HARMONIZATION OF MULTI-SCANNER MAGNETIC RESONANCE IMAGING DATA

by

Mahbaneh Eshaghzadeh Torbati

BS in Computer Engineering, Sadjad University, 2009

MS in Computer Engineering, Sharif University of Technology, 2012

MS in Intelligent Systems, University of Pittsburgh, 2019

Submitted to the Graduate Faculty of

the School of Computing and Information, Intelligent Systems Program in partial

fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Mahbaneh Eshaghzadeh Torbati

It was defended on

August 15, 2024

and approved by

Dana L. Tudorascu, Intelligent Systems Program, University of Pittsburgh

Shyam Visweswaran, Intelligent Systems Program, University of Pittsburgh

Ahmad P. Tafti, Intelligent Systems Program, University of Pittsburgh

Davneet S. Minhas, Department of Radiology, University of Pittsburgh

Seong Jae Hwang, Department of Artificial Intelligence, Yonsei University

HARMONIZATION OF MULTI-SCANNER MAGNETIC RESONANCE IMAGING DATA

Mahbaneh Eshaghzadeh Torbati, PhD

University of Pittsburgh, 2024

The integration of datasets from multiple sites or scanners in neuroimaging studies has become increasingly prevalent. However, the presence of substantial technical variability associated with scanners poses a challenge that can introduce unintended biases in downstream analyses. Moreover, this scanner-related variability, known as scanner effects, can manifest in longitudinal neuroimaging data due to potential scanner upgrades or replacements at sites. Harmonization methods have emerged as techniques to address scanner effects on multi-scanner neuroimaging data, encompassing both brain images and image-derived summary measures. Harmonization can be accomplished through various approaches, including the estimation and removal of scanner effects, as well as adapting the multi-scanner data to a scanner-middle-ground space or a target scanner domain. In these approaches, matched data can serve as additional labeled dataset to uncover scanner effects in the multi-scanner data. Harmonization methods that utilize matched data are referred to as supervised harmonization methods, leading many sites to collect additional matched data to facilitate harmonization. However, the current availability of neuroimaging data often lacks such matched data. Consequently, a thorough understanding of scanner effects and the development of both supervised and unsupervised harmonization methods are imperative.

This dissertation contributes to the field of harmonization of T1-weighted MRIs. Firstly, scanner effects and two harmonization methods for mitigating scanner effects in both images and image-derived measures are investigated. Secondly, MISPEL, a novel supervised image harmonization method is developed. MISPEL leverages matched data to learn a mapping to a scanner-middle-ground space. Third, a novel unsupervised image harmonization method, ESPA, is proposed. ESPA simulates scanner effects as augmentations on images and learns to harmonize images by adapting them to a scanner-middle-ground space. These contributions aim to enhance the understanding and effectiveness of harmonization techniques.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation owes its realization to the remarkable support and guidance of my advisors, Dana Tudorascu and Seong Jae Hwang. I am sincerely grateful to them for their unwavering patience, insightful inputs, contagious enthusiasm, and invaluable editing advice. I consider myself privileged to have had the opportunity to collaborate with them. I also wish to extend my heartfelt thanks to my committee member, Davneet Minhas, for his valuable insights into every aspect of my research. Additionally, I express my gratitude to the other members of my dissertation committee, Shyam Visweswaran, and Ahmad Tafti, for their insightful comments and critiques, which have greatly contributed to the refinement of this dissertation.

I extend my gratitude to the ISP directors: Peter Brusilovsky, Vanathi Gopalakrishnan, Diane Litman, and all other ISP faculty members for their dedicated efforts in fostering a distinctive environment for ISP students to engage in multidisciplinary research in artificial intelligence. I also want to acknowledge the invaluable assistance and support provided by the administrative staff at ISP: Michele Thomas, and Heidi Davis.

One of the most enriching aspects of my academic journey has been the incredible friendships I've cultivated along the way. I'm deeply thankful for the enduring support of my longtime friends, Azita Hashemnia and Fatemeh Afghahi. I also want to express my heartfelt appreciation to my friends in Pittsburgh, whose presence has made my time here truly memorable and enjoyable: Afsoon Afzal, Amin Tajgardoon, Daniel Petrov, Faezeh Movahedi, Fattaneh Jabbari, Jana Savelka, Jaromir Savelka, Jeya Balaji Balasubramanian, Majid Mahzoon, Martin Michelini, Mina Akhondzadeh, Mostafa Mirshekari, Neda Mirzaeian, Peter Curtis, Rakan Alseghayer, Sadaf Tayefeh, Salim Malakouti, Saba Dadsetan, Sareh Yousefzadeh, Sina Malakouti, Zahra Ebrahimi, and Zuha Agha. Their friendship has been a source of joy and support throughout my academic journey.

Lastly, I extend my deepest and most heartfelt gratitude to my family for being my constant source of inspiration in pursuing my dreams. To my beloved parents, Mina and Abbas, I owe an immeasurable debt of gratitude for the countless sacrifices they have made

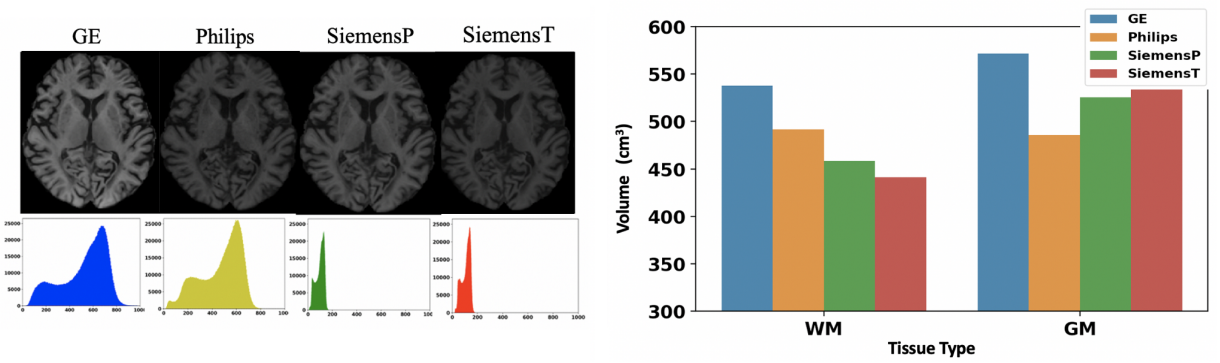
for me throughout my life. The distance that separates us by thousands of miles, knowing that we may not reunite for several years, has been the most challenging aspect of my academic journey, and indeed, of their lives as well. I am also immensely thankful to my dear sister and closest confidante, Atefeh, for her unwavering love, unwavering support, and unshakable belief in me. Her presence has been a pillar of strength, and I am truly blessed to have her by my side.

This research received financial support from grants from the National Institutes of Health (NIH) and a fellowship awarded by the University of Pittsburgh.

1.0 Introduction

There is a growing interest in the neuroimaging community to combine imaging data from a variety of diverse datasets so as to enable having large-scale multi-study data with desirable breadth of biological variability, and conducting analyses that have high statistical power, reliability, and robustness (Madan, 2017, 2021; Mar et al., 2013; Milham et al., 2018). Despite the promise of massive data aggregation initiatives, large-scale neuroimaging analyses from such data collections often suffer from issues of technical variability due to scanner heterogeneity across studies, which may introduce bias in neuroimaging measures (Clark et al., 2022; Kruggel et al., 2010; Potvin et al., 2019), as well as alterations of the biological signals of clinical interest (Shinohara et al., 2017, 2014a), among other unwanted and unexpected artifacts. Other than data aggregation, such technical variability can be observed in longitudinal data collected in sites with multiple scanners and/or sites with scanner upgrades or replacements (Beer et al., 2020). Figure 1 is an example of such technical variability among the axial slices of 3T T1-weighted (T1-w) MRIs taken from an individual by different scanners and with short time gap. Although these images were expected to be identical, the cross-scanner technical variability can be observed as discrepancy in contrast and histogram of these slices in Figure 1a, as well as differences in their volume distributions of gray matter (GM) and white matter (WM) tissue types, Figure 1b.

This technical variability is primarily attributed to *intensity unit effects* and *scanner effects* (Wrobel et al., 2020). Intensity unit effects arise from the arbitrary nature of the image intensity scale, which can lead to variations in the interpretation of intensity units and make direct quantitative analysis of image intensities challenging (Nyúl and Udupa, 1999; Shinohara et al., 2011, 2014b; Wrobel et al., 2020). Intensity unit effects have been long recognized and addressed by intensity standardization and normalization methods (Shah et al., 2011). Scanner effects refer to any post-normalization/standardization inter- or intra-scan variation that is not biological in nature (Fortin et al., 2016) and stems from scanner and acquisition differences (Dinsdale et al., 2021). The group of methods that aim to remove scanner effects is referred to as harmonization. Unlike normalization, harmonization is a complex and



(a) **Technical variability in images.** (b) **Technical variability in volumes.**

Figure 1: **Example of technical variability in pooled multi-scanner T1-w MRIs.** For this example, images are axial slices of T1-w MRIs taken with short time gap from an individual on four different 3T scanners: General Electric (GE), Philips, Siemens Prisma (SiemensP), and Siemens Trio (SiemensT). Specifications of these scanners can be found in Table 3 in section 4.1.1 . Figure (a) depicts the technical variability of the slices as dissimilarity in their contrast, as well as discrepancy among histograms of their whole brain. Figure (b) shows the technical variability of the slices in terms of their tissue type volumetric dissimilarity. Histograms of matched images have identical axes and correspond (from left to right) to GE, Philips, SiemensP, and SiemensT scanners.

challenging task due to (1) a lack of thorough understanding of scanner effects, (2) a lack of standardized criteria for assessment of scanner effects and evaluation of harmonization, and (3) a limited number of available harmonization methods.

In this dissertation, our primary focus lies in understanding and mitigating cross-scanner technical variability, particularly scanner effects. Thus far, we are aware that cross-scanner discrepancies leading to scanner effects have been recognized in scanner manufacturer (Takao et al., 2014), scanner upgrade (Han et al., 2006), scanner drift (Takao et al., 2011), scanner strength (Han et al., 2006), and gradient non-linearities (Jovicich et al., 2006). However, despite the recent noticeable growth in the number of studies dedicated to scanner effects

and harmonization (Cackowski et al., 2021; Dewey et al., 2019, 2020; Liu et al., 2021; Zuo et al., 2021b), there is a lack of understanding regarding how these scanner effects manifest in images. One primary reason for this could be the absence of a ground truth for these studies, which leaves them without standardized evaluation criteria and consequently renders their observations partially inconsistent and challenging to compare. Based on the findings corroborated by several of these studies, it is now established that scanner effects can vary across individual's image voxels (Chen et al., 2020a) and consequently brain regions (Beer et al., 2020). Furthermore, it is also recognized that scanner effects alter brain tissue contrast and consequently impact the outcomes of tissue segmentations (Meyer et al., 2019).

The best experimental design setup to understand scanner effects and evaluate harmonization is to conduct a *matched study*, in which traveling subjects are scanned on different scanners, collecting a *matched image dataset* (Dewey et al., 2019; Zuo et al., 2021b). A matched image dataset comprises a set of *matched images* using *more than two* scanners. These matched images are expected to depict biologically similar brains, with differences solely attributable to scanner effects. A *matched dataset* can encompass matched image datasets or datasets of any brain measures derived from matched images. By utilizing the matched dataset, scanner effects and harmonization can be estimated through evaluating dissimilarities and similarities within matched images/measures, respectively. Such dataset with only two scanners is referred to as a *paired* dataset and shares the same characteristics as a matched dataset.

The currently available harmonization methods exhibit two prevalent drawbacks: (1) over-correction, and (2) brain structural modifications. Over-correction occurs when biological or clinical variables are corrected in addition to or instead of scanner effects (Liu et al., 2023). This can happen if scanner effects are correlated with other variables in the data. Structural modifications involve alterations to the brain's structure during the image harmonization process (Zuo et al., 2021b). To deal with structural modifications, a group of methods restricts harmonization to the contrast or style of images. For example, CALAMITI (Zuo et al., 2021b) harmonizes images by adapting them to the contrast of images in a target scanner. This method can thus suffer from over-correction if data across scanners differ biologically in their populations. As a solution, a style-transfer harmonization method was

designed to adapt images to the style of an *individual* target image (Liu et al., 2023). Although this method addresses population-wise over-correction, it may over-correct images to the style of the target image, which may still convey biological information. One example could be the white matter hyper-intensity which appears on images as their style (Debette and Markus, 2010).

To mitigate the issues of over-correction and brain structural modification, one approach is to utilize *matched data* (Torbati et al., 2023). The distinctive design of matched datasets enables the identification of scanner effects, manifested as dissimilarities among its matched images or measures. Matched data serves as labeled data for harmonization, making it an optimal source for learning harmonization techniques (Dewey et al., 2019). Depending on their reliance on this data, harmonization methods can be categorized into three traditional groups: (1) supervised, (2) semi-supervised, and (3) unsupervised harmonization methods (Zuo et al., 2021b). Supervised harmonization methods are less susceptible to over-correction and brain structural modifications since they can directly and exclusively address scanner effects. However, the applicability of these methods is limited to datasets for which matched data is available. Additionally, these methods may lack robustness due to the typically limited collection of matched data for a restricted number of individuals (Dewey et al., 2019; Modanwal et al., 2020). To overcome these issues, many sites have begun to collect such additional data on a larger scale (Duchesne et al., 2019; Hawco et al., 2022; Magnotta et al., 2020; Maikusa et al., 2021).

A novel and more straightforward perspective on harmonization to tackle the associated issues and limitations with matched data involves simulating scanner effects through augmentation methods. Such scanner-specific augmentation methods could be used in self-supervised augmentation-based frameworks (Chen et al., 2020b) for generating scanner-free pretext, or they can be used for simulating matched data to pretrain harmonization methods of any type. Any of these usages help harmonization models with their possible lack of robustness due to data size. This approach also deals with the over-correction and structural brain modification issues. Over-correction is addressed by data stratification and population matching strategies, which are feasible in the scanner effects simulation process. Structural modifications can be also addressed by limiting the augmentation methods to appearance-

based modifications of images.

The remainder of this chapter outlines the three primary contributions of this dissertation. Firstly, it delves into the investigation of two cross-scanner technical variability removal methods for harmonizing image-derived measures. Secondly, it introduces a supervised harmonization method named MISPEL. Lastly, it presents the development of ESPA, an unsupervised harmonization framework utilizing scanner-specific augmentation methods.

1.1 Investigating two methods of cross-scanner technical variability removal in harmonization of image-derived measures

The cross-scanner technical variability that exists in multi-scanner data could significantly bias any down-stream analysis that is being conducted on neuroimaging data. A good testbed for estimating scanner effects and evaluating harmonization could be any of such analyses. We therefore selected derivation of biomarkers of Alzheimer's disease (AD) from aggregated neuroimaging data as our downstream task and used it to study scanner effects and investigate two harmonization methods. We hypothesized that *the pipeline of cross-scanner technical variability removal from both images and image-derived measures would result in better removal of unwanted variability and consequently would improve harmonization of our image-derived biomarkers of AD.* Accordingly, we selected RAVEL (Fortin et al., 2016) and ComBat (Johnson et al., 2007) for removing technical variability from images and image-derived measures, respectively. RAVEL is a framework for normalizing and subsequently harmonizing images by removing their inter-subject variability. ComBat is a location and scale adjustment method for harmonization of image-derived measures.

Additionally, we assumed that scanner effects and harmonization can be estimated as dissimilarity and similarity within paired neuroimaging data, respectively. Accordingly, we collected a paired dataset consisting of 16 healthy subjects scanned on General Electric (GE) 1.5T and Siemens 3T MRI scanners and designed and evaluated a set of similarity and dissimilarity metrics on this data.

1.2 Developing image harmonization methods for T1-weighted MRIs

As mentioned earlier, harmonization methods can target either images or image-derived measures and can be categorized as either task-agnostic or task-specific methods. Among the various harmonization scenarios, task-agnostic image harmonization methods represent a more generalized and interpretable category. They are generalized in that they can serve as an independent preprocessing step for any downstream tasks and are considered interpretable because harmonization accuracy can be assessed directly on images. Therefore, we chose to focus on developing task-agnostic image harmonization approaches. In this work, we have introduced both supervised and unsupervised image harmonization methods tailored specifically for T1-weighted MRIs.

1.2.1 MISPEL: Multi-scanner Image harmonization via Structure Preserving Embedding Learning

We have devised a supervised image harmonization method named MISPEL (Multi-scanner Image Harmonization via Structure Preserving Embedding Learning). Our hypothesis posits that *harmonization can be achieved for scanners within a matched dataset by constructing a model that maps matched images from the dataset to a scanner-middle-ground space, where matched images lose scanner effects by becoming similar to each other.* To address this, MISPEL was developed with several key objectives: (1) generalization to multiple (more than two) scanners, (2) preservation of the structural (anatomical) information of the original brains, and (3) learning harmonization on a matched dataset. Subsequently, MISPEL can harmonize unmatched images from the scanners for which the matched dataset was collected.

To train and validate our model, we collected a matched image dataset comprising 18 subjects scanned across four 3T scanners. We proceeded with the assumption that scanner effects and harmonization could be inferred through the measurement of dissimilarities and similarities within the matched images, respectively. Enhancing our metrics for image similarity and dissimilarity, we rigorously assessed the harmonization efficacy of MISPEL by

evaluating (1) visual similarity of images, (2) similarity in gray matter-white matter contrast, (3) consistency in volumetric and segmentation attributes, and (4) similarity in biological attributes. Additionally, we focused on small vessel disease (SVD) as the clinical marker of interest within the matched dataset, exploring whether MISPEL harmonization could maintain or even enhance group distinctions associated with SVD.

1.2.2 ESPA: An unsupervised harmonization framework via Enhanced Structure Preserving Augmentation

We developed an unsupervised image harmonization framework, named ESPA, in addition to an extensive set of experiments for evaluating this framework. Our hypothesis posits that *harmonization for scanners can be acquired through mappings to their scanner-middleground domain via a framework that concurrently simulates matched data for the scanners using appearance-based augmentation methods and learns the corresponding mappings from this simulated data.* For this hypothesis, we developed the ESPA framework with the following objectives: (1) generalizing to multiple scanners (more than two), (2) addressing the over-correction issue during harmonization, (3) preserving the structural (anatomical) information of brains, and (4) enhancing the robustness of harmonization methods, particularly the supervised harmonization methods. ESPA represents an extension of MISPEL with a significant modification: rather than relying on matched data, we employ two novel structure-preserving augmentation methods to simulate matched data. These methods, namely tissue-type contrast augmentation and GAN-based residual augmentation, focus on modifying the appearance and contrast of images.

We additionally formulate a comprehensive set of evaluation criteria based on the matched data gathered for MISPEL. Our evaluation encompasses five key analyses: (1) validation of domain adaptation in augmentation methods, (2) validation of brain structure preservation in augmentation methods, (3) validation of augmentation removal in ESPA, (4) validation of ESPA harmonization, and (5) an ablation study. These criteria serve as the basis for comparing ESPA with the current state-of-the-art supervised and unsupervised harmonization methods.

2.0 Background

In this chapter, Section 2.1 delves into brain MRI and its associated biomarkers. Section 2.2 elaborates on brain MRI artifacts and outlines preprocessing steps to mitigate them. Section 2.3 provides background information on multi-scanner neuroimaging data and the impact of scanner effects. Section 2.4 discusses various approaches used for harmonization and outlines the goals that should be set for this task. Furthermore, Section 2.5 describes the limitations and challenges inherent in current harmonization methods. Lastly, Section 2.6 provides an overview of existing harmonization methods.

2.1 Brain MRI and brain MRI biomarkers

Magnetic Resonance Imaging (MRI) has revolutionized the field of neuroimaging by providing detailed anatomical and functional information about the brain. MRI utilizes a powerful magnetic field and radio waves to generate high-resolution images of brain structures and functions. These images are produced based on the interaction between hydrogen atoms in water molecules and the magnetic field, allowing for exquisite visualization of brain anatomy and pathology (Brown et al., 2014). MRI scans are performed at various imaging sites equipped with different types of MRI scanners. MRI scanners vary in magnetic field strength, with higher-field scanners (e.g., 3 Tesla) providing improved signal-to-noise ratio and spatial resolution compared to lower-field scanners (e.g., 1.5 Tesla). Additionally, advanced MRI scanners may be equipped with specialized coils and sequences that enable enhanced imaging capabilities, such as ultra-high-resolution structural imaging, and functional connectivity mapping. These technological advancements continue to expand the diagnostic and research potential of brain MRI (Jezzard and Clare, 1999; Setsompop et al., 2012).

MRI imaging encompasses various modalities, each tailored to different aspects of brain examination. T1-weighted MRI offers distinct contrasts between brain tissues, rendering cerebrospinal fluid (CSF) dark, gray matter (GM) medium gray, and white matter (WM)

bright. In contrast, T2-weighted MRI is highly sensitive to tissue water content, resulting in CSF appearing bright, GM medium gray, and WM darker than in T1-weighted images. Diffusion-weighted imaging (DWI) captures the random motion of water molecules within tissues, making it particularly useful at detecting acute stroke. Diffusion tensor imaging (DTI) delves into white matter microstructure by mapping water diffusion in multiple directions, offering valuable insights into WM integrity and connectivity (Basser et al., 1994; Le Bihan et al., 2001). Additionally, functional MRI (fMRI) measures brain activity by detecting changes in blood flow and oxygenation levels, providing insights into brain function during tasks or at rest.

MRI serves as a cornerstone in clinical neurology, playing an essential role in diagnosing and monitoring plenty of neurological conditions. Its utility spans from detecting and characterizing brain tumors, vascular abnormalities, neurodegenerative diseases, to inflammatory disorders such as multiple sclerosis. Moreover, MRI stands as a pivotal tool in preoperative planning, treatment surveillance, and post-treatment evaluation. Complementing its diagnostic abilities, functional MRI (fMRI) offers clinicians a non-invasive means to map brain activity, aiding in the localization of critical brain regions prior to surgical interventions and facilitating the assessment of neurological function in patients with brain injuries or disorders (Filippi et al., 2013; Raichle, 2009).

Concurrently, MRI biomarkers emerge as quantifiable metrics derived from MRI data, shedding light on specific facets of brain structure, function, or pathology. Within the realm of neurological disorders, notably Alzheimer's disease (AD), these biomarkers serve as invaluable tools for diagnosis, monitoring disease progression, and advancing our understanding of the underlying mechanisms. In AD research, MRI biomarkers offer insights into brain atrophy, white matter integrity, and alterations in functional connectivity, thereby enriching our comprehension of the disease trajectory (Filippi et al., 2013; Raichle, 2009).

2.2 Brain MRI artifacts and preprocessing steps

Like any imaging modality, MRI of the brain is susceptible to various artifacts that can degrade image quality and compromise diagnostic accuracy. An artifact, in the context of medical imaging, refers to any anomaly or distortion present in the image that is not representative of the true anatomy or pathology being imaged (Mahesh, 2013). Understanding the causes of these artifacts, their identification, and appropriate preprocessing steps is crucial for obtaining reliable brain MRI images for clinical interpretation and research analysis. Brain MRI artifacts can arise from a variety of sources, including hardware imperfections, physiological factors, patient motion, and environmental interference. Common types of artifacts encountered in brain MRI include:

- **Motion artifact:** Patient motion during image acquisition can lead to blurring or ghosting of brain images, particularly problematic in studies involving pediatric or restless patients (Maclaren et al., 2018).
- **Susceptibility artifact:** Variations in magnetic susceptibility between tissues can cause signal loss or distortion, particularly at tissue-air interfaces such as the sinuses or near metallic implants (Brown et al., 2014).
- **Gradient non-linearity artifact:** Imperfections in the magnetic field gradients can lead to spatial distortions and misregistration of brain structures (Nacher, 2007).
- **Chemical shift artifact:** Differences in precession frequencies between fat and water molecules can result in misregistration and signal misinterpretation, particularly problematic in spectroscopic imaging (Dixon, 1984).
- **RF interference artifact:** External radiofrequency (RF) interference from nearby electronic devices can introduce spurious signals, manifesting as bright or dark bands in the brain image (Lustig et al., 2007).

Accurate identification of brain MRI artifacts is essential for implementing appropriate mitigation strategies. Visual inspection by trained radiologists remains a primary method for artifact identification. Additionally, various software tools and algorithms have been

developed to automatically detect and correct specific types of artifacts (Newton, 2016). Mitigation strategies may include:

- **Motion correction techniques:** Utilizing motion tracking and retrospective image registration algorithms to correct for patient motion during brain image acquisition (Maclaren et al., 2013).
- **Gradient distortion correction:** Calibration and correction algorithms to compensate for gradient non-linearity and spatial distortions in brain MRI (Nacher, 2009).
- **Susceptibility artifact reduction:** Employing specialized sequences such as susceptibility-weighted imaging (SWI) or multi-echo gradient echo sequences to minimize susceptibility artifacts in brain imaging (Haacke et al., 2009).
- **Parallel imaging:** Utilizing parallel imaging techniques to accelerate brain MRI acquisition and reduce susceptibility to motion artifacts (de Zwart et al., 2006).
- **Post-processing filtering:** Application of image filtering techniques, such as spatial and temporal filtering, to reduce noise and enhance brain MRI image quality (Castleman, 1996).

Preprocessing of brain MRI data is essential for optimizing image quality and preparing the data for further analysis. Common preprocessing steps include:

- **Noise reduction:** Applying noise reduction techniques, such as Gaussian smoothing or wavelet denoising, to improve signal-to-noise ratio (SNR) in brain MRI images (Manjón et al., 2008).
- **Intensity normalization:** Normalizing brain MRI image intensities to correct for intensity unit effects possibly stem from variations in scanner parameters and acquisition protocols (Shinohara et al., 2014b).
- **Motion correction:** Registration-based techniques to correct for inter-slice and intra-slice motion artifacts in brain MRI (Thesen et al., 2000).
- **Bias field correction:** Correction of intensity variations caused by non-uniformity in the RF field or gradient non-linearity in brain MRI (Tustison et al., 2010).
- **Spatial registration:** Alignment of brain MRI volumes to a common reference space for inter-subject analysis or longitudinal studies (Avants et al., 2008).

Brain MRI artifacts represent a significant challenge in clinical neuroimaging, potentially compromising diagnostic accuracy and reliability. Understanding the underlying causes of artifacts and implementing appropriate preprocessing steps are essential for obtaining high-quality brain MRI images suitable for clinical interpretation and research analysis.

2.3 Multi-scanner MRI data and scanner effects

Having a large sample size serves as a motivating factor for many neuroimaging studies to collect imaging data from a diverse range of datasets. These datasets are typically acquired from different sites, each equipped with its own set of scanners. Examples of such multi-site studies include the Adolescent Brain Cognitive Development (ABCD) (Jernigan et al., 2018), the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005), and the Australian Imaging, Biomarkers and Lifestyle Flagship Study of Aging (AIBL) (Ellis et al., 2009). The presence of a large sample size brings several advantages to both hypothesis-driven and exploratory analyses. One such benefit is the increase in statistical power during hypothesis testing, leading to heightened confidence in rejecting the null hypothesis and uncovering true effects (Suresh and Chandrashekara, 2012). Another advantage is the enhanced generalizability of data and study outcomes. Aggregating data across multiple datasets allows for more comprehensive coverage of biological and clinical variables, such as age, race, gender, and health status, within neuroimaging studies. This comprehensive data makes it feasible to conduct studies that necessitate variables with a wide range of values. For instance, the study of the association between age and brain volumes requires a substantial age span (Pomponio et al., 2020). Furthermore, studies utilizing multiple datasets can serve as a testbed for validating the generalizability of study outcomes on independent cohorts (Ramspek et al., 2021). Additionally, in studies employing machine learning models as outputs, the utilization of large, varied, and representative datasets can facilitate the development of models that generalize well to unseen data (An et al., 2022; Aslani et al., 2020; Dinsdale et al., 2021).

The benefits previously mentioned can potentially be undermined by *scanner effects*.

These effects, which refer to variations occurring post-normalization/standardization that are unrelated to biological factors (Fortin et al., 2016), stem from differences in scanners and acquisition methods (Dinsdale et al., 2021). When dealing with datasets involving multiple scanners, scanner effects emerge as a significant source of variability, surpassing MRI artifacts due to the intricate differences among scanners, including hardware, software, and acquisition protocols (Han et al., 2006; Jovicich et al., 2006; Takao et al., 2011). Research suggests that scanner effects persist prominently even after within-scanner preprocessing steps aimed at mitigating MRI artifacts, even gradient non-linearity as a scanner-related artifact (Fortin et al., 2016). Scanner effects appear to be unavoidable during the process of aggregating or even collecting neuroimaging data (Dewey et al., 2021). The only scenario that can eliminate scanner effects in data collection for a study is to exclusively collect the data at a single site using a single scanner. However, this is far from the typical data collection process. Even if we limit ourselves to a single site, neuroimaging datasets are usually acquired using the available set of scanners at that site. As a result, scanner effects become an inherent phenomenon in most of the available neuroimaging datasets. Moreover, there is always the possibility of scanner upgrades or replacements at a site. While this may rarely occur during the data collection phase of cross-sectional studies, it can be highly probable in longitudinal studies (Sederevicius et al., 2022). Similarly, any other scenario involving data collection and aggregation across multiple sites leads to the presence of scanner effects too.

It has been widely demonstrated that scanner effects can have an adverse impact on downstream analyses in neuroimaging data (Fortin et al., 2018, 2017, 2016). These effects can introduce bias, overshadow the intended biological or clinical signal of interest, and consequently render the results unreliable. Several studies have highlighted the presence of scanner effects in derived measures of regional healthy tissue and brain lesion volumes (Jovicich et al., 2013; Schnack et al., 2010; Schwartz et al., 2019). Shinohara et al. (2014a) also provided evidence of striking differences in raw image intensities across different sites in the ADNI (Mueller et al., 2005) and AIBL (Ellis et al., 2009) studies. In addition, Heinen et al. (2016) investigated scanner effects on brain volume measures extracted from pooled MRIs acquired using both 1.5T and 3T scanners. They showed that scanner effects can manifest as variations in volumetric accuracy when manual tissue segmentations were

available for evaluation. This evidence highlights the importance of harmonization pipelines as a means of addressing scanner effects. Such pipelines play a crucial role in ensuring the reproducibility of analyses and fostering trust in the results of neuroimaging studies (Karayumak et al., 2019; Ning et al., 2020; Yu et al., 2018).

2.4 Harmonization approaches and goals

There are two prevailing perspectives regarding harmonizing neuroimaging data. The first perspective involves harmonizing either (1) images or (2) image-derived measures, leading to two broad categories of harmonization methods. Harmonizing images presents greater challenges due to the complexity of neuroimages, yet offers greater interpretability when assessing harmonization accuracy at the image level (Torbati et al., 2021). Alternatively, harmonization can be viewed as either an independent preprocessing step providing harmonized data for downstream tasks (Dewey et al., 2019), or as an integral component of methods targeting specific tasks (Dinsdale et al., 2021). Methods falling into the former category are termed task-agnostic, while those in the latter are referred to as task-specific harmonization. In the task-specific approach, harmonization is embedded within a model designed for a specific downstream task, allowing for leveraging task-related signals during the harmonization process but potentially limiting generalizability (An et al., 2022).

Concerning the harmonization approach, task-specific methods are heavily influenced by the task they aim to address. However, in task-agnostic harmonization, two primary approaches are commonly employed: (1) removal of scanner effects from the data (Fortin et al., 2016; Johnson et al., 2007), and (2) adaptation of the data to a *target* scanner domain or a scanner-variant component of a *target* individual's data (Dewey et al., 2019; Liu et al., 2023; Zuo et al., 2021b). In the former approach, scanner effects are regarded as unwanted variability that can be estimated and eliminated from the neuroimaging data. Conversely, the latter approach considers scanner effects as causing domain shift, treating data from different scanners as distinct domains. Harmonization is achieved by adapting the data to (1) a scanner-middle-ground domain, (2) the domain of a target scanner, or (3) the scanner-

variant component of data for a target individual, resulting in harmonized images with similar scanner characteristics.

Regardless of the chosen harmonization perspective and approach, harmonization methods are generally expected to achieve three main objectives: (1) addressing scanner effects, (2) preserving the biological and clinical information within the data, and (3) enhancing downstream tasks by mitigating the impact of scanner effects (Beer et al., 2020; Dewey et al., 2019, 2020).

2.5 Limitations and challenges of harmonization approaches

The current harmonization approaches and methods may present certain limitations and challenges.

- **Over-correction.** This phenomenon refers to the correction of biological or clinical variables, potentially in addition to or instead of scanner effects (Liu et al., 2021, 2023). Depending on the chosen harmonization approach, there is a risk of over-correction if scanner effects are statistically correlated with other variables in the data (Bayer et al., 2022a). This occurrence has the potential to disrupt downstream studies conducted on harmonized data, as it can impact the desired biological and clinical variables or confounding factors that are necessary for data modeling (Solanes et al., 2021).
- **Brain structural modifications.** This phenomenon can occur in image harmonization methods and involves the modification of the brain's structure during the harmonization process. This effect is predominantly observed in methods that utilize Image-to-Image translation approaches for harmonization (Torbati et al., 2023; Zuo et al., 2021b).
- **Matched data requirement.** *Matched images* are images of the same individual captured by more than two scanners within a short time period. A *matched image dataset* consists of such matched images for multiple individuals. Matched images are expected to be images of biologically similar brain with differences due to solely scanner effects. *Matched data* can be matched image datasets or datasets of any measures derived from matched images. Matched data serves as a means to identify and quantify scanner ef-

fects. Dissimilarities observed within matched images enable the detection of scanner effects. Matched data essentially serves as *labeled* data for the harmonization task, and methods that utilize such data are known as *supervised* methods. Supervised image harmonization methods could be designed to leverage matched data to learn harmonization with least possibility of biological, clinical, and structural modifications and therefore tackle the two former challenges. However, the availability of matched data restricts the applicability of supervised harmonization methods to the data they can harmonize. Furthermore, supervised harmonization methods are susceptible to a lack of robustness due to the inherently limited size of matched data.

- **Additional labeled data requirement.** Some harmonization methods utilize additional labeled data from desired downstream tasks during the harmonization process. These methods typically leverage this data to preserve various aspects of the original data, such as brain anatomy, throughout the harmonization procedure.
- **Target domain determination.** There exists a group of harmonization methods that learns mappings for adapting data of scanners to a selected domain, called *target* domain. Using these mappings, data of all scanners could get the same scanner characteristics as that of the target domain. A target domain could be the domain of data (image or image-derived measures) in a target scanner or the scanner-variant component of a single target image. Even though adapting data to a target domain seems as a straightforward harmonization approach, it introduces the new challenge of determining the "best" domain. Selecting such domain is not a trivial task when, for example, motion artifacts in images could be of concern (Alexander-Bloch et al., 2016; Torbati et al., 2021). For instance, Tian et al. (2022) employed visual screening to select their target scanner. However, this approach may not be the most reliable strategy since factors that lead to errors or inefficiency in downstream tasks could be imperceptible to the human eye.
- **Number of scanner limitation.** The effectiveness of harmonization methods in accommodating multiple scanners depends on the specific methodology employed. Consequently, certain harmonization approaches may encounter limitations regarding the number of scanners they can effectively harmonize within a given dataset (Torbati et al., 2023).

2.6 Related work

A significant number of harmonization methods have been specifically developed for the harmonization of diffusion MRI data. In a study by Pinto et al. (2020), these methods were classified into two categories based on the type of data they aim to harmonize: diffusion parametric map harmonization and diffusion weighted image harmonization. The diffusion parametric map harmonization category includes works such as those by Jahanshad et al. (2013); Kochunov et al. (2014); Prohl et al. (2019); Salimi-Khorshidi et al. (2009); Teipel et al. (2012); Timmermans et al. (2019); Zhu et al. (2019). On the other hand, the diffusion weighted image harmonization approaches consist of methods proposed by Fortin et al. (2017); Hansen et al. (2022); Karayumak et al. (2019); Mirzaalian et al. (2015, 2016, 2018). It is worth noting that these methods are specifically designed for diffusion MRI data and are data-dependent, which limits their applicability to other imaging modalities. Therefore, we have excluded this group from our literature review.

In this section, we delve into harmonization methods with broader applicability across various data modalities. We begin by categorizing these methods according to the type of data they aim to harmonize, distinguishing between the harmonization of (1) images and (2) image-derived measures. Further categorization includes (1) task-agnostic and (2) task-specific harmonization methods. We then provide a detailed exploration of methods within each category, focusing on their specific harmonization approaches: (1) removal of scanner effects, (2) adaptation of data to a scanner-middle-ground domain, (3) adaptation of data to a target scanner domain, (4) adaptation of data to a target image contrast or style, and (5) task-related approach. Additionally, we summarize this classification in Table 1.

2.6.1 Harmonizing images

In this section, we will discuss the methods specifically designed for harmonizing images.

2.6.1.1 Task-agnostic harmonization

- **Removal of scanner effects**

  In this group of methods, the primary objective is to achieve image harmonization by estimating and removing scanner effects. One notable approach within this category is RAVEL (Removal of Artificial Voxel Effect by Linear Regression) (Fortin et al., 2016), which provides a framework for intensity normalization and harmonization. The process begins with a White Stripe normalization step (Shinohara et al., 2014b), followed by a voxel-wise harmonization strategy applied to the images. Within this strategy, RAVEL utilizes singular value decomposition on CSF voxels, which are known to be unaffected by disease status and clinical covariates. By doing so, RAVEL estimates the components associated with scanner effects. RAVEL then employs these estimated components to harmonize the images by removing the scanner effects from the voxel intensities using a linear regression model. For a more comprehensive understanding, please refer to Algorithm 1 in Section 3.2.1.

  This harmonization approach appears to be the most straightforward strategy for image harmonization; however, the estimation of scanner effects can pose challenges. For instance, when dealing with images containing motion artifacts, the application of RAVEL may lead to inconsistent image harmonization across subjects. These artifacts can affect the CSF area of the brain, potentially resulting in their extraction as scanner effects.

- **Adaptation of data to a scanner-middle-ground domain**

  This group of methods typically consists of supervised harmonization approaches that utilize additional matched data provided for multi-scanner datasets. The purpose of these methods is to learn mappings that facilitate the adaptation of matched scanner domains to a scanner-middle-ground domain. Within this domain, the matched images become more similar across scanners, resulting in a reduction in scanner effects. Since the domain was learned using matched images, the harmonization mapping is expected to minimize biological, clinical, and structural modifications to the least extent possible. The learned mappings can then be applied to harmonize images from each individual scanner within the multi-scanner dataset, even if the images are not necessarily matched. However, it is

important to note that these methods have limitations and can only harmonize images from scanners with available matched data.

DeepHarmony (Dewey et al., 2019) is a supervised deep-learning harmonization framework that utilizes two U-Net networks (Ronneberger et al., 2015) to learn mappings between two scanners. The first U-Net is employed to learn a mapping from the domain of the first scanner to the domain of the second scanner. Subsequently, the second U-Net learns a mapping from the images of the second scanner to the domain of the output images in the first U-Net. By applying these learned mappings to the images from both scanners, all the images are transformed into a shared middle-ground domain, resulting in harmonization. However, it is important to note that this methodology, which focuses on learning the mappings between two scanners, restricts the applicability of DeepHarmony to multi-scanner data with only two scanners.

- **Adaptation of data to a target scanner domain**

  This group of methods involves learning mappings that facilitate the adaptation of images from the scanners within multi-scanner datasets to the domain of a target scanner. By applying these mappings, the images from all scanners can effectively lose scanner effects and achieve similar scanner characteristics, specifically those of the target scanner. However, harmonization methods using this approach present the user with the challenge of selecting the target scanner.

  Mica (Multi-site Image harmonization by cumulative distribution functions Alignment) (Wrobel et al., 2020) and RIDA (Robust Intensity Distribution Alignment) (Sederevicius et al., 2022) are two supervised harmonization methods that utilize additional matched data to learn mappings from images of one scanner to those of a target scanner. These methods operate based on the assumption that the mappings between scanners can be represented as transformations in the cumulative distribution functions (CDFs) of their respective matched images. Specifically, they learn these CDF transformations for all matched images and subsequently utilize their average to derive the ultimate mapping for harmonization. This mapping can then be utilized to harmonize images of the scanners, even if they are not necessarily matched. Additionally, RIDA introduces the hypothesis that scanner effects may have distinct impacts on different brain regions. As a result,

RIDA incorporates image segmentation and considers CDF transformations for specific brain regions rather than the entire brain image. It is important to note that both methods were designed specifically for harmonizing data involving only two scanners.

