Intensity warping for multisite MRI harmonization

J Wrobel[1*], ML Martin[2], R Bakshi[3, 4], PA Calabresi[5], M Elliot[6], D Raolf[7], RC Gur[6, 7, 8], RE Gur[6, 7, 8], RG Henry[9], G Nair[10], J Oh[5, 11], N Papinutto[9], D Pelletier[9], DS Reich[5, 10], W Rooney[12], TD Satterthwaite[7], W Stern[9], K Prabhakaran[7], N Sicotte[13], RT Shinohara[2], and J Goldsmith[1]
on behalf of the NAIMS Cooperative [14]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University
[2]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Laboratory for Neuroimaging Research, Partners Multiple Sclerosis Center, 7Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[4]Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[5]Department of Neurology, the Johns Hopkins University School of Medicine, Baltimore, MD, USA
[6]Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA, 19104, USA
[7]Brain Behavior Laboratory, Department of Psychiatry, Neuropsychiatry Section, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA
[8]Lifespan Brain Institute (LiBI) at the University of Pennsylvania and Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA
[9]Department of Neurology, University of California - San Francisco, San Francisco, CA, USA
[10]Translational Neuroradiology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
[11]St. Michael's Hospital, University of Toronto, Toronto, Ontario, Canada
[12]Advanced Imaging Research Center, Oregon Health & Science University, Portland, OR, USA
[13]Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[14]The North American Imaging in Multiple Sclerosis (NAIMS) Cooperative (https://journals.sagepub.com/doi/abs/10.1177/1352458517739990)
*jw3134@cumc.columbia.edu

May 14, 2019

Abstract

In multisite neuroimaging studies there is often unwanted technical variation across scanners and sites. These "scanner effects" can hinder detection of biological features of interest, produce inconsistent results, and lead to spurious associations. We assess scanner effects in two brain magnetic resonance imaging (MRI) studies where subjects were measured on multiple scanners within a short time frame, so that one could assume any differences between images were due to technical rather than biological effects. We propose *mica* (**m**ultisite **i**mage harmonization by **C**DF **a**lignment), a tool to harmonize images taken on different scanners by identifying and removing within-subject scanner effects. Our goals in the present study were to (1) establish a method that removes scanner effects by leveraging multiple scans collected on the same subject, and, building on this, (2) develop a technique to quantify scanner effects in large multisite trials so these can be reduced as a preprocessing step. We found that unharmonized images were highly variable across site and scanner type, and our method effectively removed this variability by warping intensity distributions. We further studied the ability to predict intensity harmonization results for a scan taken on an existing subject at a new site using cross-validation.

Key Words: intensity normalization, image harmonization, warping, curve registration, image densities, multisite imaging

1 Introduction

Medical imaging has become an established practice in clinical studies and medical research, leading to situations where images must be compared across site locations, scanners, or scanner types. Upgrades in scanner technology within a site may render old data not comparable to data collected on a newer machine, and this presents challenges in studies where acquisition techniques change over time. Multisite studies have become common as well; examples include large neuroimaging studies such as the Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005) and the Human Connectome Project (Van Essen et al., 2013), as well as targeted clinical trials studying multiple sclerosis (MS) interventions such as Kappos et al. (2006) and Hauser et al. (2017).

Measurement across multiple sites and scanners introduces unwanted technical variability in the images (Schnack et al., 2004). Going forward we will refer to technical artifacts introduced across either sites or scanners as "scanner effects." Scanner effects in imaging studies can reduce power to detect true differences across images and distort downstream measurements of regional volumes, brain lesions, and other biological features of interest (Schnack et al., 2010; Jovicich et al., 2013; Cannon et al., 2014; Keshavan et al., 2016; Schwartz et al., 2019). In structural magnetic resonance imaging (MRI) studies, detection of scanner effects is particularly challenging because images are collected in arbitrary units of voxel intensity; as a result, raw MRI intensities are often not comparable across study visits even within the same subject and scanner. We refer to unwanted technical variability within the same scanner and subject that are due to arbitrary unit intensity values as "intensity unit effects." Though often conflated, intensity unit effects and scanner effects are distinct sources of unwanted technical variation and should be treated separately. We refer to methods intended to address intensity unit effects as "normalization" methods to distinguish them from methods intended to reduce scanner effects, which we term "harmonization" methods. Both scanner effects and unit effects are present in multisite MRI studies, and in practice they can be challenging to separate.

