ComBat Harmonization: Empirical Bayes versus Fully Bayes Approaches

Maxwell Reynolds\textsuperscript{a}, Tigmanshu Chaudhary\textsuperscript{a}, Mahbaneh Eshaghzadeh Torbati\textsuperscript{b}, Dana L. Tudorascu\textsuperscript{c,d}, Kayhan Batmanghelich\textsuperscript{a}, for the Alzheimer’s Disease Neuroimaging Initiative\textsuperscript{1}

\textsuperscript{a} Department of Biomedical Informatics, University of Pittsburgh School of Medicine, 5607 Baum Blvd. Suite 500, Pittsburgh, PA, 15206, USA
\textsuperscript{b} Intelligent System Program, University of Pittsburgh School of Computing and Information, 210 South Bouquet Street, Pittsburgh, PA 15260, USA
\textsuperscript{c} Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O’Hara Street, Pittsburgh, PA 15213, USA
\textsuperscript{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)

\textsuperscript{1} 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, introduced by (Johnson et al., 2007), 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. Our experiments show that our new fully Bayesian approach offers more accurate harmonization, unconstrained posterior distributions, and representative uncertainty quantification at the expense of higher computation costs for the inference. This fully Bayesian approach generates a rich posterior distribution, which is also 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) and generative adversarial networks (Liu et al., 2021) to harmonize images to a specific reference scanner.

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), 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. While the FB method maintains the robustness of the EB approach, we also found that FB ComBat yields harmonized features with greater retention of biologically-relevant information and smaller differences in test-retest subjects. FB ComBat produces 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 use to augment a dataset to improve Alzheimer’s Disease classifier performance. We also use 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.

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 was launched in 2003 as a public-
private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see www.adni-info.org. All subjects gave informed consent in accordance with local Institutional Review Board Regulations.

We obtain 3894 initial images from patients grouped as either cognitively normal (CN), mild cognitive impaired (MCI), or Alzheimer’s disease (AD). 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 were dropped due to poor image quality, 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 were 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:

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

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 \( j \). \( \gamma_{iv} \) is the additive scanner factor for scanner \( i \) and feature \( v \). \( \delta_{iv}^2 \) is the scaling scanner factor from scanner \( i \) and feature \( v \). A description of all variables is given in Table 1.

In EB ComBat, \( \alpha_v, \beta_v, \eta_{jv}, \sigma_v, \) and priors for \( \gamma_{iv} \) and \( \delta_{iv}^2 \) are estimated using restricted maximum likelihood (REML) and method of moments, then conditional posteriors for \( \gamma_{iv} \) and \( \delta_{iv}^2 \) are identified using an expectation-maximization (EM) algorithm.

Harmonized adjusted feature values are obtained by the equation:

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

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 chose 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 added the following constraints to scanner additive and multiplicative factors to ensure identifiability:
where $n$ is the total number of images, and $n_i$ is the number of images from scanner $i$.

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. The entire FB ComBat model is inferred jointly, as opposed

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:
\[ y_{ijv}^F(t) = \frac{\gamma_{ijv} - \delta_{iv} - \hat{\gamma}_{ijv}}{\delta_{iv}} + \alpha_v + X_j^T(t)\beta_v + \hat{\gamma}_{jv} \]

where \( \alpha_v, \beta_v, \hat{\gamma}_{jv}, \delta_{iv}, \) are the posterior parameter estimates. We perform this transformation on the joint posterior parameter distribution to obtain the posterior harmonized data distribution.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>Scanner index</td>
</tr>
<tr>
<td>( j )</td>
<td>Patient index</td>
</tr>
<tr>
<td>( v )</td>
<td>Feature index</td>
</tr>
<tr>
<td>( \gamma_{ijv} )</td>
<td>Unharmonized feature value</td>
</tr>
<tr>
<td>( \gamma_{ijv}^F )</td>
<td>Harmonized feature value using EB ComBat</td>
</tr>
<tr>
<td>( \gamma_{ijv}^{FB} )</td>
<td>Harmonized feature value using FB ComBat</td>
</tr>
<tr>
<td>( \alpha_v )</td>
<td>Feature mean</td>
</tr>
<tr>
<td>( \chi )</td>
<td>Covariate term</td>
</tr>
<tr>
<td>( \beta_v )</td>
<td>Covariate Coefficients</td>
</tr>
<tr>
<td>( \eta_{ijv} )</td>
<td>Subject-specific intercept</td>
</tr>
<tr>
<td>( \gamma_{iv} )</td>
<td>Additive scanner factor</td>
</tr>
<tr>
<td>( \delta_{iv} )</td>
<td>Multiplicative scanner factor</td>
</tr>
<tr>
<td>( \sigma_v^2 )</td>
<td>Feature-specific average variance</td>
</tr>
<tr>
<td>( \gamma_{i}, \tau_i )</td>
<td>Hyperparameters for additive scanner factor</td>
</tr>
<tr>
<td>( m_i, s_i )</td>
<td>Hyperparameters for multiplicative scanner factor</td>
</tr>
<tr>
<td>( \rho_v )</td>
<td>Hyperparameter for subject-specific intercept</td>
</tr>
</tbody>
</table>

**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. Posterior distributions for all parameters are gathered from these 40000 MCMC samples.