The remaining methods in this category are unsupervised methods. A significant portion of these methods employs CycleGAN (Zhu et al., 2017) to learn the mappings between images from two different scanners. While Zhao et al. (2019) apply CycleGAN to harmonize cortical thickness maps, Modanwal et al. (2020) and Bashyam et al. (2020) use this framework for harmonizing dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) of breasts and T1-weighted MRIs of brains, respectively. Regardless of the targeted data type, these methods rely on the CycleGAN framework illustrated in Figure 2 to achieve image harmonization. As depicted in the figure, CycleGAN consists of two generative adversarial networks (GANs). Given the input image from ScannerA (X_A), generator $G_B : ScannerA \rightarrow ScannerB$ is used to generate harmonized images of ScannerA ($X_A^{'} = G_B(X_A)$). Discriminator D_B was designed to classify harmonized images of ScannerA, i.e. $X_A^{'}$, from images of ScannerB ($X_B \in ScannerB$). The other GAN is designed to learn the mapping from ScannerB to ScannerA. For this, the generator $G_A : ScannerB \rightarrow ScannerA$ is designed to generate harmonized images of ScannerB ($X_B^{'} = G_A(X_B)$). The discriminator D_A was designed to classify harmonized images of ScannerB, i.e. $X_B^{'}$, from images of ScannerA ($X_A \in ScannerA$). The purpose of having the second GAN is to provide a corresponding reconstruction loss for images. This loss for X_A is equal to $||X_A - X_A^{''}||$ in which $X_A^{''} = G_A(X_A^{'}) = G_A(G_B(X_A))$. The same reconstruction loss can be defined for X_B as $||X_B - X_B^{''}||$ in which $X_B^{''} = G_B(X_B^{'}) = G_B(G_A(X_B))$. These reconstruction losses help the whole harmonization network to learn a one-to-one mapping between images of scanners. Either of the trained generators (G_A or G_B) could then be used to map images of their source scanner to the domain of their target scanner. These methods that utilize the CycleGAN network may have several limitations. Firstly, they are restricted to harmonizing multi-scanner data involving only two scanners. Secondly, the learned mappings in these methods lack interpretability, meaning that these mappings can introduce changes to image variabilities beyond the scanner effects. This can lead to two significant drawbacks. Firstly, the mappings have the potential to alter

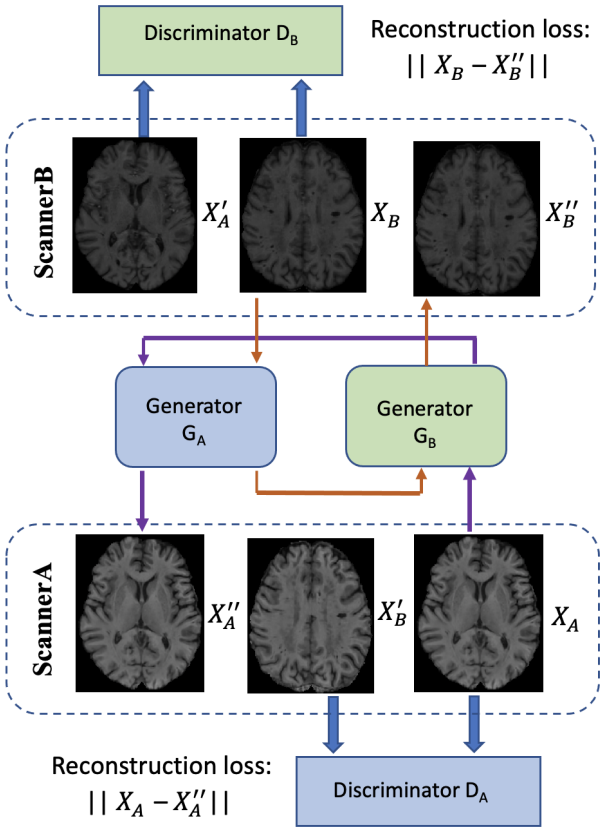


Figure 2: **Harmonization with CycleGAN method.** This figure was modified and borrowed from (Zhong et al., 2020).

the anatomy of images, particularly when GANs are susceptible to the mode collapse problem (Thanh-Tung and Tran, 2020). Secondly, the images from the source scanner may acquire population-wise characteristics of the target scanner if the domains (image sets of the two scanners) significantly differ in that aspect. We refer to this phenomenon as over-correction.

A group of harmonization methods has been developed to address these limitations. Bashyam et al. (2022) propose the use of StarGAN (Choi et al., 2018) as an alternative to CycleGAN, enabling harmonization across more than two scanners. StarGAN employs a conditional GAN (cGAN) framework, which incorporates an additional input: a one-hot vector representing the scanner labels. This vector specifies the target scanner for

the cGAN. The cGAN is trained to learn mappings between all pairs of scanners by conditioning each iteration on learning a mapping from the input image to a randomly selected target scanner. To ensure cycle-consistency, i.e., the reconstruction loss, the reverse mapping is also performed in each iteration. During the harmonization process, the generator in the trained StarGAN requires input images from all scanners, along with the label indicating the target scanner.

A group of methods has focused on modifying the CycleGAN framework to control and enhance the interpretability of the learned mappings. Robinson et al. (2020) propose that scanner effects can be interpreted as appearance and spatial transformations in images. To incorporate this idea, they modify the CycleGAN framework by introducing image-and-spatial transformation networks (Lee et al., 2019) as the generator modules. These networks help constrain the mappings to meaningful appearance and shape transformations. In another approach, Ren et al. (2021) adopted a multi-task learning strategy to enhance the harmonization of the CycleGAN network while preserving anatomical information during image generation. They considered two tasks: harmonization and brain segmentation. Simultaneously, they trained two supervised deep learning segmentation models alongside the CycleGAN harmonization network. Each segmentation module was assigned to one of the two scanners, and they were trained on both the original images from their respective scanner and the images that underwent harmonization to match the domain of the assigned scanner. To preserve the anatomy during harmonization, the authors incorporated the modified output of the segmentation models as image structural priors within the layers of the generators in the CycleGAN framework. The rationale behind this transfer was to retain the anatomical information of the images during the mapping process for harmonization. Consequently, the transfer was made from the segmentation model of each scanner to the generator responsible for mapping the images of that specific scanner. However, it is important to note that this harmonization approach is limited to datasets that possess segmentation labels for images from both scanners. In a different method, Chang et al. (2022) employed a two-stage framework consisting of a CycleGAN for learning the mappings and a histogram matching module to recover any anatomical information that may have been lost during the mapping process. While this

approach was initially applied to pelvic MRIs, it could potentially be extended to brain MRIs as well.

Another group of methods also aims to map images to the domain of a target scanner, but their approach to harmonization differs from learning cross-scanner mappings. These methods employ adversarial autoencoders (AD-AEs) or variational autoencoders (VAEs) to learn latent embeddings that are invariant to scanner variations. These embeddings capture essential image information while remaining independent of the specific scanner used. During network training, the models learn two key tasks: (1) extracting these scanner-invariant embeddings from the images, and (2) reconstructing the images back to their original domain using the generated embeddings. During the harmonization process, instead of reconstructing the images in their original scanner's domain, the model constructs the images in the domain of the target scanner. This approach enables harmonization across multiple scanners, making these methods suitable for datasets involving more than two scanners.

HarMOnAE (Fatania et al., 2022) utilizes an AD-AE network to implement this harmonization strategy for images. It employs a single AD-AE for harmonizing images from all scanners. During training, the encoder takes an input image X (from any scanner) and produces its latent embedding Z. This embedding is then passed to a pre-trained classifier that predicts the scanner of image X. The prediction loss from this classifier is used to adversarially update the encoder, generating scanner-invariant embeddings Zs. Finally, the decoder uses the scanner ID of image X and the generated Z to reconstruct the image in its original scanner domain. The trained encoder-decoder can be used to harmonize images from all scanners, with the decoder requiring the scanner ID of the target scanner for reconstructing images in that specific domain. In a similar harmonization framework, Moyer et al. (2020) employ a VAE network instead of an AD-AE. They propose that learning scanner-invariant embeddings can be achieved by penalizing the network to minimize the mutual information between Z and the scanner ID of Z's corresponding image. This strategy serves as a substitute for the scanner classifier used in the AD-AE. Although initially implemented for diffusion MRIs, this method can be adapted for other modalities as well.

Both of these methods may encounter the over-correction phenomenon. Moreover, they are susceptible to information loss during harmonization, as the generated latent embeddings have been shown to be biased towards the least informative scanner (Moyer and Golland, 2021).

- **Adaptation of data to a target image contrast or style**

  This group of methods aims to adapt images of scanners to the scanner-variant components of a target image, rather than domain of a target scanner. During the model training phase, these methods learn two key tasks: (1) extracting or generating the scanner-variant components of the images, and (2) reconstructing the original images by incorporating these components into the process. When it comes to harmonization, the trained models utilize the scanner-variant component of a target image to reconstruct the images. By employing this harmonization strategy, these methods can overcome the main challenges faced by harmonization methods that target the scanner domain. Specifically, they can address the limitations of being restricted to harmonizing data from only two scanners and the lack of interpretable harmonization transformations. Such advantages can be realized because these methods have a framework that is applicable to images from any number of scanners, and their harmonization transformation is intentionally designed to focus on the scanner-variant components. The primary challenge in these methods is the precise extraction of the scanner-variant components from the images. Any inaccuracies in this extraction process can lead to changes in the harmonized images. For instance, if the scanner-variant components still contain structural information of the target image, the harmonized images may undergo brain structural modifications. Additionally, if the scanner-variant components encompass biological or clinical information of the target image, the harmonization process may result in an over-correction of these variabilities in the images.

  Methods in this group focus on the contrast-based components or style of images as the scanner-variant components. The CALAMITI method was initially proposed in (Dewey et al., 2020) and later improved upon in (Zuo et al., 2021a,b). CALAMITI considers the contrast-based components of images as its scanner-variant component. It comprises an encoder-decoder network and requires additional inter-modality paired data for the

harmonization of multi-scanner data. The inter-modality paired image dataset comprises images from two predetermined modalities, captured from each subject using the same scanner with a short time gap. In this dataset, individuals undergo scanning using different scanners, while their inter-modality paired images are acquired on the same scanner. This dataset assists CALAMITI in disentangling images into their scanner-variant components (scanner effects) and scanner-invariant components (anatomical information). The encoder in CALAMITI takes an inter-modality image pair from any scanner and learns to encode these images into their scanner-variant and scanner-invariant components. The two extracted scanner-invariant components of the image pair are expected to be similar as they originate from similar brains. The decoder is then trained to reconstruct the input image pair using their respective scanner-variant components and one of the two decomposed scanner-invariant components, selected randomly during model training. The incorporation of randomness and a designed image reconstruction loss encourages CALAMITI to learn the disentanglement of images into their scanner-variant and -invariant components. For image harmonization, CALAMITI first decomposes the image into its components and then reconstructs the image using its scanner-invariant component and the scanner-variant component of a target image. Later, Tian et al. (2022) developed a supervised version of CALAMITI using paired image datasets instead of inter-modality paired data.

The other methods in this group do not use any type of paired data for extracting the scanner-variant component of images. They consider the style of images as the scanner-variant component and used unsupervised style transfer strategies for harmonization. For example, Liu and Yap (2021) proposed a content-style disentangled cycle translation framework for harmonization. In their framework, they first use two individual encoders to disentangle images into their scanner-variant and -invariant components. They then use a cycle-consistent GAN framework to learn style-based transformations between images of scanners. For such transformation, they modify the generators to learn mappings from the two components of images, instead of the images. This way, the generator takes the scanner-invariant component of an image as well as the scanner-variant component of a target image and learn to map the image to style of the target image. The learning

strategy in this cycle-consistent GAN network resembles that of the CycleGAN network. This cycle-consistent network can also provide the two encoders with two separate consistency losses. These losses are used to force the encoders to learn disentangling images into their scanner-variant and -invariant components.

In another approach, Liu et al. (2021, 2023) also employ the cycle-consistent style translation framework. However, unlike other methods, they do not disentangle images into scanner-variant and -invariant components. Instead, they focus solely on extracting the style of images as the scanner-variant component and utilize it to map images to the style of a target image. For extracting style of images, they train an encoder on images. For the style transformation, they employ a modified CycleGAN network with generators that take images and the target style as input to generate the transformed images. This modified CycleGAN network incorporates consistency losses for both the images and the styles. The style consistency loss plays a crucial role in compelling the encoder to learn the extraction of image styles.

2.6.1.2 Task-specific harmonization

This group of methods approaches harmonization as a task-specific problem. In this approach, harmonization is not treated as a standalone preprocessing task; rather, it is integrated into a model that is specifically designed for a desired downstream task. This approach offers several advantages, such as (1) the ability to leverage the downstream task for harmonization, and (2) the ability to account for confounding factors during data harmonization. However, it is important to note that this approach may limit the generalizability of the harmonization method, as it is tailored to be task-specific rather than task-agnostic.

Methods in this group (Aslani et al., 2020; Dinsdale et al., 2021) utilize a traditional encoder-decoder network along with an adversarial classifier to predict the domain (scanner) of images. While the encoder-decoder network handles the main task, the adversarial classifier's role is to eliminate the scanner effects from the encoder-decoder. The network training in this framework consists of three main stages. In the first stage, the encoder-decoder network is optimized for the main task. The encoder extracts latent embeddings from the input

image, and the decoder learns to perform the main task using these embeddings. In the second stage, the adversarial classifier is individually trained to take the latent embeddings from the encoder and predict the scanner of the image. In the third stage, the two encoder-decoder networks and the classifier are trained simultaneously and adversarially. During this stage, the encoder-decoder network should remain optimized for the main task, while the encoder generates embeddings that confuse the classifier. The trained encoder-decoder network then serves as a scanner-invariant framework for the main task. Aslani et al. (2020) proposed such a framework for the brain segmentation task. Dinsdale et al. (2021) extended this idea with three similar frameworks, each designed for regression, classification, and segmentation tasks respectively. They also expanded their network to account for confounders other than the scanner effects by adding an adversarial classifier for each confounder.

These methods employ a similar strategy to the adversarial and variational autoencoder models, which aim to learn scanner-invariant latent embeddings for harmonization. Likewise, these methods may encounter the issue of over-correction and the potential loss of image information during the harmonization process. It has been demonstrated that the generated latent embeddings of these methods tend to be biased towards the least informative scanner (Moyer and Golland, 2021).

2.6.2 Harmonizing image-derived measures

In this section, we will outline the methods specifically developed for harmonizing image-derived measures.

2.6.2.1 Task-agnostic harmonization

- **Removal of scanner effects**

  ComBat (Johnson et al., 2007) and its extensions are widely recognized methods in this category. ComBat is a location and scale adjustment method that utilizes an empirical Bayes (EB) framework to harmonize image-derived measures in multi-scanner data, particularly when there are only a few images available for each scanner. It addresses

harmonization individually for each image-derived measure while leveraging information from all other measures in the process. ComBat operates on three key assumptions. Firstly, it assumes that each measure can be modeled as a linear combination of the overall mean of the measure, individual biological/clinical variables, scanner effects, and Gaussian noise. Secondly, it assumes that scanner effects manifest as additive and multiplicative factors on the measures. Lastly, it assumes that scanner effects tend to impact all of the derived measures in a similar manner.

ComBat proposes to remove scanner effects from the distribution (the linear model) of a derived measure by correcting the location and scale of its distribution. Accordingly, it proposes to harmonize the measure across scanners by standardizing it, using additive and multiplicative parameters of scanner effects specific to that measure. The ComBat harmonization process begins by normalizing all imaging measures to ensure comparable distributions across measures. This normalization effectively removes measure scales as a source of variability, ensuring consistency in the data. Subsequently, ComBat estimates the scanner effect parameters and applies the standardization procedure for each measure within each scanner individually. For estimating the scanner effect parameters of a particular measure of a scanner, ComBat employs an EB framework to estimate the distribution of the overall mean scanner effects parameters across all imaging measures of the corresponding scanner. It then derives the scanner effects parameters specific to the desired measure from this overall distribution that was estimated for all measures. This approach enables ComBat to handle cases with a limited number of images per scanner by utilizing the pooled information from all measures of that scanner.

ComBat has been applied to derive measures from various modalities, including DTI (Fortin et al., 2017), MRI (Fortin et al., 2018), and fMRI (Nielson et al., 2018; Yu et al., 2018). Several extensions of ComBat have also been proposed to address its limitations. Chen et al. (2020a) investigated the impact of scanner effects on cortical thickness measures obtained from ADNI data (Mueller et al., 2005). They demonstrated that scanner effects can influence the correlation between derived measures across scanners. To address this, they introduced CovBat, which adjusts for scanner effects in both the covariance and the mean and standard deviation of the measures. Pomponio et al. (2020)

explored the association between age (a biological variable) and brain volumes (derived measures) using a large dataset spanning different age groups. To account for scanner effects, they developed ComBat-GAM, a generalized additive model version of ComBat that allows for nonlinear associations between age and brain volumes during the harmonization process. Beer et al. (2020) proposed longitudinal-ComBat, a mixed effects variant of ComBat designed for longitudinal data. This extension considers within-study participant variability, specifically the correlation among imaging measures collected over time for each subject. Maikusa et al. (2021) introduced TS-ComBat, which incorporates individual effects estimated from data obtained from traveling subjects (matched data). This enhancement allows ComBat to account for unknown individual effects in addition to known effects included as biological variables in its linear model.

Other extensions of ComBat have been developed to enhance its harmonization performance. Chen et al. (2022a) introduced distributed ComBat (d-ComBat) for harmonizing distributed data, addressing the challenges of data sharing when restricted by privacy policies. Unlike ComBat, which requires aggregated data for its initial measure-wise standardization step, d-ComBat proposes an estimation process to calculate the necessary parameters, enabling harmonization without sharing data. The remaining harmonization process follows the standard ComBat approach. Da-Ano et al. (2021) proposed bootstrap ComBat (B-ComBat) as a more robust extension. B-ComBat utilizes a Monte Carlo method to repetitively estimate the parameters of scanner effects and subsequently averages them to improve robustness.

Even though ComBat and its extensions have demonstrated acceptable success in harmonizing imaging measures in numerous studies (Foy et al., 2020; Radua et al., 2020; Yu et al., 2018), evaluating their accuracy of harmonization at the image level remains challenging. This lack of transparency makes these methods less interpretable and difficult to improve, as it hinders the investigation of scanner effects and potential causes of harmonization failures. To address this issue, Neuroharmony (Garcia-Dias et al., 2020) was proposed. It utilizes a random forest model to translate the behavior of ComBat into image quality metrics (IQMs). However, this approach adds complexity by indirectly targeting IQMs instead of directly harmonizing the images themselves. Furthermore,

Neuroharmony's harmonization performance is limited to that of ComBat, leading to potential failures in harmonizing measures where ComBat falls short as well.

Another major risk of using ComBat and its extensions, except TS-ComBat, is the over-correction of biological variability (Obenauer et al., 2019). ComBat is prone to removing the biological variability that is correlated to scanner effects and was not known to be considered in the linear model of ComBat. TS-ComBat was designed to address this issue using matched data, however, such data are not available for all studies. There is another group of harmonization methods addressing this issue without requiring matched data (An et al., 2022; Bayer et al., 2022a; Wang et al., 2021). Wang et al. (2021) proposed a normalizing-flow-based method to learn the bijection (one-to-one correspondence mapping) from a neuroimaging measures of a source scanner to that of a target scanner. These mappings are used for harmonizing measures. For coming up with these mappings, they apply a counterfactual inference upon a structural causal model. In their model, the neuroimaging measure is modeled as the result of known confounders (site, gender, and age), and exogenous noise variables of the measure as well as each of the known variables. The mappings are learned through an inference step that addresses the counterfactual question of the form: "what would the value of the measures from source and target scanners would be, if they had been acquired from the same scanner?" This harmonization method can then address the over-correction by preserving the unknown confounders through capturing them as the exogenous noise variables. The two other methods in this category are task-specific harmonization approaches.

2.6.2.2 Task-specific harmonization

Bayer et al. (2022a) introduce harmonization for normative modeling, which involves mapping the variability in biological response variables (e.g., cortical thickness measures) to covariates (e.g., age) in order to redefine the response variable variation explained by the new covariates (e.g., scanner effects). Normative modeling aims to capture individual-level variation, aligning with the principles of personalized medicine. As an example, consider a normative model with cortical thickness measures as the response variable and age as

the covariate. By knowing an individual's age, this model can estimate the deviation of their cortical thickness measures from the normative age curve of cortical thickness (Bayer et al., 2022b), considering other covariates. Bayer et al. (2022a) employed a hierarchical Bayesian method for normative modeling and proposed incorporating the scanner as one of the covariates in the model to achieve harmonization. Instead of removing scanner effects from the model and potentially encountering over-correction issues, they incorporate it within the process of estimating variation in their model.

In another approach, An et al. (2022) proposed incorporating the downstream task into the harmonization process to address the issue of over-correction that may occur for the task. They developed a framework consisting of two networks: (1) a variational autoencoder (VAE) proposed in (Moyer et al., 2020) for harmonization, and (2) a task-specific network. To prevent over-correction, they utilize the performance of the task-specific network as a regularizer during the training of the VAE network for harmonization. By this strategy, they penalize their network if their signal of interest is being over-corrected along with scanner effects.

Table 1: Categorization of harmonization methods

Harmonization Category		**Harmonization Approach**	**Harmonization Method**
Harmonization of images	Task agnostic	Removal of scanner effects	(Fortin et al., 2016)
		Domain adaptation: Scanner middle ground	(Dewey et al., 2019)*
		Domain adaptation: Target domain	(Wrobel et al., 2020)*, (Sederevicius et al., 2022)*, (Zhao et al., 2019), (Modanwal et al., 2020), (Bashyam et al., 2020), (Bashyam et al., 2022), (Robinson et al., 2020), (Ren et al., 2021), (Chang et al., 2022), (Fatania et al., 2022), (Moyer et al., 2020)
		Domain adaptation: Target image component	(Tian et al., 2022)*, (Zuo et al., 2021a), (Zuo et al., 2021b), (Dewey et al., 2020), (Liu and Yap, 2021), (Liu et al., 2023), (Liu et al., 2021)
	Task specific	Task-related approach	(Dinsdale et al., 2021), (Aslani et al., 2020)
Harmonization of image-derived measures	Task agnostic	Removal of scanner effects	(Maikusa et al., 2021)*, (Fortin et al., 2017), (Fortin et al., 2018), (Nielson et al., 2018), (Yu et al., 2018), (Chen et al., 2020a), (Pomponio et al., 2020), (Beer et al., 2020), (Johnson et al., 2007), (Chen et al., 2022a), (Da-Ano et al., 2021), (Garcia-Dias et al., 2020), (Wang et al., 2021)
	Task specific	Task-related approach	(Bayer et al., 2022a), (An et al., 2022)

* denotes supervised harmonization methods.

3.0 Investigating two methods of cross-scanner technical variability removal in harmonization of image-derived measures

In this section, we investigated two methods of cross-scanner technical variability removal in harmonization of image-derived measures. As a testbed for studying these methods, we selected deriving biomarkers of Alzheimer's disease (AD) from aggregated neuroimaging data and study harmonization of these measures. We hypothesized that *the pipeline of technical variability removal from both images and image-derived measures would result in better removal of unwanted variability and consequently would improve harmonization of our image-derived biomarkers of AD.*

For neuroimaging data, we used *paired dataset* of T1-weighted (T1-w) MRIs on General Electric (GE) 1.5T and Siemens 3T MRI scanners. We elaborated on this dataset and the derived biomarkers in Sections 3.1.1 and 3.3, respectively. For removing the technical variability *from images*, we selected RAVEL (Fortin et al., 2016), which is a framework for normalizing and subsequently harmonizing images by removing their inter-subject variability. RAVEL will be explained in Section 3.1.2. We also used ComBat (Johnson et al., 2007) for harmonizing *image-derived measures*; i.e., our selected biomarkers of AD. ComBat will be elaborated in Section 3.2.2. Lastly, we investigated the pipeline of these two methods, called RAVEL-ComBat in Section 3.2.2 to test our hypothesis. Figure 3 shows the experimental setup that we conducted for RAVEL, ComBat, and RAVEL-ComBat. For assessing technical variability and evaluating the three methods, we used the metrics estimating dissimilarity and similarity within *paired images*. These metrics will be explained in Section 3.3 and the results will be reported in Section 3.4.

3.1 Paired data

A *paired image dataset* is a set of *paired images* that are the images of each individual scanned on *two* scanners with short time gap. Paired images are expected to be images

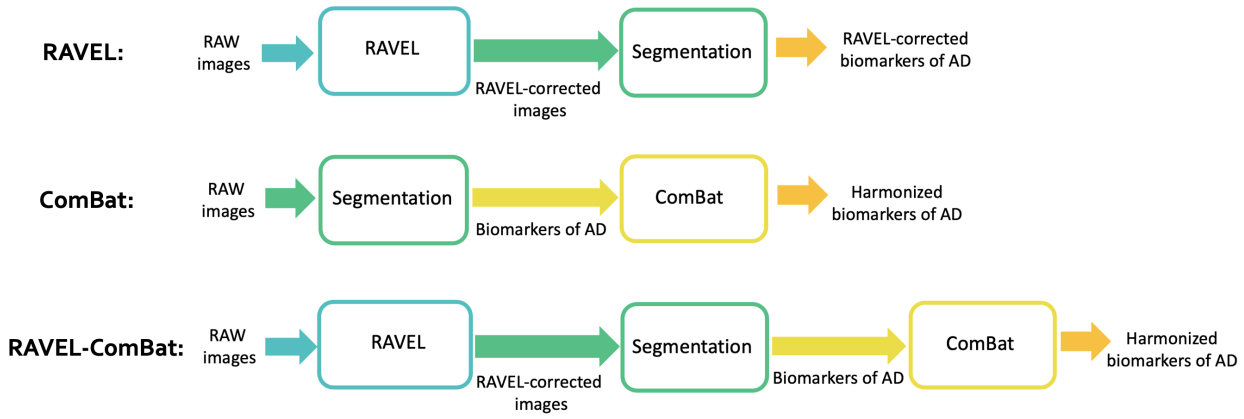


Figure 3: Experimental setup for RAVEl, ComBat, and RAVEL-ComBat. RAW images and image segmentation will be explained in Sections 3.1.2 and 3.3, respectively.

of biologically similar brain with differences solely due to scanner effects. We can provide a paired dataset of the summary measures pertinent to AD by applying a segmentation pipeline to the paired image dataset. We then assume that using the paired dataset of these summary measures, we can estimate the scanner effects and assess the harmonization by metrics of dissimilarity and similarity within the paired measures, respectively.

3.1.1 Study population and image acquisition

The sample used for collecting paired image dataset consists of 16 subjects that are part of an ongoing project (Normal aging, RF1 AG025516 to W.E. Klunk). These 16 subjects were scanned on both GE 1.5T and Siemens 3T MRI scanners, separated by at most 3 months. The median age in the sample was 77.5 years (range=70-79 years) and 25% (n=4) were males. T1-weighted MRIs were acquired coronally on a GE Signa 1.5T MRI scanner with a birdcage volume coil (TE = 5 ms; TR = 25 ms; Flip Angle = 40°; Pulse Sequence = SPGR) and sagittally on a Siemens MAGNETOM Prisma 3T MRI scanner (TE = 2.22 ms; TI = 1000 ms; TR = 2400 ms; Flip Angle = 8°; Pulse Sequence = MPRAGE). No scanner-specific non-uniformity correction was applied to the 1.5T MRI. Siemen's Prescan

Normalize was applied to the 3T MRI. Image matrix and voxel sizes were $256 \times 256 \times 124$ mm and $0.94 \times 0.94 \times 1.5$ mm, respectively, for the 1.5T T1-w MRI and $240 \times 256 \times 160$ mm and $1.0 \times 1.0 \times 1.2$ mm, respectively, for the 3T T1-w MRI.

3.1.2 Image preprocessing

The paired image dataset were preprocessed in R (R Core Team, 2020) following the exact preprocessing steps[1] prescribed before using RAVEL. Accordingly, all images were first registered to a high-resolution T1-w image atlas (Oishi et al., 2009) using the non-linear symmetric diffeomorphic image registration algorithm proposed in (Avants et al., 2008). Then, the N4 bias correction (Tustison et al., 2010) was applied to each of the images to correct for spatial intensity inhomogeneity. Images were then skull-stripped using the brain mask provided in (Fortin et al., 2016). Throughout Section 3, these preprocessed but not intensity normalized images will be referred to as *RAW* images.

3.2 Methods

3.2.1 RAVEL (Removal of Artificial Voxel Effect by Linear regression)

RAVEL (Fortin et al., 2016) is a voxel-wise normalization and harmonization framework that is applied to images. Figure 3 shows the experimental setup for RAVEL. This technique takes the RAW set of MRIs as input and: (1) applies the White Stripe intensity normalization (Shinohara et al., 2014b); and (2) *estimates* and *removes* the remaining inter-subject unwanted intensity variation, detailed in Algorithm 1. The first step is an individual-level intensity normalization method for removing discrepancy of intensities across subjects within tissue types (Shinohara et al., 2014b). It first extracts the normal-appearing white matter voxels of the image and estimates moments of their intensity distribution. It then uses these moments in the z-score transformation for normalizing the voxels of all brain tissue types.

[1] `https://github.com/Jfortin1/RAVEL`

The second step extracts the singular value decomposition of the observed variability in *control voxels* (e.g., cerebrospinal fluid voxels) from the population of participants and selects the first b components of the decomposition as an estimation of the unwanted intensity variation. These steps can be found in Algorithm 1. Once the unwanted variation is estimated from the control voxels, RAVEL then models the intensity values of all the image voxels as a linear combination of the unwanted variables and the clinical covariates (Algorithm 1). Using this model, the technical variability is estimated for each voxel and then removed from the original voxel intensity values (Algorithm 1). Henceforth, the set of RAVEL-corrected images and their derived biomarkers will be referred to as *RAVEL-corrected*.

Although adjusting for clinical covariates are optional in RAVEL correction, we investigated the effects of age and gender on density plots of the tissue types. While controlling for gender did not change the plots, age widened them. It was observed that the higher rank resulted in greater overlap. Since better overlap and narrower density plots are desired, we fit RAVEL to our data by setting the decomposition rank (b) to three and controlling for no clinical variables. More details on the fitting process are provided in Appendix A.1.

Algorithm 1 RAVEL intensity correction.

Results: RAVEL-corrected voxels, $\mathbf{V}^{RAVEL}$.

-p, b: number of clinical covariates, and unwanted variables (decomposition rank), respectively.
-$\mathbf{V}^{WS}$: $k \times m$ matrix of White-Striped RAW voxel intensities.
-$\mathbf{V}_c^{WS}$: $k_c \times m$ matrix of control voxels in White-Striped RAW images.
-$\mathbf{X}$: $m \times p$ matrix of clinical covariates.
-$\mathbf{Z}$: $m \times b$ matrix of unwanted variables.
-$\mathbf{R}$: $k \times m$ matrix of residuals.
-α: $k \times 1$ vector of baseline intensities (average intensities).
-β: $k \times p$ coefficient matrix corresponding $\mathbf{X}$.
-γ: $k \times b$ coefficient matrix corresponding $\mathbf{Z}$.

Algorithm:

The unwanted intensity variation estimation.

1. Centering the $\mathbf{V}_c^{WS}$: $\mathbf{V}_c^* = \mathbf{V}_c^{WS} - v_c 1^T$.
2. Estimating unwanted variables, $\mathbf{Z}$, via singular value decomposition: $\mathbf{Z} = W_b$ where $\mathbf{V}_c^* = \mathbf{U}_b \mathbf{D}_b \mathbf{W}_b^T + \mathbf{R}_c$ is the truncated singular value decomposition of rank b for $\mathbf{V}_c^*$.

The unwanted inter-subject intensity variation removal.

3. Modeling all voxels as $\mathbf{V}^{WS} = \alpha 1^T + \beta \mathbf{X}^T + \gamma \mathbf{Z}^T + \mathbf{R}$.
4. Estimating the coefficients: $\hat{\beta}, \hat{\gamma} \leftarrow \text{Solve}(\mathbf{V}^{WS} = \alpha 1^T + \beta \mathbf{X}^T + \gamma \mathbf{Z}^T + \mathbf{R})$.
5. Computing RAVEL-corrected voxels: $\mathbf{V}^{RAVEL} : \mathbf{V}^{WS} - \hat{\gamma} \mathbf{Z}^T$.

3.2.2 ComBat (Combating Batch effects)

ComBat[2] (Johnson et al., 2007) is an empirical location (mean) and scale (variance) adjustment based on empirical Bayes (EB) estimation for harmonizing numerical data. Data is first modeled as a linear combination of the biological variables of interest and scanner effects. Scanner effects appear as additive and multiplicative effects. Data adjustments are then made to harmonize across scanners. Here, our focus is on analyzing image-derived biomarkers of AD, henceforth called *features*. As shown in Figure 3, these biomarkers are extracted from RAW images. Using ComBat, the value for each feature f, i.e. Y_{ijf}, for subject j for site/scanner i is first modeled as follows:

$$Y_{ijf} = \alpha_f + X_{ij}\beta_f + \gamma_{if} + \delta_{if}\epsilon_{ijf}. \tag{1}$$

Here α_f is the average for feature f, X_{ij} is the design vector of biological variables, β_f is the vector of regression coefficients corresponding to X_{ij}, and γ_{if} and δ_{if} are the additive and multiplicative terms for site/scanner i and feature f, respectively. The error terms, ϵ_{ijf}, are assumed to be independent with distribution $N(0, \sigma_f^2)$. The estimated parameters of scanner effects are γ^*_{if} and δ^*_{if}, respectively. Data are harmonized as follows:

$$Y^*_{ijf} = \frac{Y_{ijf} - (\hat{\alpha}_f + X_{ij}\hat{\beta}_f + \gamma^*_{if})}{\delta^*_{if}} + \hat{\alpha_f} + X_{ij}\hat{\beta}_f, \tag{2}$$

where $\hat{\alpha}_f$ represents the average over the values of feature f for all subjects, and $\hat{\beta}_f$ is estimated using a feature-wise ordinary least-squares approach. See (Johnson et al., 2007) for details on derivation of Equation 2 and the non-parametric ComBat framework. Henceforth, the image-derived biomarkers harmonized using ComBat will be referred to as *ComBat-harmonized*.

Here we have used the parametric EB framework and did not adjust ComBat for age and gender. We tested the addition of age, gender or age and gender to our model using F-tests. None of the F-tests were significant, therefore no age, gender or age and gender effects were

[2]Used the public code from `https://github.com/ncullen93/neuroCombat`

added to the ComBat model. The biomarkers of AD (will be explained in Section 3.3) consist of cortical thickness and volume values, for which separate ComBat models were prepared.

3.2.3 RAVEL-Combat (Pipeline of RAVEL and ComBat)

RAVEL and ComBat were used together in this order to harmonize image-derived biomarkers based on intensity corrected images. Figure 3 summarizes the pipeline, which we refer to as *RAVEL-ComBat*. Henceforth, the biomarkers harmonized using RAVEL-ComBat will be referred to as *RAVEL-ComBat-harmonized*.

3.3 Data analysis

We first assessed the intensity normalization effect of RAVEL. For this, we compared voxel intensity density plots for RAW, White Stripe (WS)-normalized, and RAVEL-corrected images for three brain tissue types, using the tissue mask provided in the EveTemplate package (Oishi et al., 2009): cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM).

For biomarkers of AD, we used the cortical thickness for entorhinal, fusiform, inferior parietal, inferior temporal, and middle temporal regions, as well as the volume measure of the entorhinal, inferior temporal, middle temporal, amygdala, and hippocampus. For segmentation of these measures, we used FreeSurfer 7.1.1 (FS) (Fischl, 2012) and applied it to RAW and RAVEL-corrected images wherever needed in setups in Figure 3. FS was consistently run on native-space MRIs. For the RAVEL pipeline, non-linear registration to template space was performed specifically for skull-stripping and RAVEL processing, and inverse transformations were consistently applied to RAVEL images to return them to native space prior to running FS. In examining RAW FS vs RAVEL FS volume and cortical thickness measures, with and without ComBat, we did not want to confound comparisons with inconsistent preprocessing steps, e.g., bias correction and skull-stripping. As such, preprocessing involved: (1) nonlinear registration to a common template; (2) N4 bias correction;

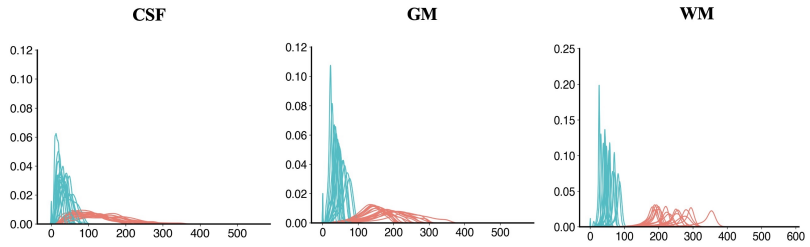
and (3) skullstripping for RAW, RAVEL, ComBat, and RAVEL-ComBat pipelines. Subsequently, as stated previously, for both the RAW and RAVEL pipelines, preprocessed MRIs were returned to native space prior to running FS, and subsequently ComBat. We extracted these 10 biomarkers for both hemispheres of images an resulted in having 20 image-derived summary measures.

We then estimated the scanner effects of the extracted summary measures and evaluated harmonization effect of RAVEL as well as ComBat and RAVEL-ComBat. We measured scanner effects and evaluated harmonization using metrics measuring dissimilarities and similarities within summary measures of paired images, respectively. For this, we designed two metrics: 1) bias: the mean of cross-scanner differences (Siemens 3T - GE 1.5T), compared using paired t-tests with $p < 0.05$ indicating statistical significance, and 2) variance: the root mean square deviation (RMSD) of measures across scanners.

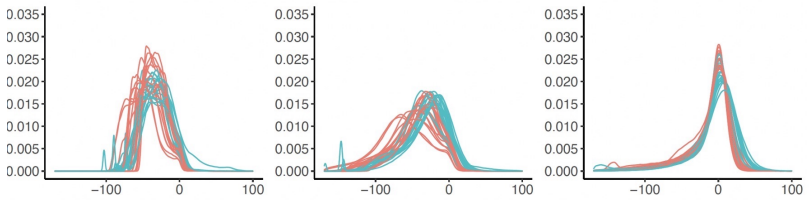
Lastly, to evaluate segmentation accuracy, a neuroradiologist visually rated FS-derived hippocampal segmentations of RAW and RAVEL-corrected MRIs. The segmentations of RAW and RAVEL-corrected MRIs were overlaid on RAW images for segmentation evaluation by our neuroradiologist. A four-point scale was used for rating the accuracy of the segmentations (1 = poor, 2 = some errors, 3 = good, 4 = excellent). The rater was blinded to subject demographics, segmentation method, and image preprocessing, normalization, and correction. We did not have the ground truth segmentation, thus we presented the relative accuracy.

3.4 Results

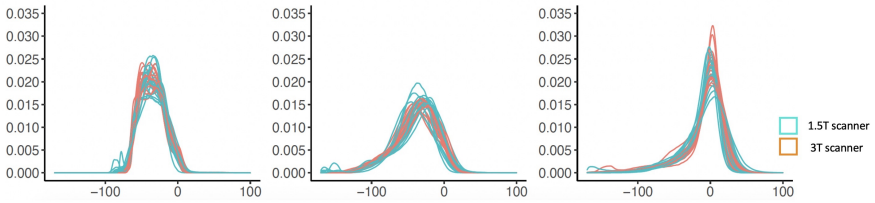
We first show the technical variability in RAW and corrected images in Section 3.4.1. We then present the results of applying other methods, including RAVEL, ComBat, and the RAVEL-ComBat pipeline to 20 image-derived measures (10 AD biomarkers for both hemispheres) in Sections 3.4.2, 3.4.3, and 3.4.4, respectively.



(a) RAW.



(b) White Striped.



(c) RAVEL corrected

Figure 4: Density plots of MRI voxel intensities by tissue type (cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM)) across scanners (GE 1.5T (cyan) and Siemens 3T (orange)) for (a) RAW, (b) White-Striped, and (c) RAVEL-corrected MRIs. Note that White Stripe increases the overlap of the densities greatly for WM, which was the intent of the method, but there is still some non-overlapping regions for GM and CSF, which RAVEL improves. Initially referenced in section 3.4.1.

3.4.1 Technical variability in RAW data

Figure 4 displays the intensity density plots for each of the three brain tissues (CSF in column one, GM in column two, and WM in column three) and different levels of data processing (RAW: panel a, White Stripe: panel b, and RAVEL: panel c). Densities are shown

in cyan for the 1.5T scanner and in orange for the 3T scanner. Results indicate that: (1) the distribution of RAW image intensities vary substantially between- and within-scanners; (2) White Stripe substantially improves the distance between densities, especially in white matter; and (3) RAVEL further improves the distance between densities, especially in CSF and GM.

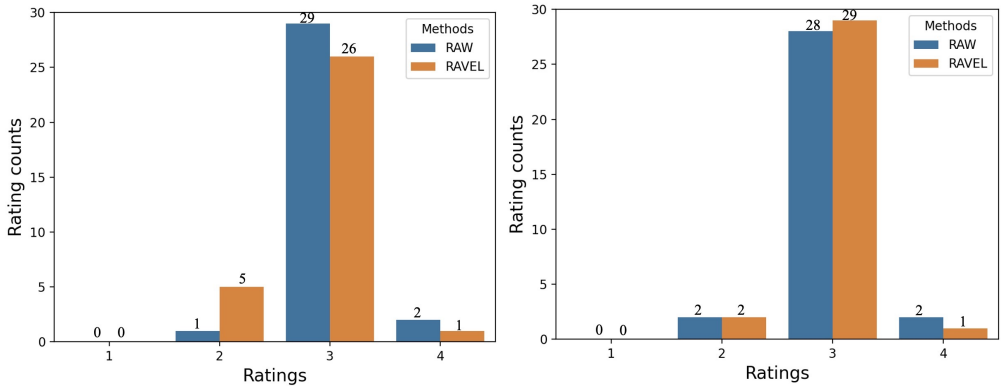
Table 2, provides bias (mean of cross-scanner differences) and variance (RMSD values) of the 20 measures extracted for RAW, RAVEL, ComBat, and RAVEL-ComBat. Also, for each method in Table 2, the statistically significant biases and increased RMSDs (compared to their corresponding values in RAW data) were highlighted and presented in bold, respectively. Focusing on these two metrics for RAW data in Table 2, we observed scanner effects as: (1) statistically significant bias for 11 summary measures, and (2) deviation of values across scanners for all summary measures. We also showed the within-scanner mean and SD of the summary measures for the 4 methods in Appendix A.2.

3.4.2 RAVEL

3.4.2.1 Segmentation accuracy

Neuroradiological ratings comparing hippocampal segmentation of RAW to RAVEL-corrected images revealed that RAVEL neither significantly improved nor deteriorated the FS segmentations. In fact, ratings of the left hemisphere segmentation in Figure 5a showed that RAVEL performed slightly worse than RAW, by having a greater number of erroneous (5 to 1) and fewer good (26 to 29) and excellent (1 to 2) segmentations. However, for the right hemisphere segmentations (Figure 5b), RAVEL performed similarly to RAW, by having similar erroneous (2 to 2), one more good (29 to 28), and one less excellent segmentations (1 to 2). The Wilcoxon hypothesis testing on collected ratings of RAW and RAVEL for the left and right hemispheres resulted in (W-value = 10.0, p-value = 0.096) and (W-value =12.0, p-value = 0.705), respectively.