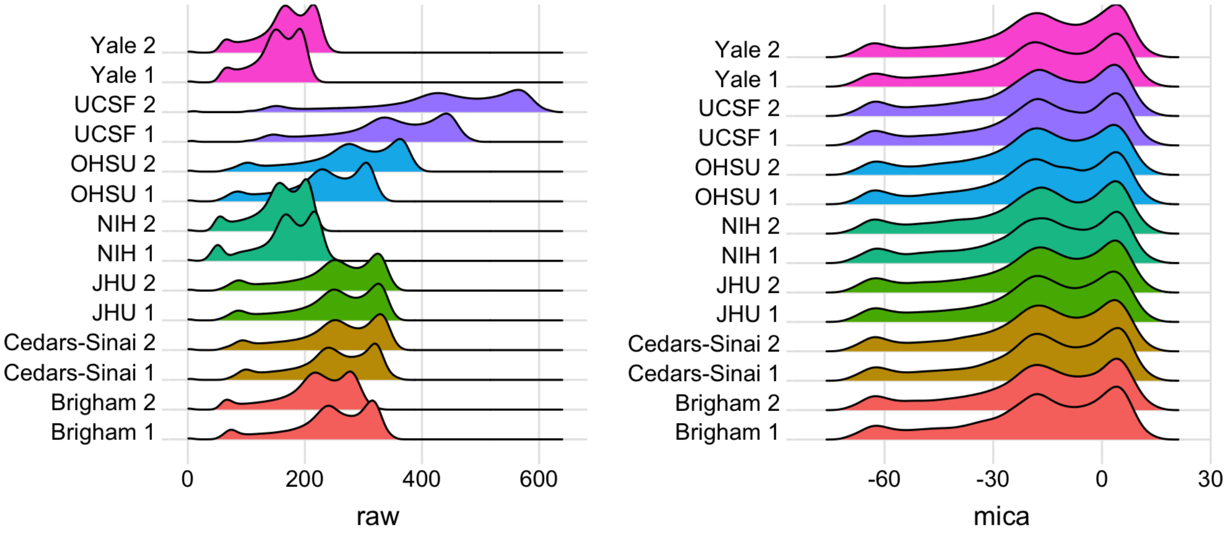


Figure 1: Histograms of voxel intensities for scan-rescan data across seven sites in the NAIMS pilot study: Brigham and Women's Hospital (Brigham), Cedars-Sinai, Johns Hopkins University (JHU), National Institutes of Health (NIH), Oregon Health & Sciences University (OHSU), University of California San Francisco (UCSF), and Yale University (Yale). Left panel shows raw voxel intensities; right panel shows densities after *mica* harmonization and White Stripe normalization. At each site two scans were collected; a 1 or 2 after site name indicates the first or second scan, respectively.

Scanner effects can be due to differences in scanner hardware, scanner software, scan acquisition protocol, or other unknown sources. When present, images collected at different sites may have systematically different distributions of intensity values. For example, Shinohara et al. (2017) showed substantial differences in volumetrics across sites and scanner types even for a single, biologically stable subject measured under standardized protocols at the same field strength on platforms produced by the same vendor. The left panel of Figure 1 shows histograms of intensity values for this single subject, who was scanned twice at each of seven sites across the U.S. Large scanner effects are evident; smaller but visible differences within site show that intensity unit effects are present as well. In subsequent analyses, scanner effects produced inconsistent measurements of MS lesion volume both when lesions were segmented manually or by a variety of automated software pipelines (Shinohara et al., 2017) .

The issue of arbitrary units has long been recognized and is the subject of a large literature on intensity normalization (Nyúl et al., 1999; Shinohara et al., 2011, 2014; Ghassemi et al., 2015). Intensity normalization methods facilitate comparability across subjects measured on the same scanner and standardize voxel intensity values; for a review of several methods see Shah et al. (2011). Histogram matching is an early approach that aligns densities of voxel intensities to quantiles of an image template constructed from several control subjects. Though popular, histogram matching often fails to preserve biological characteristics of individual scans and removes useful information regarding variation among subjects. Shinohara et al. (2014) formalized the principles of image normalization and introduced the White Stripe method. White Stripe normalizes images using patches of normal appearing white matter (NAWM), so that rescaled intensity values are biologically interpretable as units of NAWM. White Stripe can effectively normalize white matter across subjects and is a useful preprocessing step for automated lesion segmentation in MS (Sweeney et al., 2013a,b; Valcarcel et al., 2018), but technical variability can remain in the gray matter.

Unlike intensity normalization methods, which target intensity unit effects, harmonization methods aim to reduce scanner effects so that downstream analyses are more comparable across sites and scanners (Fortin et al., 2017; Yu et al., 2018). Fortin et al. (2018) described a voxel-wise regression method, based on tools from genomics, that harmonizes cortical thickness measurements from MRI scans. This method succeeds in removing scanner effects for measurements extracted from each image; in contrast, our goal in the present study was to develop an effective harmonization method that can be applied to the entire brain. Similar tools from genomics are used to correct for scanner effects in multisite diffusion tensor imaging data (Fortin et al., 2017) and multisite functional MRI data (Yu et al., 2018). However, these harmonization methods require spatial registration to a population template, which can lower image resolution and make it challenging to detect important disease features such as MS lesions. Ideally, an all-purpose harmonization method would remove scanner effects from the whole brain without requiring that all subjects be spatially registered to the same template image.

In the past, "normalization" has been used to simultaneously address the problems we characterize as unit and scanner effects, although these are more correctly viewed as distinct problems. As a result, intensity normalization techniques such as histogram matching and White Stripe are often used to address harmonization issues (Schnack et al., 2004; Shinohara et al., 2014; Fortin et al., 2016). Unlike harmonization techniques mentioned previously, these normalization techniques can be applied to the whole brain, do not require spatial registration, and reduce intensity unit effects. When scanner effects are due to the same voxel intensity transformations used to reduce unit effects, the normalization techniques will reduce scanner effects as well. However, often they fail to reduce much of the variability across sites, especially when large nonlinear scanner effects are present. Additionally, histogram matching normalizes voxel intensities across images at the cost of removing biological variability across subjects which can distort structures and mask inter-subject differences of interest.

Here, we introduce a new image intensity harmonization framework for multisite studies. We use data in which a subject was scanned on multiple scanners closely enough in time that any image differences can be attributed to differences across acquisition platforms (scanner effects) rather than biological effects. Our objectives in this study were to (1) establish a method that removes scanner effects by leveraging multiple scans collected on the same subject, and, building on this, (2) develop a technique to estimate scanner effects in large multisite trials so these can be reduced with preprocessing steps. The first objective establishes a framework for understanding harmonization, and the second relates to the practical use of this framework in multisite studies. We propose **m**ultisite **i**mage harmonization by **C**DF **a**lignment (*mica*), which harmonizes images by aligning cumulative distribution functions (CDFs) of voxel intensities. Our approach estimates nonlinear, monotonically increasing transformations of the voxel intensity values in one scan such that the resulting intensity CDF perfectly matches the intensity CDF from a second ("target") scan. CDFs can be perfectly aligned using standard approaches to curve registration in the functional data analysis literature (Srivastava et al., 2011; Tucker et al., 2013; Wrobel et al., 2018). Although these intensity transformations, called warping functions, are defined using CDFs, they can be applied to voxel-level intensity values to produce a harmonized image. For a subject measured on different scanners in close succession, this allows us to identify and remove scanner effects; mappings established in this way can be used to reduce the impact of scanner effects in multisite studies.