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. Finally, we explore posterior uncertainty as a tool for dataset augmentation, overall regional measurement uncertainty, and uncertainty-aware association tests between brain regional measures and AD.

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 joint density for FB ComBat model 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.
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.

3.3 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 greater test MAE using unharmonized data is not significantly greater than using FB ComBat ($p = 0.39$).

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE (years)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>Dummy Classifier</td>
<td>5.38 (0.19)</td>
<td>5.38 (0.19)</td>
</tr>
<tr>
<td>Unharmonized</td>
<td>1.05 (0.01)</td>
<td>2.79 (0.14)</td>
</tr>
<tr>
<td>EB ComBat</td>
<td>1.09 (0.01)</td>
<td>2.92 (0.15)</td>
</tr>
<tr>
<td>FB ComBat</td>
<td>1.05 (0.01)</td>
<td>2.80 (0.13)</td>
</tr>
</tbody>
</table>

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 parenthesis. 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.

3.4 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) values of scanner strength prediction from the random forest classification model trained on unharmonized, EB ComBat, and FB ComBat data are 0.949 (±0.018), 0.694 (±0.040), and 0.735 (±0.034) respectively, shown in Table 3. Cross-validation AUROC curves are shown in Figure 4. Low 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 4: 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 similarly reduce the AUROC closer to random chance, but there is still some scanner signal left in the harmonized data.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUROC</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>Unharmonized</td>
<td>0.980 (0.003)</td>
<td>0.949 (0.018)</td>
</tr>
<tr>
<td>EB ComBat</td>
<td>0.868 (0.015)</td>
<td>0.694 (0.040)</td>
</tr>
<tr>
<td>FB ComBat</td>
<td>0.887 (0.009)</td>
<td>0.735 (0.034)</td>
</tr>
</tbody>
</table>

*Table 3: Scanner Strength Prediction Results*

Scanner strength prediction performance using various harmonization methods and unharmonized data. Lower accuracy in the task indicates better harmonization performance due to more effective removal of non-biological (scanner magnetic field strength) information. EB ComBat performs best on this task (lowest test AUROC).

3.5 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.

Before harmonization, significant biases are found in all regional measurements except for right inferior parietal thickness and left hippocampus volume, shown in Figure 5. Both EB and FB ComBat harmonization remove any significant bias. RMSD values are shown in Figure 6. All regions had the lowest RMSD after FB ComBat harmonization. Both FB ComBat and EB ComBat improve 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 7. 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 5: 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 (*). Bias is present in all regions except Left Hippocampus and Right Inferior Parietal Cortex 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 6: 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 7: 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.

3.6 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.

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 patient in the training split in these two datasets, we draw 100 additional random samples from the posterior distribution of the patients’ harmonized imaging 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 4. Using FB ComBat harmonization with posterior distribution resampling augmentation results in the highest AUROC (0.821). This was significantly higher than FB harmonization without augmentation \(p<0.0001\) and higher than EB harmonization with augmentation \(p<0.0001\).
<table>
<thead>
<tr>
<th></th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unharmonized</td>
<td>0.772 (0.028)</td>
</tr>
<tr>
<td>EB ComBat</td>
<td>0.777 (0.026)</td>
</tr>
<tr>
<td>FB ComBat</td>
<td>0.796 (0.030)</td>
</tr>
<tr>
<td>EB Combat with posterior resampling</td>
<td>0.799 (0.027)</td>
</tr>
<tr>
<td>FB ComBat with posterior resampling</td>
<td>0.821 (0.022)</td>
</tr>
</tbody>
</table>

Table 4: 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.

3.7 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 the brain, it is necessary to normalize posterior variance, due to the difference in measurement scales (thickness vs. volume) and size of different regions. We devise a normalized uncertainty value as the ratio of variance within all individual measurements (MSW- Mean Squared Within) to the variance of individual posterior means among the population (MSA- Mean Squared Among) for each imaging feature (v):

\[
Uncertainty_v = \frac{MSW_v}{MSA_v}
\]
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 8. 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.
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 a) cortical thicknesses and b) 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.