Two single cases are illustrated in Figures 6 and 7, showing the hippocampal segmentation on RAW brain images by method with arrows pointing to the erroneous segmented



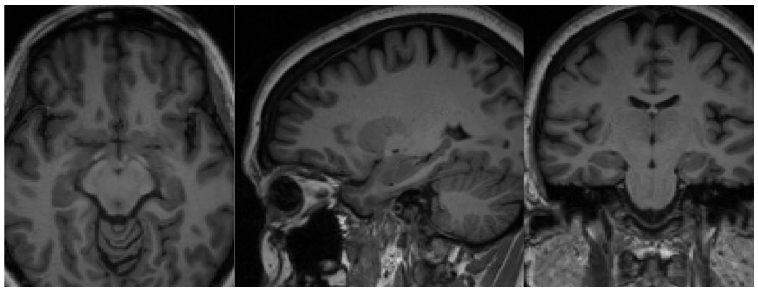
(a) Left hemisphere (b) Right hemisphere

Figure 5: Visual ratings of FS-based hippocampal segmentations for RAW and RAVEL-corrected (RAVEL) MRIs, using a four-point rating scale (1 = poor, 2 = some errors, 3 = good, and 4 = excellent). Initially referenced in section 3.4.2.1.

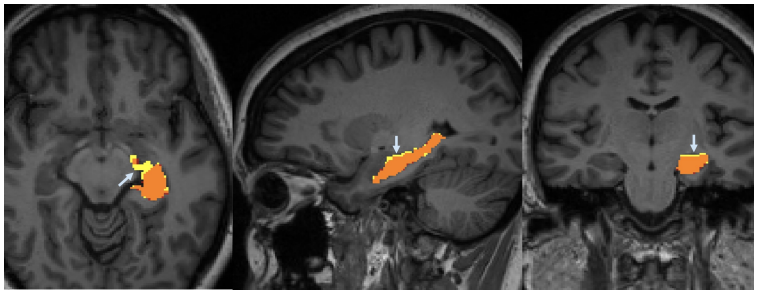
voxels. Figure 6 presents one single case in which RAVEL results in a more accurate hippocampal segmentation than RAW. The RAW brain image, without any segmentation on, is depicted in Figure 6a. Figure 6b shows that the segmentation based on RAW images (in yellow) has extraneous segmented voxels over the CSF and adjacent white matter areas, when compared to segmentation based on RAVEL (in red). The orange area shows the overlap of the two segmentations and the remained yellow and red areas are for RAW and RAVEL segmentations, respectively. Figure 7a depicts the RAW brain image for the second case. Figure 7b presents this case in which RAVEL (in red) results in a less accurate hippocampal segmentation than RAW (in yellow), by not capturing the entire hippocampus, pointed by arrows. The overlapped area of these two segmentations is in orange.

3.4.2.2 Harmonization

For most of the derived imaging measures reported in Table 2, RAVEL decreased bias (13 decreases versus 7 increases) and increased variability (SD) of differences (6 decreases versus 14 increases), when compared to the measures from the RAW data in column 1. The



(a) The non-preprocessed image with no segmentation overlaid on.

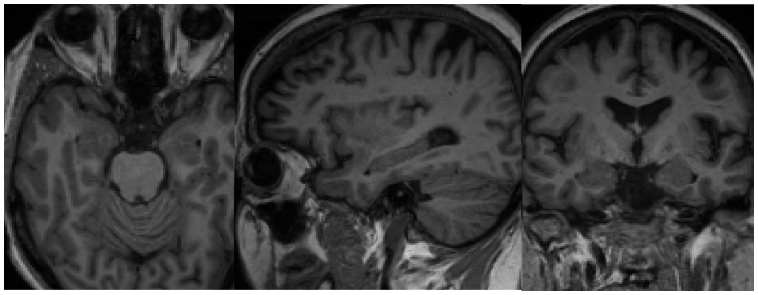


(b) Hippocampal segmentation for RAVEL-corrected images (in red) is more accurate than that of RAW images (in yellow). The overlapped area is depicted in orange. Arrows show that extraneous segmented voxels over the CSF and adjacent white matter areas exist in the RAW image.

Figure 6: Axial, sagittal, and coronal slices of a single subject MRI with overlaid hippocampal segmentations generated by FS for RAW and RAVEL-corrected images. Initially referenced in section 3.4.2.1.

comparison has been done based on absolute values of bias. The number of measures with statistically significant bias decreased from 11 for RAW to 6 for RAVEL and the RAVEL-corrected images resulted in change of RMSDs as variances (11 decreases versus 9 increases), when compared to RAW. Figure 8 presents these total number of changes (decrease, increase, and no change) in (a) bias, (b) variation (SD) of cross-scanner differences, and (c) variance (RMSD), in addition to (d) the number of statistically significant biases over all 20 summary measures.

Columns fourth and fifth in Table 2 show the mean (SD) of cross-scanner differences

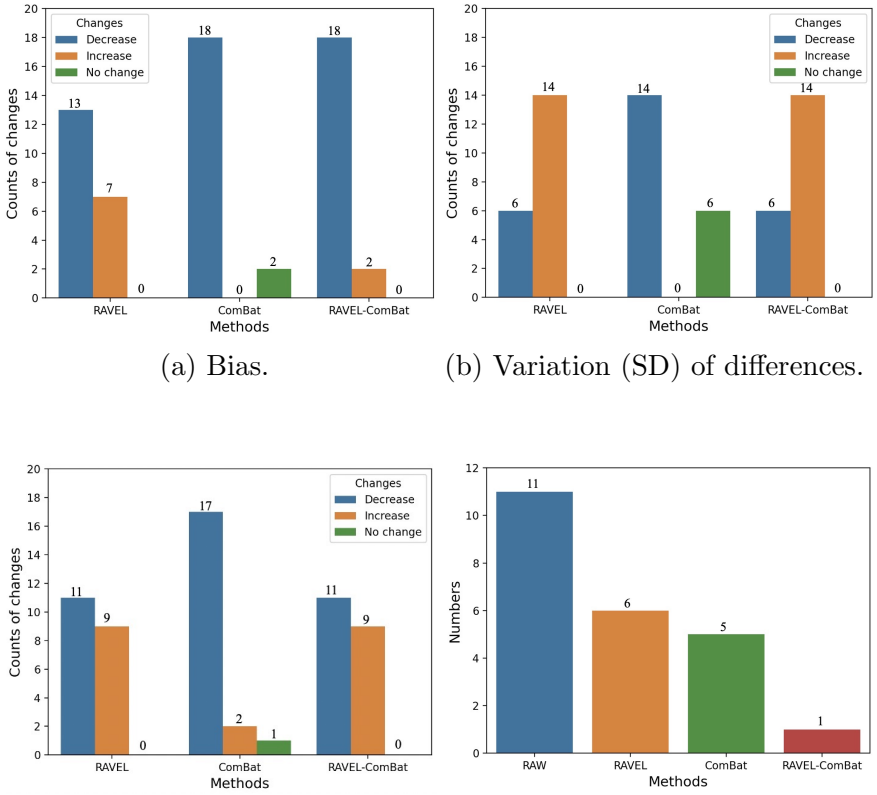


(a) The non-preprocessed image with no segmentation overlaid on.

(b) Hippocampal segmentation for RAVEL-corrected images (in red) is less accurate than that of RAW images (in yellow). The overlapped area is depicted in orange. Arrows show the missed hippocampal voxels in the segmentation for the RAVEL-corrected image.

Figure 7: Axial, sagittal, and coronal slices of two single subjects MRI with overlaid hippocampal segmentations generated by FS for RAW and RAVEL-corrected images. Initially referenced in section 3.4.2.1.

and RMSD values for all summary measures extracted from the RAVEL-corrected images. These results are complemented by the statistically significant biases (highlighted values) and increased RMSDs (values in bold). This table is accompanied by Figure 9, visualizing the results of this table for the summary measures, including volumes of respectively inferior temporal and middle temporal for left and right hemispheres, as well as cortical thickness of entorhinal and inferior parietal in left hemisphere. In Figure 9, the cross-scanner differences of all subjects were depicted as a line plot for each method. The smoother line plots indicate methods which resulted in lower variation (SD) of cross-scanner differences. The line plots



(a) Bias. (b) Variation (SD) of differences.

(c) Variance (RMSD). (d) Statistically significant bias.

Figure 8: Bar plots showing number of summary measures with changes (classified as decrease, increase, and no change) in (a) cross-scanner bias, (b) variation (SD), and (c) variance (RMSD) for tested methods compared to RAW. Part (d) shows the number of regional summary measures with statistically significant cross-scanner bias for each method. Statistical measures were calculated over 20 FS-derived summary measures (listed in Section 3.3).

closer to x-axis depict methods, which resulted in smaller variances (cross-scanner differences are closer to zero).

Based on the results in Table 2, the volume of inferior temporal (left hemisphere) is one of the measures that RAVEL harmonized by decreasing bias, SD of differences, and variance, resulting in no statistically significant differences of bias. These results were supported in Figure 9a where the line plot for RAVEL is smoother than RAW and closer to x-axis. However, the results in Table 2 showed that RAVEL resulted in increased variance for the

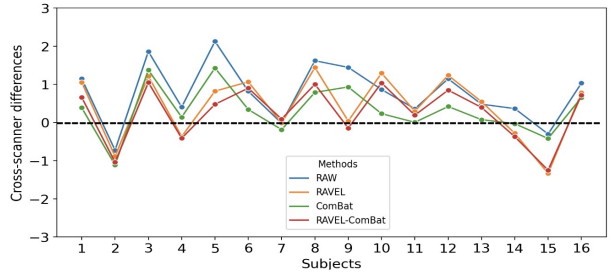
rest of the 3 summary measures. Such pattern could also be observed in Figures 9b, c, and d, as the plots for RAVEL deviated from x-axis due to increased peaks when compared to the plots for RAW.

Based on our observations, RAVEL had some potential harmonization effect on our data by showing decrease in either bias, variance, or number of summary measures with

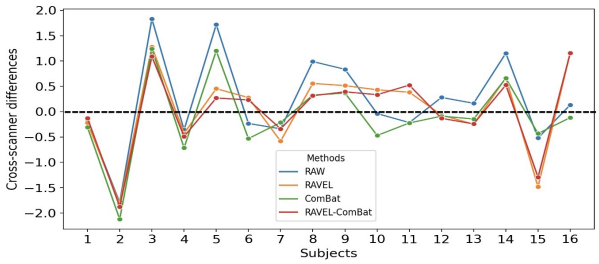
Table 2: Mean (SD) of cross-scanner differences (Siemens 3T - GE 1.5T) as well as cross-scanner RMSDs, for biomarkers relevant to AD. These statistics were prepared for each of the RAW, RAVEL, ComBat, and RAVEL-ComBat methods. For each method, the increased RMSD values (compared to RAW) were reported in bold and the statistical significant differences ($p < 0.05$) were highlighted. Information on confidence intervals of the t-tests is reported in Appendix A.3.

	RAW		RAVEL		ComBat		RAVEL-ComBat	
	Mean(SD)	RMSD	Mean(SD)	RMSD	Mean(SD)	RMSD	Mean(SD)	RMSD
ROIs	Cortical Thickness (mm)							
Left								
Entorhinal	0.22 (0.23)	0.31	0.19 (0.42)	**0.45**	0.08 (0.22)	0.23	0.18 (0.40)	**0.43**
Fusiform	0.24 (0.10)	0.26	0.10 (0.23)	0.25	0.11 (0.09)	0.14	0.10 (0.22)	0.23
Inferior Parietal	-0.05 (0.10)	0.11	-0.15 (0.20)	**0.25**	-0.04 (0.10)	0.10	-0.04 (0.19)	**0.19**
Inferior Temporal	0.25 (0.17)	0.30	0.06 (0.27)	0.27	0.11 (0.16)	0.19	0.08 (0.24)	0.25
Middle Temporal	0.08 (0.15)	0.17	-0.01 (0.19)	**0.18**	0.02 (0.15)	0.15	0.03 (0.19)	**0.18**
Right								
Entorhinal	0.23 (0.45)	0.49	0.17 (0.34)	0.37	0.08 (0.43)	0.43	0.15 (0.32)	0.35
Fusiform	0.22 (0.11)	0.25	0.05 (0.20)	0.20	0.10 (0.10)	0.14	0.06 (0.20)	0.20
Inferior Parietal	-0.02 (0.07)	0.07	-0.07 (0.15)	**0.16**	-0.02 (0.07)	0.07	-0.01 (0.14)	**0.14**
Inferior Temporal	0.26 (0.11)	0.28	0.09 (0.16)	0.18	0.11 (0.10)	0.15	0.08 (0.15)	0.17
Middle Temporal	0.05 (0.16)	0.16	-0.01 (0.13)	0.13	0.01 (0.16)	0.15	0.03 (0.13)	0.13
ROIs	Volume $(cm)^3$							
Left								
Entorhinal	0.08 (0.25)	0.26	0.19 (0.41)	**0.44**	0.01 (0.25)	0.24	0.12 (0.40)	**0.40**
Inferior Temporal	0.79 (1.09)	1.09	0.78 (0.84)	0.92	0.31 (0.65)	0.70	0.26 (0.73)	0.75
Middle Temporal	0.16 (0.84)	0.83	-0.35 (1.05)	**1.07**	-0.16 (0.74)	0.73	-0.23 (0.98)	**0.98**
Amygdala	0.13 (0.20)	0.27	0.09 (0.15)	0.17	0.06 (0.19)	0.19	0.04 (0.14)	0.14
Hippocampus	-0.08 (0.15)	1.25	-0.19 (0.19)	0.27	-0.03 (0.14)	0.14	-0.05 (0.18)	0.20
Right								
Entorhinal	0.09 (0.27)	0.91	0.12 (0.33)	0.34	0.01 (0.26)	0.25	0.07 (0.32)	0.32
Inferior Temporal	0.98 (0.80)	0.23	0.67 (0.89)	**1.09**	0.42 (0.71)	**0.80**	0.41 (0.85)	**0.92**
Middle Temporal	0.21 (0.92)	0.16	0.06 (0.85)	**0.82**	-0.10 (0.79)	**0.77**	0.02 (0.78)	**0.76**
Amygdala	0.05 (0.09)	0.10	0.02 (0.08)	0.08	0.03 (0.09)	0.09	0.01 (0.08)	0.08
Hippocampus	-0.09 (0.15)	0.17	-0.10 (0.24)	**0.26**	-0.03 (0.14)	0.14	-0.04 (0.24)	**0.24**

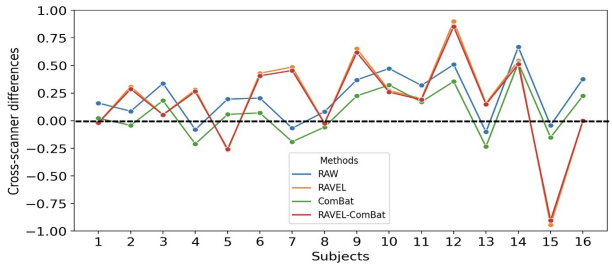
statistically significant bias. However, this effect does not seem to be consistent among (1) summary measures (increased bias and variance for some summary measures), and (2) subjects (increased SD of differences for some summary measures, for example Figures 9b, c, and d).



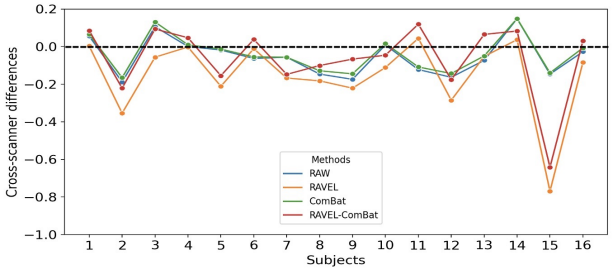
(a) Volume of inferior temporal (left).



(b) Volume of middle temporal (right).



(c) Cortical thickness of entorhinal (left).



(d) Cortical thickness of inferior parietal (left).

Figure 9: Line plots depicting cross-scanner differences (Siemens 3T - GE 1.5T) for all subjects and methods. The plots depicted for 4 summary measures which were also reported in Table 2. The plotted differences were in millimeter (mm) and cubic centimeter $(cm)^3$ for cortical thicknesses and volumes, respectively. A smoother line plot indicates a lower SD of differences and a plot closer to x-axis (zero differences) shows lower variance. The line plots showed that in (a) all three methods succeed in harmonization, in (b) all methods increased variance, in (c) and (d) RAVEL and RAVEL-ComBat increased variance while ComBat succeed in harmonization.

3.4.3 ComBat

Results for ComBat-harmonized measures in Table 2 showed that ComBat decreased bias and SD of cross-scanner differences for most of the measures. Considering absolute values of bias, these statistics were 18 and 14 decreases for bias and SD of differences, respectively, while no changes have been seen for the rest of measures. These results were supported by the decreased number of statistically significant biases (11 for RAW decreased to 5 for ComBat) and RMSD values (17 decreases, 2 increases, and 1 no change). These statistics were depicted in Figure 8.

Based on the results in Table 2, ComBat successfully harmonized volume of inferior temporal and cortical thickness of entorhinal (both for left hemisphere), by decreasing bias, SD of differences, and variance as well as removing statistical significance of bias. These results were supported by the corresponding line plots of ComBat in Figures 9a and c, where they were similar to RAW but smoother and closer to x-axis. However, the results for cortical thickness of inferior parietal (left hemisphere) did not change noticeably (ComBat almost overlapped RAW in Figure 9d) and the volume of middle temporal (right) still retained the increase in variance (Figure 9b).

Based on our observations, ComBat had potential harmonization effect on our data by showing decrease in bias, variance, SD of differences, or number of summary measures with statistically significant bias. This effect seems to be more consistent across summary measures and subjects which makes ComBat to be preferred over RAVEL for the task of harmonizing image-derived measures.

3.4.4 RAVEL-ComBat pipeline

Results of comparing RAVEL-ComBat to RAW in Table 2 showed that this method decreased bias for most of the summary measures (18 decreases versus 2 no changes), when the absolute values of bias were compared. The number of summary measures with statistically significant bias decreased from 11 for RAW to 1 for this pipeline. However, RAVEL-ComBat almost increased the SD of differences (6 decreases versus 14 increases) as well as the variance

for almost half of the summary measures (11 decreases versus 9 increases). These results were summarized in Figure 8.

Results in Table 2 showed that RAVEL-ComBat followed almost similar pattern with RAVEL in harmonization of the 4 selected summary measures. RAVEL-ComBat successfully harmonized volume of inferior temporal (left), while increased the variance for the other 3 measures. These results were also visualized in Figure 9 where all the line plots for RAVEL-ComBat closely followed RAVEL's. However, comparing these two methods, minor improvements have been observed for RAVEL-ComBat which were (1) decreasing the number of biases and (2) resulting smaller increases in RMSD values (the number of changes in variance are similar between the methods). Such differences could be seen for cortical thickness of inferior parietal (left) in Figure 9d.

Even though RAVEL-ComBat was improved by ComBat and showed decreased number of biases, it was still more similar to RAVEL in terms of harmonizing image derived measures. Although the number of statistically significant biases decreased noticeably using RAVEL-ComBat, the harmonized measures still suffer from increased SD of differences and variance when compared to RAW and ComBat. In conclusion, ComBat would be preferred over RAVEL and RAVEL-ComBat as these two methods are inconsistent among subjects and summary measures.

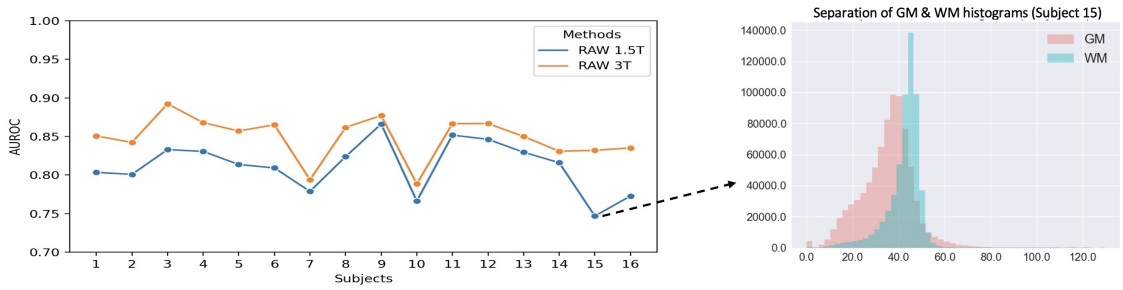
3.5 Discussion

In this study, we focused on harmonization of 10 image-derived biomarkers of AD. We hypothesized that *the pipeline of technical variability removal from images, using RAVEL, and image-derived measures, using ComBat, would result in better removal of unwanted variability and consequently would improve harmonization of our image-derived biomarkers of AD*. Accordingly, we collected a paired cohort of 16 healthy elderly study participants scanned on two different MRI scanners, GE 1.5T and Siemens 3T. We assumed that technical variability manifests as within-subject differences for summary measures and reducing these differences would achieve harmonization appearing as lowered bias and variance for measures

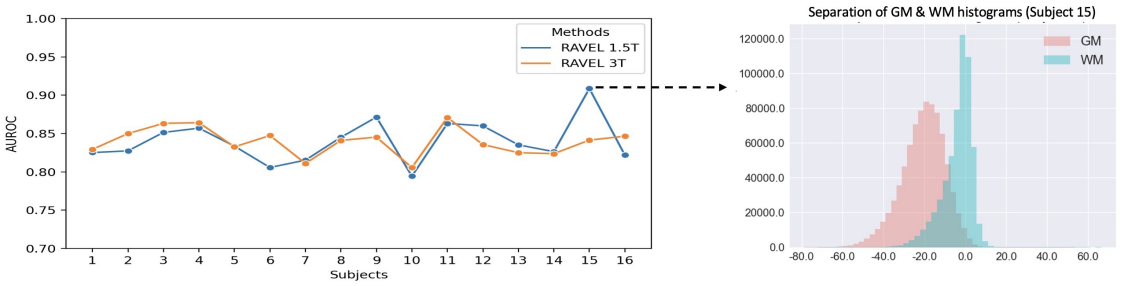
of these biomarkers for both hemispheres.

Consistent with previous reports, our results showed that RAVEL further normalized the White-Stripe-normalized images, specifically CSF and GM areas (Fortin et al., 2016). Moreover, RAVEL preserved the anatomical information of images. For example, it preserved the segmentation accuracy for the hippocampus. These results are consistent with the previous findings from the group's previously reported results for a multi-site Down Syndrome study (Minhas et al., 2020). Regarding the harmonization of our summary measures, RAVEL, ComBat, and RAVEL-ComBat effectively harmonized the 1.5T and 3T MRI summary measures in this study, in that all techniques reduced the number of statistically significant biases across the regional cortical thicknesses and volumes examined. ComBat, however, demonstrated a more consistent harmonization effect across subjects and summary measures as compared to RAVEL and RAVEL-ComBat. Based on the results on our data, ComBat would be preferred to RAVEL and RAVEL-ComBat for the task of harmonizing image-derived measures, to avoid the inconsistency across subjects and summary measures that were observed with the other two pipelines.

Despite demonstrating an overall reduction in the number of statistically significant biases between FreeSurfer outcome measures from 1.5T and 3T MRI scans (Table 2 and Figure 8d), the application of RAVEL introduced a significant difference in the left inferior parietal cortical thickness and increased RMSD across multiple regional cortical thickness and volume measures (Table 2). There are multiple possibilities as to why RAVEL introduced unwanted differences and variability. The quality of FreeSurfer segmentations, and therefore outcome measures, is dependent on the GM-WM contrast in the T1 MR image, with increased contrast resulting in more accurate FreeSurfer segmentations. To examine the effects of RAVEL on GM-WM contrast for the 1.5T and 3T scans, we calculated the area under the receiver operating characteristic (AUROC) for classification of voxel intensity values as GM relative to WM. For this, we first extracted the histograms of GM and WM using the tissue mask in the EveTemplate package (Oishi et al., 2009). We then looked at the classification of GM voxels from WM as the problem of estimating the separation of their histograms. For classifying the GM and WM voxels, we set the voxel intensity thresholds from one end of the union of histograms to the other. Every threshold position generated



(a) Line plots and histogram plots for RAW.



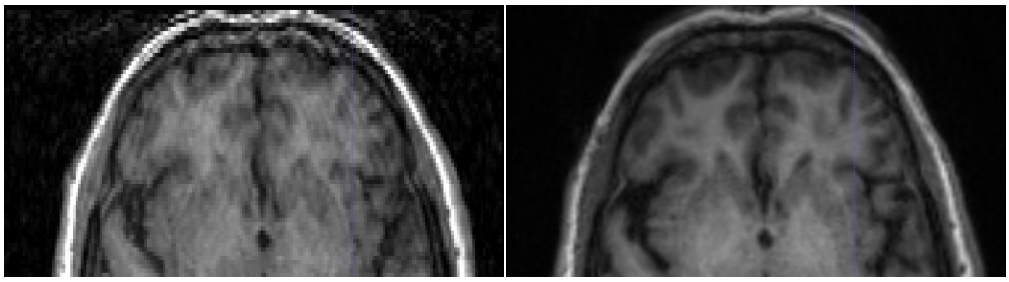
(b) Line plots and histogram plots for RAVEL.

Figure 10: AUROC for classification of voxel intensity values as GM relative to WM. The AUROCs were estimated as the separation of the GM and WM histograms. On the left, the line plots depict the AUROC of the classification for all subjects. On the right, the plots show the overlap/separation of the histograms of tissues for one single subject. The two plots were depicted separately for each scanner (GE 1.5T and Siemens 3T), for RAW (a) and RAVEL (b).

a point on the AUC curve. A complete separation of histogram would result in AUROC = 1 and completely overlapped histograms would give AUROC = 0.5. AUROC values across subjects for 1.5T and 3T scans before and after RAVEL are shown in Figure 10. RAW 3T scans consistently had better GM-WM contrast than RAW 1.5T scans (mean(SD) RAW 3T AUROC: 0.849(0.028); mean(SD) RAW 1.5T AUROC: 0.812(0.033)). The application of RAVEL reduced the mean of absolute differences between 3T and 1.5T AUROC values from 0.037 ± 0.021 to 0.017 ± 0.018. As such, just as RAVEL effectively corrected MRI voxel intensity distributions across scanners, it also corrected GM-WM contrasts. However,

in doing so, RAVEL reduced GM-WM contrast, on average, across 3T MRIs (mean(SD) RAW 3T AUROC: 0.849(0.028); mean(SD) RAVEL 3T AUROC: 0.840(0.018)). This reduction in contrast may have reduced quantitative accuracy and increased variability in cortical thickness and volume FreeSurfer measures.

Differences in motion artifacts may have also affected and possibly confounded harmonization of FreeSurfer outcome measures via RAVEL. Previous studies have demonstrated that motion artifacts, including blurring, ghosting, and ringing, reduce FreeSurfer measures of regional GM cortical thickness and volume (Alexander-Bloch et al., 2016; Backhausen et al., 2016). The effect of motion artifacts on RAVEL is not well understood, and characterizing it is beyond the scope of this study. Nevertheless, motion artifacts may introduce variability in CSF regions, from which the unwanted scanner-associated variation component is estimated for RAVEL. In this study, motion artifacts, or the lack thereof, were not consistent across 1.5T and 3T MRI acquisitions. Significant motion artifacts were observed in the frontal cortex of the 1.5T scan but not 3T scan for a single subject, as shown in Figure 11.



(a) 1.5T scan. (b) 3T scan.

Figure 11: Inconsistent motion artifact across scanners for a single subject in our data. More significant motion artifacts were observed in the frontal cortex of 1.5T scan (a) relative to the 3T scan (b).

Further investigation into ComBat led to some additional insight with respect to effects of preprocessing before applying ComBat. In this study we applied the preprocessing steps

recommended for RAVEL to our images for all four methods (RAW, RAVEL, ComBat, and RAVEL-ComBat) in order to avoid confounding the comparisons with inconsistent pipeline. However, the choice of preprocessing could affect the results of RAW and ComBat, which do not necessarily need any preprocessing before the segmentation with FreeSurfer. We investigated this issue by skipping the preprocessing step in the process of preparing RAW and ComBat-harmonized images and generated two new sets of data. We then compared RAW and ComBat with their corresponding new data using paired t-test. Our results showed that preprocessing could be a source of variability and resulted in statistically significant different values for summary measures within each scanner in our experiment: RAW (1.5T: 9 measures, 3T: 10 measures) and ComBat (1.5T: 9 measures, 3T: 7 measures). This significant effect of the preprocessing step on results should be considered in studies when ComBat is used for the purpose of data harmonization and not for method comparison. The details of the experiments were reported in Appendix A.4. Moreover, ComBat could be modified for handling the dependence for within-subject scans which is the case for our paired cohort. Thus, we added the subjects as a fixed effect to ComBat which resulted in a non-significant F-test when tested versus the original ComBat. We also handled the dependence by adapting the longitudinal Combat (Beer et al., 2020): we ran the model without the inclusion of time but, we included a random intercept. We present these results in Appendix A.5.

4.0 Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning

In this section, we developed a supervised image harmonization method: MISPEL (Multi-scanner Image harmonization via Structure Preserving Embedding Learning). We hypothesized that *harmonization can be achieved for scanners within a matched dataset by constructing a model that maps matched images from the dataset to a scanner-middle-ground space, where matched images lose scanner effects by becoming similar to each other.* For such model, we designed MISPEL to (1) generalize to multiple (more than two) scanners, (2) preserve the structural (anatomical) information of the original brains, (3) learn harmonization on a matched dataset, and (4) later harmonize unmatched images of the scanners for which the matched dataset was collected.

For the matched dataset, we collected T1-w matched images of 18 subjects for four 3T scanners. We elaborate more on this dataset and its preprocessing pipeline in Section 4.1. We also cover MISPEL and its training and harmonization strategy in Section 4.2. We compare MISPEL with one method of normalization, White Stripe, and two methods of harmonization, RAVEL and CALAMITI (Zuo et al., 2021b). We elaborate on these methods in Section 4.3. Moreover, we investigate MISPEL and our competing methods using our evaluation criteria described in Section 4.4. Lastly, we report the results of our comparisons in Section 4.5 and discuss them in Section 4.6.

4.1 Matched data

A matched image dataset is a set of matched images. Matched images are the images of each individual scanned on more than two scanners with short time gap. Matched images are expected to be images of biologically similar brain with differences solely due to scanner effects. We thus can estimate the scanner effects and assess the harmonization by metrics of dissimilarity and similarity within the matched images, respectively.

4.1.1 Study population and image acquisition

The sample used in this study consists of 18 participants which are part of an ongoing project (UH3 NS100608 grant to J. Kramer and C. DeCarli). The median age of the participants was 72 years (range 51-78 years) and 44% (N = 8) were males. All participants were cognitively unimpaired with either a low or high degree of small vessel disease (SVD) as previously defined (Wilcock et al., 2021)). T1-weighted (T1-w) images were acquired for each participant on each of four different 3T scanners [GE, Philips, SiemensP, and SiemensT (Table 3)]. For each participant, these matched images were taken at most four months apart, a time period over which we assume no biological changes could occur in the brain and differences observed between any pairs of scans are solely due to the scanner effects. In a matched dataset, the scanner and harmonization effects can be estimated based on the dissimilarity and similarity of matched images, respectively. The details of estimation of scanner effects and evaluation of harmonization methods are provided in Section 4.4.

Table 3: Scanner specifications

Scanner Name	**GE**	**Philips**	**SiemensP**	**SiemensT**
Manufacturer	General Electrics	Philips	Siemens	Siemens
Scanner Hardware	DISCOVERY-MR750w 3T	Achieva-dStream 3T	Prisma-fit 3T	TrioTim 3T
Scanner software	27-LX-MR-Software-release: DV26.0-R03-1831.b	5.6.1-5.6.1.0	syngo-MR-E11	syngo-MR-B17
Receive Coil	32Ch-Head	MULTI-COIL	BC	32Ch-Head
T1-w Sequence Type	BRAVO	ME-MPRAGE	ME-MPRAGE	ME-MPRAGE
Resolution (mm)	$1.0 \times 1.0 \times 0.5$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$
TE/ΔTE (ms)	3.7	1.66/1.9	1.64/1.86	1.64/1.86
TR (ms)	9500	2530	2530	2530
TI (ms)	600	1300	1100	1200

4.1.2 Image preprocessing

We use RAVEL as one of our harmonization methods in this study. In order to prevent confounding our evaluation with inconsistent preprocessing steps, we preprocessed all images using the pipeline prescribed for RAVEL (Fortin et al., 2016). Therefore, we first used a

non-linear symmetric diffeomorphic image registration algorithm (Avants et al., 2008) to register images to a high-resolution T1-w image atlas (Oishi et al., 2009). We then applied the N4 bias correction method (Tustison et al., 2010) to the registered images to correct them for spatial intensity inhomogeneity. As the last step of the pipeline, we skull-stripped the images using the mask provided in (Fortin et al., 2016). We also scaled images in one additional step, in which intensity values of each image were divided by their within-mask average intensity value. Throughout this manuscript, these preprocessed images are referred to as *RAW* and used as input to our methods.

4.2 MISPEL

Our proposed framework, MISPEL, is a convolutional deep neural network for harmonizing images from multiple scanners, for which a *matched* dataset is available. Although it is more desirable to train a harmonization method on the whole images rather than slices, this is not possible due to our current GPU limitations. Accordingly, we designed a two-step training framework for MISPEL which consists of units of 2D encoder and decoder modules for each of the scanners. The 2D network is trained on axial slices, since this orientation has the highest resolution in our images. Algorithm 2 and Figure 12 describe our framework.

Notations and Assumptions. We consider M scanners for the matched data where each subject is scanned on all M scanners. The axial slices across all the subjects are combined for a total of N scans for each scanner. The dataset thus consists of $X_{i=1:M}^{j=1:N}$ where X_i^j is the axial slice j from scanner i, and $i = 1 : M$ denotes $i \in \{1, \ldots, M\}$. We note that for each subject, the scans are *coregistered* across the scanners to the mean template. Thus, for each j, we assume the scans $X_1^j, X_2^j, \ldots, X_M^j$ are anatomically similar and have the same image size of H by W. The goal is to learn a framework which derives harmonized slices $\bar{X}_{i=1:M}^{j=1:N}$, where $\bar{X}_1^j \approx \cdots \approx \bar{X}_m^j \approx \cdots \approx \bar{X}_M^j$.

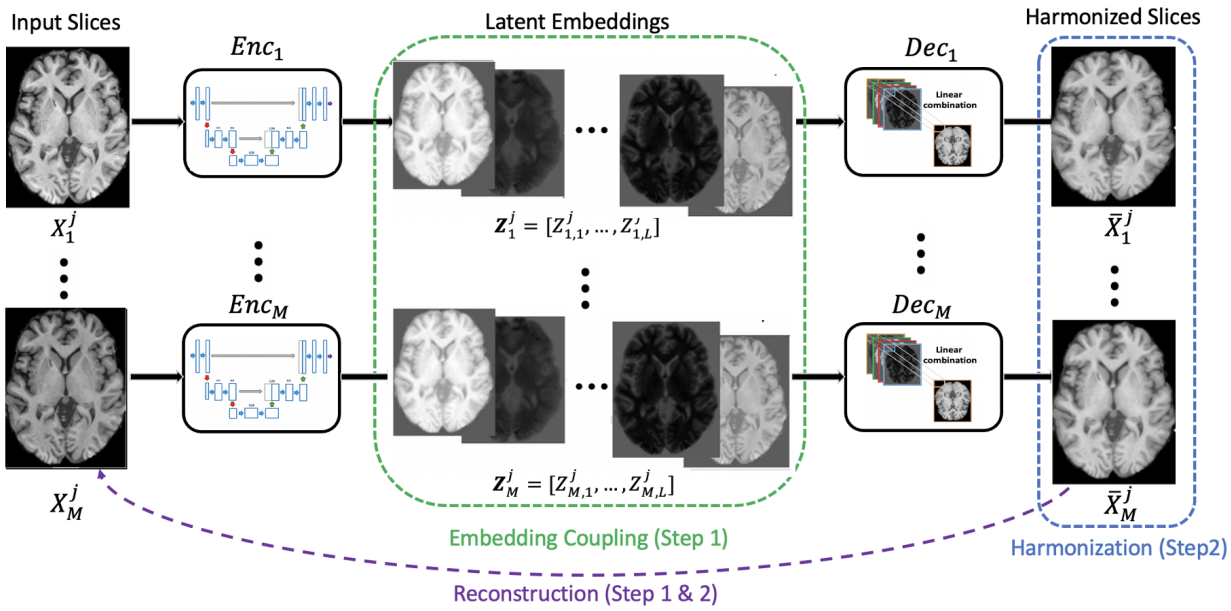


Figure 12: **Illustration of MISPEL.** For each of $j = 1 : N$ input scans and for each of $i = 1 : M$ scanners, Enc_i (U-Net) outputs the corresponding latent embeddings: $\mathbf{Z}_i^j = Enc_i(X_i^j)$. The corresponding Dec_i (linear function) maps the embeddings to the output: $\bar{X}_i^j = Dec_i(\mathbf{Z}_i^j)$. **Step 1:** Embedding Learning: $Enc_{i=1:M}$ and $Dec_{i=1:M}$ are updated using the embedding coupling loss ($\mathcal{L}_{coup}$) and the reconstruction loss ($\mathcal{L}_{recon}$). **Step 2:** Harmonization: Only $Dec_{i=1:M}$ are updated using the harmonization loss ($\mathcal{L}_{harm}$) and the reconstruction loss ($\mathcal{L}_{recon}$). Refer to Algorithm 1 for details on training.

4.2.1 Encoder-Decoder unit

Encoder. For each scanner i, its encoder network Enc_i decomposes each scan X_i^j to its set of *latent embeddings* $\mathbf{Z}_i^j = [Z_{i,1}^j, \ldots, Z_{i,L}^j]$ where $Z_{i,l}^j$ is the lth latent embedding of X_i^j. The number of embeddings L is heuristically chosen and fixed. We use a 2D U-Net (Ronneberger et al., 2015) for *each* Enc_i, and the latent embedding $Z_{i,l}^j \in \mathbb{R}^{H \times W}$ is of size identical to X_i^j.

Decoder. After each Enc_i, its corresponding decoder network Dec_i maps the latent embeddings $\mathbf{Z}_i^j$ to the image space $\bar{X}_i^j$. Since $\mathbf{Z}_i^j$ and X_i^j have the same sizes, we let Dec_i to

be a linear function:

$$\bar{X}_i^j = \sum_{l=1}^{L} \gamma_{i,l} Z_{i,l}^j, \tag{3}$$

where $\gamma_{i,l}$ is the coefficient for $Z_{i,l}^j$. Thus, each Dec_i learns the set of linear combination coefficients $\gamma_{i,1}, \ldots, \gamma_{i,L}$, which is essentially a 1×1 convolution.

Algorithm 2 MISPEL

Variables:
- i: Scanner index
- j: Slice index
- l: Embedding's component index
- T_1, T_2: Max training iterations for Step 1 and Step 2
- H, W: Height and width of each scan
- $X_i^j \in \mathbb{R}^{H \times W}$: Axial slice j from scanner i
- $Z_{i,l}^j \in \mathbb{R}^{H \times W}$: Latent embedding l of X_i^j
- $\mathbf{Z}_i^j = [Z_{i,1}^j, \ldots, Z_{i,L}^j]$: L latent embeddings of X_i^j
- $\bar{X}_i^j \in \mathbb{R}^{H \times W}$: Harmonized X_i^j

Networks:
- Enc_i: Encoder U-Net for $X_i^j \to \mathbf{Z}_i^j$ & Dec_i: Decoder linear map for $\mathbf{Z}_i^j \to \bar{X}_i^j$

Algorithm:

procedure STEP 1: EMBEDDING LEARNING
  for $t = 1, \ldots, T_1$ or until $X_i^j \approx \bar{X}_i^j$ **do**
    for each slice j **do**
      for each scanner i **do**
        $\mathbf{Z}_i^j \leftarrow Enc_i(X_i^j)$ (embeddings)
        $\bar{X}_i^j \leftarrow Dec_i(\mathbf{Z}_i^j)$ (reconstruction)
      end for
      Update $Dec_{i=1:M}$ and $Enc_{i=1:M}$ (Equation (6))
    end for
  end for
end procedure (end Step 1)
procedure STEP 2: HARMONIZATION
  for $t = 1, \ldots, T_2$ or until $\bar{X}_1^j \approx \cdots \approx \bar{X}_M^j$ **do**
    for each slice j **do**
      for each scanner i **do**
        $\mathbf{Z}_i^j \leftarrow Enc_i(X_i^j)$ (embeddings)
        $\bar{X}_i^j \leftarrow Dec_i(\mathbf{Z}_i^j)$ (harmonization)
      end for
      Update only $Dec_{i=1:M}$ (Equation (8))
    end for
  end for
end procedure (end Step 2)

4.2.2 Two-step training for harmonization

Note that each Enc_i-Dec_i setup achieves $X_i^j \rightarrow \mathbf{Z}_i^j \rightarrow \bar{X}_i^j$ only with respect to each scanner i and cannot achieve harmonization by itself. Thus, producing $\bar{X}_{i=1:M}^{j=1:N}$ which are harmonized across M scanners requires a mechanism to enforce such similarity. For instance, one may naïvely train all $Enc_{i=1:M}$ and $Dec_{i=1:M}$ to directly impose $\bar{X}_1^j \approx \cdots \approx \bar{X}_M^j$ with a loss function. However, in practice, the coregistered scans exhibit small structural differences, and this may not guarantee preserving the brain structure. Recall that the desired harmonization we seek must preserve the structure while matching the intensities. As we show next, we implement a two-step training which addresses such issues: (1) first learning the embeddings with structural information, and (2) harmonizing the intensities with the embeddings without altering the structures.

Step 1: Embedding Learning. Algorithm 1 lines 1:11 show Step 1. For slice j and scanner i, we first use the corresponding Enc_i for the input scan X_i^j to compute its embeddings $\mathbf{Z}_i^j$. Then, using Dec_i, we also compute the output $\bar{X}_i^j$. Then, we update Enc_i and Dec_i via two loss functions.