We outline our harmonization approach using two data sets with distinct but related problems. The North American Imaging in Multiple Sclerosis (NAIMS) pilot study (Shinohara et al., 2017; Dworkin et al., 2018; Oh et al., 2018; Papinutto et al., 2018; Schwartz et al., 2019) found large scanner effects in a single subject with biologically stable MS, and we use these data to show that *mica* can reduce technical variability across sites while preserving the ability to detect MS lesions. A second study, which we refer to as the trio2prisma study, scanned ten healthy subjects on two different machines and found systematic nonlinear differences between the scanners. We used *mica* to harmonize images from the first scanner so that they are comparable to images collected on the second scanner; this demonstrates how our method can be used to create a mapping between scanners, and that scanner effects can be removed when data are available from both scanners for all subjects. Since scan-rescan data are often only available for a subset of study subjects, we also employed a leave-one-scan-out cross-validation approach to assess the utility of our harmonization method in this common setting. For both studies, we used *mica* to understand and, to the extent possible, remove scanner variability. We paired our method with White Stripe to remove intensity unit effects as well as scanner effects, though other intensity normalization methods could be used instead.

In the next section, we describe our data and the *mica* methodology. We then present the results of our technique in different settings, followed by a discussion.

2 Materials and Methods

2.1 Data and processing

2.1.1 NAIMS dataset

The NAIMS steering committee developed a brain MRI protocol relevant to MS lesion quantification (Shinohara et al., 2017). Using this protocol, two scans were collected at each of seven sites across the United States on a 45-year-old man with clinically stable relapsing-remitting MS. All scans were performed on 3T Siemens scanners (four Skyra, two TimTrio, and one Verio). At each site, scan-rescan imaging was performed on the same day, with the subject exiting the machine between scans. The participant was also assessed at the beginning and end of the study on the same scanner to confirm disease stability by clinical and MRI measures .

Each image was bias-corrected using the N4 inhomogeneity correction algorithm (Tustison et al., 2010), then brain extraction was performed using the FSL BET skull-stripping algorithm (Smith, 2002). After performing *mica* harmonization as described in Section (2.2), T1-weighted (T1-w) and fluid attenuated inversion recovery (FLAIR) images were White Stripe normalized (Shinohara, R T and Muschelli, J, 2018) to remove intensity unit effects and enable automated MS lesion detection using the MIMoSA (Valcarcel et al., 2018) software pipeline.

2.1.2 trio2prisma dataset

The trio2prisma data were collected from ten healthy subjects ages 19 to 29 at the University of Pennsylvania. For each subject, brain MRI scans were obtained on both a Siemens Trio machine and a Prisma scanner. Scans were performed between 2 and 11 days apart for each subject (mean

4.2 days), a time window in which we expect no significant structural changes in the brain. We focused on T1-w images for the trio2prisma data, though our method can be applied to other modalities as well. Images were bias-corrected, skull-stripped, and White Stripe normalized using the same algorithms described for the NAIMS data. Because normalization methods have often been used for harmonization in the past, we compared *mica* to White Stripe and histogram matching normalization. To assess method performance on this data, we compared white and gray matter segmentations for *mica*-harmonized images to White Stripe and histogram matching normalized images. All white and gray matter segmentations were obtained using multi-atlas Joint Label Fusion (Wang et al., 2013).

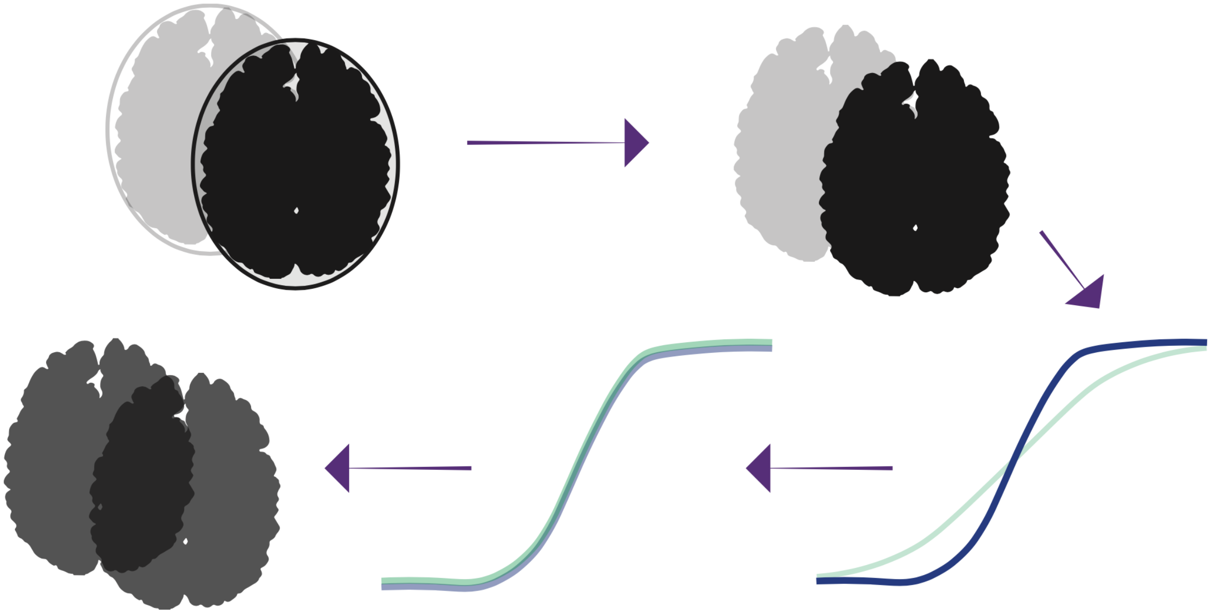


Figure 2: Harmonization pipeline. Raw images are N4 bias-corrected, skull-stripped, voxel intensities are converted to CDFs, CDFs are aligned by warping intensity values. The transformation of intensity values that produces this alignment is called a warping function, and the nonlinear transformation is applied to the raw images to produce harmonized images.

2.2 Methodology

Our framework for image harmonization uses non-linear transformations of image intensity values to remove scanner effects. The transformations were calculated by aligning distribution functions of intensity values. For a particular imaging modality (for example, T1-w), $Y_{ijk}(v)$ represents the intensity at a given voxel v for scan j of subject i measured at site k. Then $f_{ijk}(x)$ and $F_{ijk}(x)$ represent the probability density function (PDF) and CDF, respectively, for the voxel intensities of image Y_{ijk} measured over intensities x. Within each subject we assumed variability in voxel intensities across visits j and sites k is due to scanner and intensity unit effects rather than biological change, and that non-biological differences could be removed by aligning all CDFs for the i^{th} subject to a subject-specific "template CDF," $F_{it}(x)$, for template t; template choices for our motivating studies are described below.

