., Rasser, P., Michie, P., Catts, S., Gur, Raquel, Gur, Ruben, Yang, F., Fan, F., Chen, J., Guo, H., Tan, S., Wang, Z., Xiang, H., Piras, Federica, Assogna, F., Salvador, R., McKenna, P., Bonvino, A., King, M., Kaiser, S., Nguyen, D., Pineda-Zapata, J., 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218. https://doi.org/10.1016/j.neuroimage.2020.116956

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61, 1402–1418. https://doi.org/10.1016/j.neuroimage.2012.02.084

Richter, S., Winzeck, S., Correia, M.M., Kornaropoulos, E.N., Manktelow, A., Outtrim, J., Chatfield, D., Posti, J.P., Tenovuo, O., Williams, G.B., Menon, D.K., Newcombe, V.F.J., 2022. Validation of cross-sectional and longitudinal ComBat harmonization methods for magnetic resonance imaging data on a travelling subject cohort. Neuroimage: Reports 2, 100136. https://doi.org/10.1016/j.ynirp.2022.100136

Sun, D., Rakesh, G., Haswell, C.C., Logue, M., Baird, C.L., Leary, B.M.O., Cotton, A.S., Xie, H., Tamburrino, M., Chen, T., Emily, L., Jahanshad, N., Salminen, L.E., Thomopoulos, S.I., Rashid, F., 2021. A Comparison of Methods to Harmonize Cortical Thickness Measurements Across Scanners and Sites.

Torbati, M.E., Minhas, D.S., Ahmad, G., O'Connor, E.E., Muschelli, J., Laymon, C.M., Yang, Z., Cohen, A.D., Aizenstein, H.J., Klunk, W.E., Christian, B.T., Hwang, S.J., Crainiceanu, C.M., Tudorascu, D.L., 2021a. A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. Neuroimage 245, 118703.

https://doi.org/10.1016/j.neuroimage.2021.118703

Torbati, M.E., Tudorascu, D.L., Minhas, D.S., Maillard, P., Decarli, C.S., Jae Hwang, S., 2021b. Multi-scanner Harmonization of Paired Neuroimaging Data via Structure Preserving Embedding Learning. Proc. IEEE Int. Conf. Comput. Vis. 2021-Octob, 3277–3286. https://doi.org/10.1109/ICCVW54120.2021.00367

van de Wiel, M.A., Te Beest, D.E., Münch, M.M., 2019. Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. Scand. J. Stat. 46, 2–25. https://doi.org/10.1111/sjos.12335

Venkatraman, V.K., Gonzalez, C.E., Landman, B., Goh, J., Reiter, D.A., An, Y., Resnick, S.M., 2015. Region of interest correction factors improve reliability of diffusion imaging measures within and across scanners and field strengths. Neuroimage 119, 406–416. https://doi.org/10.1016/j.neuroimage.2015.06.078

Wachinger, C., Rieckmann, A., Pölsterl, S., 2021. Detect and correct bias in multi-site neuroimaging datasets. Med. Image Anal. 67, 101879. https://doi.org/10.1016/j.media.2020.101879

Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L., 2011. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. Proc. IEEE Int. Conf. Comput. Vis. 557–562. https://doi.org/10.1109/ICCV.2011.6126288

Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D., 2016. Understanding Data Augmentation for Classification: When to Warp? 2016 Int. Conf. Digit. Image Comput. Tech. Appl. DICTA 2016. https://doi.org/10.1109/DICTA.2016.7797091

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Tanaka, S.C., Kawato, M., Yamashita, O., Imamizu, H., 2019. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias, PLoS Biology. https://doi.org/10.1371/journal.pbio.3000042

Young, A.L., Marinescu, R. V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., van Swieten, J., Borroni, B., Galimberti, D., Masellis, M., Tartaglia, M.C., Rowe, J.B., Graff, C., Tagliavini, F., Frisoni, G.B., Laforce, R., Finger, E., de Mendonça, A., Sorbi, S., Warren, J.D., Crutch, S., Fox, N.C., Ourselin, S., Schott, J.M., Rohrer, J.D., Alexander, D.C., Andersson, C., Archetti, S., Arighi, A., Benussi, L., Binetti, G., Black, S., Cosseddu, M., Fallström, M., Ferreira, C., Fenoglio, C., Freedman, M., Fumagalli, G.G., Gazzina, S., Ghidoni, R., Grisoli, M., Jelic, V., Jiskoot, L., Keren, R., Lombardi, G., Maruta, C., Meeter, L., Mead, S., van Minkelen, R., Nacmias, B., Öijerstedt, L., Padovani, A., Panman, J., Pievani, M., Polito, C., Premi, E., Prioni, S., Rademakers, R., Redaelli, V., Rogaeva, E., Rossi, G., Rossor, M., Scarpini, E., Tang-Wai, D., Thonberg, H., Tiraboschi, P., Verdelho, A., Weiner, M.W., Aisen, P., Petersen, R., Jack, C.R., Jagust, W., Trojanowki, J.Q., Toga, A.W., Beckett, L., Green, R.C., Saykin, A.J., Morris, J., Shaw, L.M., Khachaturian, Z., Sorensen, G., Kuller, L., Raichle, M., Paul, S., Davies, P., Fillit, H., Hefti, F., Holtzman, D., Mesulam, M. Marcel, Potter, W., Snyder, P., Schwartz, A., Montine, T., Thomas, R.G., Donohue, M., Walter, S., Gessert, D., Sather, T., Jiminez, G., Harvey, D., Bernstein, M., Thompson, P., Schuff, N., Borowski, B., Gunter, J., Senjem, M., Vemuri, P., Jones, D., Kantarci, K., Ward, C., Koeppe, R.A., Foster, N., Reiman, E.M., Chen, K., Mathis, C., Landau, S., Cairns, N.J., Householder, E., Taylor-Reinwald, L., Lee, V., Korecka, M., Figurski, M., Crawford, K., Neu, S., Foroud, T.M., Potkin, S., Shen, L., Faber, K., Kim, S., Nho, K., Thal, L., Buckholtz, N., Albert, Marylyn, Frank, R., Hsiao, J., Kaye, J., Quinn, J., Lind, B., Carter, R., Dolen, S., Schneider, L.S., Pawluczyk, S., Beccera, M., Teodoro, L., Spann, B.M., Brewer, J., Vanderswag, H., Fleisher, A., Heidebrink, J.L., Lord, J.L., Mason, S.S., Albers, C.S., Knopman, D., Johnson, Kris, Doody, R.S., Villanueva-Meyer, J., Chowdhury, M., Rountree, S., Dang, M., Stern, Y., Honig, L.S., Bell, K.L., Ances, B., Carroll, M., Leon, S., Mintun, M.A., Schneider, S., Oliver, A., Marson, D., Griffith, R., Clark, D., Geldmacher, D., Brockington, J., Roberson, E., Grossman, H., Mitsis, E., de Toledo-Morrell, L., Shah, R.C., Duara, R., Varon, D., Greig, M.T., Roberts, P., Albert, Marilyn, Onyike, C., D'Agostino, D., Kielb, S., Galvin, J.E., Cerbone, B., Michel, C.A., Rusinek, H., de Leon, M.J., Glodzik, L., De Santi, S., Doraiswamy, P.M., Petrella, J.R., Wong, T.Z., Arnold, S.E., Karlawish, J.H., Wolk, D., Smith, C.D., Jicha, G.,