Reconstruction Loss. To derive our embeddings, we train Enc_i and Dec_i to accurately reconstruct the input: $\bar{X}_i^j = Enc_i(Dec_i(X_i^j))$. We use the following reconstruction loss which enforces each output $\bar{X}_i^j$ to be similar to its input X_i^j:

$$\mathcal{L}_{recon}(X_{i=1:M}^j, \bar{X}_{i=1:M}^j) = \sum_{i=1}^{M} MAE(X_i^j, \bar{X}_i^j), \tag{4}$$

where $MAE(X_i^j, \bar{X}_i^j)$ is the pixel-wise mean absolute error. Since each Dec_i is a linear combination of the embeddings, this reconstruction process forces the embeddings to hold structural information as shown in Figure 12.

Embedding Coupling Loss. We also incorporate a coupling mechanism to ensure that the embeddings across the scanners roughly capture similar characteristics of the scans. Namely, we seek $Z_{1,l}^j \approx \cdots \approx Z_{M,l}^j$ for each l:

$$\mathcal{L}_{coup}(Z_{1,l}^j, \ldots, Z_{M,l}^j) = \frac{1}{LP} \sum_{l=1}^{L} \sum_{p=1}^{P} var(Z_{1,l}^j(p), \ldots, Z_{M,l}^j(p)), \tag{5}$$

where $Z^j_{i,l}(p)$ denotes the p'th element of $Z^j_{i,l}$ and var computes the variance. Minimizing this loss "couples" the l'th embeddings of M scanners. In practice, this loss only needs to be weakly imposed throughout training without degrading the embedding quality.

The combined loss for Step 1 is

$$\mathcal{L}_{step1} = \lambda_1 \mathcal{L}_{recon}(X^j_{i=1:M}, \bar{X}^j_{i=1:M}) + \lambda_2 \mathcal{L}_{coup}(Z^j_{1,l}, \ldots, Z^j_{M,l}), \tag{6}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the weights. For each of $j = 1 : N$ slices, we update $Enc_{i=1:M}$ and $Dec_{i=1:M}$. We repeat this for either T_1 times or until the model accurately reconstructs (i.e., $X^j_i \approx \bar{X}^j_i$ for all j).

Step 2: Harmonization. After Step 1, we continue with the Step 2 training (Algorithm 1 lines 12:22.) Similar to Step 1, for each slice j and scanner i, we derive the embeddings $\mathbf{Z}^j_i$ and then the output $\bar{X}^j_i$. In this particular training step, we update only $Dec_{i=1:M}$ to achieve harmonization with the following loss.

Harmonization Loss. We finally impose the image similarity across the outputs $\bar{X}^j_{i=1:M}$ across the scanners. Specifically, we consider all pairwise similarities:

$$\mathcal{L}_{harm}(\bar{X}^j_{i=1:M}) = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{k=i+1}^{M} MAE(\bar{X}^j_i, \bar{X}^j_k), \tag{7}$$

which computes the MAE for all combinations of pairs. One may concern about how a pixel-wise loss such as MAE may inadvertently alter the structures to maximize the similarity. We stress that only $Dec_{i=1:M}$ are updated while $Enc_{i=1:M}$ are fixed. Thus, the intensities will be harmonized by updating $\gamma_{i,l}$ of the embeddings in Equation (3), but the structures are guaranteed to make no further changes since the embeddings are fixed.

The final loss for Step 2 also incorporates the reconstruction loss $\mathcal{L}_{recon}$ to ensure the harmonized slices do not overly deviate from their originals:

$$\mathcal{L}_{step2} = \lambda_3 \mathcal{L}_{recon}(X^j_{i=1:M}, \bar{X}^j_{i=1:M}) + \lambda_4 \mathcal{L}_{harm}(\bar{X}^j_{i=1:M}), \tag{8}$$

where $\lambda_3 > 0$ and $\lambda_4 > 0$. Similar to Step 1, for each of $j = 1 : N$ slices, we update $Dec_{i=1:M}$. We repeat this for either T_2 times or until the harmonized images are similar enough (i.e., $\bar{X}^j_1 \approx \cdots \approx \bar{X}^j_M$ for all j.). Once the training ends, the resulting outputs $\bar{X}^{j=1:N}_{i=1:M}$ will be the desired harmonized slices.

4.2.3 Harmonization practicality

The typical approach for *supervised* harmonization methods involves utilizing matched data to train models that capture scanner effects specific to the scanners from which the matched data originated (Dewey et al., 2019; Wrobel et al., 2020). Once trained, these models can be applied to harmonize images acquired from any of the scanners involved in the training dataset. Notably, the images undergoing harmonization do not necessarily have to be matched, and the harmonization process can be applied independently to images from each scanner. To demonstrate the practicality of MISPEL in harmonization, we conducted a 6-fold cross-validation at the subject level, employing a 12/3/3 split for training, validation, and testing, respectively. In this setup, the images of validation and test sets are treated as unmatched images and are harmonized individually. Moreover, these images are harmonized by models that have not seen them during their training.

We used RAW images as the input of MISPEL. As explained in Section 4.2.2, we started by training each of the 6 models (i.e. datasets) with Step 1 and then continued with Step 2. For tuning the hyper-parameters of the models, we used the images of the validation sets. In Step 1, we fixed $\lambda_1 = 1$ and trained models for $\lambda_2 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $L \in \{4, 6, 8\}$. We then selected appropriate values of these hyperparameters for each of the 6 models based on the $\mathcal{L}_{step1}$ values for their validation sets. In Step 2, we fixed the models for $\lambda_3 = 1$ and trained the models for $\lambda_4 = \{1, 2, 3, 4, 5, 6\}$. We selected appropriate values of λ_4 for each model based on the $\mathcal{L}_{step2}$ for their validation sets. The training was conducted on NVIDIA RTX5000 for $T_1 = 100$ and $T_2 = 100$ with the batch size of 4. For both steps, we used ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 0.01. Training of each model took approximately 200 and 30 minutes for Step 1 and Step 2, respectively.

We then used the tuned models for harmonizing their corresponding test sets. In the next section, we explain that two of our competing methods, WS and RAVEL, were designed to be applied to all images at once. For ease of comparing MISPEL to these methods, we pooled all the MISPEL-harmonized test sets as one harmonized set. This is the dataset that is used in Section 4.5 for reporting the harmonization performance of MISPEL.

4.3 Competing methods

We compared MISPEL with one method of intensity normalization, White Stripe (WS), and two methods of harmonization, RAVEL and CALAMITI. We selected WS and RAVEL as they (1) are widely applied to MRI neuroimaging data, (2) can be applied to multiple (more than two) scanners, and (3) do not require specifications of a *target* scanner. We considered CALAMITI as our main competing method since it can be slightly modified and applied to matched data, and could be regarded as one of the state-of-the-art methods in harmonization. We emphasize that determining the ultimate state-of-the-art harmonization method is not trivial as harmonization lacks standardized evaluation criteria.

4.3.1 White Stripe

It is an individual-level intensity normalization method for removing discrepancy of intensities across subjects within tissue types (Shinohara et al., 2014b). It first extracts the normal-appearing white matter voxels of the image and estimates moments of their intensity distribution. It then uses these moments in the z-score transformation for normalizing the voxels of all brain tissue types.

4.3.2 RAVEL

It is an intensity normalization and harmonization framework (Fortin et al., 2016). It initializes with a WS normalization step and then applies a voxel-wise harmonization strategy to images. In the harmonization strategy, RAVEL first estimates the components of scanner effects by applying singular value decomposition to cerebrospinal fluid (CSF) voxels of images. These voxels are known to be unassociated with disease status and clinical covariates and are representative of scanner effects. RAVEL then uses these voxels to estimate scanner effects and harmonizes the images by removing the estimated scanner effects from the voxel intensities. Throughout the estimation of the scanner effects, we considered the status of the subjects (cognitively normal with low or high degree of SVD) as the biological/clinical

covariates. We also set the components of scanner effects to 1, as suggested in the original work (Fortin et al., 2016). For further details on the biological/clinical covariates and components of scanner effects, see Algorithm 1 in Section 3.2.1.

4.3.3 CALAMITI

It is an unsupervised deep-learning method for harmonizing multi-scanner inter-modality paired dataset (Zuo et al., 2021b). CALAMITI maps images of scanners to the contrast of a *target* image. Inter-modality paired dataset consists of images of two predetermined modalities taken from one individual on the *same* scanner with a short time gap. This dataset can have paired images of multiple scanners. For simplicity, we refer to these images as *paired* in the description of this method. CALAMITI should be first trained on paired images of two scanners, one of which should be the *target* scanner. It could then be fine-tuned to map images of other scanners to the target domain. During the training, CALAMITI first gets the paired images as inputs and generates a disentangled representation that captures the mutual scanner-invariant anatomical information (β) of images as well as the scanner-variant contrast information (θ)s of their modalities and scanner. It then synthesizes the input paired images using their generated mutual β and θs. For harmonizing an input image, the trained model is used to generate the β of the image and θ of the target image. The model then synthesizes the harmonized image (adapted image to the target domain) using these two components.

We used CALAMITI as a supervised method by simply training it on our *inter-scanner* paired data. Like MISPEL, we used the 6-fold cross-validation strategy for training and testing the models. We also pooled the harmonized test sets to have one set of data to report the harmonization performance of CALAMITI in Section 4.5. Following its original paper, we went through one step of normalization and trained CALAMITI using the WS-normalized RAW images. Instead of conducting fine-tuning, we went for a simpler approach and trained 3 individual models to map GE, Philips, and SiemensP to SiemensT as our target scanner. We used the same machines used for MISPEL and trained CALAMITI with the hyper-parameters reported in its original paper. For being comparable and fair to

other methods, we trained CALAMITI on 2D axial slices and skipped its super-resolution preprocessing step and post-harmonization slice-to-slice consistency enhancement step.

Among the competing methods, we regard CALAMITI as a state-of-the-art harmonization method to compare against MISPEL, and we emphasize that WS and RAVEL were not designed to use matched data in their technical variability removal process. Specifically, WS is an intensity normalization method, which does not account for scanner information. However, it is beneficial to study scanner effects and harmonization on the WS-normalized data to emphasize the importance of harmonization for neuroimaging data. On the other hand, RAVEL was designed to remove the inter-subject technical variability of images after intensity normalization. Although RAVEL does not account for scanner information either, scanner effects may appear in the singular value decomposition component extracted individually for each of the subjects from their CSF tissue in this framework. As such, we regard RAVEL as a normalization and harmonization framework that can be compared to CALAMITI and MISPEL to evaluate the advantages of using and accounting for matched data in harmonization methodology.

4.4 Data analysis

A harmonization method is expected to remove scanner effects while preserving the biological variables of interest in the data. In our specific matched dataset, the matched images are assumed to be biologically identical but differ due to scanner differences. Thus, the scanner effects can be estimated as dissimilarity of the matched images, and removing the scanner effects can be regarded as increasing their similarity. We investigated the dissimilarity and similarity of matched images using four evaluation criteria: (1) image similarity, (2) GM-WM contrast similarity, (3) volumetric and segmentation similarity, and (4) biological similarity. We also selected SVD as the clinical signal of interest in our data and investigated whether we could preserve or even enhance the SVD group differences in our data after harmonization.

We performed our evaluation metrics for all five methods: RAW, White Stripe, RAVEL,

CALAMITI, and MISPEL. The entire matched dataset was used in evaluating each method unless otherwise mentioned. Many of our evaluation metrics require pairwise image-to-image comparison, for which we considered all possible combinations of *scanner pairs*: {(GE, Philips), (GE, SiemensP), (GE, SiemensT), (Philips, SiemensP), (Philips, SiemensT), and (SiemensP, SiemensT)}. Throughout this manuscript, the two matched images of each scanner pair are referred to as *paired* images. To determine the statistical significance of any comparisons, we used paired t-test with $p < 0.05$ denoting the significance.

Scanner effects could appear as contrast dissimilarity across images of different scanners (Dewey et al., 2019, 2020; Liu et al., 2021). More specifically, such dissimilarity could appear as tissue-specific contrast differences in images (Meyer et al., 2019). We, therefore, assessed scanner effects and evaluated harmonization using an **image similarity** metric to measure the similarity of cross-scanner images in their appearance, as well as a **GM-WM contrast similarity** metric to assess the tissue contrast similarity of images.

We first investigated the **image similarity**. For this, we assessed the visual quality of the matched *slices* for all methods. We also quantified the similarity of *all* paired images using the structural similarity index measure (SSIM). SSIM is a pairwise metric that compares two images in terms of their luminance, contrast, and structure. A harmonization method is expected to increase the visual and structural similarity of paired images.

Second, we investigated the **GM-WM contrast similarity** of images. The GM-WM contrast can highly influence the quality of segmentation methods, and increased contrast is expected to result in more accurate segmentation. The GM-WM contrast of an image can be estimated as the separability of its histograms of GM and WM voxels. This separability was conducted as the classification of GM and WM voxels of an image in (Torbati et al., 2021) and reported as the area under the receiver operating characteristic (AUROC) values, with AUROC = 100% denoting perfect classification (complete separation of histograms) and AUROC = 50% showing random classification (complete overlap of histograms). For calculating AUROC, we first labeled GM and WM voxels of the image using the tissue mask in the EveTemplate package (Oishi et al., 2009). We then classified these voxels using intensity thresholds selected from the range of intensity values of the GM and WM voxels. Lastly, we formed the AUC curve of the image using the result of each classification. A

harmonization method is expected to increase the GM-WM contrast similarity.

Third, we investigated the **volumetric and segmentation similarity** criterion for images. The most practical benefit of harmonization is to enable unbiased multi-scanner neuroimaging analyses with minimal scanner effects. Tissue-specific regional neuroimaging measures are the basis of these analyses, and therefore, the volumetric and segmentation similarity of these measures across paired images is a crucial metric for evaluating harmonization. We segmented and measured the volumes of the two brain tissue types: GM and WM. We then analyzed the similarity of *each of* these two tissue types *separately* in four ways: (1) volume distributions, (2) volumetric bias, (3) volumetric variance, and (4) segmentation overlap. For volumetric distributions, we compared the distributions of volumes of each tissue type across their four scanners. These plots show the harmonization performance of methods as the similarity of the distributions of their harmonized measures across scanners. Most of the metrics used in the three other criteria are pairwise comparisons, thus we applied them *separately* to all of the 6 *scanner pairs*. Volumetric bias and variance are two metrics assessing the similarity of measures across scanners in two different ways. For volumetric bias, we calculated the absolute differences between volumes of paired images of each scanner pair and evaluated the harmonization based on the mean of these differences over all individuals of the scanner pair. We used root-mean-square deviation (RMSD) for estimating the volumetric variance of paired images of all individuals within each scanner pair. RMSD of a scanner pair denotes the deviation of volumes of one scanner from that of the other scanner. Lastly, we used Dice similarity score (DSC) to estimate the overlap of tissue segmentation of paired images of each scanner pair. The mean of these DSC values over paired images of all subjects was used as an evaluation metric for harmonization. A harmonization method is expected to result in (1) similar distribution of volumes across scanners, (2) minimal (ideally zero) bias, (3) minimal (ideally zero) variance, and (4) maximal (ideally complete) segmentation overlap; for both tissue types and all scanner pairs.

We conducted the volumetric and segmentation similarity evaluation using two segmentation tools: (1) FSL FAST (version 6.0.3) (Zhang et al., 2001), and (2) segmentation in Statistical Parametric Mapping (SPM12) (Ashburner and Friston, 2005). These frameworks are widely used for tissue segmentation in neuroimaging studies, however, the results of these two

segmentation algorithms could have moderate to large differences (Tudorascu et al., 2016). We, therefore, assessed volumes from each segmentation tool independently. Originally, the output of WS, RAVEL, CALAMITI, and MISPEL methods were images in template space, as all methods used RAW images as input. The RAW images were non-linearly registered to a T1-w image atlas (Oishi et al., 2009) in the preprocessing step, Section 4.1.2. Using their inverse transformations, processed images of all methods were transferred to their native space and then used as inputs of the two segmentation tools for tissue volume extraction and then volumetric similarity evaluation. On the other hand, for having a meaningful tissue segmentation overlap, segmentations and accordingly their images should remain in their template space. Thus, we also ran FSL and SPM frameworks on the template-space images to generate the segmentations and then evaluate the segmentation overlap similarity. For all runs of the segmentation frameworks, we set the tissue class probability thresholds to 0.8.

Fourth, we investigated the **biological similarity** of images using biomarkers of Alzheimer's disease (AD). We studied the bias (mean of cross-scanner absolute differences) and variance (RMSD) for these biomarkers. For bias, we calculated the cross-scanner absolute differences of all scanner pairs and reported their mean (SD). For variance, we calculated the mean of RMSDs across all scanner pairs. We report these metrics for all 5 methods and all biomarkers of AD. As biomarkers of AD, we investigated cortical thickness measures of the entorhinal and inferior temporal cortices, as well as volume measures of the hippocampus and amygdala. These summary measures are the sum of measures over both hemispheres, and they were extracted using FreeSurfer 7.1.1 (FS) (Fischl, 2012). These regions have previously been found to be most relevant to AD (Schwarz et al., 2016). We extracted these measures across all harmonization methods for 17 of the 18 total subjects. RAVEL-harmonized images of a single subject failed FS segmentation due to an error in the corpus callosum segmentation step. Thus, for a fair comparison across methods, we omitted this subject from the experiments on biomarkers of AD. We also skipped skull stripping and bias correction steps in the FS processing pipeline, as RAW images had already gone through skull-stripping and N4 bias correction during image preprocessing (Section 4.1.2).

Fifth and last, we investigated whether each harmonization method **preserved or even enhanced a biological/clinical signal of interest** in our matched data. We selected

SVD as our clinical signal of interest and investigated the effect size between two groups of low and high SVD in our data. For this experiment, we calculated Cohen's d effect sizes of the two SVD groups for each of our FS-derived biomarkers of AD individually. For each of the biomarkers, we calculated the size effects of the scanners separately and reported the mean (SD) of these values across scanners. A harmonization method is expected to not deteriorate the effect sizes of groups after harmonization.

4.5 Results

In this section, we report our evaluation criteria on RAW, WS-normalized, RAVEL-, CALAMITI-, and MISPEL-harmonized images. For a more convenient comparison with RAW, WS and RAVEL, we pooled harmonized images of each of CALAMITI and MISPEL as one dataset.

4.5.1 Image similarity

The similarity of images across normalization and harmonization methods is depicted in Figures 13 and 14. Visual assessment of processed images in Figure 13 revealed that (1) scanner effects are present in the matched RAW images and appear most significantly as differences in image contrast, (2) White Stripe made matched images more similar, but at the expense of decreased contrast, (3) RAVEL improved upon WS by increasing contrast relative to WS-normalized images, (4) CALAMITI improved similarity of the matched images by adapting contrast across all scanners to that of the RAW SiemensT, and (5) MISPEL improved the similarity of images similarly to CALAMITI but visually smoothed images to some extent.

For a quantitative understanding of similarity of images, we explored the SSIM distribution of the matched images of all subjects for the 6 *scanner pairs* enumerated in Section 4.4. These distributions are depicted as violin plots for the five methods: RAW, WS, RAVEL, CALAMITI, and MISPEL in Figure 14. The violin plots with the smallest SSIM mean

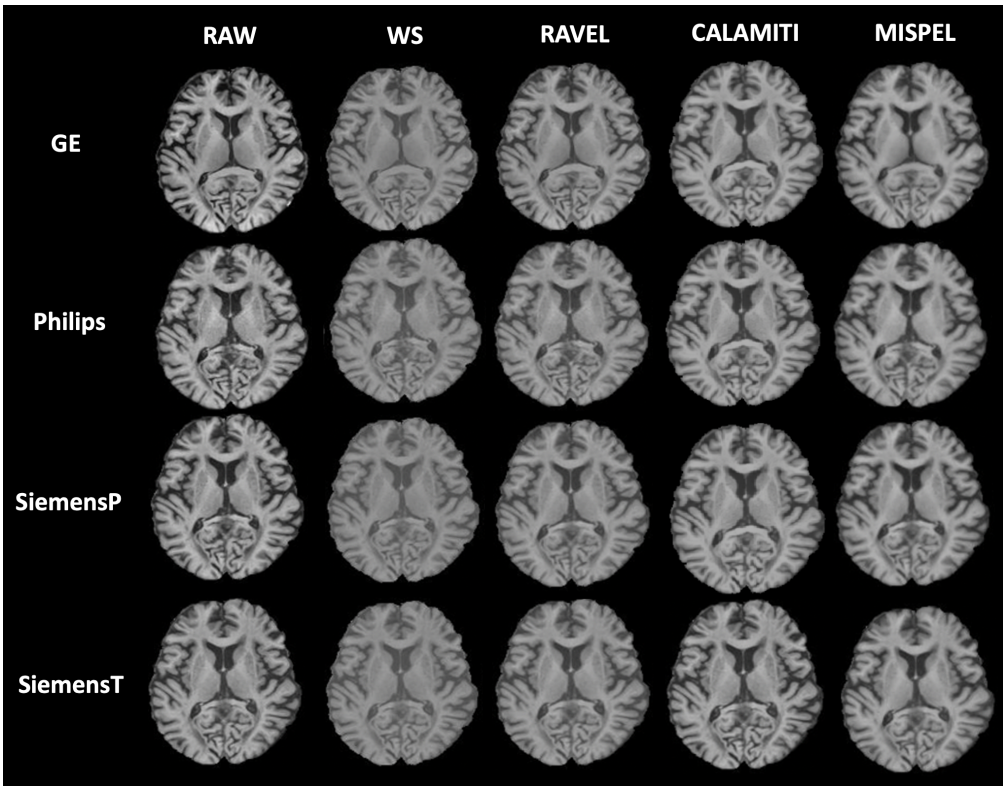


Figure 13: **Visual assessment of matched images of a slice.** Rows and columns correspond to methods and scanners, respectively. All four methods: WS, RAVEL, CALAMITI, and MISPEL made the matched slices of RAW more similar, with CALAMITI and MISPEL preserving their contrast the most.

belong to RAW, indicating scanner effects exist in our matched dataset as dissimilarity of images. Scanner pairs including GE have long-tailed distributions, which indicates that GE images are most dissimilar to others. Moreover, the SiemensP-SiemensT scanner pair had the largest SSIM mean, indicating that these two are the most similar scanners.

We observed that WS, RAVEL, CALAMITI, and MISPEL improved SSIM of RAW for all of its scanner pairs, except for CALAMITI for the SiemensP-SiemensT scanner pair. Lastly, we observed that MISPEL outperformed the other three methods. All comparisons were statistically significant (assessed using paired t-tests), except for CALAMITI for the Philips-SiemensP and SiemensP-SiemensT pairs.

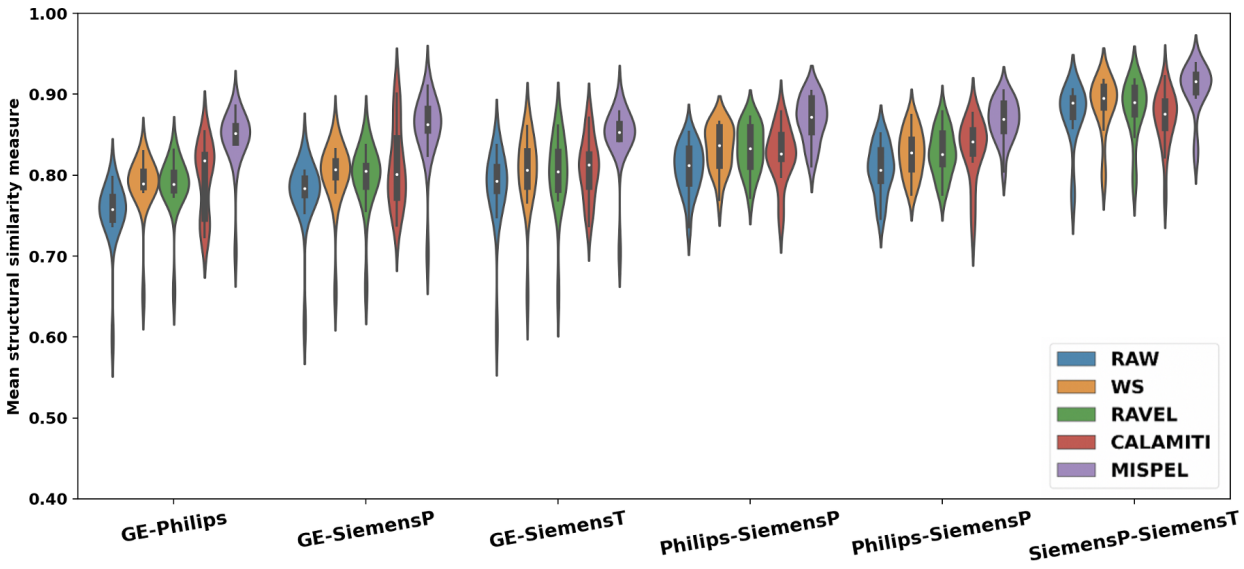


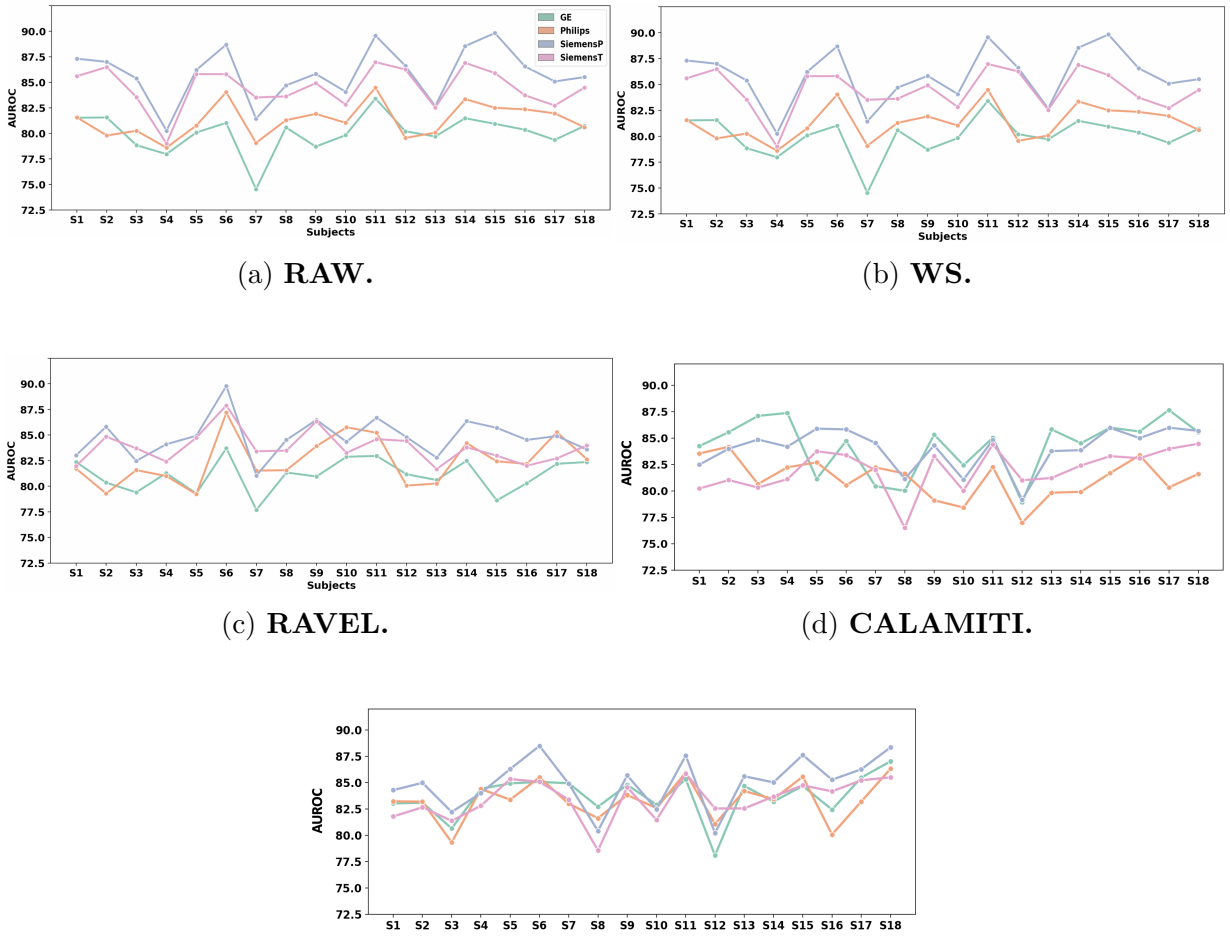
Figure 14: **Structural similarity index measures (SSIM) for the matched dataset.** The SSIM distributions of images of scanner pairs were depicted as violin plots for each of the methods. A harmonization method is expected to have the highest mean of SSIM. MISPEL improved SSIMs of RAW and outperformed the other three methods.

4.5.2 GM-WM contrast similarity

We quantified the GM-WM contrast of an image using the AUROC values denoting the separation of histograms of GM and WM voxel intensities. High AUROC indicates higher contrast, with 100% the highest. In Figure 15, we depicted the spaghetti plots of AUROC values of images of all subjects across the four scanners. A harmonization method is expected to (1) make the AUROC of matched images similar, i.e., results in overlapped lines, and (2) not deteriorate the AUROC of images.

Figure 15a shows that scanner effects exist in RAW data and appeared as dissimilarity of GM-WM contrast in matched dataset, i.e., distant lines in this plot. Figure 15b shows that WS does not change AUROCs of RAW. On the other hand, Figures 15c, 15d, and 15e show respectively that RAVEL, CALAMITI, and MISPEL resulted in more overlapped lines, with MISPEL having the highest overlap.

Figure 16 shows the bar plots indicating the mean AUROC of images of each scanner.



(a) **RAW.** (b) **WS.**

(c) **RAVEL.** (d) **CALAMITI.**

(e) **MISPEL.**

Figure 15: **GM-WM contrast spaghetti plots.** The GM-WM contrast was estimated as AUROC values and was depicted for images of all subjects as spaghetti plots. In these plots, each line corresponds to one scanner. A harmonization method that performs well should show overlap of the lines. Plots showed that MISPEL outperformed WS, RAVEL, and CALAMITI with the highest overlap of the lines.

MISPEL is the only method that increased the mean AUROC of RAW images for all scanners. We also observed that: (1) WS did not change the mean AUROC value of RAW, (2) RAVEL improved the contrast for GE and Philips, but made it worse for SiemensP and SiemensT, and (3) CALAMITI improved the mean AUROC of GE and Philips and did not

affect that of other scanners. In addition to these results, MISPEL seems to be the most successful method in bringing the mean AUROC of the scanners closer to each other. In summary, we show that MISPEL is the only method that satisfied both harmonization criteria determined for GM-WM contrast similarity.

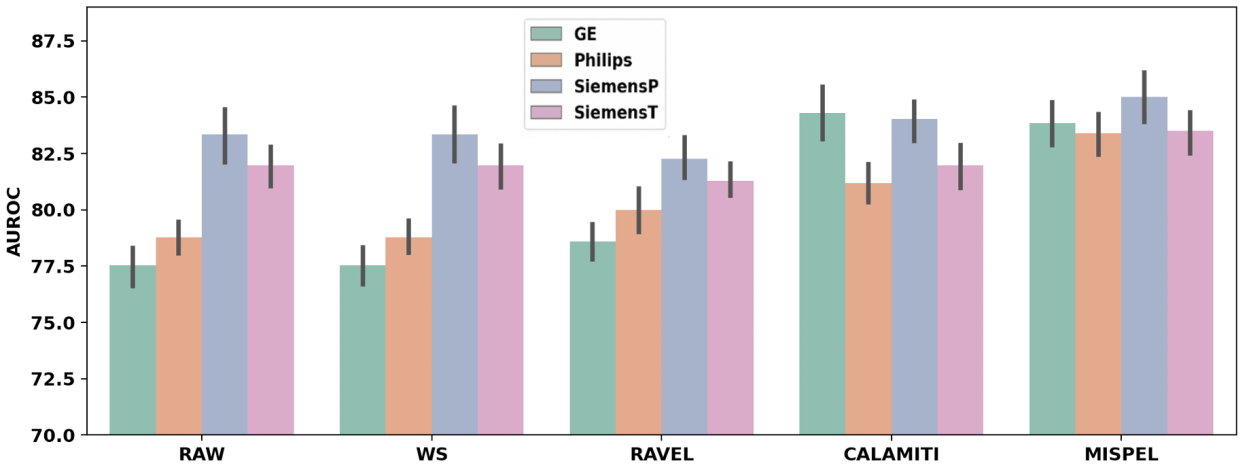
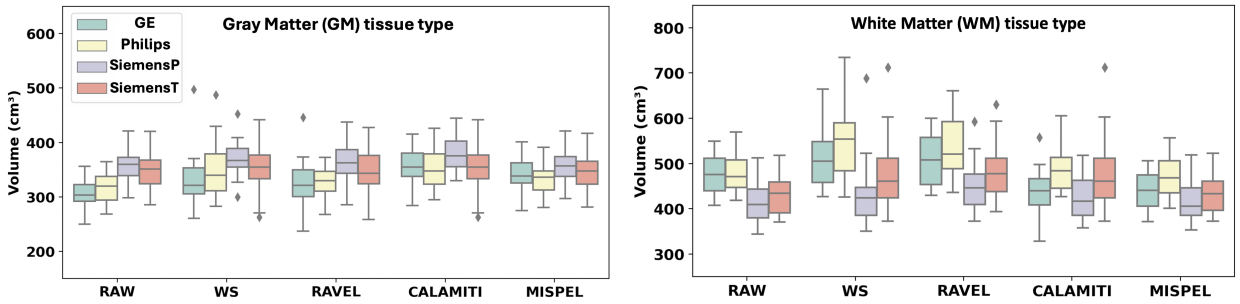


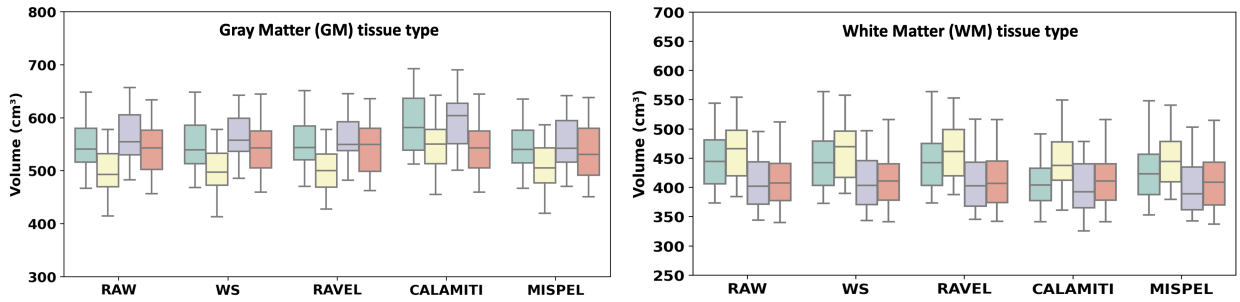
Figure 16: **GM-WM contrast bar plots.** Each bar indicates the mean AUROC of images of each scanner, with error bars denoting the standard deviation for each method. A harmonization method is expected to not deteriorate the GM-WM contrast of images. Plots show that MISPEL outperformed WS, RAVEL, and CALAMITI reflected in the similarity of the boxplots.

4.5.3 Volumetric and segmentation similarity

We estimated the volumetric and segmentation similarity of GM and WM tissue types based on four criteria: (1) volume distributions, (2) volumetric bias, (3) volumetric variance, and (4) segmentation overlap. We performed our evaluation for FSL and SPM segmentation frameworks and expected the harmonization methods to result in: (1) similar volume distributions across scanners, (2) minimal bias, (3) minimal variance, and (4) maximal segmentation overlap; for both tissue types and both segmentation frameworks.



(a) **FSL framework.** MISPEL outperformed WS, RAVEL, and CALAMITI by resulting in more similar volume distributions across scanners for both tissue types.



(b) **SPM framework.** No *visually significant* noticeable harmonization was observed for any of the methods.

Figure 17: **Volume distribution boxplots.** Boxplots denote the volume distribution of GM and WM tissue types for images of each scanner. These boxplots were depicted for all five methods and explored for two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to result in similar distributions of volumes across scanners.

4.5.3.1 Volume distributions

Figure 17 shows boxplots of volumes of the two tissue types, GM and WM, across the four scanners for all five methods, with Figures 17a and 17b depicting these boxplots for volumes extracted by FSL and SPM frameworks, respectively. Plots in Figure 17a showed that scanner effects exist in the matched volumes derived through FSL and appeared as dissimilar boxplots for RAW across scanners. When compared to RAW, WS and RAVEL resulted in

more dissimilar boxplots for FSL-derived volumes of both GM and WM. On the other hand, we noticed that the use of CALAMITI and MISPEL helped towards harmonization of data. CALAMITI made GE and Philips more similar to SiemensP and SiemensT for both GM and WM, but increased variance for distributions of all scanners for WM. Similarly, MISPEL made GE more similar to SiemensP and SiemensT for both GM and WM volumes.

Figure 17b showed that scanner effects exist in RAW volumes extracted by SPM too. Our normalization and harmonization methods though resulted in relatively minor changes in SPM-derived GM and WM volumes, with CALAMITI and MISPEL showing the most noticeable changes. Both CALAMITI and MISPEL made Philips closer to SiemensP and SiemensT for GM volumes. They also made GE closer to these two scanners for WM.

In summary, MISPEL and CALAMITI outperformed WS and RAVEL in harmonizing FSL-derived volumes and none of the methods resulted in *visually significant* assessed harmonization for the SPM-derived volumes, when volumetric distribution similarity of *both* GM and WM volumes were used as the evaluation metric. Results for the statistical assessment of harmonization of FSL- and SPM-derived GM and WM volumes are presented in the next section.

4.5.3.2 Volumetric bias

Table 4 shows mean and standard deviation (SD) of cross-scanner absolute differences of all paired volumes in each scanner pair. We calculated these statistics for volumes of GM and WM tissue types extracted using FSL and SPM segmentation frameworks, for all five methods. We also presented the distributions of these differences as violin plots in Figure 18. Using paired t-test, we compared each of these distributions to their equivalent distributions in RAW.

A harmonization method is expected to result in minimal (ideally zero) mean of absolute differences (bias), with no major increase in SD of the differences. The SD values indicate the consistency of harmonization across subjects. A harmonization method should harmonize images of all subjects to a comparable degree, and thus should not increase the SDs drastically. Likewise visually, the violin plots in Figure 18 for harmonized images are

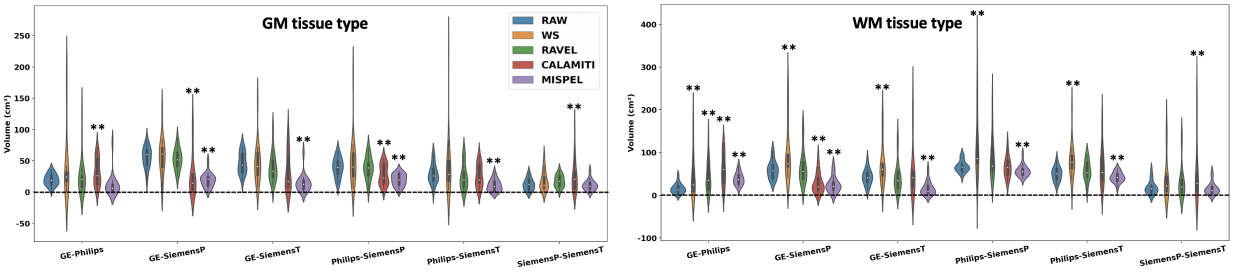
expected to be centered as close as possible to zero.

We observed that scanner effects exist in the RAW volumes extracted through FSL framework and appeared for all scanner pairs as non-zero bias values. We also observed that MISPEL resulted in the largest number of smallest biases for FSL-derived volumes, when compared to the other three methods. This number was 11 out of a total of 12 cases, which are the 6 scanner pairs of the 2 tissue types. 8 out of these 11 biases were significantly different than their equivalents in RAW. Moreover, we noticed that MISPEL did not significantly increase the SD of distributions, just 2 increases out of 12, in which only the SD of GM for the GE-Philips pair had a major increase. On the other hand, WS,

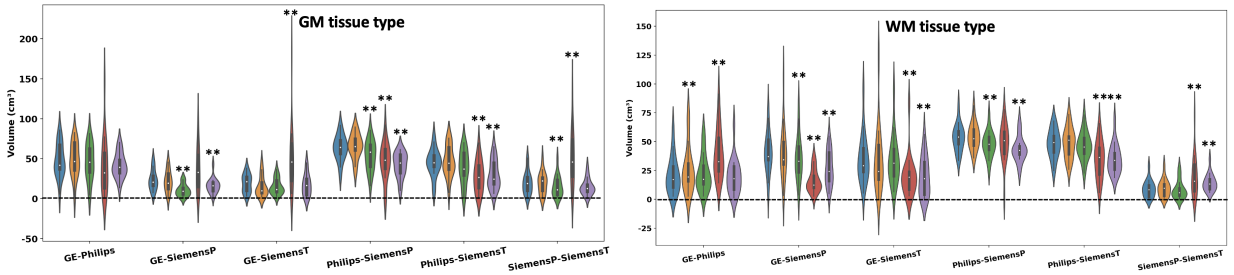
Table 4: **Mean absolute differences for GM and WM volumes.** Mean (SD) of cross-scanner absolute differences of volumes for all scanner pairs and all methods. The volumes are for GM and WM tissue types and were extracted through two segmentation frameworks: FSL and SPM. A harmonization method that works is expected to have low values of mean and SD for all paired scanners. MISPEL outperformed WS, RAVEL, and CALAMITI by having the largest number of smallest means and not significantly increasing the values of SD, for both FSL and SPM. The distributions with the smallest means are in bold. Also, the distributions that showed statistically significant t-statistics when compared to RAW were marked by *.