For image Y_{ijk} we estimate the nonlinear monotonic transformation of the intensity values, or *warping function*, $h^{-1}_{ijk}(x) = \widetilde{x}$, which aligns the CDF $F_{ijk}(x)$ to its template via

$$F_{ijk}\left\{h^{-1}_{ijk}(x)\right\} = F_{ijk}(\widetilde{x}) = F_{it}(x). \tag{1}$$

After alignment, the CDF of the original images becomes identical to the CDF of the template. For this reason, we use the notation $F_{it}(x)$ to represent the *mica*-harmonized CDF as well as the template for alignment. We further denote $f_{ijk}\left\{h^{-1}_{ijk}(x)\right\} = f_{it}(x)$ and $Y_{it}(v)$ to be the *mica*-harmonized PDFs and images, respectively. The aligned PDFs, $f_{it}(x)$, can be recovered from CDFs by differentiation. The warping functions $h^{-1}_{ijk}(x) = \widetilde{x}$ define a new intensity value, $\widetilde{x}$, for each original intensity value in x. Since each $Y_{ijk}(v)$ is a voxel intensity in x, harmonized images Y_{it} take values in $\widetilde{x}$ and are obtained by $h^{-1}_{ijk}\left\{Y_{ijk}(v)\right\} = Y_{it}(v)$. Figure (2) shows a schematic of this process: images were bias corrected and skull-stripped, voxel-intensities were converted to CDFs, CDFs were aligned, and warping functions from CDF alignment were used to generate harmonized images.

Given this framework for quantifying scanner effects, we now address objectives (1) and (2) stated in Section (1). Our first objective, to establish a method that removes scanner effects, is illustrated using both the NAIMS and the trio2prisma data. For NAIMS data, we obtained empirical CDFs of T1-w and FLAIR images from the NAIMS dataset. Within an imaging modality, each CDF is given by $F_{ijk}(x), i = 1, j \in \{1, 2\}, k \in \{1, ..., 7\}$. We used the Karcher mean as the common template $F_{it}(x)$ to which all CDFs within a modality are aligned, though in principle other templates could be used. For the trio2prisma data, we obtained empirical CDFs of T1-w images. Each CDF is given by $F_{ijk}(x), i \in \{1, ..., 10\}, j = 1, k \in \{\text{Trio, Prisma}\}$. For each subject, we used the CDF from the Prisma image, $F_{i\text{Prisma}}(x)$, as the template to which we align the CDF from the Trio image, $F_{i\text{Trio}}(x)$. Functions from the `fdasrvf` `R` package (Tucker, 2017) were used to perform alignment.

Our second objective was to develop a technique to estimate scanner effects in large multisite trials; to illustrate this, we used warping functions from the trio2prisma data. In such studies, most subjects are only measured on a single scanner. At best, only a subset of subjects will have scans collected at all locations in the study. In order to harmonize scans for all subjects in this real-world setting, we propose to use *mica* to estimate warping functions for the subset of subjects who have multiple scans, average these warping functions across subjects; and use the resulting mean to harmonize images for subjects with only a single scan available. We assessed the performance of this approach using leave-one-scan-out cross validation in the trio2prisma data. Specifically, we removed the Prisma scan for one subject and computed the *mica* warping functions $\{h_i^{-1}(x)\}$ for the remaining subjects. We then computed the pointwise mean of these warping functions; using this as the warping function for the removed subject, we obtained a predicted Prisma scan from the known Trio scan. This process was repeated for each of the ten subjects. In the subsequent sections, scans harmonized using this leave-one-scan-out (*loso*) approach will be referred to below as *loso*-harmonized images and Trio scans harmonized using the full data will be referred to as *mica*-harmonized scans.

2.3 Statistical performance

All analyses were performed in the R software environment.

2.3.1 NAIMS data

To assess the performance of our method on the NAIMS data we quantified T2-hyperintense lesion volume from the 3D FLAIR and T1-w images in both the White Stripe normalized and *mica*-harmonized images using MIMoSA (Valcarcel et al., 2018) for automated lesion segmentation. Because the number and volume of lesions are important metrics for monitoring MS disease progression (Bakshi et al., 2008) and the evaluation of therapeutic efficacy (Filippi et al., 2006), eliminating non-biological variability in detected lesion volumes will help clinicians deliver the best possible care to their patients.

We quantified mean and variance of lesion volumes within and across sites after applying White Stripe alone and after applying *mica* followed by White Stripe.

2.3.2 trio2prisma data

For the trio2prisma data, we compared *mica* and *loso* to the histogram matching algorithm proposed by Nyúl et al. (1999), as implemented in Fortin et al. (2016). For better performance we first removed background voxels before running the histogram matching algorithm. To quantify performance of the methods we computed Hellinger distance of images before and after normalization, both within and across subjects. The Hellinger distance operates on PDFs of intensities, and its square is given by

$$h^2(f_l, f_k) = \frac{1}{2} \int \left(\sqrt{f_l(x)} - \sqrt{f_k(x)} \right)^2 dx \tag{2}$$

for PDFs $f_l(x)$ and $f_k(x)$. We visualized CDFs and PDFs and calculated Hellinger distances (Figures 4, 5, and 6, respectively) using images that had been *mica* or *loso*-harmonized but not yet White Stripe normalized. This is to isolate and visualize the effects of our method. For downstream analyses, including automated white and gray matter segmentation, we applied White Stripe normalization to the *mica* and *loso*-harmonized images to remove any residual intensity unit effects. We then estimated gray and white matter volumes and compare these across harmonization methods.

3 Results

For the NAIMS pilot data, we compared White Stripe normalized images to images processed using the *mica* approach outlined in section (2). For the trio2prisma data, we compared four harmonization strategies: no harmonization, histogram matching, *mica*, and *loso*. The main findings from these comparisons are summarized in the following two sections.