3.8 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:

$$y_{ijv}^{FB}(t) = X_j^{AD}(t)\beta_v^{AD} + X_j^{covar}(t)\beta_v^{covar} + \varepsilon$$
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 \( t \), \( \beta_v^{AD} \) is the association of Alzheimer’s disease with feature \( v \), \( x_j^{covar}(t) \beta_v^{covar} \) is the covariate term, and \( \varepsilon \) is the random error term. Imaging features \( y_{ij}(t) \) are adjusted by mean thickness or overall brain segmentation volume for regional thickness and regional volume measurements, respectively. A significant \( \beta_v^{AD} \) 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 between brain regions and Alzheimer’s disease are shown in Figure 9 and Figure 10. Several regions show differing significance when using WLS versus OLS regression including right insula thickness, left inferior parietal thickness, and left thalamus volume.
Figure 9: Cortical Thickness Associations with AD

Cortical thickness 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 insula and left inferior parietal thickness are highlighted as regions with different associations using OLS and WLS models.

Figure 10: Subcortical Volume Associations with AD

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. Left thalamus volume is highlighted as a region with different associations using OLS and WLS models.
4. Discussion

Scanner harmonization is an important step of brain MRI pre-processing to reduce noise and potential biases. Statistical methods such as ComBat (Fortin et al., 2018) harmonize on the image-derived feature level, while some deep learning harmonization methods adjust the image directly (Liu et al., 2021; Modanwal et al., 2020; Torbati et al., 2021b). Deep learning-based approaches are increasingly adopted for various applications in the medical imaging domain, including harmonization. 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 focused on image-derived feature harmonization and provide a fully Bayesian extension of a popular feature-based harmonization, namely ComBat. Fully Bayesian approaches tend to provide more accurate estimations of uncertainty (Gelman et al., 2021; van de Wiel et al., 2019). We investigated the value of uncertainty for data augmentation, association testing, and overall brain measurement reliability, thus adding novel findings and harmonization model use cases to the existing literature.

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 a random forest model to predict age and evaluated its performance on datasets with unharmonized versus harmonized data. Harmonizing brain imaging feature data using FB ComBat resulted in stronger age classification performance than EB ComBat (shown in Table 2), suggesting that FB ComBat was able to effectively retain biologically-relevant brain structural information. 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 (a non-biological variable), was removed during harmonization. EB ComBat performed best on this task, while FB ComBat performed similarly. Both harmonization methods significantly outperformed unharmonized data, indicating that
non-biological information related to scanner strength was effectively removed in both methods.

We additionally 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 also presented the 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. We demonstrated that our FB harmonization model’s posterior distribution can be used as a rich data generation tool that can improve classifier performance. Additionally, 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 greatly 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.

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 period are our best 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, these tasks are only approximations and do not include all possible biological and non-biological variables of interest, many of which are unobserved. Confounding between biological and scanner factors would also limit the usefulness of these metrics. 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 are straightforward to implement. The probabilistic programming approach used for FB ComBat just requires specification of the model, then sampling is done automatically. 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. However, harmonization is generally performed just once before imaging features are analyzed, so we expect that the improved harmonization performance and uncertainty quantification outweigh the efficiency drawback in most cases.

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
The research reported in this publication was supported by the National Institutes of Health under award numbers 1R01HL141813-01 (K Batmanghelich), 1R01-AG063752 (DL Tudorascu), 5T15LM007059 (M. Reynolds), the National Science Foundation under award 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.; 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**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**


system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980. http://doi.org/10.1016/j.neuroimage.2006.01.021


Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J,


Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q.,

Brain Graph Data Synthesis from a Single Timepoint. Lect. Notes Comput. Sci. (including
https://doi.org/10.1007/978-3-030-59354-4_14

Probabilistic Programming in NumPyro 1–10.


Confounding in the Alzheimer’s Continuum.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V.,
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.,
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.,


https://doi.org/10.1016/j.media.2020.101879

https://doi.org/10.1109/ICCV.2011.6126288


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.,
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., Swerdlov, 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,
R.E., MacAvoy, M.G., Chertkow, H., Bergman, H., Hosein, C., Stefanovic, B., Caldwell, C.,
Hsiung, G.Y.R., Feldman, H., Mudge, B., Assały, 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.,
Sabbagh, M.N., Belden, C.M., Jacobson, S.A., Sirrel, S.A., Kowall, N., Killiany, R., Budson,
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., Katakai, 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,