Framework	**Tissue**	**Method**	**Mean (SD) of Volumetric Absolute Differences for Paired Dataset**					
			GE-Philips	GE-SiemensP	GE-SiemensT	Philips-SiemensP	Philips-SiemensT	SiemensP-SiemensT
FSL	GM	RAW	19.82 (9.10)	55.84 (16.54)	46.53 (16.94)	39.70 (15.28)	29.00 (15.75)	12.14 (9.17)
		WS	43.53 (56.27)	56.66 (29.56)	46.34 (31.84)	49.31 (40.21)	43.09 (49.92)	18.00 (15.37)
		RAVEL	27.53 (32.28)	52.88 (16.60)	39.22 (22.41)	38.87 (17.79)	24.65 (20.50)	17.53 (9.35)
		CALAMITI	32.29 (21.87)*	26.07 (33.88)*	32.18 (31.05)	28.06 (16.00)*	26.07 (17.98)	26.02 (26.47)*
		MISPEL	**11.10** (17.81)	**19.04** (10.91)*	**14.26** (14.14)*	**19.48** (9.59)*	**10.73** (9.28)*	**11.10** (8.31)
	WM	RAW	**15.39** (11.29)	59.59 (20.92)	42.45 (18.37)	67.30 (13.41)	50.16 (16.43)	17.89 (15.48)
		WS	46.99 (54.09)*	100.63 (64.18)*	71.35 (47.72)*	119.73 (79.23)*	81.41 (41.29)*	41.03 (50.11)
		RAVEL	43.95 (36.46)*	65.02 (37.96)	42.42 (34.98)	89.59 (49.34)	57.60 (23.77)	32.18 (39.53)
		CALAMITI	61.39 (37.88)*	28.57 (23.69)*	59.84 (64.36)	65.60 (24.08)	68.64 (46.78)	58.07 (74.88)*
		MISPEL	38.56 (14.31)*	**24.95** (18.39)*	**15.73** (15.60)*	**57.09** (13.30)*	**42.27** (11.87)*	**15.34** (14.01)
SPM	GM	RAW	48.22 (20.82)	23.45 (12.67)	19.37 (11.23)	63.57 (15.90)	44.65 (16.94)	19.86 (13.77)
		WS	48.60 (21.35)	21.75 (13.15)	**14.94** (12.70)	65.72 (15.54)	46.84 (18.99)	19.46 (13.53)
		RAVEL	46.12 (22.48)	**10.44** (7.57)*	15.22 (9.77)	53.82 (18.48)*	39.14 (20.99)	15.26 (12.85)*
		CALAMITI	42.22 (36.06)	37.85 (28.09)	49.62 (41.39)*	46.81 (23.06)*	**28.44** (19.67)*	51.51 (34.64)*
		MISPEL	42.74 (15.04)	16.28 (10.44)	18.06 (15.41)	**41.07** (15.77)*	30.28 (17.58)*	**14.65** (10.15)
	WM	RAW	21.06 (15.98)	40.40 (18.08)	35.45 (20.87)	53.16 (11.74)	48.22 (12.32)	9.06 (7.31)
		WS	25.97 (20.29)*	40.18 (23.46)	34.18 (27.71)	54.43 (11.43)	48.80 (13.16)	9.69 (7.53)
		RAVEL	22.49 (15.69)	35.60 (19.36)*	34.02 (21.03)	47.64 (10.95)*	46.48 (12.14)	**8.41** (8.00)
		CALAMITI	40.14 (21.52)*	**16.24** (9.80)*	**20.13** (18.07)*	49.31 (16.99)	**34.67** (16.21)*	20.21 (17.92)*
		MISPEL	**19.82** (15.10)	27.34 (14.73)*	**19.88** (17.08)*	**43.61** (11.00)*	35.05 (12.66)*	14.43 (7.34)*

RAVEL, and CALAMITI showed increases in SD of differences for all 12 distributions, with WS showing the most drastic increases (Figure 18). In general, RAVEL and CALAMITI harmonized FSL-derived volumes to some extent. Compared to RAW, RAVEL resulted in 5 decreased biases and CALAMITI resulted in 6 decreases. However, CALAMITI also resulted in drastically increased biases for the WM volumes of 5 of the scanner pairs (Figure 18a).



(a) **FSL framework.**



(b) **SPM framework.**

Figure 18: **Absolute difference violin plots.** The distributions of absolute differences of paired volumes as violin plots for all scanner pairs. The volumes are for GM and WM tissue types and extracted using two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to result in short and fat (wide) violins, with mean values centered at zero. MISPEL outperformed WS, RAVEL, and CALAMITI by having largest number of these violin plots for both FSL and SPM. The distributions that showed statistically significant t-statistics when compared to RAW were marked by **.

Results of RAW volumes extracted by SPM show that SPM is also sensitive to scanner effects. MISPEL and CALAMITI decreased bias for 11 and 7 cases, respectively. They

resulted in the largest numbers of smallest biases for SPM: 5 and 4 out of 12 cases for MISPEL and CALAMITI, respectively. Among these cases, 3 for each of MISPEL and CALAMITI showed statistically significant differences when compared to RAW. On the other hand, CALAMITI increased SD for 8 out of 12 cases, while other methods did not show any major increases. This can be observed in Table 4 as well as Figure 18b. WS and RAVEL harmonized the SPM-derived volumes to some extent by decreasing the biases of 5 and 11 cases, respectively. They also resulted in a few smallest biases: 1 case for WS and 2 cases for RAVEL.

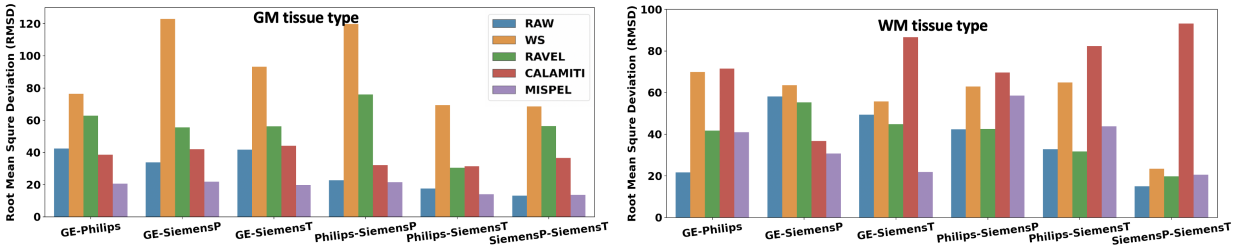
Summarizing Table 4 and Figure 18, we observed that MISPEL outperformed WS, RAVEL, and CALAMITI when FSL and SPM were used for extracting volumes and volumetric bias and SD of differences were used as harmonization evaluation metrics.

4.5.3.3 Volumetric variance

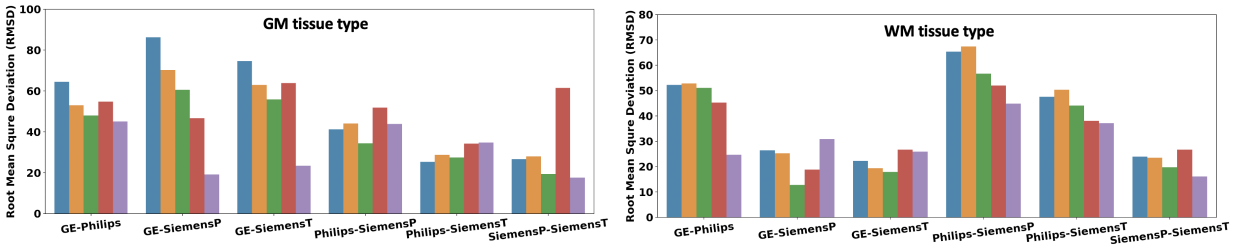
Figure 19 shows bar plots that indicate the RMSD of paired volumes in each of the scanner pairs. We calculated these values for volumes of GM and WM tissue types and depicted them for all five methods. Figure 19 contains these sets of bar plots for volumes extracted through FSL and SPM frameworks in Figures 19a and 19b, respectively. Ideal harmonization would result in a zero RMSD for each scanner pair.

We observed that scanner effects exist in RAW volumes for both segmentation frameworks and appeared as non-zero RMSD values. Also, MISPEL outperformed WS, RAVEL, and CALAMITI, showing the smallest RMSD values: 6 and 8 out of 12 cases for FSL and SPM, respectively. These statistics are 0 and 1 for CALAMITI as well as 0 and 3 for RAVEL. We also observed that WS did not improve the RMSD values of any 12 scanner pairs for FSL, when compared to RAW. However, it performed better for SPM by decreasing the number of worse cases to 6. MISPEL, CALAMITI, and RAVEL deteriorated some of the RMSDs too. Among these methods, MISPEL deteriorated the least number of cases, 4 for each of the FSL- and SPM-derived volumes.

In summary, we observed that MISPEL outperformed WS, RAVEL, and CALAMITI when FSL and SPM were used for deriving volumes and volumetric variance was used as the



(a) **FSL framework.**



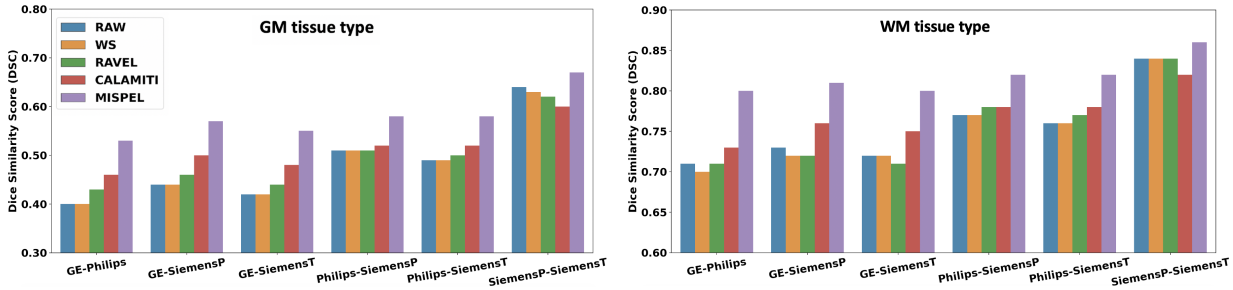
(b) **SPM framework.**

Figure 19: **Root-mean-square deviation (RMSD) bar plots for GM and WM volumes.** Bar plots indicate the RMSD of paired volumes in scanner pairs. These values were calculated for volumes of GM and WM tissue types and depicted for all five methods. These set of bar plots were depicted for volumes extracted through two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to lower values of RMSDs. MISPEL outperformed WS, RAVEL, and CALAMITI by having the largest number of smallest RMSD values for volumes of both FSL and SPM.

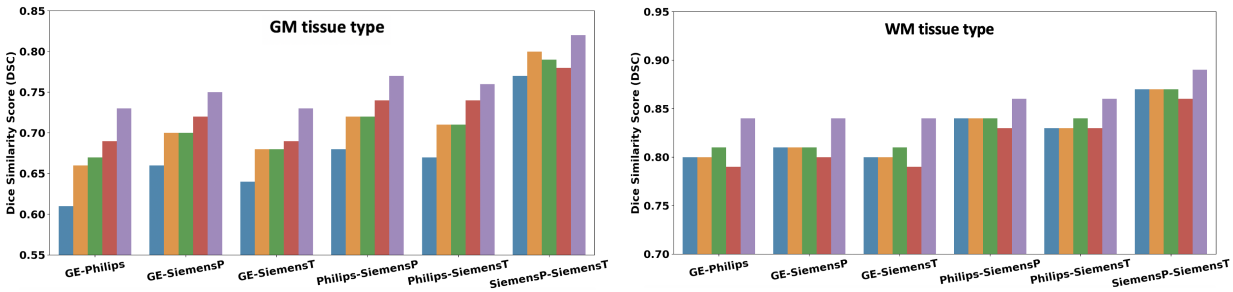
harmonization evaluation metric.

4.5.3.4 Segmentation overlap

Figure 20 shows bar plots that indicate the mean DSC of all paired segmentations in each scanner pair. We calculated the means of DSCs for segmentations of GM and WM tissue types and depicted them for all five methods. Figure 20 contains these sets of bar plots for segmentations extracted through FSL and SPM frameworks in Figures 20a and 20b,



(a) **FSL framework.**



(b) **SPM framework.**

Figure 20: **Dice similarity score (DSC) bar plots.** Bar plots indicate the means of DSCs of all paired segmentations in scanner pairs. These values were calculated for segmentations of GM and WM tissue types and depicted for all four methods. These set of bar plots were depicted for volumes extracted through two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to result in high mean of DSCs. MISPEL outperformed WS, RAVEL, and CALAMITI by having the largest DSC means for all scanner pairs in both FSL and SPM.

respectively. DSC shows the overlap of two paired segmentations. A good harmonization method would result in an increased mean of DCSs for all scanner pairs, with 1 indicating the highest.

We observed in Figure 20 that scanner effects exist in RAW segmentations of both FSL and SPM and appeared as relatively low means of DSC values. MISPEL outperformed WS, RAVEL, and CALAMITI in harmonization by having the largest means of DSC for all scanner pairs for both FSL and SPM. We compared the DSC distributions of MISPEL

with their equivalents in RAW using paired t-test and all improvements of MISPEL over RAW were statistically significant. Results also showed that while WS decreased the DSC for two scanner pairs for FSL, it did better for SPM by increasing the means for 6 of the cases. RAVEL performed slightly better than WS by increasing 6 and decreasing 3 of the DSC means for FSL and improved 9 cases for SPM. CALAMITI showed 10 and 6 increases for FSL and SPM, respectively, while decreasing the rest of the cases. Using paired t-tests, we observed that these DSCs were statistically significantly larger than that of their RAW equivalents.

In summary, MISPEL outperformed WS and RAVEL, when FSL and SPM were used as segmentation frameworks and segmentation overlap was used as the harmonization evaluation metric.

4.5.4 Biological similarity

We investigated biological similarity of images over several biomarkers of AD: cortical thickness values of the entorhinal and inferior temporal cortices, as well as volume measures of the hippocampus and amygdala. As the evaluation criteria, we selected (1) biomarker bias, and (2) biomarker variance. A harmonization method is expected to result in minimal bias and variance for the biomarkers.

Table 5 shows the biomarker bias for each of the AD biomarkers. We reported this metric for all 5 methods. For each method, we first calculated the absolute differences between paired measures of all the scanner pairs and then reported their overall mean (SD). We also compared the distribution of differences for each of the methods to that of RAW, using paired t-test. Moreover, Figure 21 shows the mean of RMSDs across all scanner pairs for each of the methods. These means were calculated for each of the AD biomarkers separately.

We observed in Table 5 and Figure 21 that scanner effects appeared as non-zero bias and variance values for the biomarker measures in the RAW data, respectively. We also noticed that MISPEL resulted in the largest number of statistically-significant smallest biases: 3 out of 4. MISPEL did not harmonize hippocampus. It slightly increased cross-scanner volumetric differences for hippocampus, but this increase is not statistically significant. On the other

hand, WS and RAVEL statistically significantly increased the distribution of differences for all biomarkers, except for amygdala. CALAMITI showed similar performance. This method resulted in increase in distribution of differences for 3 biomarkers while being statistically significant for 2 of them. The same trend of results was also seen for the mean of RMSD values in Figure 21.

In summary, we observed that MISPEL outperformed WS, RAVEL, and CALAMITI when harmonization was investigated as bias and variance across scanners in FS-derived biomarkers of AD.

4.5.5 Analysis on biological variables of interest

We investigated whether harmonization could succeed in preserving or strengthening SVD-related group differences in our data. For this, we studied the Cohen's d effect sizes of SVD groups in each of the scanners. We calculated these values for each of the biomarkers and methods separately. Table 6 shows mean (SD) of these Cohen's d values over all scan-

Table 5: **Mean absolute differences for biomarkers of AD.** Mean (SD) of cross-scanner absolute differences were calculated for paired measures across all scanner pairs. The measures are the FS-derived cortical thicknesses for the entorhinal and inferior temporal cortices, as well as volumes for the hippocampus and amygdala. A harmonization method is expected to decrease mean and SD of differences in RAW. MISPEL showed the best harmonization performance by having the largest number of smallest mean of differences. The distributions with the smallest means are in bold. Also, the distributions that showed statistically significant t-statistics when compared to RAW were marked by *.

Mean (SD) of absolute differences over all scanner pairs				
	Cortical Thickness (mm)		**Volume (cm^3)**	
Method	Entorhinal	Inferior Temporal	Hippocampus	Amygdala
RAW	0.62 (0.42)	0.46 (0.36)	0.30 (0.23)	0.25 (0.20)
WS	1.00 (0.73)*	0.63 (0.48)*	0.43 (0.52)*	0.23 (0.30)
RAVEL	0.84 (0.57)*	0.56 (0.41)*	0.41 (0.29)*	0.24 (0.21)
CALAMITI	0.87 (0.60)*	0.45 (0.32)	0.71 (0.54)*	0.30 (0.26)
MISPEL	**0.46** (0.34)*	**0.25** (0.24)*	0.32 (0.26)	**0.19** (0.18)*

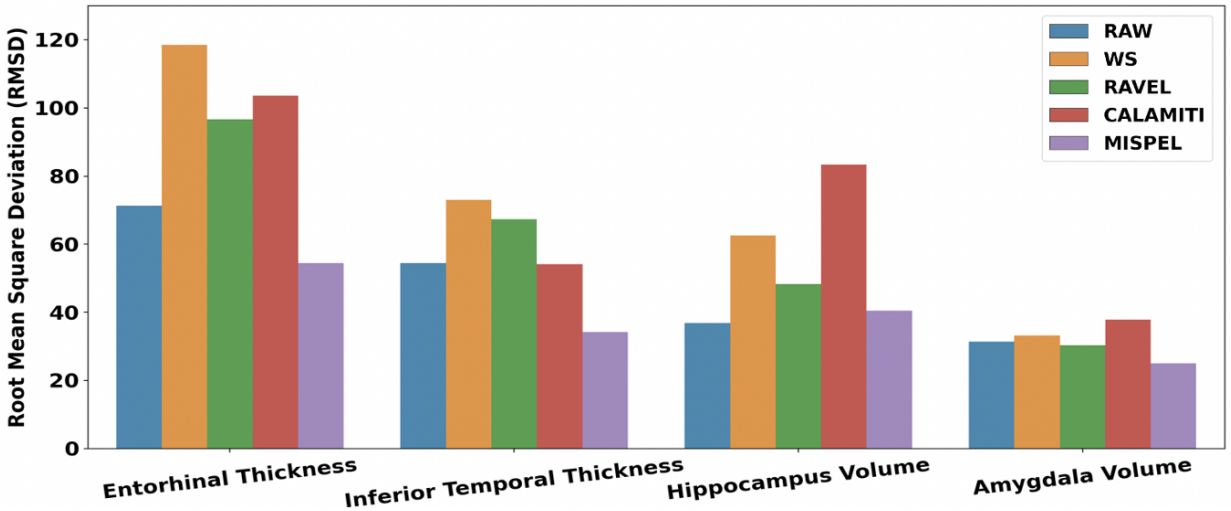


Figure 21: **Root-mean-square deviation (RMSD) bar plots for biomarkers of AD.** Each bar indicates the mean RMSD of paired measures of all scanner pairs for each of the methods. The RMSDs were reported for each of the FS-derived biomarkers of AD. A harmonization method is expected to lower values of RMSDs. MISPEL outperformed WS, RAVEL, and CALAMITI by having the largest number of smallest RMSD values..

ners. A harmonization method is expected to not reduce these means of Cohen's d after harmonization, that is to preserve group differences. We observed that MISPEL increased

Table 6: **Mean (SD) of Cohen's d measures for biomarkers of AD.** Mean (SD) of Cohen's d values were calculated over all scanners for biomarkers of AD and all methods. A harmonization method is expected to preserve or increase the effect sizes calculated relative to RAW. Increased effect sizes relative to RAW are in bold.

Mean (SD) of Cohen's d measures over all scanners				
	Cortical Thickness		**Volume**	
Method	Entorhinal	Inferior Temporal	Hippocampus	Amygdala
RAW	0.46 (0.14)	0.66 (0.38)	0.76 (0.20)	0.74 (0.26)
WS	0.50 (0.11)	0.62 (0.39)	0.29 (0.13)	0.40 (0.11)
RAVEL	0.49 (0.18)	0.62 (0.34)	0.26 (0.11)	0.30 (0.22)
CALAMITI	0.50 (0.51)	0.57 (0.65)	0.31 (0.13)	0.28 (0.10)
MISPEL	**0.71** (0.09)	**0.73** (0.14)	0.73 (0.20)	**0.80** (0.17)

effect sizes for all of the biomarkers, except for hippocampus. MISPEL resulted in a minor decrease in Cohen's d of hippocampus. On the other hand, WS, RAVEL, and CALAMITI resulted in major decreases for hippocampus and amygdala, a minor decrease for inferior temporal, and a minor increase for entorhinal. In summary, we observed that MISPEL succeeded in preserving our biological signal of interest and outperformed other methods in this respect.

4.6 Discussion

In this study, we presented MISPEL, a supervised deep harmonization technique for removing scanner effects from images of multiple scanners, while preserving their biological and anatomical information. Unlike other supervised or unsupervised methods, MISPEL is a multi-scanner method mapping images to a scanner *middle-ground* space in which images are harmonized. We evaluated MISPEL against commonly used intensity normalization and harmonization methods (White Stripe, RAVEL, and CALAMITI) using a set of evaluation criteria including image similarity, GM-WM tissue contrast, tissue volumes and segmentation similarity, and biological similarity in a dataset of matched T1 MR images acquired from 4 different 3T scanners. We also investigated whether these methods could preserve or even enhance the SVD group differences as a biological signal of interest. We found that (1) scanner effects appear in our dataset as dissimilarity in image appearance/contrast, GM-WM contrast, tissue type volumetric and segmentation distributions, and distributions of regional measures of AD; (2) White Stripe normalized images, but did not achieve harmonization; (3) RAVEL and CALAMITI achieved harmonization to some extent; and (4) MISPEL outperformed all other methods in harmonization.

Based on the evaluated harmonization metrics, we observed that images of GE were more similar to those of Philips and images of SiemensP showed more similarity to SiemensT's. We also observed that scanner effects appeared mainly as the dissimilarity between pairs of GE or Philips and SiemensP or SiemensT. We observed that removing intensity unit effects using White Stripe successfully normalized images (Appendix B.1) and resulted in

improved image similarity, but did not majorly enhance other metrics we used for evaluating harmonization. The relative failure to harmonize may be due to the fact that WS is an intensity normalization method, which does not account for scanner information. We also observed that WS increased the variability of image-derived measures across subjects. Such behavior was observed in bias and variance metrics for GM and WM volumes, as well as biomarkers of AD. This was expected as WS is an individual-level method. This property of WS makes the normalization of any new unseen image more convenient but may also result in inconsistent normalization across images. WS also decreased the effect size for volumetric biomarkers of AD, when SVD group differences were studied. In fact, scaling and centering the intensity distributions does not necessarily remove scanner effects; on the contrary, over-matching distributions could result in the removal of other sources of variability that could be of interest (Fortin et al., 2016). These results show that scanner effects are not addressed solely through intensity normalization and a more comprehensive harmonization method is necessary.

RAVEL is an unsupervised normalization and harmonization framework that could extract components of scanner effects for each of the subjects as inter-subject variability across their CSF area. Our results show that RAVEL achieved harmonization to some extent relative to White Stripe, but was outperformed by MISPEL. RAVEL increased the similarity of images in their appearance/contrast, GM-WM contrast, and tissue type volumes and segmentation overlap when the SPM framework was used. However, RAVEL could not achieve harmonization for FSL-derived GM and WM volumes. Moreover, it deteriorated the bias and variance for biomarkers of AD, except for volumes of the Amygdala. RAVEL also did not preserve the SVD group differences when *volumetric* biomarkers were investigated. These relative failures could be due to several reasons. First, RAVEL uses neither the information of scanners nor the matched data during its harmonization process. Second, RAVEL is prone to remove some biological variability across subjects, if such variability is not accounted for in RAVEL modeling. RAVEL also showed large variability and inconsistent harmonization across subjects, especially for FSL-derived volumes. Such results have been also reported in (Torbati et al., 2021) when RAVEL was used for harmonizing paired images of GE 1.5T and Siemens 3T scanners and FreeSurfer was used. Similar results were observed for WS. Thus,

such behavior of RAVEL could be due to using WS in its normalization step.

For a fair comparison with CALAMITI, we used it in a supervised manner by applying it to our inter-scanner paired dataset instead of inter-modality paired data as discussed in (Zuo et al., 2021b). Results showed that CALAMITI achieved harmonization to some extent relative to White Stripe. However, it did not perform better than RAVEL and was outperformed by MISPEL. CALAMITI improved similarity of images in their appearance/contrast, GM-WM contrast, and tissue type volumes and segmentation overlap when the SPM framework was used. CALAMITI did not show consistent harmonization for FSL-derived volumes. It resulted in both increased and decreased biases for these measures. Moreover, CALAMITI showed large variability and inconsistent harmonization across subjects for both FSL- and SPM-derived volumes. This method did not achieve harmonization for AD biomarkers either. It deteriorated the bias and variance for the entorhinal and hippocampus measures. It also deteriorated the SVD group differences for all biomarkers, except for the entorhinal. These failures in harmonization could be due to CALAMITI's harmonization approach. CALAMITI encodes paired images into their mutual scanner-invariant anatomical components, and their individual contrast and scanner-variant components. For harmonizing an image, it synthesizes the harmonized image by using its anatomical component and the target scanner's contrast component. Such methodology is prone to losing some anatomical information of images, if it could not segregate the anatomical and contrast components properly. Similar harmonization failures were observed for CALAMITI in (Zuo et al., 2021b) when image-derived summary measures were investigated.

MISPEL outperformed White Stripe, RAVEL, and CALAMITI based on all harmonization evaluation criteria. MISPEL mapped images to a middle-ground harmonized space, in which matched images were made more similar in contrast by removing scanner effects. For our data, GE and Philips images were more similar to those of SiemensP and SiemensT, in terms of GM-WM contrast and tissue type volumetric distributions. It should be noted that no directed mapping or a *target* scanner was selected for MISPEL harmonization, and MISPEL does not require a selected *target*. In fact, MISPEL naturally finds this middle-ground space. GE and Philips images were made more similar to SiemensP and SiemensT, with relatively minimal change made to SiemensP and SiemensT by MISPEL, likely due to

SiemensP and SiemensT images being most similar and therefore biasing the middle-ground space found by MISPEL. For this scenario of data, not requiring a target scanner could be regarded as an advantage for MISPEL over other deep-learning based harmonization frameworks. Other widely used statistical harmonization methods, including WS, RAVEL, and ComBat, also do not require a target scanner. However, harmonizing to a middle-ground rather than a specified target could be problematic in other scenarios, such as if the data were collected on a majority of lower-quality scanners. This may bias MISPEL to learn a lower-quality middle-ground space for harmonizing images and degrade the quality of images from more advanced scanners. In such cases, MISPEL could potentially be modified to map images to a target scanner.

Additionally, our volumetric and segmentation evaluations demonstrate that MISPEL image-based harmonization significantly enhances downstream image analysis results across different frameworks. Notably, improvements were observed across both FSL and SPM segmentation platforms, which have previously shown considerable discrepancies in segmentation outcomes, even among healthy volunteers (Tudorascu et al., 2016). MISPEL also showed success in harmonization of biomarkers of AD and enhancing the SVD group differences when these biomarkers were used. The improved performance of MISPEL compared to RAVEL and CALAMITI could be due to the design choices for MISPEL. First, U-Net (Ronneberger et al., 2015) units were used as the encoder-decoder units in MISPEL. The U-Net could preserve the structure of brain by transferring the information of images from encoder layers to the decoder layers. Second, the loss functions for MISPEL were selected cautiously to tackle the contrast discrepancy within paired images and preserve their anatomy. Even so, MISPEL is far from perfect. We observed that MISPEL showed better harmonization for cortical thickness biomarkers relative to volumetric measures. MISPEL improved volumetric bias and variance for the amygdala and preserved the SVD group differences in amygdala volumes, but MISPEL also slightly reduced the SVD group differences in hippocampal volumes.

One possible reason for the suboptimal performance of MISPEL in hippocampal-derived harmonization metrics could be related to its 2D network. Such a network may result in slice-to-slice inconsistency for harmonized images. To evaluate this, we assessed slice-to-slice

consistency measures for each of the RAW and MISPEL-harmonized images. We collected an array of SSIM measures between each adjacent axial slice of each image. We then paired each of the harmonized images with their equivalent RAW image and calculated the correlation between SSIM consistency measures of images of each pair. A harmonization method that preserves the slice-to-slice consistency of RAW images should have a statistically significant correlation near 1 over all pairs. We conducted this experiment for slices of each brain orientation separately and observed 0.994 (ranges: [0.969, 0.999]), 0.992 (ranges: [0.962, 0.999]), and 0.991 (ranges: [0.973, 0.998]) mean of correlations across subjects for axial, sagittal, and coronal slices, respectively. These high correlations demonstrate that slice-to-slice inconsistency is not a significant concern for MISPEL when trained exclusively on axial slices. As such, further investigation is necessary to optimize MISPEL for multi-scanner studies where focal regional volumes are of interest.

5.0 ESPA: An unsupervised harmonization framework via Enhanced Structure Preserving Augmentation

In this section, we introduce ESPA, an unsupervised image harmonization framework, alongside an extensive array of experiments to assess its efficacy. Our hypothesis suggests that *harmonization for scanners can be acquired through mappings to their scanner-middle-ground domain via a framework that concurrently simulates matched data for the scanners using appearance-based augmentation methods and learns the corresponding mappings from this simulated data.* To test this hypothesis, we developed ESPA with the following objectives: (1) accommodating multiple scanners (more than two), (2) mitigating over-correction issues during harmonization, (3) preserving the structural (anatomical) integrity of brain images, and (4) enhancing the robustness of harmonization methods, particularly supervised ones.

ESPA expands upon MISPEL with a notable adjustment: instead of depending on matched data, we employ two novel structure-preserving augmentation methods—tissue-type contrast augmentation and GAN-based residual augmentation—to simulate matched data, limiting modifications to image appearance and contrast. Further details on ESPA and our augmentation methods are provided in Section 5.1. In configuring augmentations, we utilize two sets of data referred to as *source* and *multi-scanner* data, elaborated along with their preprocessing steps in Section 5.2. ESPA is compared against state-of-the-art (SOTA) supervised and unsupervised methods, detailed in Section 5.3, with the model training procedures for ESPA and the competing methods outlined in Section 5.4. Furthermore, we assess ESPA and the competing methods using the evaluation criteria detailed in Section 5.5. Finally, the results of our comparisons are presented in Section 5.6, followed by a discussion of these findings in Section 5.7.

5.1 ESPA

ESPA, crafted as an unsupervised task-agnostic image-harmonization framework, adapts images to a scanner-middle-ground domain. In achieving this adaptation, we employed our harmonization technique, MISPEL (detailed in Section 4.2), albeit with a significant modification. Instead of relying on matched data, we introduce a novel approach wherein we simultaneously generate and utilize simulated matched images during the training of MISPEL. This approach equips MISPEL with simulated data of substantial size, offering a solution to the model robustness challenge in harmonization, particularly concerning supervised harmonization methods. The simulated matched images are generated using our novel structure-preserving augmentation methods. Initially, each augmentation method is individually configured to adapt images of the source scanner to those of the target scanners. During this adaptation process, over-correction can be mitigated through population matching strategies between source and and target domains. Subsequently, the configured augmentations are integrated into two distinct ESPA frameworks to generate simulated matched images and learn two harmonization frameworks. Integrating appearance-based augmentation methods into MISPEL as a structure-preserving framework considers the brain's structure more effectively during harmonization. The ESPA harmonization process is depicted in Figure 22.

In Section 5.1.1, we delve into the notations and assumptions used in training ESPA. Following that, Section 5.1.2 provides a brief overview of MISPEL. Finally, the configuration steps for our augmentation methods are elucidated in Sections 5.1.3 and 5.1.4.

5.1.1 Notations and Assumptions

We refer to the data targeted for harmonization as *multi-scanner data*. This data contains images of M scanners. We consider another set of data with images of one arbitrary scanner and refer to it as *source data*. Throughout this chapter, we refer to scanners of the source and multi-scanner data as the *source scanner* and *target scanners*, respectively. Source data, $X_{n=1:N}$, consists of N scans with a total of $X_{n=1:N}^{k=1:K}$ slices where K is the number of axial slices of an image. Our goal is to design augmentation methods to adapt images of the *source data*

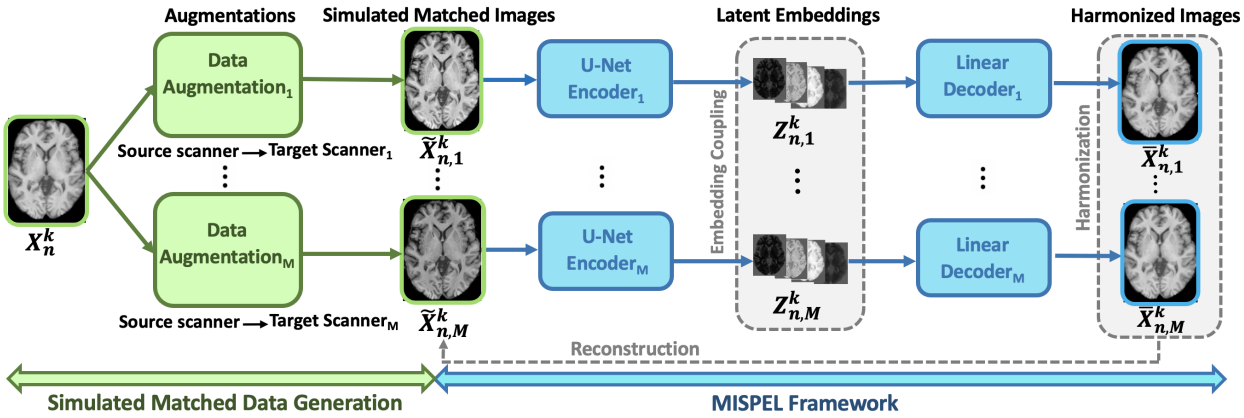


Figure 22: **Illustration of ESPA.** For each of the $m = 1 : M$ target scanners, a data augmentation model, *Data Augmentation*$_m$, is first configured. All M configured augmentation models are then individually applied to each of the $k = 1 : K$ axial slices of the $n = 1 : N$ source scans, $X^{k=1:K}_{n=1:N}$, to generate the simulated matched data: $\tilde{X}^{k=1:K}_{n=1:N,m=1:M}$. For each of the $m = 1 : M$ scanners, a unit of an encoder (*U-Net Encoder*$_m$) and a linear decoder (*Linear Decoder*$_m$) is considered in MISPEL. All M units of encoder-decoder are applied to each of their corresponding $k = 1 : K$ axial slices of the $n = 1 : N$ simulated matched data: $\tilde{X}^{k=1:K}_{n=1:N,m=1:M}$. *U-Net Encoder*$_m$ maps its input slice $\tilde{X}^k_{n,m}$ to its corresponding latent embeddings $Z^k_{n,m}$. The corresponding *Decoder*$_m$ maps its input embeddings to the output: $\bar{X}^k_{n,m}$. During these mappings, which incorporate loss functions such as Embedding Coupling, Harmonization, and Reconstruction, harmonized matched images are generated as $\bar{X}^{k=1:K}_{n=1:N,m=1:M}$. These images are mapped to a scanner-middle-ground domain, where $\bar{X}^k_{n,1} \approx \cdots \approx \bar{X}^k_{n,m} \approx \cdots \approx \bar{X}^k_{n,M}$ (for all n images and k axial slices).

to those of the M scanners in the *multi-scanner data*. These methods are applied to the slices in *source data*, $X^{k=1:K}_{n=1:N}$, to generate our desired simulated matched data. We refer to this simulated set as $\tilde{X}^{k=1:K}_{n=1:N,m=1:M}$ in which $\tilde{X}^k_{n,m=1:M}$ are simulated matched slices for X^k_n. ESPA uses the augmented methods to sample variations of such data during its training to learn generating their harmonized images $\bar{X}^{k=1:K}_{n=1:N,m=1:M}$. Harmonized images $\bar{X}^{k=1:K}_{n=1:N,m=1:M}$ are images mapped to a scanner-middle-ground domain, where $\bar{X}^k_{n,1} \approx \cdots \approx \bar{X}^k_{n,m} \approx \cdots \approx \bar{X}^k_{n,M}$ (for all n images and k axial slices).

5.1.2 MISPEL

MISPEL (Figure 22) specializes in harmonizing images of scanners for which matched data is available. It uses encoder-decoder units for each scanner, translating input slices into latent embeddings using a U-Net (Ronneberger et al., 2015) as encoder. Linear decoding combines these embeddings to reconstruct the input image, ensuring similarity between embeddings and reconstructed images across scanners for harmonization. Additionally, MISPEL maintains brain structure by ensuring similarity between reconstructed and original images. These operations were respectively referred to as Embedding Coupling, Harmonization, and Reconstruction in Figures 12 and 22. For further information about MISPEL, see Section 4.2.

5.1.3 Tissue-type contrast augmentation

Scanner effects can impact brain tissue contrast, as demonstrated by Meyer et al. (2019). To address this issue, we utilize a three-step augmentation approach aimed at adjusting tissue contrast from a source scanner to a target scanner while maintaining brain structure. This method builds upon previous work by Meyer et al. (2021) initially developed for brain segmentation purposes. It is important to note that this augmentation technique adapts images from the source scanner to a *single* target scanner. Therefore, for each of the M target scanners, we should configure M distinct tissue-type contrast augmentation methods.

Step 1: Estimating the distributions of tissue types. In this step, we apply the Gaussian Mixture Model (Reynolds et al., 2009) to the intensity values of the brain voxels in source image X_n. The intensity set $\{v_1, \ldots, v_P\}$, where P is the total number of brain voxels in the image, is modeled as

$$p(v_p) = \Sigma_{t=1}^{T=3} \pi_t \mathcal{N}(v_p | \mu_t, \sigma_t^2), \tag{9}$$

with t denoting brain tissue types, $\mathcal{N}(\mu_t, \sigma_t^2)$ representing a Gaussian distribution with mean μ_t and variance σ_t^2, and π_t as the mixing coefficient. Using Bayes' rule, we compute the

probability of each class label C as

$$p(C = t|v) = \frac{\pi_t \mathcal{N}(v|\mu_t, \sigma_t^2)}{\Sigma_{t'=1}^{3} \pi_{t'} \mathcal{N}(v|\mu_{t'}, \sigma_{t'}^2)} \tag{10}$$

Step 2: Modifying tissue type distributions. We adapt this method step to align images from our *source* data with those of a *single* target scanner in our *multi-scanner* data. To achieve this, we adjust the tissue type distributions of images in the source data by sampling from estimated normal distributions capturing directional differences in tissue-type parameters between the source data images and those of the target scanner. The desired modified tissue type distribution of the source image X_n is determined as

$$\mathcal{N}(\mu_t^{'}, \sigma_t^{'2}) = (\mu_t - q_{\mu_t}, \sigma_t^2 - q_{\sigma_t^2}), \tag{11}$$

where q_{μ_t} and $q_{\sigma_t^2}$ are adaptation terms sampled from the determined distributions of differences. To calculate these terms, we first compute directional differences of distribution parameters from all source images to all target images. One instance of such differences is denoted as

$$D_{\mu_t}^f = \mu_{n,t} - \mu_{l,t}, \tag{12}$$

and

$$D_{\sigma_t^2}^f = \sigma_{n,t}^2 - \sigma_{l,t}^2, \tag{13}$$

where $(\mu_{n,t}, \sigma_{n,t}^2)$ and $(\mu_{l,t}, \sigma_{l,t}^2)$ are distribution parameter pairs for images n and l in the source data and target scanner, respectively, and f denotes the f^{th} difference in a total of F calculated differences. Finally, we compute the adaptation terms as

$$q_{\mu_t} = Mean(D_{\mu_t}^{f=1:F}) + r_\mu, \tag{14}$$

and

$$q_{\sigma_t^2} = Mean(D_{\sigma_t^2}^{f=1:F}) + r_\sigma, \tag{15}$$

where r_μ and r_σ are sampled from the uniform distributions $U(-Std(D_{\mu_t}^{f=1:F}), Std(D_{\mu_t}^{f=1:F}))$ and $U(-Std(D_{\sigma_t^2}^{f=1:F}), Std(D_{\sigma_t^2}^{f=1:F}))$, and $Mean(\cdot)$ and $Std(\cdot)$ denote the mean and standard deviation functions.

Step 3: Reconstructing augmented image. For reconstructing the augmented image for source image X_n, we adapt each voxel value v_p for each tissue type as

$$v'_{p,t} = \mu'_t + d_{pt}\sigma'_t, \tag{16}$$

where $d_{pt} = (v_p - \mu_t)/\sigma_t$ maintains the original relative distance of voxel intensity from the mean intensity of tissue type t in the images, preserving structural brain information. We then compute the augmented intensity voxel v'_p as

$$v'_p = \Sigma_{t=1}^{T=3} p(C = t|v_p) v'_{p,t} \tag{17}$$

After calculating all v'_p for $p \in \{1, \ldots, P\}$, we obtain the adapted image of X_n aligned with the tissue-type distribution of the *single* target scanner.

5.1.4 GAN-based residual augmentation

Scanner effects can be more intricate than tissue-type modifications and can vary across brain regions. Thus, we develop a GAN-based augmentation method to generate and sample scanner effects as images (*residuals*) added to the original images. By limiting scanner effects as additive components to images, we consider brain structure during augmentation. For this purpose, we introduce Residual StarGAN, which performs image-to-image translation between all pairs of our scanner domains (source and target scanners) using a single generator and discriminator pair (Figure 23(a)). We elaborated on the interactions between these two modules within a GAN network for harmonization under the topic of *Adaptation of data to a target scanner domain*, covered in Section 2.6.1.1. Residual StarGAN is a modification of StarGAN (Choi et al., 2018), where we replace the generator with a *Residual Generator*, and include noise as input to this generator for sampling. The Residual Generator comprises the StarGAN Generator followed by the *Additive Module* (Figure 23(b)). The StarGAN Generator generates the residuals to be added to the image in the Additive Module for domain adaptation. The process of adapting an image from the source scanner to the domain of a target scanner in Residual StarGAN is depicted in Figure 23(b). These steps mirror those outlined in StarGAN, with details provided in (Choi et al., 2018). We utilize the trained Residual Generator as our residual augmentation method which is used for adapting images of the source scanner to any of the target scanners.