3.1 *mica* reduces variation in lesion volumes across sites in the NAIMS study

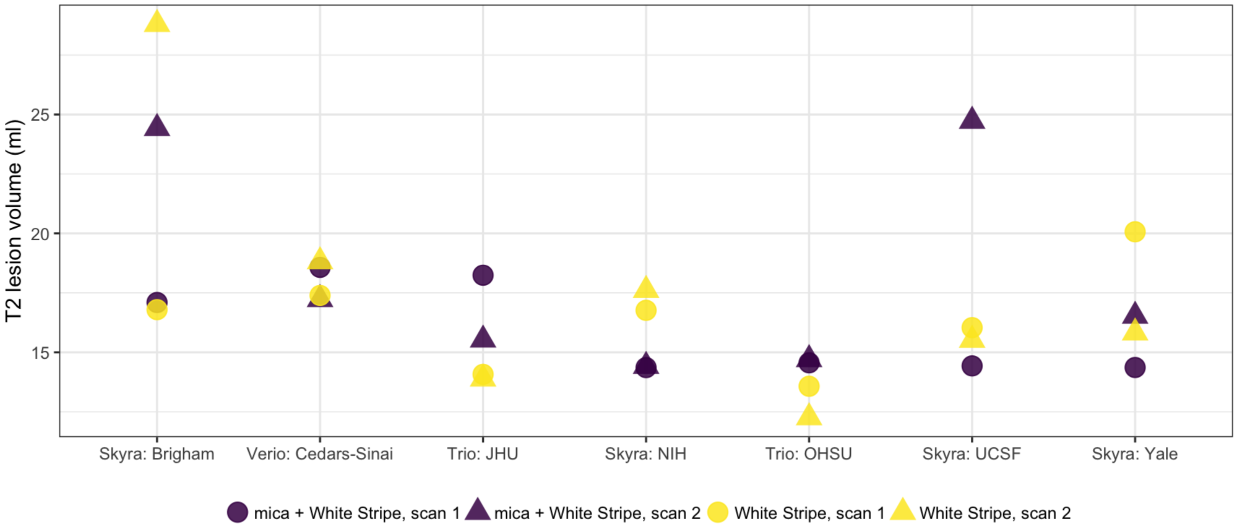


Figure 3: Estimated T2 lesion volumes for scan-rescan pairs at each of 7 sites in the NAIMS study. Circles indicate scan 1 and triangles indicate scan 2. Light and dark colors are volumes for White Stripe normalized images and *mica* normalized images, respectively.

We *mica*-harmonized then White Stripe normalized the NAIMS scans, and then quantified MS lesion volume to assess the effect of scanner variability on a common downstream analysis before and after *mica* harmonization. The left panel of Figure 1 shows PDFs of raw voxel intensities from the NAIMS study images, and the right panel shows PDFs of images that have been *mica*-harmonized then White Stripe normalized. The raw PDFs show small differences within site, which are attributable to intensity unit effects, and larger differences across site, which are attributable to scanner effects. Scanner effects are particularly large between the UCSF site and other sites. After *mica* harmonization, the images across and within site have the same distributions of voxel intensities.

Figure 3 shows estimated T2-hyperintense lesion volume across sites for both White Stripe alone and White Stripe in conjunction with *mica* for scan-rescan pairs across the seven NAIMS sites. Compared to White Stripe alone, *mica* in conjunction with White Stripe yielded less variable lesion volume measurements across sites (variance 11.8 ml^2 vs. 17.1 ml^2) and similar lesion volume measurements within sites (variance 12.4 ml^2 vs. 11.9 ml^2). We see a larger impact across sites than within sites, suggesting that our method decreases site-to-site variance as expected and, together with White Stripe, performs comparably to existing methods for within site variance.

3.2 *mica* preserves variation across subjects in the trio2prisma study

An appropriate harmonization method for multisite studies should reduce variability across scanners within the same subject but preserve biological differences across subjects. Here, we evaluate results from the trio2prisma data with these goals in mind. We compared *mica* and *loso*-harmonized images to images processed by histogram matching.

Figures 4 and 5 show CDFs and PDFs, respectively, under different harmonization scenarios. Visual inspection of intensity PDFs and CDFs in untransformed images suggests differences across scanners: the Prisma scans tend to have lower intensity values and higher peaks than the Trio scans. For both *mica* and histogram matching, within-subject technical variability is reduced because PDFs of Trio scans and Prisma scans are aligned. *mica* accomplishes this by mapping the Trio scan to the original Prisma scan, thus preserving the original features of the Prisma scans including variability across subjects. Histogram matching must be applied to scans from both the Trio and Prisma scanners, and reduces within-subject variability at the expense of eliminating desired differences across subjects. *loso* provides reasonable harmonization in that it maps Trio scans into the same range of intensity values as Prisma scans, but has less accuracy in reducing