Hardy, P., Sinha, P., Oates, E., Conrad, G., Lopez, O.L., Oakley, M.A., Simpson, D.M., Porsteinsson, A.P., Goldstein, B.S., Martin, K., Makino, K.M., Ismail, M.S., Brand, C., Mulnard, R.A., Thai, G., Mc-Adams-Ortiz, C., Womack, K., Mathews, D., Quiceno, M., Diaz-Arrastia, R., King, R., Weiner, M., Martin-Cook, K., DeVous, M., Levey, A.I., Lah, J.J., Cellar, J.S., Burns, J.M., Anderson, H.S., Swerdlow, R.H., Apostolova, L., Tingus, K., Woo, E., Silverman, D.H., Lu, P.H., Bartzokis, G., Graff-Radford, N.R., Parfitt, F., Kendall, T., Johnson, H., Farlow, M.R., Hake, A.M., Matthews, B.R., Herring, S., Hunt, C., van Dyck, C.H., Carson, R.E., MacAvoy, M.G., Chertkow, H., Bergman, H., Hosein, C., Stefanovic, B., Caldwell, C., Hsiung, G.Y.R., Feldman, H., Mudge, B., Assaly, M., Kertesz, A., Rogers, J., Bernick, C., Munic, D., Kerwin, D., Mesulam, Marek Marsel, Lipowski, K., Wu, C.K., Johnson, N., Sadowsky, C., Martinez, W., Villena, T., Turner, R.S., Johnson, Kathleen, Reynolds, B., Sperling, R.A., Johnson, K.A., Marshall, G., Frey, M., Lane, B., Rosen, A., Tinklenberg, J., Sabbagh, M.N., Belden, C.M., Jacobson, S.A., Sirrel, S.A., Kowall, N., Killiany, R., Budson, A.E., Norbash, A., Johnson, P.L., Allard, J., Lerner, A., Ogrocki, P., Hudson, L., Fletcher, E., Carmichael, O., Olichney, J., DeCarli, C., Kittur, S., Borrie, M., Lee, T.Y., Bartha, R., Johnson, S., Asthana, S., Carlsson, C.M., Potkin, S.G., Preda, A., Nguyen, D., Tariot, P., Reeder, S., Bates, V., Capote, H., Rainka, M., Scharre, D.W., Kataki, M., Adeli, A., Zimmerman, E.A., Celmins, D., Brown, A.D., Pearlson, G.D., Blank, K., Anderson, K., Santulli, R.B., Kitzmiller, T.J., Schwartz, E.S., Sink, K.M., Williamson, J.D., Garg, P., Watkins, F., Ott, B.R., Querfurth, H., Tremont, G., Salloway, S., Malloy, P., Correia, S., Rosen, H.J., Miller, B.L., Mintzer, J., Spicer, K., Bachman, D., Pasternak, S., Rachinsky, I., Drost, D., Pomara, N., Hernando, R., Sarrael, A., Schultz, S.K., Ponto, L.L.B., Shim, H., Smith, K.E., Relkin, N., Chaing, G., Raudin, L., Smith, A., Fargher, K., Raj, B.A., Neylan, T., Grafman, J., Davis, M., Morrison, R., Hayes, J., Finley, S., Friedl, K., Fleischman, D., Arfanakis, K., James, O., Massoglia, D., Fruehling, J.J., Harding, S., Peskind, E.R., Petrie, E.C., Li, G., Yesavage, J.A., Taylor, J.L., Furst, A.J., 2018. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nat. Commun. 9, 1–16. https://doi.org/10.1038/s41467-018-05892-0

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, 4213–4227. https://doi.org/10.1002/hbm.24241

Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., Glymour, C., 2017. Causal Discovery in the Presence of Measurement Error: Identifiability Conditions.

Zuo, L., Dewey, B.E., Liu, Y., He, Y., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A., 2021. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. Neuroimage 243, 118569. https://doi.org/10.1016/j.neuroimage.2021.118569