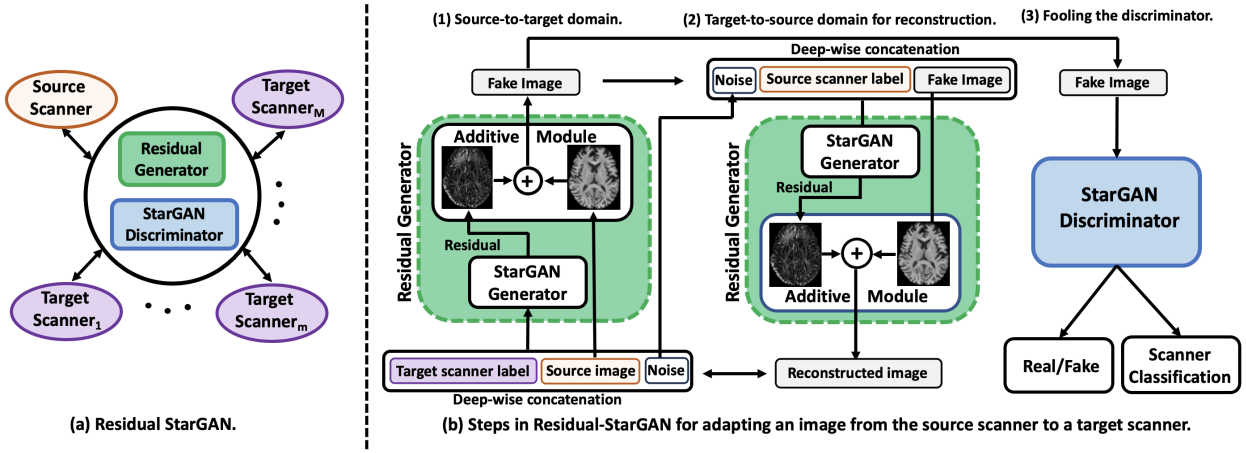


Figure 23: **Illustration of Residual StarGAN.** (a) Overview of Residual StarGAN, comprising two modules: a StarGAN Discriminator and a Residual Generator. These modules facilitate domain adaptation between all pairs of scanner domains (source and target scanners). (b) Steps in Residual StarGAN for scanner domain adaptation, illustrated for adapting an image from the source to one target scanner. (1) The Residual Generator receives the source image (referred to as *real* image), target scanner label, and random noise as input. It generates the residual, which is added to the source image to produce the adapted image to the target scanner domain (referred to as *fake* image). The input source image is concatenated with spatially replicated target scanner labels and random noise before being processed by the Residual Generator. (2) Residual Generator takes the fake image, source scanner label, and the same random noise generated in (1) as input. It then tries to reconstruct the real (source) image from the fake (adapted) image given the source scanner label. (3) Residual Generator tries to fool the StarGAN Discriminator. It tries to generate fake images that are not only indistinguishable from real images, but also classifiable as images of the target scanner by the StarGAN Discriminator.

5.2 Data

5.2.1 Study populations and image acquisition

We opted for *source* data comprising 192 T1-w images obtained from a 3T Siemens Trio scanner within the OAISIS-3 dataset (LaMontagne et al., 2019). Additionally, our *multi-scanner* data is the matched dataset detailed in Section 4.1.1. This dataset encompasses 3T T1-w images across four scanners: General Electrics (GE), Philips, Siemens Prisma (SiemensP), and Siemens Trio (SiemensT). In ESPA, we treated this data as unmatched for the *multi-scanner* data, utilizing its matched aspect for our evaluation. Participants in the matched dataset had a median age of 72 years (with a range of 51-78 years), with 44% being male and all being cognitively unimpaired, except for 10 individuals showing a high degree of small vessel disease (SVD). We specifically selected the source data from the OAISIS-3 dataset to align with the demographics of the matched data. We refer to this process of population matching as our approach to mitigating over-correction during augmentation configuration. Detailed scanner specifications for both the source and multi-scanner data are provided in Table 7. For further insights into the matched data, refer to Section 4.1.1.

Table 7: Scanner specifications for source and multi-scanner data

	Multi-scanner (matched) data				**Source data**
Scanner Name	**GE**	**Philips**	**SiemensP**	**SiemensT**	–
Manufacturer	General Electrics	Philips	Siemens	Siemens	Siemens
Scanner Hardware	DISCOVERY-MR750w 3T	Achieva-dStream 3T	Prisma-fit 3T	TrioTim 3T	TrioTim 3T
Resolution (mm)	$1.0 \times 1.0 \times 0.5$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$
TE (ms)	3.7	1.66	1.64	1.64	3.16
TR (ms)	9500	2530	2530	2530	2400

5.2.2 Image preprocessing

We conducted preprocessing on both the source and multi-scanner datasets, following a pipeline similar to that outlined in Section 4.1.2. This pipeline includes non-linear registration (Avants et al., 2008) to a T1-w atlas (Oishi et al., 2009), N4 bias correction (Tustison

et al., 2010), skull-stripping via brain masking, and image scaling by dividing images by their mean intensity. The preprocessed matched data is referred to as *RAW*. For additional information on the preprocessing pipeline, please see Section 4.1.2.

5.3 Competing methods

In our pursuit of SOTA methods, we employed MISPEL and modified CALAMITI as supervised techniques. MISPEL leverages matched scanner data to learn harmonization, enabling the adaptation (harmonization) of unseen unmatched scanner images to a scanner-middle-ground space. The modified CALAMITI approach utilizes matched data to initially disentangle images into their scanner-invariant anatomical component and scanner-variant contrast component. Subsequently, it learns reconstructing images using these two components. During harmonization, all images are mapped to the contrast of a target image using the scanner-invariant component of the image and the scanner-variant component of the target image. Detailed explanations of these methods can be found in Sections 4.2 and 4.3.3, respectively.

We also utilized the unsupervised SOTA method known as style-transfer harmonization (Style-Trans) for our study, as introduced by Liu et al. (2023). This method considers the style of images as the scanner-variant component and operates within a content-style disentangled cycle translation framework for harmonization. Within this framework, two individual encoders are employed to initially disentangle images into their scanner-variant and -invariant components. Subsequently, a cycle-consistent GAN framework is utilized to learn style-based transformations between images from different scanners. Specifically, the generators are modified to learn mappings from the two components of images as input, rather than directly from the images themselves. This modification enables the generator to take the scanner-invariant component of an image and the scanner-variant component of a target image, allowing it to learn to map the image to the style of the target image for harmonization.

5.4 Training setup

In this chapter, our setup was devised to implement harmonization on the preprocessed multi-scanner data (RAW), enabling evaluation on matched data. For MISPEL and CALAMITI, we conducted 6-fold cross-validation on the matched data, partitioning subjects into 12/3/3 for train/validation/test sets. Subsequently, we combined the harmonized test sets into one harmonized set for our evaluation step. Further details on the training strategies for these methods can be found in Sections 4.2.3 and 4.3.3, respectively.

Liu et al. (2023) trained the Style-Trans model on T1-weighted images of 718 subjects from 5 diverse sites, including the UK Biobank (Sudlow et al., 2015), Parkinson's Progression Markers Initiative (Marek et al., 2018), Alzheimer Disease Neuroimaging Initiative (Mueller et al., 2005), Adolescent Brain Cognitive Development (Jernigan et al., 2018), and International Consortium for Brain Mapping (Mazziotta et al., 2001). They made their model publicly available along with their target image in a Github repository[1]. Leveraging their trained model, we applied it to the RAW data to achieve harmonization by aligning it with their target image style.

ESPA follows the same cross-validation approach for RAW: partitioning subjects into 12/3/3 for train/validation/test sets. We treated this folded RAW data as unmatched for the multi-scanner data in ESPA. ESPA has a 3-steps training and harmonization process for which we split source data into 12/20/20 and 100/20/20 splits of train/validation/test sets for its first two steps, respectively. **(1)** In the initial step, two augmentation methods are configured individually to adapt images of 12 source images to 12 training images within each of the 4 scanners in each fold of the multi-scanner dataset. **(2)** The second step involves training ESPA by creating variations of simulated matched data, individually applying augmentations to 100 source training images. Separate sets of models for folds, referred to as ESPA_{TC} and ESPA_{Res}, are trained for tissue-type contrast augmentation and GAN-based residual augmentation, respectively. **(3)** In the final step, these models are individually applied to images of 3 test subjects in their corresponding folds in the multi-scanner data. Harmonized test sets are combined across folds as harmonized version of RAW for evaluation.

[1] `https://github.com/USC-IGC/style_transfer_harmonization`

Model training and hyper-parameter tuning for all methods were conducted on NVIDIA RTX5000. These procedures for MISPEL and modified CALAMITI were detailed in Sections 4.2.3 and 4.3.3, respectively. We omitted this step for Style-Trans, as we utilized the authors' pre-trained model for our experiments. In ESPA_{TC}, the tissue-type contrast augmentation does not have any hyper-parameter for tuning. Regarding ESPA_{Res}, we trained Residual StarGAN using the hyper-parameters recommended in (Choi et al., 2018). For training the MISPEL framework in either ESPA_{TC} or ESPA_{Res}, we selected hyper-parameters from the ranges outlined in Section 4.2.3. Additionally, we set T_1, T_2, and batch size parameters to 100, 400, and 32, respectively. T_1 and T_2 are training iterations for Step 1 and Step 2 in MISPEL, respectively.

5.5 Data analysis

Our evaluation analyses are five-fold: (1) validation on domain adaptation in augmentation methods, (2) validation on brain structure preservation in augmentation methods, (3) validation on augmentation removal in ESPA, (4) validation on harmonization, and (5) ablation study. Several of our evaluation metrics necessitate pairwise image-to-image comparisons, requiring us to consider all potential combinations of *scanner pairs*: {(GE, Philips), (GE, SiemensP), (GE, SiemensT), (Philips, SiemensP), (Philips, SiemensT), and (SiemensP, SiemensT)}. Our results are presented for the pool of all these image pairs. When we mention target scanners, we are referring to GE, Philips, SiemensP, and SiemensT, in that specific sequence.

First, we assessed whether our augmentation methods achieved domain adaptation. This involved testing the performance of a scanner classifier on augmented images generated by each method. Domain adaptation occurs when the classifier accurately predicts the target scanner to which the augmented image has adapted. To conduct this analysis, we trained, optimized, and evaluated 6 scanner classifiers separately for the 6 folds in the multi-scanner data. For this procedure, we used images from both the corresponding fold of the multi-scanner data and the source data designated to the first step of ESPA. To ensure balanced

classification, we utilized images from only 3 subjects from each of the source validation and test sets. For the classifier, we employed the discriminator network as introduced in (Bashyam et al., 2022). For the augmented images, we applied the configured augmentations of each fold to the 20 source test images designated for the first step in ESPA.

Second, we assessed whether brain structure is being preserved in augmentation methods. We expect augmentations to result in minor structural modifications which can be evaluated by visualizing augmentations and augmented images as well as estimating the structural similarity of augmented images to their original images. To accomplish this, we initially generated the augmented images by separately applying the configured augmentations of each fold to the 20 source test images designated for the first step in ESPA. For visual assessment, we presented one original source slice alongside its applied augmentation and resulting augmented image, repeating this process for all three brain orientations. Additionally, we quantified the structural similarity between augmented and original images using structural similarity index measure (SSIM). We reported these values as the cross-fold mean and cross-fold standard deviation (SD) of SSIM scores for all of the augmented images adapted to target scanners. We anticipate that high SSIM values will corroborate the preservation of brain structure.

Third, we evaluated the capability of our cross-validated ESPA_{TC} and ESPA_{Res} models to eliminate the simulated augmentations applied to the images. Removal aimed to decrease dissimilarity between augmented images of a source image. Mean Average Error (MAE) and Jensen–Shannon Divergence (JD) metrics were used to assess this dissimilarity, reported as mean$\pm$SD for images of all scanner pairs and folds. Initially, we augmented images of the source test set designated for the second step in ESPA, using the configured augmentation of each fold. Then, the trained ESPA models of each fold were applied to their corresponding augmented image sets to obtain augmented-free (harmonized) images.

Fourth, we estimated scanner effects and evaluated harmonization on RAW and harmonized RAW, respectively. A harmonization method aims to eliminate scanner effects while maintaining the biological variables of interest present in the data. In our particular matched dataset, the matched images are presumed to be biologically identical but exhibit differences due to variations in scanners. Consequently, the scanner effects can be inferred by analyz-

ing the dissimilarity among the matched images, and the removal of these effects can be viewed as enhancing their similarity. Our investigation into the dissimilarity and similarity of matched images involved three evaluation criteria: (1) image similarity, (2) GM-WM contrast similarity, and (3) biological similarity. Additionally, we identified SVD as the clinical signal of interest in our dataset and explored whether we could retain or potentially improve the differentiation between SVD groups following harmonization.

Scanner effects may manifest as variations in contrast across images captured by different scanners (Dewey et al., 2019, 2020; Liu et al., 2021). Specifically, these differences might present as variations in tissue-specific contrast within images (Meyer et al., 2019). To address this, we conducted an evaluation of scanner effects and harmonization effectiveness. We employed an **image similarity** metric to gauge the visual consistency of images across scanners and a **GM-WM contrast similarity** metric to assess the similarity in tissue contrast across images. In evaluating *image similarity*, we examined the visual quality of matched *slices* across all methods. Additionally, we measured the structural similarity of paired images for *all* scanner pairs. For this, we measured the mean $\pm$ SD of SSIMs for all of these images. An effective harmonization method is anticipated to enhance both the visual and structural resemblance of paired images.

Furthermore, we explored the *GM-WM contrast similarity* of the images. The contrast of GM and WM can significantly impact the performance of segmentation techniques, with heightened contrast expected to yield more precise segmentation results. This contrast can be assessed by estimating the separability of the histograms of GM and WM voxels within an image. This separability was quantified through voxel classification as GM or WM, and represented by the area under the receiver operating characteristic (AUROC) curve. An AUROC value of 100% indicates perfect classification (complete separation of histograms), while a value of 50% suggests random classification (complete overlap of histograms). To compute AUROC, we initially labeled the GM and WM voxels using the tissue mask provided in the EveTemplate package (Oishi et al., 2009). Subsequently, we classified these voxels based on intensity thresholds selected from the range of intensity values of GM and WM voxels. Finally, we generated the AUROC curve for the image using the outcomes of each classification. A harmonization method is anticipated to enhance the GM-WM contrast

similarity across scanners resulting in comparable AUROCs.

We further examined the **biological similarity** of images by analyzing biomarkers of Alzheimer's disease (AD). Specifically, we assessed bias (mean of cross-scanner absolute differences) and variance (root mean square deviation, RMSD) for these biomarkers. To calculate bias, we determined the cross-scanner absolute differences for all scanner pairs and reported their mean (SD). For variance, we computed the mean of cross-scanner RMSDs for all scanner pairs. These metrics were evaluated for all 6 methods and all AD biomarkers. The biomarkers we investigated included cortical thickness measures of the entorhinal, inferior temporal, middle temporal, inferior parietal, and fusiform cortices, as well as volume measures of the amygdala, hippocampus, entorhinal, middle temporal, and inferior temporal regions. These summary measures represent the total measures over both hemispheres and were derived using FreeSurfer 7.1.1 (FS) (Fischl, 2012). These regions have been previously identified as highly relevant to AD (Schwarz et al., 2016). We omitted the skull stripping and bias correction steps in the FS processing pipeline, as RAW images had already undergone skull-stripping and N4 bias correction during image preprocessing (Section 5.2.2). A harmonization method is expected to decrease both bias and variance.

We concluded our harmonization evaluation by examining whether each harmonization method **preserved or potentially enhanced a biological/clinical signal of interest** in our matched data. Our chosen clinical signal of interest is SVD, and we explored the effect size between two groups representing low and high SVD in our dataset. To conduct this analysis, we computed Cohen's d effect sizes for the two SVD groups across each of our FreeSurfer (FS)-derived biomarkers of AD individually. For each biomarker, we calculated the effect sizes of the scanners separately and reported the mean (SD) of these values across scanners. It is expected that a harmonization method does not diminish the effect sizes of the groups following harmonization.

Fifth and last, we conducted an ablation study to demonstrate the effectiveness of our augmentation methods. Specifically, we trained ESPA with random contrast and brightness augmentations as described in (Chaitanya et al., 2021). These techniques involve contrast transformation $(X_n - E(X_n)) * b + E(X_n)$ and brightness transformation $X_n + c$, where b and c are uniformly sampled from $[0.8, 1.2]$ and $[-0.1, 0.1]$, respectively, with $E(X_n)$ representing

the mean brain intensity values in the original scan X_n. To evaluate the effectiveness of these augmentation methods, we *replicated* our experiments to validate augmentation removal and assessed structural image similarity for harmonization using combination of these two augmentation techniques. We aimed to ensure that the ESPA employing these augmentation methods resulted in augmentation removal but no harmonization, thereby highlighting the effectiveness of our proposed methods: tissue-type contrast and GAN-based residual augmentations.

For evaluating augmentation removal, we assumed that a decrease in dissimilarity among augmented images of an original image would indicate successful removal. To measure this dissimilarity, we calculated the MAE and JD measures for both the augmented and augmented-free (ESPA-harmonized) images of the source test set designated for the second step in ESPA. The mean±SD of each of these measures was reported as an estimate of the desired dissimilarity. Additionally, for harmonization evaluation, we assessed image similarity before and after harmonization for RAW data. For image similarity assessment, we computed the mean±SD of SSIM scores for paired images of all scanner pairs in each of the two sets: RAW and ESPA-harmonized. By ESPA-harmonized, we mean the ESPA model trained with simulated data generated through random contrast and brightness augmentations. An increase in SSIM after harmonization indicates successful harmonization.

5.6 Results

5.6.1 Validation on domain adaptation in augmentation methods

Initially, we assessed the ability of our trained classifiers to predict the scanner origin in the cross-folded multi-scanner data. The cross-fold accuracy of the classifiers for all scanners averaged $78.6 \pm 1.9\%$, with cross-fold accuracies of [85.2 ± 5.7, 81.0 ± 2.5, 73.9 ± 4.1, 74.4 ± 4.2]% for the target scanners: [GE, Philips, SiemensP, SiemensT], respectively. Subsequently, we generated a set of augmented images for evaluation purposes. This involved individually applying the configured augmentations of each fold to our 20 source test images.

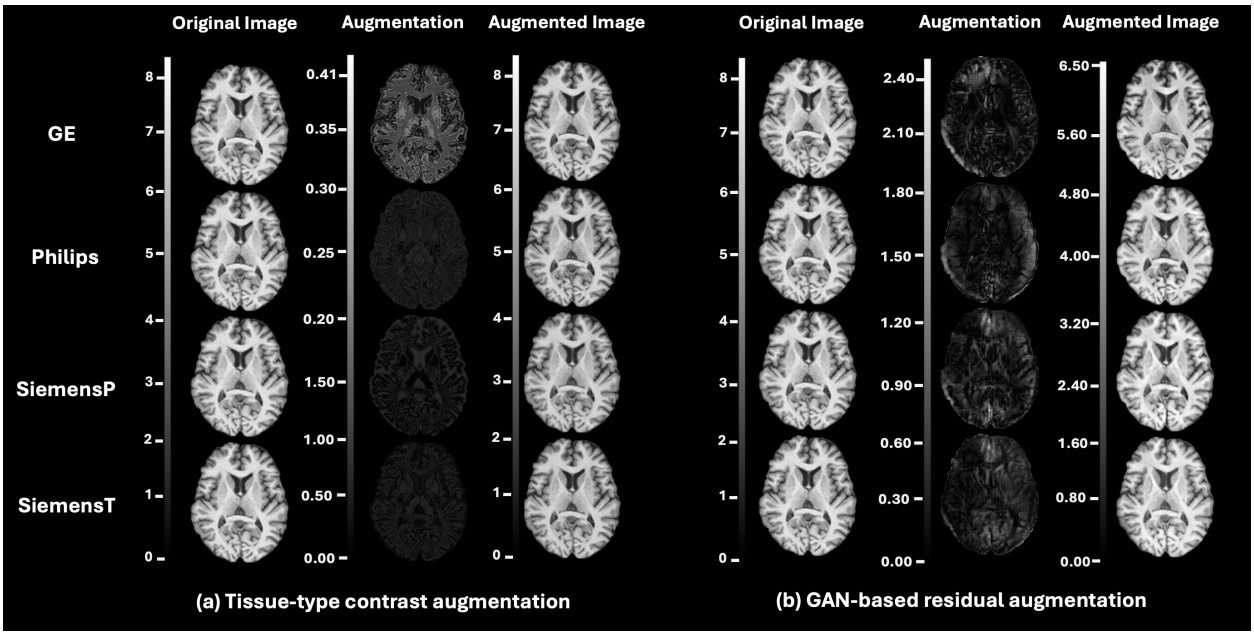


Figure 24: Visualizing augmentations and augmented images for one axial slice. A single axial slice, designated as the Original Image, was replicated across all scanners.

The classifiers were then utilized to classify the augmented images, resulting in an average cross-fold accuracy of $88.2 \pm 3.9\%$ for tissue-type contrast augmentation, with cross-fold accuracies of $[86.1 \pm 6.9,\ 86.5 \pm 7.2,\ 91.3 \pm 2.8,\ 89.0 \pm 3.4]\%$ for the target scanners. Likewise, for residual augmentation, the cross-fold accuracy averaged $88.1 \pm 3.9\%$, with accuracies of $[86.4 \pm 5.7,\ 84.3 \pm 2.5,\ 92.3 \pm 4.1,\ 89.4 \pm 4.2]\%$ for the target scanners. Despite the classifiers' limited performance attributable to the small training image size, these findings underscore the effectiveness of our augmentation techniques in facilitating domain adaptation.

5.6.2 Validation on brain structure preservation in augmentation methods

In Figures 24, 25, and 26, we depicted both augmentations and augmented images, each featuring axial, sagittal, and coronal slices, respectively. Augmentations are images that may contain both negative and positive intensity values. Our depiction presents the absolute values of these intensities. Upon observation, we noted no apparent structural changes in the resulting augmented images, a conclusion supported by our SSIM comparisons between

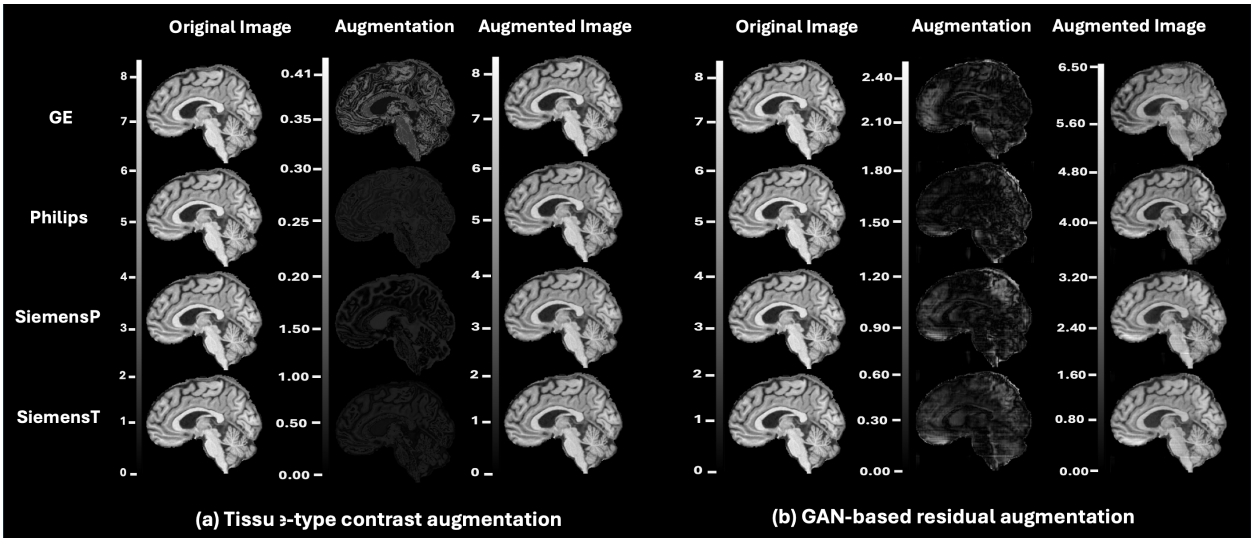


Figure 25: Visualizing augmentations and augmented images for one sagittal slice. A single sagittal slice, designated as the Original Image, was replicated across all scanners.

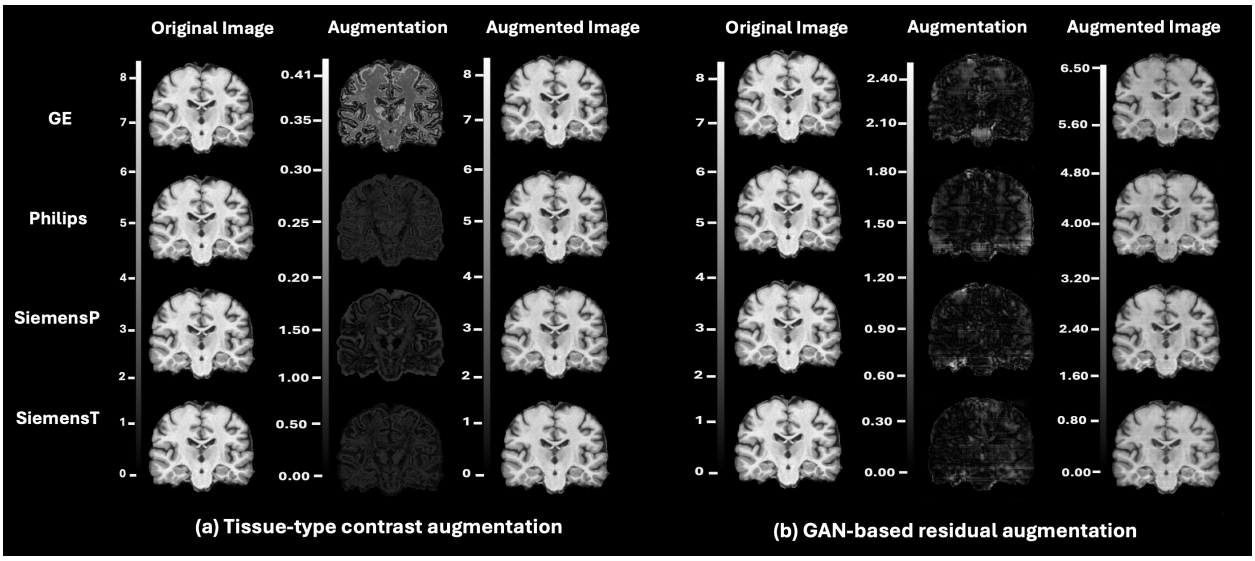


Figure 26: Visualizing augmentations and augmented images for one coronal slice. A single coronal slice, designated as the Original Image, was replicated across all scanners.

augmented and original images. For the tissue-type contrast augmentation, our analysis revealed a mean±SD of mean SSIMs across folds as $\mathbf{0.999} \pm 0.0008$. Across folds, the

mean±SD of mean SSIMs for the target scanners: [GE, Philips, SiemensP, SiemensT] were $[0.9980 \pm 0.0009,\ 0.9993 \pm 0.0010,\ 0.9993 \pm 0.0008,\ 0.9994 \pm 0.0006]$. These statistics for SD of SSIMs are $\mathbf{0.0009} \pm 0.0001$: $[0.0013 \pm 0.0006,\ 0.0004 \pm 0.0001,\ 0.0003 \pm 0.0001,\ 0.0002 \pm 0.0001]$. Regarding the GAN-based residual augmentation, such cross-fold statistics for mean and SD of SSIMs were observed as $\mathbf{0.941} \pm 0.009$: $[0.941 \pm 0.010,\ 0.938 \pm 0.014,\ 0.943 \pm 0.006,\ 0.942 \pm 0.009]$ and $\mathbf{0.006} \pm 0.0023$: $[0.003 \pm 0.0005,\ 0.004 \pm 0.0006,\ 0.005 \pm 0.0005,\ 0.005 \pm 0.0010]$, respectively.

The resulting mean and SD of SSIMs were notably smaller than cross-scanner SSIMs observed for scanner pairs in RAW, measuring **0.81±0.05**. As a result, this indicates that our augmentation methods led to minimal structural modifications, even smaller than the structural modifications caused by scanner effects observed in the matched data.

Table 8: Validation on augmentation removal in ESPA_{TC} for scanner pairs.

Mean (SD) of MAE						
Method	GE-Philips	GE-SiemensP	GE-SiemensT	Philips-SiemensP	Philips-SiemensT	SiemensP-SiemensT
Augmented	0.085 (0.046)	0.073 (0.025)	0.076 (0.024)	0.065 (0.055)	0.068 (0.038)	0.061 (0.022)
Harmonized	0.035 (0.008)*	0.028 (0.005)*	0.028 (0.005)*	0.032 (0.011)*	0.032 (0.010)*	0.024 (0.005)*
Mean (SD) of JD						
Method	GE-Philips	GE-SiemensP	GE-SiemensT	Philips-SiemensP	Philips-SiemensT	SiemensP-SiemensT
Augmented	0.036 (0.035)	0.038 (0.037)	0.036 (0.037)	0.010 (0.014)	0.010 (0.012)	0.009 (0.007)
Harmonized	0.018 (0.019)*	0.021 (0.018)*	0.018 (0.019)*	0.006 (0.005)*	0.004 (0.005)*	0.006 (0.005)

Statistically significantly changes (paired t-test, $p < 0.05$) were marked *.

5.6.3 Validation on augmentation removal in ESPA

For ESPA_{TC}, the mean ± SD of MAE and JD decreased from 0.071 ± 0.037 to $\mathbf{0.030} \pm \mathbf{0.009}$, and 0.023 ± 0.030 to $\mathbf{0.012} \pm \mathbf{0.015}$ before and after harmonization, respectively. Similarly, for the ESPA_{Res}, MAE values decreased from 0.403 ± 0.107 to $\mathbf{0.135} \pm \mathbf{0.023}$, and JD values decreased from 0.012 ± 0.009 to $\mathbf{0.007} \pm \mathbf{0.006}$. All changes were statistically significant (paired t-test, $p < 0.05$), indicating successful augmentation removal from images. These statistics were also reported for ESPA_{TC} and ESPA_{Res} in Tables 8 and 9, respectively.

5.6.4 Validation on harmonization

5.6.4.1 Image similarity

Scanner effects manifest *visually*, as depicted in Figure 27, through noticeable cross-scanner contrast discrepancies in RAW slices, which were mitigated post-harmonization across all methods. Nevertheless, certain drawbacks were observed with specific methods. For instance, CALAMITI disrupted image contrast, while MISPEL and Style-Trans introduced slight smoothing effects. On the contrary, ESPA_{TC} and ESPA_{Res} exhibited superior visual quality. The structural similarity, as measured by the mean±SD of SSIMs for all images of scanner pairs, exhibited a notable increase across all methods. Specifically, it rose significantly from 0.81 ± 0.05 for RAW to 0.83 ± 0.04, $\mathbf{0.87 \pm 0.04}$, $\mathbf{0.87 \pm 0.05}$, 0.83 ± 0.05, and 0.85 ± 0.05 for CALAMITI, MISPEL, Style-Trans, ESPA_{TC}, and ESPA_{Res}, respectively. Notably, MISPEL and Style-Trans demonstrated the most substantial increase. All observed enhancements compared to RAW were statistically significant ($p < 0.05$), as confirmed by paired t-tests.

Table 9: Validation on augmentation removal in ESPA_{Res} for scanner pairs.

Mean (SD) of MAE						
Method	GE-Philips	GE-SiemensP	GE-SiemensT	Philips-SiemensP	Philips-SiemensT	SiemensP-SiemensT
Augmented	0.446 (0.087)	0.450 (0.095)	0.502 (0.097)	0.344 (0.068)	0.389 (0.067)	0.286 (0.061)
Harmonized	0.150 (0.019)*	0.148 (0.019)*	0.148 (0.019)*	0.127 (0.016)*	0.127 (0.016)*	0.110 (0.016)*
Mean (SD) of JD						
Method	GE-Philips	GE-SiemensP	GE-SiemensT	Philips-SiemensP	Philips-SiemensT	SiemensP-SiemensT
Augmented	0.007 (0.004)	0.022 (0.004)	0.025 (0.005)	0.007 (0.001)	0.009 (0.001)	0.002 (0.000)
Harmonized	0.008 (0.008)	0.008 (0.007)*	0.007 (0.006)*	0.010 (0.007)	0.004 (0.005)*	0.006 (0.004)*

Statistically significantly changes (paired t-test, $p < 0.05$) were marked *.

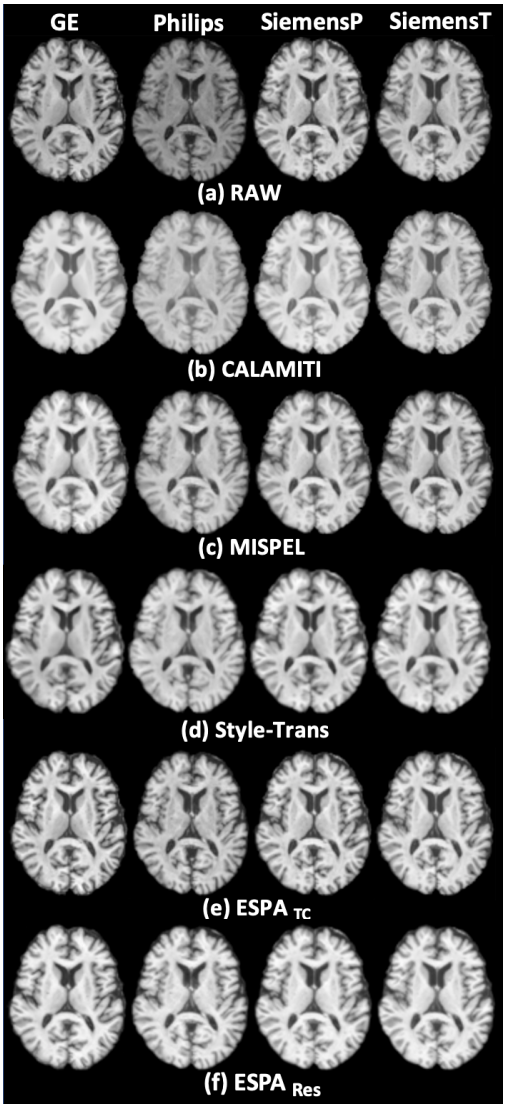


Figure 27: Visual assessment of scanner effects and harmonization across matched images.

5.6.4.2 GM-WM contrast similarity

We measured the GM-WM contrast of an image using AUROC values, which represent the separation of histograms of GM and WM voxel intensities. A high AUROC indicates higher contrast, with 100% being the highest achievable value. A harmonization method is expected to achieve two objectives: (1) not deteriorate the AUROC of images, and (2) make the AUROC of matched images similar. Figure 28 displays bar plots indicating the mean AUROC of images for each scanner, before and after harmonization. All harmonization

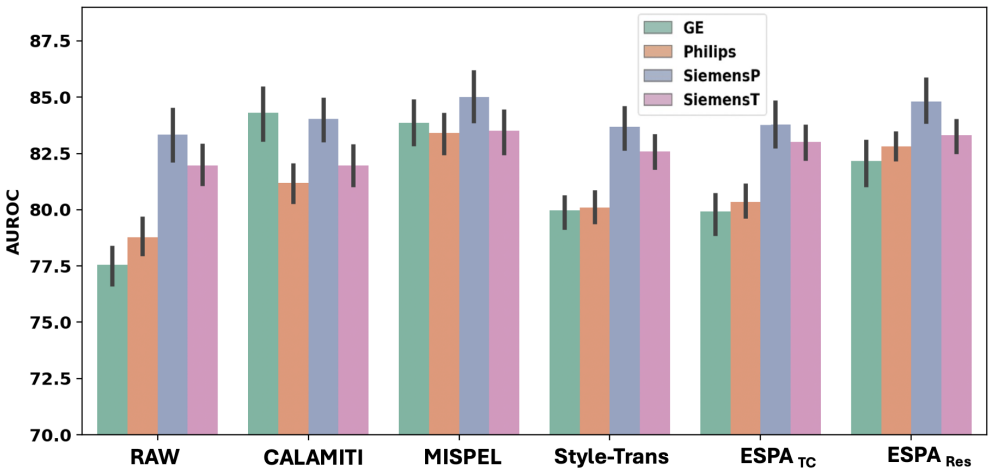


Figure 28: **GM-WM contrast bar plots.** Each bar indicates the mean AUROC of images of each scanner, with error bars denoting the standard deviation for each method.

methods improved the AUROC values of the scanners and made them similar across scanners. MISPEL demonstrated the best performance in this regard, followed by ESPA_{Res}.

5.6.4.3 Biological similarity

We quantified bias for biomarkers of AD as mean±SD of cross-scanner differences for all scanner pairs, as outlined in Table 10. Interestingly, our findings indicated that an increase in SSIM (seen in Section 5.6.4.1) does not necessarily equate to improved harmonization for biological variables. For instance, CALAMITI worsened bias for 9 biomarkers, while it achieved improved SSIM. Moreover, Style-Trans was surpassed by ESPA_{TC} and ESPA_{Res} in assessing biological similarity, despite achieving the highest SSIM. Specifically, ESPA_{TC} and ESPA_{Res} outperformed MISPEL and Style-Trans, reducing bias for **7** and **9** biomarkers, respectively, compared to 5 for MISPEL and Style-Trans. Notably, ESPA_{TC} demonstrated the largest reductions in bias for **4** cases, while MISPEL, ESPA_{Res}, and Style-Trans exhibited decreases for 3, 2, and none, respectively. Paired t-tests ($p < 0.05$) revealed significant decreases in 5, 5, and 4 cases for ESPA_{TC}, ESPA_{Res}, and MISPEL, respectively, whereas none were significant for Style-Trans. These statistics were also depicted in Figure 30.

We also visualized the root mean square deviation (RMSD) of the biomarkers using bar plots in Figure 29. The interpretation of RMSD results closely resembles that of bias, as depicted in Figure 30. Specifically, CALAMITI worsened RMSD for 9 biomarkers. Additionally, ESPA_{TC} and ESPA_{Res} outperformed MISPEL and Style-Trans with each achieving 6 decreases compared to 4 for each of MISPEL and Style-Trans. Notably, ESPA_{Res} achieved the most largest decreases with 4 cases, whereas MISPEL, Style-Trans, and ESPA_{TC} achieved decreases in 2, 1, and 2 cases, respectively.

5.6.4.4 Analysis on biological variables of interest

We also explored whether harmonization preserved or enhanced biological signals by comparing Cohen's d effect sizes between low and high SVD groups for each AD biomarker. Cohen's d was computed separately for each scanner, and the mean±SD across scanners was reported in Table 10. Harmonization success was determined by an increase in Cohen's d compared to RAW. Our findings in Figure 30 revealed CALAMITI and Style-Trans's failure, possibly due to deteriorated contrast and over-correction. ESPA_{TC} and ESPA_{Res} each surpassed MISPEL with **7** increases, while yielding the best Cohen's d values for 2 and **6** biomarkers, respectively, compared to MISPEL's 5 increases and 0 best increases.

5.6.5 Ablation study

To demonstrate the efficacy of our augmentation methods (tissue-type contrast and GAN-based residual augmentations), we trained ESPA with the combination of random contrast and brightness augmentation (Chaitanya et al., 2021). We repeated experiments for validation on augmentation removal, confirming reduction in MAE and JD from 0.164 ± 0.088 and 0.028 ± 0.025 to $\mathbf{0.099 \pm 0.037}$ and $\mathbf{0.013 \pm 0.016}$, respectively. All changes were statistically significant (paired t-test, $p < 0.05$), indicating successful augmentation removal from images. However, our structural similarity analysis for harmonization yielded SSIMs similar to that of RAW, suggesting no significant modification in harmonized images across scanners and thus no harmonization.

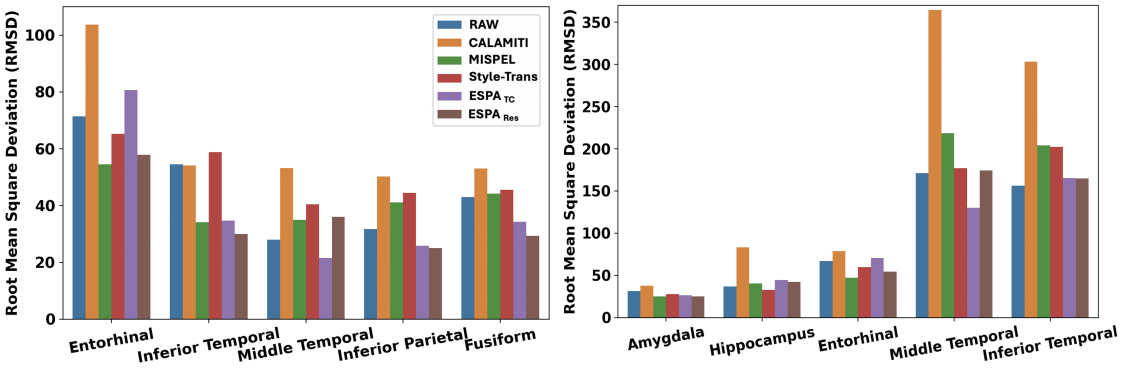
5.7 Discussion

In this study, we introduced ESPA, an unsupervised image harmonization framework designed to learn mappings from images of scanners to their scanner-middle-ground domain for

Table 10: **Bias and Cohen's d values for biomarkers of AD.** Mean (SD) of cross-scanner absolute differences were calculated as bias for biomarkers of all scanner pairs. The distributions with the smallest bias are in bold. Also, the distributions that showed statistically significant t-statistics when compared to RAW were marked by *. Mean (SD) of Cohen's d values were calculated over all scanners for the biomarkers. Largest increased effect sizes relative to RAW are in bold.