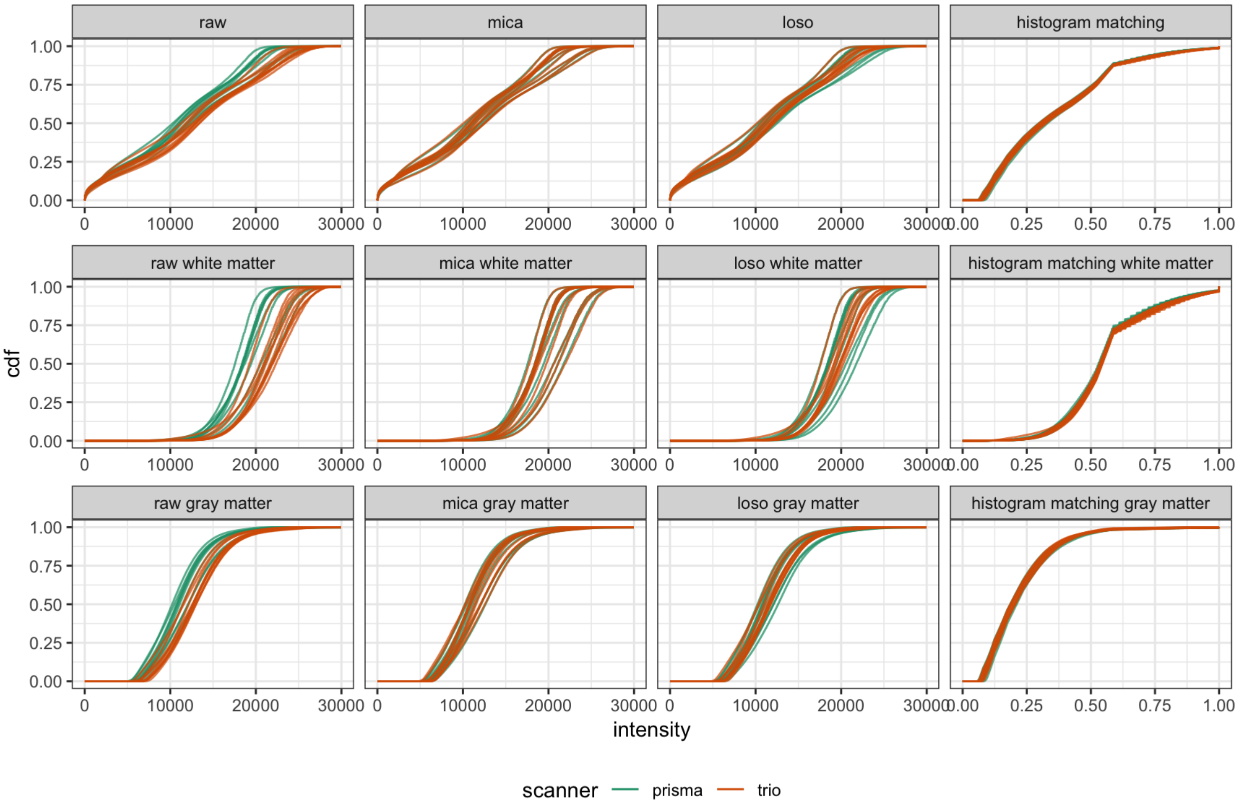


Figure 4: CDFs of intensities before and after harmonization by tissue type in the trio2prisma study. Rows indicate tissue type, with whole brain, white matter, and gray matter shown in rows 1, 2, and 3, respectively. Columns correspond to different harmonization methods.

within-subject variability than *mica* or histogram matching. However, much of the desired across-subject variability is retained.

We quantified the variability across subjects using the Hellinger distance from equation (2) on PDFs of voxel intensities. Figure 6 displays boxplots of these pairwise distances for the original Trio scans, original Prisma scans, and scans processed by histogram matching, *loso*, and *mica*. The figure is divided into distances calculated on the full skull-stripped images (left column), white matter (middle column), and gray matter (right column). The *mica*-harmonized Trio scans have similar across-subject variability to the Prisma scans. The *loso* scans have variability comparable to the original Trio scans but smaller than the Prisma scans. Histogram matching virtually eliminates inter-subject variability, including that which is presumably biological.

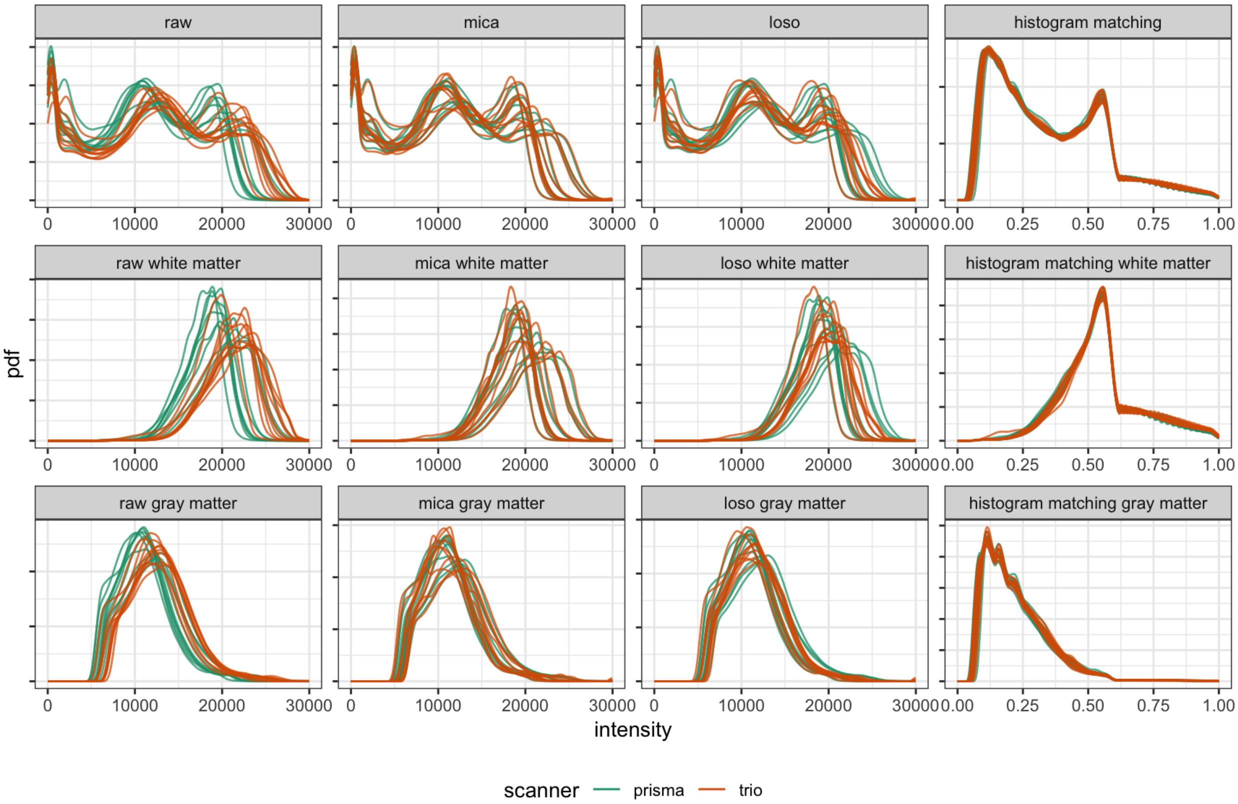


Figure 5: Histograms of intensities before and after harmonization by tissue type in the trio2prisma study. Rows indicate tissue type, with whole brain, white matter, and gray matter shown in rows 1, 2, and 3, respectively. Columns correspond to different harmonization methods.

Figure 7 shows an axial slice of the Trio image for one subject from the trio2prisma dataset. The slice is shown for raw intensity values (center), intensity values after *mica* harmonization (left), and intensity values after histogram matching (right). Here, *mica*-harmonization brightens the contrast between white and gray matter but does not distort the shape of biological features in the tissue. Histogram matching, however, drastically changes the appearance of the image, converting some gray matter to CSF and some white matter to gray matter.

Finally, neither harmonization nor normalization methods should bias assignment of tissue type. After harmonization or normalization, we expect that segmentation volumes from harmonized Trio scans should be similar to segmentation volumes from unharmonized and unnormalized (raw) Trio scans. We estimated white and gray matter volumes on original Trio scans and after histogram

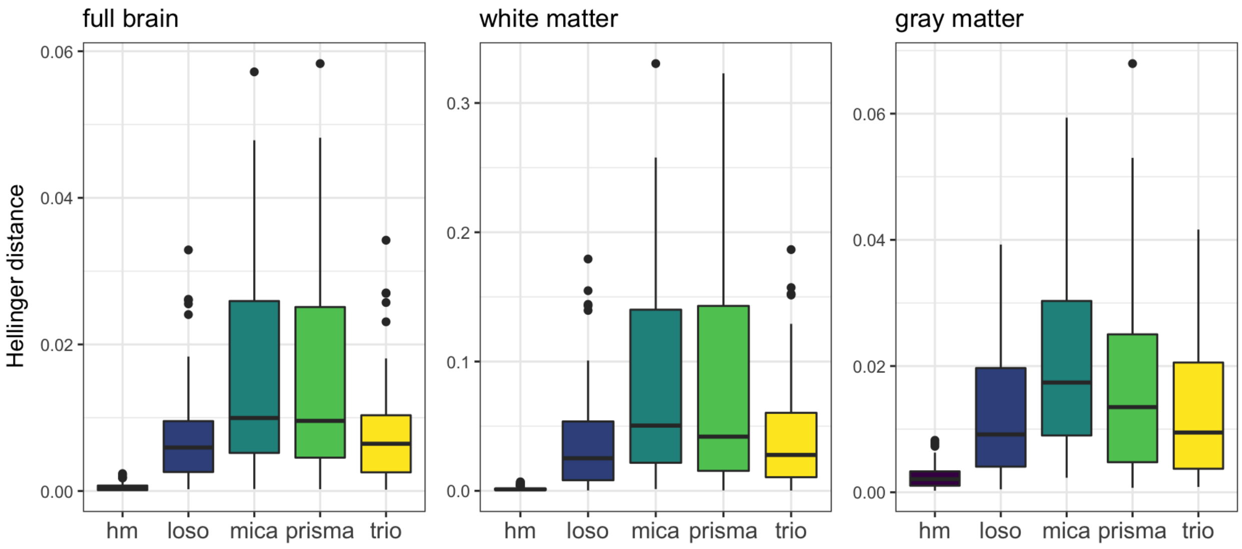


Figure 6: Boxplots of Hellinger distances across subjects, shaded by method. Columns show results for full brain (left), white matter (middle), and gray matter (right).

matching, White Stripe, *mica* followed by White Stripe, and *loso* followed by White Stripe. Figure 8 shows these volumes for each subject and tissue type. All methods have at least some difference in segmentation volume compared to the raw data. The *mica*, *loso*, and White Stripe methods all performed similarly, with volumes that are close to those of the raw images but slightly lower for the gray matter and slightly higher for the white matter. Histogram matching, however, had much lower segmentation volumes in both the gray matter and the white matter than either the raw data or any other method. As shown in Figure 7, histogram matching severely distorts the image; we believe this distortion causes the segmentation algorithm to convert some gray matter to CSF and some white matter to gray matter, which explains the consistently lower volumes.

4 Discussion

Unwanted technical variability due to scanner effects in multisite clinical trials and observational studies is an increasingly common problem; to mitigate these scanner effects we introdce *mica*, a method that harmonizes structural MRI images by defining nonlinear transformations between