		Mean $_{\pm SD}$ of absolute differences over all scanner pairs (Bias)									
		Cortical Thickness (mm)					**Volume (cm^3)**				
Method		Entorhinal	Inferior Temp.[1]	Middle Temp.	Inferior Parietal	Fusiform	Amygdala	Hippo.[2]	Entorhinal	Middle Temp.	Inferior Temp.
	RAW	$0.62_{\pm 0.4}$	$0.46_{\pm 0.4}$	$0.22_{\pm 0.2}$	$0.25_{\pm 0.2}$	$0.38_{\pm 0.3}$	$0.25_{\pm 0.2}$	$0.30_{\pm 0.2}$	$0.56_{\pm 0.4}$	$1.48_{\pm 1.1}$	$1.19_{\pm 1.1}$
Sup.[3]	CALAMITI	$0.87_{\pm 0.6}$*	$0.45_{\pm 0.3}$	$0.40_{\pm 0.4}$*	$0.37_{\pm 0.4}$*	$0.43_{\pm 0.3}$	$0.30_{\pm 0.3}$	$0.71_{\pm 0.5}$*	$0.62_{\pm 0.5}$	$2.78_{\pm 2.6}$*	$2.31_{\pm 2.1}$*
	MISPEL	$\mathbf{0.46}_{\pm 0.3}$*	$\mathbf{0.25}_{\pm 0.2}$*	$0.25_{\pm 0.3}$	$0.34_{\pm 0.3}$*	$0.36_{\pm 0.3}$	$0.19_{\pm 0.2}$*	$0.32_{\pm 0.3}$	$\mathbf{0.37}_{\pm 0.3}$*	$1.55_{\pm 1.7}$	$1.50_{\pm 1.6}$
Unsup.[4]	Style-Trans [5]	$0.54_{\pm 0.4}$	$0.51_{\pm 0.4}$	$0.31_{\pm 0.3}$*	$0.37_{\pm 0.3}$*	$0.39_{\pm 0.3}$	$0.22_{\pm 0.2}$	$\mathbf{0.25}_{\pm 0.2}$	$0.47_{\pm 0.4}$	$1.36_{\pm 1.2}$	$1.62_{\pm 1.3}$*
	$ESPA_{TC}$	$0.66_{\pm 0.5}$	$0.28_{\pm 0.2}$*	$\mathbf{0.17}_{\pm 0.1}$*	$\mathbf{0.21}_{\pm 0.2}$	$\mathbf{0.29}_{\pm 0.2}$*	$0.21_{\pm 0.2}$*	$0.34_{\pm 0.3}$	$0.53_{\pm 0.5}$	$\mathbf{1.10}_{\pm 0.8}$*	$1.21_{\pm 1.2}$
	$ESPA_{Res}$	$0.57_{\pm 0.4}$	$0.28_{\pm 0.2}$*	$0.26_{\pm 0.2}$	$0.22_{\pm 0.2}$	$0.32_{\pm 0.2}$*	$\mathbf{0.18}_{\pm 0.2}$*	$0.29_{\pm 0.2}$	$0.45_{\pm 0.3}$*	$1.31_{\pm 0.9}$	$\mathbf{1.04}_{\pm 0.8}$*
		Mean $_{\pm SD}$ of Cohen's d measures over all scanner pairs									
		Cortical Thickness (mm)					**Volume (cm^3)**				
Method		Entorhinal	Inferior Temp.	Middle Temp.	Inferior Parietal	Fusiform	Amygdala	Hippo.	Entorhinal	Middle Temp.	Inferior Temp.
	RAW	$0.46_{\pm 0.1}$	$0.66_{\pm 0.4}$	$1.14_{\pm 0.2}$	$0.97_{\pm 0.3}$	$\mathbf{0.74}_{\pm 0.2}$	$0.74_{\pm 0.3}$	$\mathbf{0.76}_{\pm 0.2}$	$0.51_{\pm 0.1}$	$0.61_{\pm 0.4}$	$0.81_{\pm 0.3}$
Sup.	CALAMITI	$0.50_{\pm 0.5}$	$0.57_{\pm 0.6}$	$0.31_{\pm 0.2}$	$0.40_{\pm 0.3}$	$0.54_{\pm 0.5}$	$0.28_{\pm 0.1}$	$0.31_{\pm 0.1}$	$-0.31_{\pm 0.6}$	$0.14_{\pm 0.1}$	$0.18_{\pm 0.2}$
	MISPEL	$0.71_{\pm 0.1}$	$0.73_{\pm 0.1}$	$1.21_{\pm 0.3}$	$0.92_{\pm 0.1}$	$0.57_{\pm 0.2}$	$0.80_{\pm 0.2}$	$0.73_{\pm 0.2}$	$0.17_{\pm 0.2}$	$0.63_{\pm 0.4}$	$0.69_{\pm 0.2}$
Unsup.	Style-Trans	$0.21_{\pm 0.4}$	$0.53_{\pm 0.5}$	$0.75_{\pm 0.2}$	$0.51_{\pm 0.4}$	$0.33_{\pm 0.4}$	$0.56_{\pm 0.1}$	$0.28_{\pm 0.2}$	$0.16_{\pm 0.2}$	$0.60_{\pm 0.3}$	$0.54_{\pm 0.1}$
	$ESPA_{TC}$	$0.61_{\pm 0.3}$	$0.84_{\pm 0.2}$	$1.12_{\pm 0.2}$	$1.20_{\pm 0.1}$	$0.67_{\pm 0.1}$	$0.87_{\pm 0.2}$	$0.50_{\pm 0.2}$	$\mathbf{0.54}_{\pm 0.4}$	$0.78_{\pm 0.4}$	$\mathbf{1.26}_{\pm 0.1}$
	$ESPA_{Res}$	$\mathbf{0.73}_{\pm 0.4}$	$\mathbf{1.00}_{\pm 0.2}$	$\mathbf{1.23}_{\pm 0.3}$	$\mathbf{1.30}_{\pm 0.2}$	$0.69_{\pm 0.2}$	$\mathbf{1.00}_{\pm 0.3}$	$0.55_{\pm 0.1}$	$0.24_{\pm 0.1}$	$\mathbf{0.92}_{\pm 0.3}$	$1.08_{\pm 0.2}$

[1]Temp.: Temporal, [2]Hippo.: Hippocampus, [3]Sup.: Supervised, [4]Unsup.: Unsupervised, [5]Style-Trans: Style Transfer Harmonization

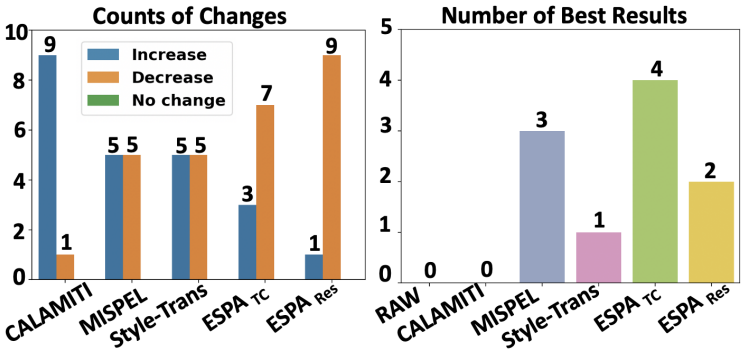


(a) **Cortical Thickness.** (b) **Volume.**

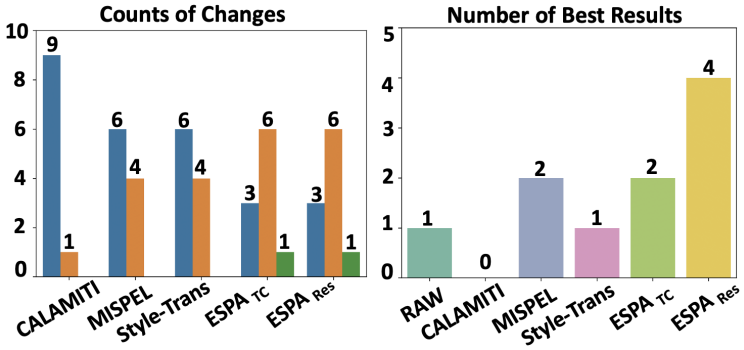
Figure 29: **Root-mean-square deviation (RMSD) bar plots for biomarkers of AD.** Each bar indicates the mean RMSD of paired measures of all scanner pairs for each of the methods

harmonization. Within ESPA, we employ MISPEL as our chosen harmonization framework, utilizing simulated matched data generated through our two novel augmentation methods. These methods aim to adapt images of an arbitrary *source* scanner to those of *target* scanners, while emphasizing on preserving their brain structure. The simulated matched data generated by our augmentation methods is flexible in size, accommodating source dataset of any desired scale. This flexibility enhances the robustness of our harmonization model, which could be a challenge for supervised harmonization methods. Additionally, leveraging MISPEL alongside our structure-preserving augmentation techniques ensures that the anatomical structure of brains is appropriately accounted for in ESPA. Furthermore, the potential for over-correction is mitigated through population-matching between the source and target scanner populations during the simulation-based domain adaptation process.

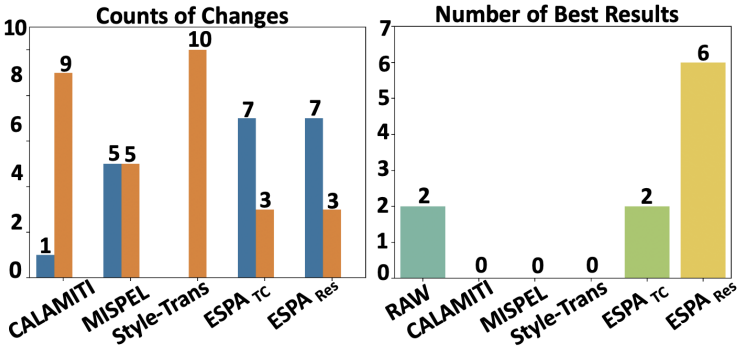
We devised two sets of augmentation methods: (1) tissue-type contrast augmentation, and (2) GAN-based residual augmentation. The first method operates under the assumption that scanner effects manifest as disparities in brain tissue type distributions, while the second method posits that these effects can vary across brain regions and can be simulated as additive augmentations to images. To assess the effectiveness of these augmentations,



(a) **Bias.**



(b) **Variance (RMSD).**



(c) **Cohen's d.**

Figure 30: Statistics on bias, variance, and Cohen's d for biomarkers of AD.

we validated their performance in domain adaptation, focusing on their capability to preserve brain structure throughout this process. Additionally, we assessed their performance in harmonization. For the harmonization evaluation, we constructed two instances of the

ESPA harmonization framework, ESPA_{TC} and ESPA_{Res}, each utilizing one of these methods for data augmentation. We then compared the performance of these frameworks against SOTA harmonization methods, including MISPEL and a modified version of CALAMITI as supervised methods, as well as Style-Trans as an unsupervised method.

Our findings indicate that: (1) scanner effects manifest in our dataset as disparities in image appearance/contrast, GM-WM contrast, and distributions of regional biomarkers of AD; (2) CALAMITI and Style-Trans achieved harmonization to some degree, with Style-Trans demonstrating superior performance compared to CALAMITI; (3) MISPEL achieved harmonization and surpassed Style-Trans in downstream tasks, particularly when analyzing effect sizes for SVD groups; (4) both ESPA_{TC} and ESPA_{Res} achieved structure-preserving domain adaptation and outperforming MISPEL in harmonization; and (5) ESPA_{Res} is preferred over ESPA_{TC} due to its superior harmonization capabilities and ability to simulate region-wise scanner effects.

Our initial analyses on the validity of augmentation demonstrated the effectiveness of both augmentation methods in adapting images from the source scanner to the target scanners. We assessed this capability by evaluating the prediction accuracy of classifiers trained to classify the scanners to which the images were adapted. Notably, both augmentation methods yielded similar accuracies, which may not adequately capture the full extent of their adaptation performance. It's possible that our assessment is hindered by the limited classification ability of the classifiers, stemming from the small sample size on which they were trained. To explore this further, additional investigation using unmatched data of larger size is warranted.

Our analysis of the brain structure preservation capability of the augmentations revealed their success in this aspect. Visualization of the harmonized images demonstrated that the structural integrity of the brain was maintained. Notably, this preservation was observed across all three brain orientations, highlighting a distinctive characteristic of ESPA. This framework was trained solely on 2D axial slices, yet it effectively preserved brain structure in orientations other than that it was specifically trained on. This stands in contrast to many existing harmonization methods, such as (Dewey et al., 2019; Liu et al., 2023; Zuo et al., 2021b), where brain modifications in orientations different from the trained one were prob-

lematic. To address this issue, these methods proposed training three separate harmonization models per orientation and averaging their outputs to obtain the final harmonized images. However, this approach significantly increases the time and cost complexity of harmonization methods.

Based on the evaluated harmonization metrics, we observed that images of GE were more similar to those of Philips and images of SiemensP showed more similarity to SiemensT's. We also observed that scanner effects appeared mainly as the dissimilarity between pairs of GE or Philips and SiemensP or SiemensT. The harmonization results revealed that while CALAMITI achieved some degree of harmonization, it was outperformed by Style-Trans. Specifically, CALAMITI improved the similarity of images in terms of appearance/contrast and GM-WM contrast. However, it failed to harmonize AD biomarkers, resulting in deteriorated bias and variance for the majority (9) of the biomarkers. Similarly, it worsened the differences observed in the SVD group for the same number of biomarkers. These shortcomings in harmonization could be attributed to CALAMITI's failure to adequately disentangle scanner-variant components from the images. Another contributing factor could be over-correction in this method. The low Cohen's d values may be indicative of over-correction of contrast in the images, potentially causing SVD signatures such as white matter hyperintensity (WMH) to be over-corrected as well. Further investigation into these hypotheses is warranted.

Further harmonization analyses revealed that Style-Trans achieved superior harmonization compared to CALAMITI, despite not utilizing matched data. Notably, Style-Trans demonstrated acceptable harmonization performance, comparable to MISPEL, across all harmonization metrics except for Cohen's d for SVD groups. Style-Trans deteriorated this metric for all AD biomarkers, potentially attributed to over-correction resulting from its style transfer approach. It is possible that WMH as signature of SVD have been corrected to match the style of the target images used by this method. Additional investigation into these hypotheses is necessary.

MISPEL showcased superior harmonization compared to Style-Trans, effectively achieving harmonization across image similarity, GM-WM contrast similarity, and biological similarity. Moreover, it enhanced differentiation among SVD groups using biomarkers of AD.

Despite these notable strengths, MISPEL was surpassed by ESPA_{TC} and ESPA_{Res}, particularly in terms of biological similarity and Cohen's d for SVD groups.

Our ablation study underscored the inadequacy of randomly selected contrast and brightness augmentations in yielding significant harmonization, emphasizing the necessity for more refined augmentation methods, as exemplified by our approach. This highlights the complexity of scanner effects for downstream tasks compared to contrast and appearance-based variability in natural images, as demonstrated by frameworks like SimCLR (Chen et al., 2020b), which successfully addressed natural image classification using randomly tuned appearance-based augmentation methods. However, our ablation study revealed that if harmonization is to be approached as a task-specific method using the SimCLR framework, more sophisticated augmentations, such as our novel ones, are imperative. For such framework, we advocate for GAN-based residual augmentation, as implemented in our ESPA_{Res}, which yielded the best harmonization results among all compared methods and better simulated scanner effects at a region-wise level. Nevertheless, our augmentation methods were not flawless. Both ESPA_{TC} and ESPA_{Res} exacerbated size effects when using the hippocampus as a biomarker of interest. Further investigation into this phenomenon is warranted, alongside exploration of our methods across more brain regions.

6.0 Conclusion and future work

In this dissertation, our main focus is on understanding, illustrating, and addressing scanner effects. We aim to address several current issues related to scanner effects, including the lack of understanding of these effects, the absence of standardized criteria for assessing them and evaluating harmonization, and the limited availability of harmonization methods. To achieve this, we conducted three studies specifically designed for T1-weighted MRIs:

(1) In the first study, we employed matched data as the best available experimental setup to assess scanner effects and evaluate harmonization. Our investigation encompassed scanner effects on images and image-derived measures. Furthermore, we utilized matched data to establish our harmonization evaluation criteria and to evaluate two SOTA harmonization methods at the time.

(2) In the second study, we utilized matched data as labeled data to develop MISPEL, a supervised harmonization method. While supervised harmonization methods, such as MISPEL, are less susceptible to the two current harmonization issues, including over-correction and brain structural modifications resulting from the use of matched data, they do have two shortcomings: not all datasets include additional matched data, and there is a possibility of lack of model robustness due to the small size of matched data.

(3) In the third study, we introduced ESPA as an unsupervised harmonization framework. ESPA is an extension of MISPEL that proposes using simulated matched data instead of actual matched data. By providing simulations of matched data of flexible size, ESPA can overcome issues associated with supervised harmonization. Additionally, it can address over-correction and mitigate brain structure modifications during its simulation process.

Our contributions to MRI image processing can be further enhanced through several avenues of future work. Evaluating methods on larger or varied matched datasets presents an opportunity for improvement, although such data is not currently publicly available. Consequently, new experimental setups utilizing larger unmatched datasets could prove beneficial. Additionally, employing matched data with phantoms has the potential to enhance the reliability of methods and experiments, thereby reducing the image artifacts to solely

scanner effects. While our methods have primarily focused on T1-w MRIs, exploring other image modalities could yield valuable insights. Expanding experiments to encompass more brain regions would offer a more comprehensive assessment of the methods' effectiveness across different anatomical areas. Furthermore, future evaluations could involve more extensive variation of hyper-parameters to potentially optimize results across all aspects of our evaluation criteria. This approach would provide a deeper understanding of the methods' performance and their adaptability to different settings.

6.1 Investigating two methods of cross-scanner technical variability removal in harmonization of image-derived measures

Chapter 3 delved into two harmonization methods: RAVEL, which focuses on normalizing and harmonizing images, and ComBat, designed for harmonizing image-derived measures. To test these methods, we selected top 10 biomarkers of AD from a paired dataset comprising T1-w MRIs of 16 subjects obtained from General Electric 1.5T and Siemens 3T MRI scanners. Our harmonization criteria were tailored using paired data, with metrics evaluating dissimilarity and similarity across paired images and measures. We assessed images for normalization and segmentation accuracy across scanners, while biomarkers were evaluated for bias and variance. Our findings revealed varying degrees of harmonization achieved by RAVEL, ComBat, and the RAVEL-ComBat pipeline. Specifically, RAVEL effectively normalized images, particularly in the GM and CSF regions, while preserving brain anatomy, as evidenced by our experiments on hippocampus segmentation. However, it exhibited low variance, indicating inconsistent harmonization across subjects, a trend also observed in the RAVEL-ComBat pipeline. In contrast, ComBat demonstrated superior harmonization when bias and variance were analyzed. While our results partially support our hypothesis that indicates *removing cross-scanner technical variability from both images and image-derived measures enhances harmonization of AD biomarkers*, we recommend a more consistent approach than RAVEL for use in this setup.

The primary objectives of this chapter were to investigate the scanner effect in paired

data and develop harmonization evaluation criteria, which were further expanded upon in our subsequent two chapters. Alongside these goals, we also aimed to assess the performance of SOTA methods at the time without necessarily intending to enhance them. Nonetheless, it's noteworthy to highlight their advancements, particularly RAVEL, as potential avenues for further research.

- One possible reason for RAVEL's inconsistent behavior across subjects could be its selected normalization strategy, referred to as White Stripe (WS). WS is an individual-level method, which makes the normalization of any new, unseen image more convenient. However, this approach may also result in inconsistent normalization across images. Scaling and centering the intensity distributions does not necessarily remove scanner effects; on the contrary, over-matching distributions could result in the removal of other sources of variability that could be of interest (Fortin et al., 2016). This possibility could be further investigated, and RAVEL could be improved by using a different normalization method.
- Another possible reason for RAVEL's inconsistency could be incorrectly capturing motion artifacts as scanner effects in the CSF area. This idea can be investigated by identifying subjects causing high values of RMSD and examining their images for motion artifacts. If motion artifacts are the issue, the identified subjects should have one image with artifacts and another scan taken by a different scanner without artifacts, resulting in high cross-scanner differences and consequently increased RMSD.
- As the final solution, RAVEL could be replaced by another image harmonization method. ComBat has shown acceptable harmonization in terms of adjusting distributions of biomarkers across scanners. This method could be adapted at the image level for harmonizing images, as initiated in (Chen et al., 2022b). However, this direction may face its own difficulties, as distribution matching must be implemented at the voxel intensity level, and errors in image registration could pose challenges.

6.2 Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning

Chapter 4 introduces a supervised image harmonization method, MISPEL: Multi-scanner Image harmonization via Structure Preserving Embedding Learning. MISPEL leverages pre-collected matched data from scanners to learn how to map their images to a scanner-middle-ground domain, effectively reducing scanner effects by making the images more similar to each other. The mappings can then be used to harmonize unmatched images from the scanners. To acquire these mappings, MISPEL employs encoder-decoder units tailored for each scanner's images, facilitating harmonization across multiple scanners. Each unit encodes images into latent embeddings (in the form of images) and decodes them back into harmonized images. The harmonization process enforces similarity between latent and decoded images across scanners. Throughout the mapping learning, MISPEL preserves brain structure, maintaining the resemblance of harmonized images to their originals. Notably, MISPEL handles over-correction, as it solely accounts for dissimilarity within matched images primarily attributed to scanner effects.

The matched dataset we used consists of T1-w matched images from 18 subjects, acquired across four 3T scanners: General Electric, Philips, Siemens Prisma, and Siemens Trio. Leveraging this dataset, we broadened our harmonization evaluation criteria to encompass (1) image similarity, (2) GM-WM contrast similarity, (3) volumetric and segmentation similarity of tissue types, and (4) biological similarity, delving into biomarkers of AD. Additionally, we singled out small vessel disease (SVD) as a key clinical signal of interest and explored the potential preservation or enhancement of SVD group differences post-harmonization. Our experimental design involved the use of various segmentation platforms to illustrate scanner effects and the effectiveness of harmonization across different platforms.

We compared MISPEL with the SOTA harmonization methods at the time, including RAVEL and CALAMITI. Our results indicated that RAVEL and supervised CALAMITI achieved harmonization to some extent. However, MISPEL outperformed all other methods based on all harmonization evaluation criteria. These findings support our hypothesis that *harmonization can be achieved for scanners within a matched dataset by constructing a model*

that maps matched images from the dataset to a scanner-middle-ground space, where matched images lose scanner effects by becoming similar to each other. There are several aspects by which MISPEL can be extended or explored, including the following:

- Selecting a target scanner to which images are mapped could be challenging and controversial. On the other hand, harmonizing images to a scanner-middle-ground space rather than a specified target scanner could be problematic in scenarios where data were collected primarily using lower-quality scanners. This may bias MISPEL to learn a lower-quality middle-ground space for harmonizing images, potentially degrading the quality of images from more advanced scanners. While this was not the case with our matched dataset, it is a possibility that should be considered. In such cases, MISPEL could be easily modified to map images to a specified target scanner.
- With the current design of MISPEL, a separate encoder-decoder unit is required for each scanner. This could potentially make running MISPEL impossible on GPUs with smaller RAM capacity for larger number of scanners. One possible solution is to use a single encoder-decoder unit for all scanners. MISPEL could be easily modified to such design.
- It has been observed that MISPEL may introduce minor blurriness to images. This blurriness could be attributed to our selection of the loss function, specifically the Embedding Coupling Loss, which aims to make embeddings similar in their characteristics across scanners. We chose to ensure that the embeddings have similar variance, which may be the reason for the blurriness. To mitigate this issue, we can explore alternative loss functions such as Mean Absolute Error (MAE) and Structural Similarity Index Measure (SSIM).
- Even though MISPEL has shown major success in harmonization, there are minor shortcomings as well. For example, it slightly reduced the differences between groups identified through small vessel disease (SVD) in hippocampal volumes. This could be related to its 2D network architecture, which may result in slice-to-slice inconsistency for harmonized images. Although our results showed no signs of such inconsistency, it may be worthwhile to develop a 2.5D version of MISPEL. In a 2.5D MISPEL setting, three separate 2D networks could be trained for each brain orientation, and the final harmonized image could be generated as the average of the three images resulting from these networks.

6.3 ESPA: An unsupervised harmonization framework via Enhanced Structure Preserving Augmentation

Chapter 5 introduces ESPA, an unsupervised harmonization framework that extends MISPEL. ESPA proposes using simulated matched data instead of actual matched data. To achieve this, ESPA employs two novel appearance-based augmentation methods designed to adapt images from an arbitrary source scanner to those of target scanners. Utilizing simulated matched data, ESPA learns how to map images to the target scanners' middle-ground domain. The first augmentation method, tissue-type contrast augmentation, assumes that scanner effects can manifest as tissue-type distribution discrepancies across scanners. The second augmentation method, GAN-based residual augmentation, assumes that augmentation can be simulated as a region-wise additive image during domain adaptation between source and target scanners. This additive image is then added to the original image to make augmented image. ESPA addresses concerns about model robustness due to the small sample size of matched data. It can generate a large simulated matched dataset of any desired size. It also accounts for brain structural changes and over-correction using structure-preserving augmentations and population matching during simulation, respectively.

We utilized the same matched dataset as in Chapter 4 for both simulated matched data generation and evaluation. This data was used in its unmatched version for the simulation. Our evaluation setup consisted of five main components: (1) validation of domain adaptation in augmentation methods, (2) validation of brain structure preservation in augmentation methods, (3) validation of augmentation removal in ESPA, (4) validation of harmonization, and (5) an ablation study. For the harmonization validation, we employed the same analyses as in Chapter 4. We compared the performance of ESPA with SOTA methods of the time, including MISPEL, supervised CALAMITI, and Style-Trans.

Our findings indicated that both ESPA_{TC} and ESPA_{Res}, trained respectively with the two augmentation methods, achieved structure-preserving domain adaptation and outperformed the selected SOTA methods in harmonization. We also concluded that ESPA_{Res} is preferred over ESPA_{TC} due to its superior harmonization capabilities and ability to simulate region-wise scanner effects. These results support our hypothesis that *harmonization for scanners*

can be achieved through mappings to their scanner-middle-ground domain via a framework that concurrently simulates matched data for the scanners using appearance-based augmentation methods and learns the corresponding mappings from this simulated data. There are several aspects by which ESPA can be extended or explored, including the following:

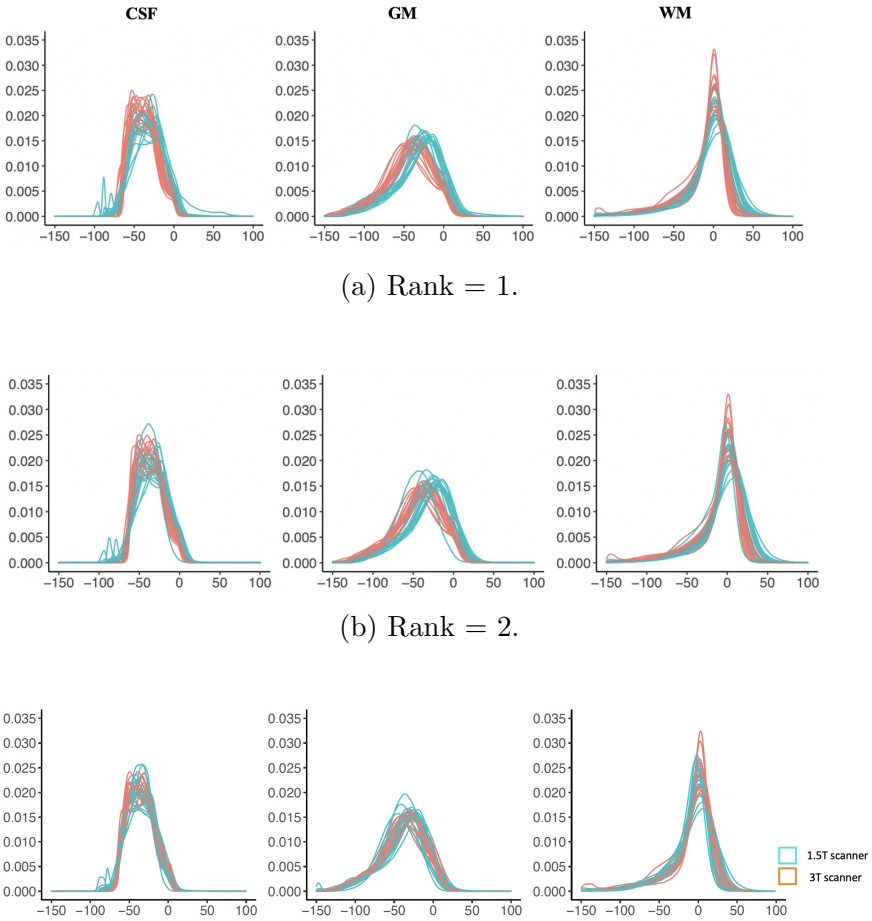
- When adapting images to a scanner-middle-ground domain likewise MISPEL, there's a risk of image quality deterioration, especially when the majority of images for harmonization are of low quality. To address this challenge, MISPEL can be easily modified for domain adaptation towards a target scanner.
- Our analyses on MISPEL revealed no slice-to-slice inconsistency post-harmonization. Furthermore, we did not visually observe such problems in our augmented methods utilizing either of our augmentation techniques. This reduces the necessity of employing 2.5D or 3D networks for ESPA. However, within the neuroimaging or vision domains, networks trained on 3D images are highly preferred, yet they have not been largely developed due to the scarcity of scans. This is why studies resort to using 2D networks to increase the volume of images for training. With our simulation strategy, we can generate simulated matched data of the desired size, enabling the training of 3D networks for harmonization.
- It is worthwhile to assess the effectiveness of pre-training existing harmonization methods using our augmentations for providing simulated data. We hypothesize that incorporating a pre-training step will enhance their harmonization performance. Additionally, our augmentation methods could offer another advantage by being integrated into task-specific harmonization frameworks. SimCLR, originally designed for natural images, augments various sources of variability in such images to generate embeddings, typically utilized in downstream tasks like image classification. SimCLR could potentially be adapted into a task-specific harmonization framework. However, selecting appropriate augmentation methods poses a challenge, as our ablation study revealed the inefficacy of simple contrast and brightness variability as augmentation for harmonization. Hence, it is worthwhile to explore whether our augmentation methods can be utilized to modify SimCLR to generate embeddings to be used in neuroimaging downstream tasks.
- While ESPA has demonstrated significant success in harmonization, it also exhibits minor

shortcomings. Both ESPA_{TC} and ESPA_{Res} showed a decline in the preservation of size effects when utilizing hippocampal volume as a biomarker of interest. Notably, none of the SOTA harmonization methods employed in this dissertation, including RAVEL, MISPEL, supervised CALAMITI, and Style-Trans, were able to improve size effect for volume of this region. We suggest the harmonization of this biomarker as a challenge for future harmonization methodologies.

Appendix A Additional Results from Section 3

A.1 Fitting RAVEL for hyper-parameters

For fitting RAVEL to our data, we explored the effects of the decomposition rank b and the biological variables age and gender on density plots of tissue types: CSF, GM, and WM. Figure 31 contains the plots for which rank was set to either one, two, or three. The results showed that the higher rank, in our case three, gave us greater overlap of the plots, which means better intensity normalization. In the second set of experiments, we controlled models for age and/or gender. The density plots for each of these settings were depicted in Figure 32, in which rank was fixed to three. We observed that while controlling for gender did not change the density plots, age widened them, specifically the plots for the WM and GM. Wider density plots could be an evidence of resulting in images with lower quality/contrast, which was the exact case for our images. Based on these observations, we decided to fix rank to three and control for no biological variables, when we fitted the final model.

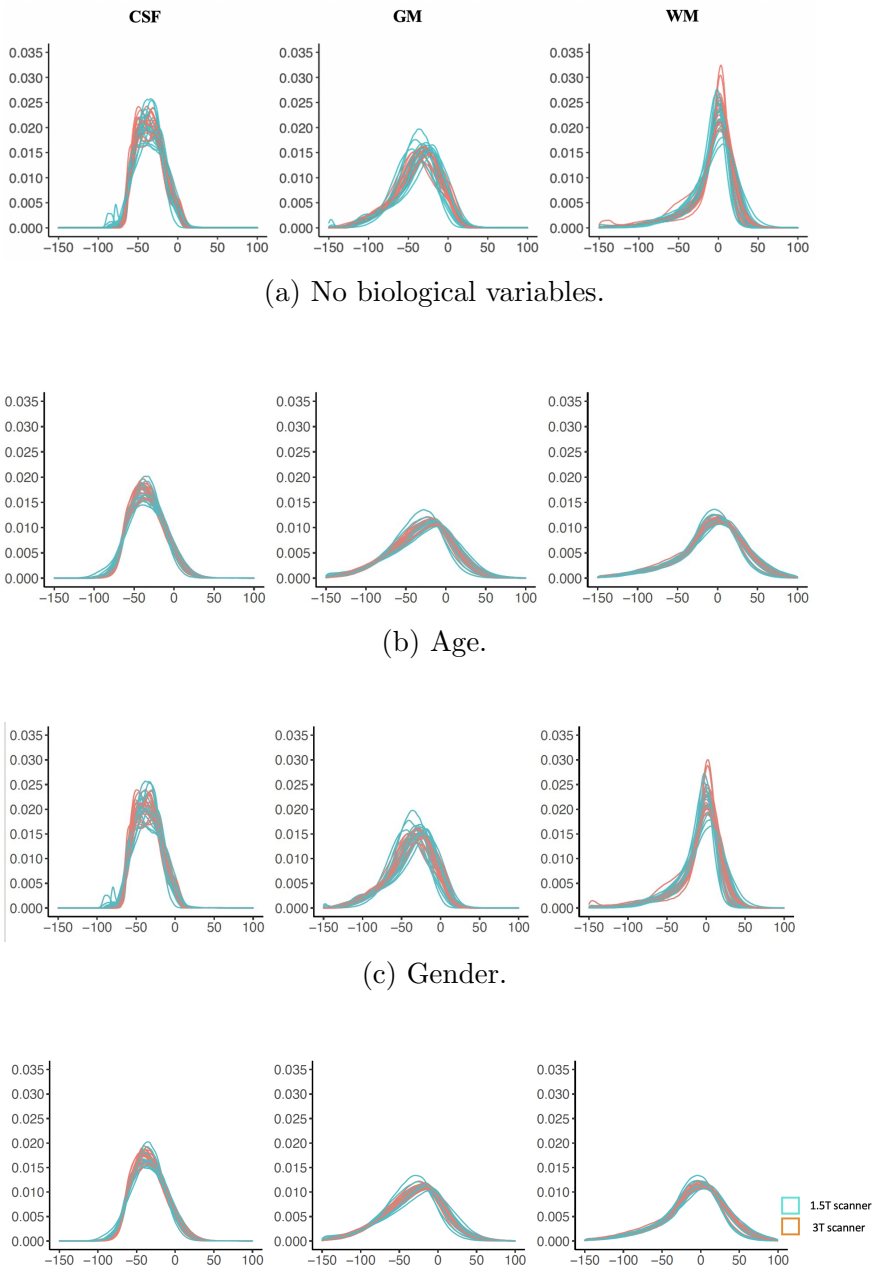


(a) Rank = 1.

(b) Rank = 2.

(c) Rank = 3.

Figure 31: Density plots of MRI voxel intensities by tissue type (cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM)) across scanners (GE 1.5T (cyan) and Siemens 3T (orange)) by setting (a) Rank = 1, (b) Rank = 2, and (c) Rank = 3.



(a) No biological variables.

(b) Age.

(c) Gender.

(d) Age and gender.

Figure 32: Density plots of MRI voxel intensities by tissue type (cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM)) across scanners (GE 1.5T (cyan) and Siemens 3T (orange)), controlling for (a) no biological variables, (b) age, (c) gender, and (d) age and gender.

A.2 Within-scanner descriptive statistics of summary measures

Table 11: Descriptive statistics, including mean and standard deviation (SD) for FreeSurfer-derived cortical thickness and volume measures relevant to AD. These statistics are reported for the RAW, RAVEL-corrected, ComBat-harmonized, and RAVEL-ComBat-harmonized measures across GE 1.5T and Siemens 3T scanners. Initially referenced in section 3.1. Values 0.00 and -0.00 indicate values < 0.005 and < -0.005, respectively.

	RAW		RAVEL		ComBat		RAVEL-ComBat	
	1.5T	3T	1.5T	3T	1.5T	3T	1.5T	3T
ROIs	Cortical Thickness (mm)							
Left								
Entorhinal	2.96 (0.29)	3.18 (0.37)	2.93 (0.29)	3.12 (0.40)	3.03 (0.28)	3.11 (0.36)	2.94 (0.28)	3.12 (0.38)
Fusiform	2.25 (0.18)	2.50 (0.15)	2.37 (0.25)	2.47 (0.17)	2.32 (0.17)	2.43 (0.16)	2.37 (0.21)	2.47 (0.19)
Inferior Parietal	2.13 (0.15)	2.08 (0.13)	2.22 (0.21)	2.07 (0.13)	2.13 (0.14)	2.09 (0.14)	2.18 (0.18)	2.12 (0.15)
Inferior Temporal	2.32 (0.21)	2.57 (0.17)	2.48 (0.30)	2.54 (0.18)	2.39 (0.19)	2.50 (0.17)	2.47 (0.26)	2.55 (0.21)
Middle Temporal	2.52 (0.24)	2.60 (0.25)	2.59 (0.22)	2.57 (0.23)	2.55 (0.23)	2.57 (0.25)	2.56 (0.21)	2.60 (0.24)
Right								
Entorhinal	3.00 (0.40)	3.24 (0.35)	3.04 (0.24)	3.22 (0.31)	3.08 (0.38)	3.16 (0.36)	3.05 (0.22)	3.21 (0.30)
Fusiform	2.32 (0.18)	2.54 (0.18)	2.43 (0.20)	2.48 (0.16)	2.38 (0.17)	2.48 (0.18)	2.42 (0.18)	2.48 (0.18)
Inferior Parietal	2.14 (0.19)	2.13 (0.15)	2.20 (0.20)	2.13 (0.14)	2.15 (0.18)	2.12 (0.15)	2.17 (0.18)	2.16 (0.16)
Inferior Temporal	2.33 (0.20)	2.59 (0.15)	2.41 (0.17)	2.51 (0.16)	2.40 (0.18)	2.52 (0.15)	2.42 (0.15)	2.50 (0.16)
Middle Temporal	2.55 (0.19)	2.60 (0.21)	2.60 (0.16)	2.59 (0.21)	2.57 (0.19)	2.58 (0.20)	2.58 (0.15)	2.61 (0.20)
ROIs	Volume $(cm)^3$							
Left								
Entorhinal	1.64 (0.32)	1.72 (0.30)	1.54 (0.34)	1.74 (0.38)	1.68 (0.32)	1.68 (0.29)	1.58 (0.34)	1.70 (0.35)
Inferior Temporal	8.05 (1.03)	8.84 (1.57)	8.48 (1.00)	8.91 (1.53)	8.29 (1.10)	8.60 (1.42)	8.56 (1.04)	8.82 (1.41)
Middle Temporal	8.80 (1.58)	8.96 (2.00)	9.26 (1.44)	8.91 (1.81)	8.96 (1.66)	8.80 (1.84)	9.21 (1.47)	8.97 (1.69)
Amygdala	1.40 (0.26)	1.53 (0.30)	1.43 (0.21)	1.52 (0.25)	1.43 (0.26)	1.50 (0.28)	1.46 (0.22)	1.50 (0.23)
Hippocampus	3.85 (0.46)	3.77 (0.48)	4.00 (0.47)	3.81 (0.49)	3.82 (0.46)	3.79 (0.46)	3.94 (0.47)	3.86 (0.46)
Right								
Entorhinal	1.60 (0.34)	1.69 (0.36)	1.55 (0.33)	1.67 (0.35)	1.65 (0.35)	1.65 (0.33)	1.57 (0.33)	1.65 (0.33)
Inferior Temporal	7.71 (1.23)	8.69 (1.62)	7.98 (1.37)	8.64 (1.52)	7.99 (1.29)	8.41 (1.49)	8.11 (1.38)	8.51 (1.43)
Middle Temporal	9.48 (1.35)	9.70 (1.87)	9.74 (1.39)	9.80 (1.73)	9.64 (1.43)	9.54 (1.71)	9.76 (1.42)	9.78 (1.61)
Amygdala	1.55 (0.21)	1.60 (0.21)	1.56 (0.19)	1.58 (0.20)	1.56 (0.21)	1.59 (0.20)	1.57 (0.19)	1.57 (0.19)
Hippocampus	3.99 (0.40)	3.90 (0.44)	4.09 (0.46)	3.99 (0.42)	3.96 (0.40)	3.93 (0.42)	4.06 (0.45)	4.02 (0.41)

A.3 Confidence intervals of bias for summary measures

Table 12: Mean (95% confidence interval) of cross-scanner differences, (Siemens 3T - GE 1.5T), for cortical thickness and volume measures relevant to AD. These statistics were prepared for each of the RAW, RAVEL, ComBat, and RAVEL-ComBat methods, using the paired t-test. The measures with statistically significant differences ($P < 0.05$) were highlighted. Values 0.00 and -0.00 indicate values < 0.005 and < -0.005, respectively.

	RAW	RAVEL	ComBat	RAVEL-ComBat
ROIs	Cortical Thickness (mm)			
Left				
Entorhinal	0.22 (0.09, 0.34)	0.19 (-0.03, 0.41)	0.08 (-0.04, 0.20)	0.18 (-0.04, 0.39)
Fusiform	0.24 (0.19, 0.30)	0.10 (-0.02, 0.23)	0.11 (0.06, 0.16)	0.10 (-0.02, 0.22)
Inferior Parietal	-0.05 (-0.11, 0.0)	-0.15 (-0.26, -0.04)	-0.04 (-0.09, 0.01)	-0.04 (-0.16, 0.04)
Inferior Temporal	0.25 (0.16, 0.34)	0.06 (-0.08, 0.20)	0.11 (0.02, 0.19)	0.08 (-0.05, 0.21)
Middle Temporal	0.08 (0.00, 0.16)	-0.01 (-0.11, 0.09)	0.02 (-0.06, 0.10)	0.03 (-0.07, 0.13)
Right				
Entorhinal	0.23 (0.00, 0.47)	0.17 (-0.01, 0.35)	0.08 (-0.15, 0.31)	0.15 (-0.02, 0.33)
Fusiform	0.22 (0.17, 0.28)	0.05 (-0.06, 0.16)	0.10 (0.04, 0.15)	0.06 (-0.04, 0.17)
Inferior Parietal	-0.02 (-0.06, 0.02)	-0.07 (-0.15, 0.01)	-0.02 (-0.06, 0.01)	-0.01 (-0.09, 0.06)
Inferior Temporal	0.26 (0.19, 0.32)	0.09 (0.01, 0.18)	0.11 (0.06, 0.17)	0.08 (0.00, 0.17)
Middle Temporal	0.05 (-0.03, 0.14)	-0.01 (-0.08, 0.06)	0.01 (-0.08, 0.09)	0.03 (-0.04, 0.10)
ROIs	Volume $(cm)^3$			
Left				
Entorhinal	0.08 (-0.05, 0.22)	0.19 (-0.03, 0.41)	0.01 (-0.12, 0.14)	0.12 (-0.09, 0.33)
Inferior Temporal	0.79 (0.37, 1.21)	0.43 (-0.01, 0.88)	0.31 (-0.03, 0.66)	0.26 (-0.13, 0.65)
Middle Temporal	0.16 (-0.29, 0.61)	-0.35 (-0.91, 0.21)	-0.16 (-0.55, 0.24)	-0.23 (-0.76, 0.29)
Amygdala	0.13 (0.03, 0.23)	0.09 (0.01, 0.17)	0.06 (-0.04, 0.16)	0.04 (-0.04, 0.11)
Hippocampus	-0.08 (-0.16, -0.01)	-0.19 (-0.29, -0.09)	-0.03 (-0.1, 0.05)	-0.05 (-0.18, 0.01)
Right				
Entorhinal	0.09 (-0.05, 0.23)	0.12 (-0.05, 0.29)	0.01 (-0.13, 0.15)	0.07 (-0.10, 0.24)
Inferior Temporal	0.98 (0.55, 1.40)	0.67 (0.19, 1.14)	0.42 (0.04, 0.80)	0.41 (-0.05, 0.86)
Middle Temporal	0.21 (-0.27, 0.70)	0.06 (-0.39, 0.51)	-0.10 (-0.52, 0.32)	0.02 (-0.40, 0.44)
Amygdala	0.05 (0.01, 0.10)	0.02 (-0.03, 0.06)	0.03 (-0.02, 0.08)	0.01 (-0.04, 0.05)
Hippocampus	-0.09 (-0.16, -0.01)	-0.10 (-0.23, 0.03)	-0.03 (-0.11, 0.04)	-0.04 (-0.17, 0.08)

A.4 Experiments on different preprocessing pipelines

To investigate the effects of different preprocessing methods on RAW and ComBat data, we created two new datasets, each with its preprocessing steps removed. Specifically, we generated RAW_{Orig} and $\text{ComBat}_{\text{Orig}}$ datasets by excluding the preprocessing steps used for RAW and ComBat data generation, respectively. We then conducted paired t-tests with a 95% confidence interval (CI) to compare RAW_{Orig} with RAW and $\text{ComBat}_{\text{Orig}}$ with ComBat. The results of these comparisons were documented in Tables 13 and 14 for each scanner. In these tables, the first two columns represent the mean (SD) of the datasets being compared, while the third column displays the results of the comparison: the mean directional differences (95% CI) obtained from the t-test. Statistically significant differences ($p < 0.05$) are highlighted in the tables.

Table 13: Comparing different preprocessing pipelines for generating RAW data using paired t-test. For each scanner, the first two columns are the mean (SD) of the corresponding data and the third column is the mean of directional differences (95% confidence interval) of data in the first two columns. The statistically significant differences ($p < 0.05$) are highlighted in the tables. Values 0.00 and -0.00 indicate values < 0.005 and < -0.005, respectively.

	1.5T			3T		
	RAW_{Orig}	RAW	RAW_{Orig} - RAW	RAW_{Orig}	RAW	RAW_{Orig} - RAW
ROIs	Cortical Thickness (mm)			Cortical Thickness (mm)		
Left						
Entorhinal	3.06 (0.39)	2.96 (0.29)	0.10 (-0.26, 0.06)	3.42 (0.39)	3.18 (0.37)	0.24 (-0.36, -0.12)
Fusiform	2.42 (0.15)	2.25 (0.18)	0.17 (-0.22, -0.11)	2.50 (0.17)	2.50 (0.15)	0.00 (-0.05, 0.05)
Inferior Parietal	2.20 (0.14)	2.13 (0.15)	0.07 (-0.13, -0.02)	2.07 (0.14)	2.08 (0.13)	-0.01 (-0.02, 0.05)
Inferior Temporal	2.43 (0.17)	2.32 (0.21)	0.11 (-0.19, -0.02)	2.55 (0.17)	2.57 (0.17)	-0.02 (-0.04, 0.09)
Middle Temporal	2.63 (0.25)	2.52 (0.24)	0.11 (-0.19, -0.03)	2.60 (0.23)	2.60 (0.25)	0.00 (-0.06, 0.05)
Right						
Entorhinal	3.17 (0.42)	3.00 (0.40)	0.17 (-0.32, -0.01)	3.58 (0.40)	3.24 (0.35)	0.34 (-0.52, -0.16)
Fusiform	2.47 (0.16)	2.32 (0.18)	0.14 (-0.22, -0.07)	2.60 (0.20)	2.54 (0.18)	0.05 (-0.11, -0.00)
Inferior Parietal	2.20 (0.16)	2.14 (0.19)	0.05 (-0.13, 0.02)	2.16 (0.14)	2.13 (0.15)	0.03 (-0.08, 0.02)
Inferior Temporal	2.41 (0.17)	2.33 (0.20)	0.08 (-0.18, 0.02)	2.65 (0.16)	2.59 (0.15)	0.06 (-0.11, -0.01)
Middle Temporal	2.68 (0.20)	2.55 (0.19)	0.13 (-0.21, -0.05)	2.67 (0.20)	2.60 (0.21)	0.07 (-0.12, -0.01)
ROIs	Volume $(\text{cm})^3$			Volume $(\text{cm})^3$		
Left						
Entorhinal	1.68 (0.27)	1.64 (0.32)	0.04 (-0.26, 0.19)	1.68 (0.34)	1.72 (0.30)	-0.05 (-0.07, 0.16)
Inferior Temporal	8.20 (1.73)	8.05 (1.03)	0.15 (-0.65, 0.35)	8.50 (1.57)	8.84 (1.57)	-0.34 (0.12, 0.56)
Middle Temporal	8.49 (1.90)	8.80 (1.58)	-0.31 (-0.00, 0.63)	8.54 (1.86)	8.96 (2.00)	-0.42 (0.14, 0.70)
Amygdala	1.39 (0.21)	1.40 (0.26)	-0.01 (-0.05, 0.08)	1.46 (0.31)	1.53 (0.30)	-0.07 (-0.01, 0.15)
Hippocampus	3.97 (0.58)	3.85 (0.46)	0.11 (-0.22, -0.01)	3.76 (0.55)	3.77 (0.48)	-0.01 (-0.08, 0.09)
Right						
Entorhinal	1.45 (0.28)	1.60 (0.34)	-0.15 (0.03, 0.27)	1.59 (0.34)	1.69 (0.36)	-0.10 (-0.01, 0.21)
Inferior Temporal	7.77 (1.49)	7.71 (1.23)	0.06 (-0.41, 0.29)	8.39 (1.51)	8.69 (1.62)	-0.30 (0.05, 0.56)
Middle Temporal	9.63 (1.71)	9.48 (1.35)	0.14 (-0.53, 0.24)	9.55 (1.75)	9.70 (1.87)	-0.15 (-0.08, 0.38)
Amygdala	1.43 (0.19)	1.55 (0.21)	-0.11 (0.05, 0.18)	1.43 (0.23)	1.60 (0.21)	-0.17 (0.10, 0.24)
Hippocampus	4.02 (0.49)	3.99 (0.40)	0.04 (-0.15, 0.07)	3.86 (0.52)	3.90 (0.44)	-0.04 (-0.06, 0.14)

Table 14: Comparing different preprocessing pipelines for generating ComBat data using paired t-test. For each scanner, the first two columns are the mean (SD) of the corresponding data and the third column is the mean of directional differences (95% confidence interval) of data in the first two columns. The statistically significant differences ($p < 0.05$) are highlighted in the tables. Values 0.00 and -0.00 indicate values < 0.005 and < -0.005, respectively.

	1.5T			3T		
	$\text{ComBat}_{\text{Orig}}$	ComBat	$\text{ComBat}_{\text{Orig}}$ - ComBat	$\text{ComBat}_{\text{Orig}}$	ComBat	$\text{ComBat}_{\text{Orig}}$ - ComBat
ROIs	Cortical Thickness (mm)			Cortical Thickness (mm)		
Left						
Entorhinal	3.14 (0.38)	3.03 (0.28)	0.11 (-0.05, 0.26)	3.35 (0.37)	3.11 (0.36)	0.23 (0.12, 0.34)
Fusiform	2.43 (0.15)	2.32 (0.17)	0.11 (0.06, 0.16)	2.48 (0.16)	2.43 (0.16)	0.05 (0.01, 0.10)
Inferior Parietal	2.16 (0.14)	2.13 (0.14)	0.04 (-0.02, 0.09)	2.11 (0.13)	2.09 (0.14)	0.02 (-0.01, 0.05)
Inferior Temporal	2.45 (0.16)	2.39 (0.19)	0.06 (-0.02, 0.14)	2.52 (0.17)	2.50 (0.17)	0.03 (-0.04, 0.09)
Middle Temporal	2.61 (0.24)	2.55 (0.23)	0.06 (-0.01, 0.14)	2.62 (0.22)	2.57 (0.25)	0.05 (-0.00, 0.10)
Right						
Entorhinal	3.26 (0.41)	3.08 (0.38)	0.18 (0.03, 0.33)	3.49 (0.38)	3.16 (0.36)	0.33 (0.15, 0.50)
Fusiform	2.49 (0.16)	2.38 (0.17)	0.11 (0.03, 0.18)	2.57 (0.18)	2.48 (0.18)	0.09 (0.04, 0.14)
Inferior Parietal	2.18 (0.16)	2.15 (0.18)	0.03 (-0.04, 0.10)	2.18 (0.13)	2.12 (0.15)	0.05 (0.0, 0.10)
Inferior Temporal	2.46 (0.16)	2.4 (0.18)	0.06 (-0.03, 0.15)	2.59 (0.15)	2.52 (0.15)	0.08 (0.03, 0.12)
Middle Temporal	2.66 (0.20)	2.57 (0.19)	0.10 (0.02, 0.18)	2.68 (0.19)	2.58 (0.20)	0.10 (0.05, 0.15)
ROIs	Volume $(\text{cm})^3$			Volume $(\text{cm})^3$		
Left						
Entorhinal	1.40 (0.22)	1.43 (0.26)	-0.03 (-0.10, 0.04)	1.44 (0.29)	1.50 (0.28)	-0.06 (-0.13, 0.02)
Inferior Temporal	3.90 (0.56)	3.82 (0.46)	0.08 (-0.02, 0.18)	3.82 (0.55)	3.79 (0.46)	0.03 (-0.06, 0.11)
Middle Temporal	1.43 (0.19)	1.56 (0.21)	-0.13 (-0.20, -0.06)	1.43 (0.22)	1.59 (0.20)	-0.15 (-0.22, -0.08)
Amygdala	3.97 (0.48)	3.96 (0.40)	0.01 (-0.09, 0.12)	3.91 (0.51)	3.93 (0.42)	-0.02 (-0.11, 0.08)
Hippocampus	1.68 (0.27)	1.68 (0.32)	0.00 (-0.22, 0.23)	1.68 (0.32)	1.68 (0.29)	-0.01 (-0.12, 0.10)
Right						
Entorhinal	8.25 (1.67)	8.29 (1.10)	-0.04 (-0.49, 0.42)	8.46 (1.54)	8.60 (1.42)	-0.15 (-0.36, 0.07)
Middle Temporal	8.50 (1.85)	8.96 (1.66)	-0.46 (-0.75, -0.17)	8.53 (1.81)	8.80 (1.84)	-0.28 (-0.53, -0.02)
Inferior Temporal	1.47 (0.27)	1.65 (0.35)	-0.17 (-0.29, -0.05)	1.57 (0.33)	1.65 (0.33)	-0.08 (-0.18, 0.02)
Amygdala	9.62 (1.67)	9.64 (1.43)	-0.02 (-0.38, 0.35)	9.55 (1.70)	9.54 (1.71)	0.01 (-0.20, 0.22)
Hippocampus	7.86 (1.45)	7.99 (1.29)	-0.13 (-0.46, 0.21)	8.30 (1.47)	8.41 (1.49)	-0.11 (-0.34, 0.12)

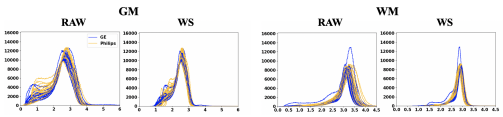
A.5 Comparing ComBat-harmonized and Longitudinal-ComBat-harmonized biomarkers of AD

Table 15: Mean (SD) of FreeSurfer-derived cortical thickness and volume measures relevant to AD. These statistics are reported for the ComBat-harmonized and Longitudinal-ComBat-harmonized measures within GE 1.5T and Siemens 3T scanners.

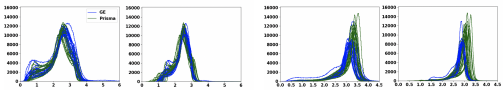
	ComBat		Longitudinal ComBat	
	1.5T	3T	1.5T	3T
ROIs	Cortical Thickness (mm)			
Left				
Entorhinal	3.03 (0.28)	3.11 (0.36)	3.07 (0.28)	3.07 (0.37)
Fusiform	2.32 (0.17)	2.43 (0.16)	2.37 (0.18)	2.38 (0.15)
Inferior Parietal	2.13 (0.14)	2.09 (0.14)	2.11 (0.15)	2.10 (0.13)
Inferior Temporal	2.39 (0.19)	2.50 (0.17)	2.44 (0.20)	2.45 (0.17)
Middle Temporal	2.55 (0.23)	2.57 (0.25)	2.56 (0.23)	2.56 (0.25)
Right				
Entorhinal	3.08 (0.38)	3.16 (0.36)	3.13 (0.39)	3.11 (0.36)
Fusiform	2.38 (0.17)	2.48 (0.18)	2.43 (0.18)	2.44 (0.18)
Inferior Parietal	2.15 (0.18)	2.12 (0.15)	2.14 (0.19)	2.13 (0.15)
Inferior Temporal	2.40 (0.18)	2.52 (0.15)	2.45 (0.19)	2.46 (0.15)
Middle Temporal	2.57 (0.19)	2.58 (0.20)	2.58 (0.19)	2.57 (0.21)
ROIs	Volume $(cm)^3$			
Left				
Entorhinal	1.68 (0.32)	1.68 (0.29)	1.68 (0.32)	1.68 (0.30)
Inferior Temporal	8.29 (1.10)	8.6 (1.42)	8.41 (1.03)	8.49 (1.52)
Inferior Parietal	8.96 (1.66)	8.80 (1.84)	8.89 (1.58)	8.87 (1.97)
Amygdala	1.43 (0.26)	1.50 (0.28)	1.46 (0.26)	1.47 (0.29)
Hippocampus	3.82 (0.46)	3.79 (0.46)	3.82 (0.46)	3.80 (0.48)
Right				
Entorhinal	9.64 (1.43)	9.54 (1.71)	1.65 (0.35)	1.65 (0.35)
Inferior Temporal	1.65 (0.35)	1.65 (0.33)	9.6 (1.35)	9.58 (1.82)
Inferior Parietal	7.99 (1.29)	8.41 (1.49)	8.15 (1.23)	8.26 (1.59)
Amygdala	1.56 (0.21)	1.59 (0.20)	1.57 (0.22)	1.58 (0.21)
Hippocampus	3.96 (0.40)	3.93 (0.42)	3.95 (0.40)	3.93 (0.43)

Appendix B Additional Results from Section 4

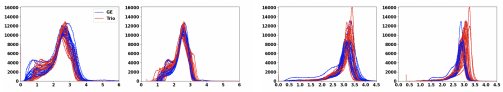
B.1 White stripe normalization in matched data



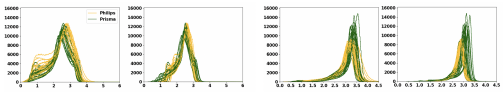
(a) **GE-Philips pair.**



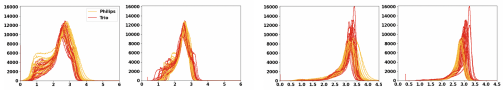
(b) **GE-SiemensP pair.**



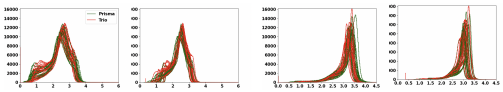
(c) **GE-SiemensT pair.**



(d) **Philips-SiemensP pair.**



(e) **Philips-SiemensT pair.**



(f) **SiemensP-SiemensT pair.**

Figure 33: Histograms of gray matter (GM) and white matter (WM) voxels for RAW and White Stripe (WS)-normalized images of all subjects. These histograms were plotted for all 6 scanner pairs. WS makes the plots more centered, overlapped, and therefore comparable across subjects. WS usually outputs images with negative intensity values. For plotting the histograms, we shifted the WS-normalized images to have positive intensity values.

Bibliography

Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J., and Raznahan, A. (2016). Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo mri. *Human brain mapping*, 37(7):2385–2397.

An, L., Chen, J., Chen, P., Zhang, C., He, T., Chen, C., Zhou, J. H., Yeo, B. T., of Aging, L. S., Initiative, A. D. N., et al. (2022). Goal-specific brain mri harmonization. *NeuroImage*, 263:119570.

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–851.

Aslani, S., Murino, V., Dayan, M., Tam, R., Sona, D., and Hamarneh, G. (2020). Scanner invariant multiple sclerosis lesion segmentation from mri. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 781–785. IEEE.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41.

Backhausen, L. L., Herting, M. M., Buse, J., Roessner, V., Smolka, M. N., and Vetter, N. C. (2016). Quality control of structural mri images applied using freesurfer—a hands-on workflow to rate motion artifacts. *Frontiers in neuroscience*, 10:558.

Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., Zhuo, C., et al. (2020). Medical image harmonization using deep learning based canonical mapping: Toward robust and generalizable learning in imaging. *arXiv preprint arXiv:2010.05355*.

Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., et al. (2022). Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *Journal of Magnetic Resonance Imaging*, 55(3):908–916.

Basser, P. J., Mattiello, J., and LeBihan, D. (1994). Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267.

Bayer, J. M., Dinga, R., Kia, S. M., Kottaram, A. R., Wolfers, T., Lv, J., Zalesky, A., Schmaal, L., and Marquand, A. (2022a). Accommodating site variation in neuroimaging data using normative and hierarchical bayesian models. *NeuroImage*, page 119699.

Bayer, J. M. M., Thompson, P., Ching, C. R., Liu, M., Chen, A., Panzenhagen, A. C., Jahanshad, N., Marquand, A., Schmaal, L., and Saemann, P. G. (2022b). Site effects how-to & when: an overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses.

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., Linn, K. A., Initiative, A. D. N., et al. (2020). Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage*, 220:117129.

Brown, R. W., Cheng, Y.-C. N., Haacke, E. M., Thompson, M. R., and Venkatesan, R. (2014). *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons.

Cackowski, S., Barbier, E. L., Dojat, M., and Christen, T. (2021). Imunity: a generalizable vae-gan solution for multicenter mr image harmonization. *arXiv preprint arXiv:2109.06756*.

Castleman, K. R. (1996). *Digital image processing*. Prentice Hall Press.

Chaitanya, K., Karani, N., Baumgartner, C. F., Erdil, E., Becker, A., Donati, O., and Konukoglu, E. (2021). Semi-supervised task-driven data augmentation for medical image segmentation. *Medical Image Analysis*, 68:101934.

Chang, X., Cai, X., Dan, Y., Song, Y., Lu, Q., Yang, G., and Nie, S. (2022). Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. *Physics in Medicine & Biology*, 67(14):145004.

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., Shou, H., Initiative, A. D. N., et al. (2020a). Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. *bioRxiv*, page 858415.

Chen, A. A., Luo, C., Chen, Y., Shinohara, R. T., Shou, H., Initiative, A. D. N., et al. (2022a). Privacy-preserving harmonization via distributed combat. *Neuroimage*, 248:118822.

Chen, C.-L., Torbati, M. E., Wilson, J. D., Minhas, D. S., Laymon, C. M., Hwang, S. J., Maillard, P., Fletcher, E., DeCarli, C., and Tudorascu, D. (2022b). Reducing mri inter-scanner variability using 3d superpixel combat. In *Alzheimer's Association International Conference, ALZ*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

Clark, K. A., O'Donnell, C. M., Elliott, M. A., Tauhid, S., Dewey, B. E., Chu, R., Khalil, S., Nair, G., Sati, P., DuVal, A., et al. (2022). Inter-scanner brain mri volumetric biases persist even in a harmonized multi-subject study of multiple sclerosis. *bioRxiv*.

Da-Ano, R., Lucia, F., Masson, I., Abgral, R., Alfieri, J., Rousseau, C., Mervoyer, A., Reinhold, C., Pradier, O., Schick, U., et al. (2021). A transfer learning approach to facilitate combat-based harmonization of multicentre radiomic features in new datasets. *Plos one*, 16(7):e0253653.

de Zwart, J. A., van Gelderen, P., Golay, X., Ikonomidou, V. N., and Duyn, J. H. (2006). Accelerated parallel imaging for functional imaging of the human brain. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, 19(3):342–351.

Debette, S. and Markus, H. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *Bmj*, 341.

Dewey, B. E. et al. (2021). *Synthesis-Based Harmonization of Multi-Contrast Structural MRI*. PhD thesis, Johns Hopkins University.

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., et al. (2019). Deepharmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64:160–170.

Dewey, B. E., Zuo, L., Carass, A., He, Y., Liu, Y., Mowry, E. M., Newsome, S., Oh, J., Calabresi, P. A., and Prince, J. L. (2020). A disentangled latent space for cross-site mri harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 720–729. Springer.

Dinsdale, N. K., Jenkinson, M., and Namburete, A. I. (2021). Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. *NeuroImage*, 228:117689.

Dixon, W. T. (1984). Simple proton spectroscopic imaging. *Radiology*, 153(1):189–194.

Duchesne, S., Dieumegarde, L., Chouinard, I., Farokhian, F., Badhwar, A., Bellec, P., Tétreault, P., Descoteaux, M., Boré, A., Houde, J.-C., et al. (2019). Structural and functional multi-platform mri series of a single human volunteer over more than fifteen years. *Scientific data*, 6(1):1–9.

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., et al. (2009). The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687.

Fatania, K., Clark, A., Frood, R., Scarsbrook, A., Al-Qaisieh, B., Currie, S., and Nix, M. (2022). Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders. *Physics and Imaging in Radiation Oncology*, 22:115–122.

Filippi, M., Agosta, F., Scola, E., Canu, E., Magnani, G., Marcone, A., Valsasina, P., Caso, F., Copetti, M., Comi, G., et al. (2013). Functional network connectivity in the behavioral variant of frontotemporal dementia. *Cortex*, 49(9):2389–2401.

Fischl, B. (2012). FreeSurfer. *Neuroimage*, 62(2):774–781.

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., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120.

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., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170.

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., Initiative, A. D. N., et al. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, 132:198–212.

Foy, J. J., Al-Hallaq, H. A., Grekoski, V., Tran, T., Guruvadoo, K., Armato III, S. G., and Sensakovic, W. F. (2020). Harmonization of radiomic feature variability resulting from differences in ct image acquisition and reconstruction: assessment in a cadaveric liver. *Physics in Medicine & Biology*, 65(20):205008.

Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W. H., Corvin, A., Redolfi, A., Nelson, B., Crespo-Facorro, B., McDonald, C., et al. (2020). Neuroharmony: A new tool for harmonizing volumetric mri data from unseen scanners. *NeuroImage*, 220.

Haacke, E. M., Mittal, S., Wu, Z., Neelavalli, J., and Cheng, Y.-C. (2009). Susceptibility-weighted imaging: technical aspects and clinical applications, part 1. *American Journal of Neuroradiology*, 30(1):19–30.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., et al. (2006). Reliability of mri-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage*, 32(1):180–194.

Hansen, C. B., Schilling, K. G., Rheault, F., Resnick, S., Shafer, A. T., Beason-Held, L. L., and Landman, B. A. (2022). Contrastive semi-supervised harmonization of single-shell to multi-shell diffusion mri. *Magnetic Resonance Imaging*, 93:73–86.

Hawco, C., Dickie, E. W., Herman, G., Turner, J. A., Argyelan, M., Malhotra, A. K., Buchanan, R. W., and Voineskos, A. N. (2022). A longitudinal multi-scanner multimodal human neuroimaging dataset. *Scientific Data*, 9(1):1–7.

Heinen, R., Bouvy, W. H., Mendrik, A. M., Viergever, M. A., Biessels, G. J., and De Bresser, J. (2016). Robustness of automated methods for brain volume measurements across different mri field strengths. *PloS one*, 11(10):e0165719.

Jahanshad, N., Kochunov, P. V., Sprooten, E., Mandl, R. C., Nichols, T. E., Almasy, L., Blangero, J., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., et al. (2013). Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the enigma–dti working group. *Neuroimage*, 81:455–469.

Jernigan, T. L., Brown, S. A., and Dowling, G. J. (2018). The adolescent brain cognitive development study. *Journal of research on adolescence: the official journal of the Society for Research on Adolescence*, 28(1):154.

Jezzard, P. and Clare, S. (1999). Sources of distortion in functional mri data. *Human brain mapping*, 8(2-3):80–85.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van Der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., et al. (2006). Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage*, 30(2):436–443.

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F., et al. (2013). Brain morphometry reproducibility in multi-center 3 t mri studies: a comparison of cross-sectional and longitudinal segmentations. *Neuroimage*, 83:472–484.

Karayumak, S. C., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., and Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion mri data acquired with different acquisition parameters. *Neuroimage*, 184:180–200.

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

Kochunov, P., Jahanshad, N., Sprooten, E., Nichols, T. E., Mandl, R. C., Almasy, L., Booth, T., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., et al. (2014). Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling. *Neuroimage*, 95:136–150.

Kruggel, F., Turner, J., Muftuler, L. T., Initiative, A. D. N., et al. (2010). Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the adni cohort. *Neuroimage*, 49(3):2123–2133.

LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A. G., et al. (2019). Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, pages 2019–12.

Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., and Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 13(4):534–546.

Lee, M. C., Oktay, O., Schuh, A., Schaap, M., and Glocker, B. (2019). Image-and-spatial transformer networks for structure-guided image registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 337–345. Springer.

Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., and Jahanshad, N. (2021). Style transfer using generative adversarial networks for multi-site mri harmonization. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 313–322. Springer.

Liu, M., Zhu, A. H., Maiti, P., Thomopoulos, S. I., Gadewar, S., Chai, Y., Kim, H., Jahanshad, N., and Initiative, A. D. N. (2023). Style transfer generative adversarial networks to harmonize multisite mri to a single reference image to avoid overcorrection. *Human Brain Mapping*, 44(14):4875–4892.

Liu, S. and Yap, P.-T. (2021). Learning multi-site harmonization of magnetic resonance images without traveling human phantoms. *arXiv preprint arXiv:2110.00041*.

Lustig, M., Donoho, D., and Pauly, J. M. (2007). Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195.

Maclaren, J., Aksoy, M., Ooi, M. B., Zahneisen, B., and Bammer, R. (2018). Prospective motion correction using coil-mounted cameras: cross-calibration considerations. *Magnetic resonance in medicine*, 79(4):1911–1921.

Maclaren, J., Herbst, M., Speck, O., and Zaitsev, M. (2013). Prospective motion correction in brain imaging: a review. *Magnetic resonance in medicine*, 69(3):621–636.

Madan, C. R. (2017). Advances in studying brain morphology: The benefits of open-access data. *Frontiers in human neuroscience*, 11:405.

Madan, C. R. (2021). Scan once, analyse many: using large open-access neuroimaging datasets to understand the brain. *Neuroinformatics*, pages 1–29.

Magnotta, V. A., Matsui, J. T., Liu, D., Johnson, H. J., Long, J. D., Jr, B. D. B., Mueller, B. A., Lim, K., Mori, S., Helmer, K. G., Turner, J. A., Reading, S., Lowe, M. J., Aylward, E., Flashman, L. A., Bonett, G., and Paulsen, J. S. (2020). "dwi traveling human phantom study".

Mahesh, M. (2013). The essential physics of medical imaging. *Medical physics*, 40(7):077301.

Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S. C., et al. (2021). Comparison of traveling-subject and combat harmonization methods for assessing structural brain characteristics. *Human brain mapping*, 42(16):5278–5287.

Manjón, J. V., Carbonell-Caballero, J., Lull, J. J., García-Martí, G., Martí-Bonmatí, L., and Robles, M. (2008). Mri denoising using non-local means. *Medical image analysis*, 12(4):514–523.

Mar, R. A., Spreng, R. N., and DeYoung, C. G. (2013). How to produce personality neuroscience research with high statistical power and low additional cost. *Cognitive, Affective, & Behavioral Neuroscience*, 13(3):674–685.

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C. M., Trojanowski, J. Q., et al. (2018). The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort. *Annals of clinical and translational neurology*, 5(12):1460–1477.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., et al. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322.

Meyer, M. I., de la Rosa, E., Pedrosa de Barros, N., Paolella, R., Van Leemput, K., and Sima, D. M. (2021). A contrast augmentation approach to improve multi-scanner generalization in mri. *Frontiers in neuroscience*, 15:708196.

Meyer, M. I., Rosa, E. d. l., Leemput, K. V., and Sima, D. M. (2019). Relevance vector machines for harmonization of mri brain volumes using image descriptors. In *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, pages 77–85. Springer.

Milham, M. P., Craddock, R. C., Son, J. J., Fleischmann, M., Clucas, J., Xu, H., Koo, B., Krishnakumar, A., Biswal, B. B., Castellanos, F. X., et al. (2018). Assessment of the impact of shared brain imaging data on the scientific literature. *Nature Communications*, 9(1):1–7.

Minhas, D. S., Yang, Z., Muschelli, J., Laymon, C. M., Mettenburg, J. M., Zammit, M. D., Johnson, S., Mathis, C. A., Cohen, A. D., Handen, B. L., et al. (2020). Statistical methods for processing neuroimaging data from two different sites with a down syndrome population application. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 367–379. Springer.

Mirzaalian, H., de Pierrefeu, A., Savadjiev, P., Pasternak, O., Bouix, S., Kubicki, M., Westin, C.-F., Shenton, M. E., and Rathi, Y. (2015). Harmonizing diffusion mri data across multiple sites and scanners. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 12–19. Springer.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C., Morey, R. A., Flashman, L., et al. (2016). Inter-site and inter-scanner diffusion mri data harmonization. *NeuroImage*, 135:311–323.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Karmacharya, S., Grant, G., Marx, C. E., Morey, R. A., et al. (2018). Multi-site harmonization of diffusion mri data in a registration framework. *Brain imaging and behavior*, 12(1):284–295.

Modanwal, G., Vellal, A., Buda, M., and Mazurowski, M. A. (2020). Mri image harmonization using cycle-consistent generative adversarial network. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 1131413. International Society for Optics and Photonics.

Moyer, D. and Golland, P. (2021). Harmonization and the worst scanner syndrome. *arXiv preprint arXiv:2101.06255*.

Moyer, D., Ver Steeg, G., Tax, C. M., and Thompson, P. M. (2020). Scanner invariant representations for diffusion mri harmonization. *Magnetic resonance in medicine*, 84(4):2174–2189.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66.

Nacher, P. (2007). Magnetic resonance imaging: From spin physics to medical diagnosis/nacher, pj. *Quantum Spaces.–Birkhauser Verlag Basel.–2007*.

Nacher, P.-J. (2009). Magnetic resonance imaging: from spin physics to medical diagnosis. In *The Spin: Poincaré Seminar 2007*, pages 159–193. Springer.

Newton, H. B. (2016). *Handbook of neuro-oncology neuroimaging*. Academic Press.

Nielson, D. M., Pereira, F., Zheng, C. Y., Migineishvili, N., Lee, J. A., Thomas, A. G., and Bandettini, P. A. (2018). Detecting and harmonizing scanner differences in the abcd study-annual release 1.0. *BioRxiv*, page 309260.

Ning, L., Bonet-Carne, E., Grussu, F., Sepehrband, F., Kaden, E., Veraart, J., Blumberg, S. B., Khoo, C. S., Palombo, M., Kokkinos, I., et al. (2020). Cross-scanner and cross-protocol multi-shell diffusion mri data harmonization: Algorithms and results. *NeuroImage*, 221:117128.

Nyúl, L. G. and Udupa, J. K. (1999). On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1072–1081.

Obenauer, J. C., Stockfisch, T. P., and Fournier, M. V. (2019). Overcorrection of batch effects by combat can be avoided by using an equal medians method. *Cancer Research*, 79(13_Supplement):1659–1659.

Oishi, K., Faria, A., Jiang, H., Li, X., Akhter, K., Zhang, J., Hsu, J. T., Miller, M. I., van Zijl, P. C., Albert, M., et al. (2009). Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. *Neuroimage*, 46(2):486–499.

Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P.-J., Jeurissen, B., Ribbens, A., den Dekker, A. J., and Sijbers, J. (2020). Harmonization of brain diffusion mri: Concepts and methods. *Frontiers in Neuroscience*, page 396.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., et al. (2020). Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450.

Potvin, O., Khademi, A., Chouinard, I., Farokhian, F., Dieumegarde, L., Leppert, I., Hoge, R., Rajah, M. N., Bellec, P., Duchesne, S., et al. (2019). Measurement variability following mri system upgrade. *Frontiers in neurology*, 10:726.

Prohl, A. K., Scherrer, B., Tomas-Fernandez, X., Flip-Dhima, R., Velasco-Annis, C., Clancy, S., Carmody, E., Dean, M., Valee, M., P Prabhu, S., et al. (2019). Reproducibility of structural and diffusion tensor imaging in the tacern multi-center study. *Frontiers in integrative neuroscience*, 13:24.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., Weickert, C. S., Weickert, T., Bruggemann, J., Kircher, T., et al. (2020). Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma. *NeuroImage*, 218:116956.

Raichle, M. E. (2009). A brief history of human brain mapping. *Trends in neurosciences*, 32(2):118–126.

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., and van Diepen, M. (2021). External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, 14(1):49–58.

Ren, M., Dey, N., Fishbaugh, J., and Gerig, G. (2021). Segmentation-renormalized deep feature modulation for unpaired image harmonization. *IEEE Transactions on Medical Imaging*, 40(6):1519–1530.

Reynolds, D. A. et al. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Robinson, R., Dou, Q., Coelho de Castro, D., Kamnitsas, K., Groot, M. d., Summers, R. M., Rueckert, D., and Glocker, B. (2020). Image-level harmonization of multi-site data using image-and-spatial transformer networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 710–719. Springer.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., and Nichols, T. E. (2009). Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage*, 45(3):810–823.

Schnack, H. G., van Haren, N. E., Brouwer, R. M., van Baal, G. C. M., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T. D., Huttunen, M., Lepage, C., et al. (2010). Mapping reliability in multicenter mri: Voxel-based morphometry and cortical thickness. *Human brain mapping*, 31(12):1967–1982.

Schwartz, D. L., Tagge, I., Powers, K., Ahn, S., Bakshi, R., Calabresi, P. A., Todd Constable, R., Grinstead, J., Henry, R. G., Nair, G., et al. (2019). Multisite reliability and repeatability of an advanced brain mri protocol. *Journal of Magnetic Resonance Imaging*, 50(3):878–888.

Schwarz, C. G., Gunter, J. L., Wiste, H. J., Przybelski, S. A., Weigand, S. D., Ward, C. P., Senjem, M. L., Vemuri, P., Murray, M. E., Dickson, D. W., et al. (2016). A large-scale comparison of cortical thickness and volume methods for measuring alzheimer's disease severity. *NeuroImage: Clinical*, 11:802–812.

Sederevicius, D., Bjornerud, A., Walhovd, K. B., Van Leemput, K., Fischl, B., and Fjell, A. M. (2022). A robust intensity distribution alignment for harmonization of t1w intensity values. *bioRxiv*.

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., and Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic resonance in medicine*, 67(5):1210–1224.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., and Arbel, T. (2011). Evaluating intensity normalization on mris of human brain with multiple sclerosis. *Medical image analysis*, 15(2):267–282.

Shinohara, R. T., Crainiceanu, C. M., Caffo, B. S., Gaitán, M. I., and Reich, D. S. (2011). Population-wide principal component-based quantification of blood–brain-barrier dynamics in multiple sclerosis. *NeuroImage*, 57(4):1430–1446.

Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J., Henry, R. G., Kim, G., Linn, K. A., Papinutto, N., et al. (2017). Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis. *American Journal of Neuroradiology*, 38(8):1501–1509.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2014a). Australian imaging biomarkers lifestyle flagship study of ageing, and alzheimer's disease neuroimaging initiative. statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin*, 6(9).

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., Crainiceanu, C. M., et al. (2014b). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19.

Solanes, A., Palau, P., Fortea, L., Salvador, R., González-Navarro, L., Llach, C. D., Valentí, M., Vieta, E., and Radua, J. (2021). Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. *Psychiatry Research: Neuroimaging*, 314:111313.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779.

Suresh, K. and Chandrashekara, S. (2012). Sample size estimation and power analysis for clinical research studies. *Journal of human reproductive sciences*, 5(1):7.

Takao, H., Hayashi, N., and Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. *Journal of Magnetic Resonance Imaging*, 34(2):438–444.

Takao, H., Hayashi, N., and Ohtomo, K. (2014). Effects of study design in multi-scanner voxel-based morphometry studies. *Neuroimage*, 84:133–140.

Teipel, S. J., Wegrzyn, M., Meindl, T., Frisoni, G., Bokde, A. L., Fellgiebel, A., Filippi, M., Hampel, H., Klöppel, S., Hauenstein, K., et al. (2012). Anatomical mri and dti in the diagnosis of alzheimer's disease: a european multicenter study. *Journal of Alzheimer's Disease*, 31(s3):S33–S47.

Thanh-Tung, H. and Tran, T. (2020). Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE.

Thesen, S., Heid, O., Mueller, E., and Schad, L. R. (2000). Prospective acquisition correction for head motion with image-based tracking for real-time fmri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 44(3):457–465.

Tian, D., Zeng, Z., Sun, X., Tong, Q., Li, H., He, H., Gao, J.-H., He, Y., and Xia, M. (2022). A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. *NeuroImage*, page 119297.

Timmermans, C., Smeets, D., Verheyden, J., Terzopoulos, V., Anania, V., Parizel, P. M., and Maas, A. (2019). Potential of a statistical approach for the standardization of multi-center diffusion tensor data: A phantom study. *Journal of Magnetic Resonance Imaging*, 49(4):955–965.

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., et al. (2021). A multi-scanner neuroimaging data harmonization using ravel and combat. *NeuroImage*, 245:118703.

Torbati, M. E., Minhas, D. S., Laymon, C. M., Maillard, P., Wilson, J. D., Chen, C.-L., Crainiceanu, C. M., DeCarli, C. S., Hwang, S. J., and Tudorascu, D. L. (2023). Mispel: A supervised deep learning harmonization method for multi-scanner neuroimaging data. *Medical image analysis*, 89:102926.

Tudorascu, D. L., Karim, H. T., Maronge, J. M., Alhilali, L., Fakhran, S., Aizenstein, H. J., Muschelli, J., and Crainiceanu, C. M. (2016). Reproducibility and bias in healthy brain segmentation: comparison of two popular neuroimaging platforms. *Frontiers in neuroscience*, 10:503.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320.

Wang, R., Chaudhari, P., and Davatzikos, C. (2021). Harmonization with flow-based causal inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 181–190. Springer.

Wilcock, D., Jicha, G., Blacker, D., Albert, M. S., D'Orazio, L. M., Elahi, F. M., Fornage, M., Hinman, J. D., Knoefel, J., Kramer, J., et al. (2021). Markvcid cerebral small vessel consortium: I. enrollment, clinical, fluid protocols. *Alzheimer's & Dementia*, 17(4):704–715.

Wrobel, J., Martin, M., Bakshi, R., Calabresi, P., Elliot, M., Roalf, D., Gur, R., Gur, R., Henry, R., Nair, G., et al. (2020). Intensity warping for multisite mri harmonization. *NeuroImage*, 223:117242.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., and Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human brain mapping*, 39(11):4213–4227.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*, 20(1):45–57.

Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Shen, D., Li, G., Consortium, U. B. C. P., et al. (2019). Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 475–483. Springer.

Zhong, J., Wang, Y., Li, J., Xue, X., Liu, S., Wang, M., Gao, X., Wang, Q., Yang, J., and Li, X. (2020). Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development. *Biomedical engineering online*, 19(1):1–18.

Zhu, A. H., Moyer, D. C., Nir, T. M., Thompson, P. M., and Jahanshad, N. (2019). Challenges and opportunities in dmri data harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 157–172. Springer.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Zuo, L., Dewey, B. E., Carass, A., Liu, Y., He, Y., Calabresi, P. A., and Prince, J. L. (2021a). Information-based disentangled representation learning for unsupervised mr harmonization. In *International Conference on Information Processing in Medical Imaging*, pages 346–359. Springer.

Zuo, L., Dewey, B. E., Liu, Y., He, Y., Newsome, S. D., Mowry, E. M., Resnick, S. M., Prince, J. L., and Carass, A. (2021b). Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, 243:118569.