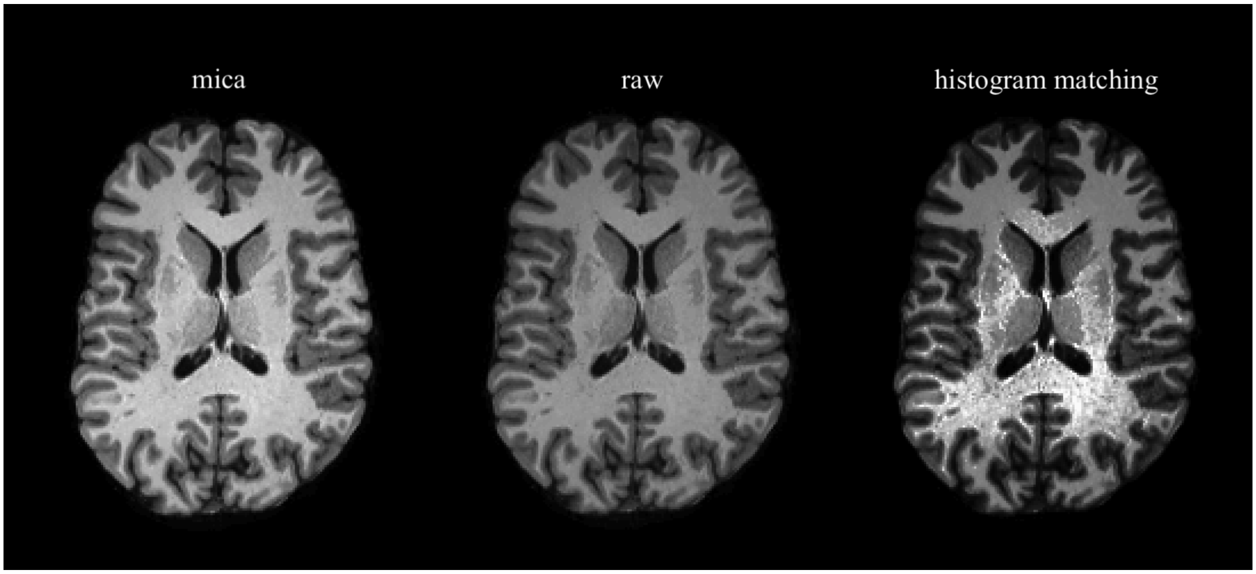


Figure 7: Axial slice of skull-stripped images from a single subject in the trio2prisma dataset. Center panel shows the raw intensity values from an image collected on the Trio scanner. Left and right panels show the same image after *mica* harmonization and histogram matching, respectively.

CDFs of voxel intensities. To specifically target scanner effects, we developed a paradigm for understanding scanner effects and intensity unit effects as related but distinct sources of technical variability in MRI scans. Intensity unit effects are due to arbitrary MRI unit intensities within a single scanner, and scanner effects are unwanted technical artifacts introduced across scanners or sites. We also distinguish between approaches targeting these sources of variability: normalization methods address intensity unit effects, and harmonization methods, the focus of our study, address scanner effects.

Our data came from two small studies, the NAIMS pilot study and the trio2prisma study, with multiple images per subject taken on multiple scanners, and nonlinear scanner effects. We found that *mica* reduced within-subject variability in whole brain scans as well as white and gray matter while preserving biological variability across subjects. We also found that *mica*, paired with White Stripe, enhanced reproducibility of measurements of MS lesion volume across sites.

Normalization methods such as histogram matching and White Stripe are sometimes used for harmonization, but they are inadequate in cases where across-site differences are much larger than

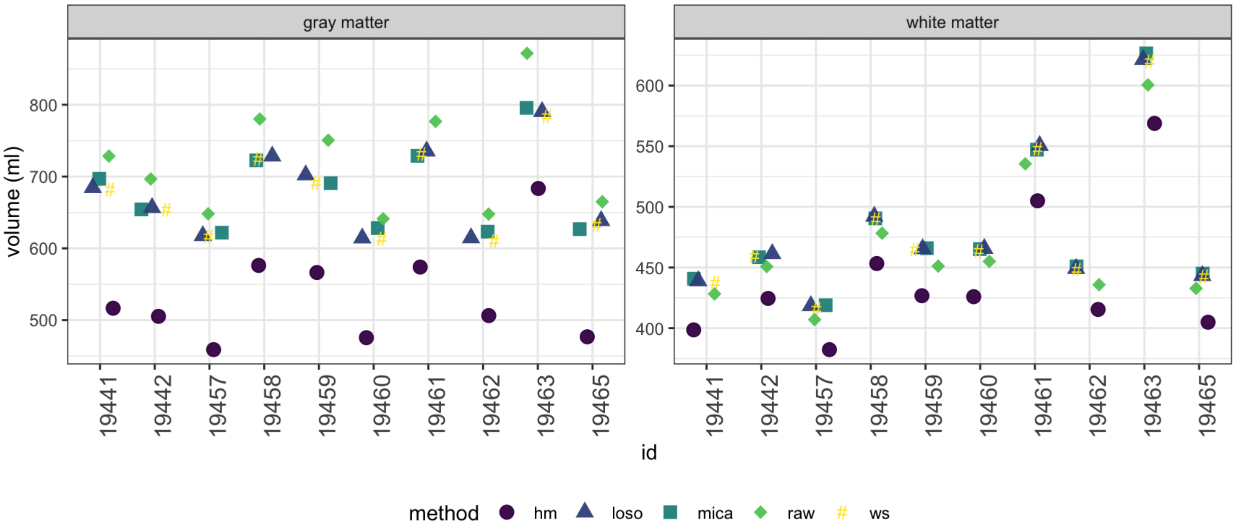


Figure 8: Segmented brain volume in the gray matter (left) and white matter (right) for each trio2prisma subject across harmonization approaches. We compare no normalization or harmonization (raw), histogram matching (hm), White Stripe normalization (ws), *mica*, and *loso*.

those within site. Additionally, histogram matching can reduce biological variability across subjects and White Stripe can leave residual technical variability in the gray matter. While we differentiate conceptually between intensity unit effects and scanner effects, we also acknowledge that in reality these artifacts can be challenging to separate. As a result, *mica* is likely to remove some intensity unit effects and intensity normalization methods are likely to remove some scanner effects when applied separately. In particular, White Stripe alone will likely perform well as a harmonization method when scanner effects are small, linear transformations. Histogram matching, however, is likely to remove desired variability across subjects and bias results.

Because our method is flexible and operates on the full brain, we can map images from one scanner to another. This mapping is only exact for a particular subject when images are available from both scanners, which is not realistic for most studies. That said, our leave-one-scan-out analysis suggests that when systematic site differences are present, *mica* can help understand scanner effects and mitigate those differences. Before conducting multisite studies, we recommend obtaining a baseline measurement of scanner variability by having a subset of patients measured at all sites.

Our method can then be applied to all images collected to remove average scanner variability. We acknowledge that this solution is imperfect in the sense that average scanner variability collected from a subset of patients in a trial will not always capture the true scanner variability for each subject. However, our simple and easy-to-apply methodology is an important step forward for an increasingly prevalent problem. There is evidence that scanner effects may vary across covariates such as gender and age, so extensions to *mica* that incorporate covariates may address some of the issues outlined above.

5 Software

To enable use of *mica* we have written an *R* software package which is available for download at https://github.com/julia-wrobel/mica.

Acknowledgements

This work was supported by the National Institutes of Health R01NS085211, R21NS093349, R01MH112847, R01NS060910, R01EB017255, R01HL123407, R01NS097423, and the National Multiple Sclerosis Society RG-1707-28586, and the Race to Erase MS Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Bakshi, R., Neema, M., Healy, B. C., Liptak, Z., Betensky, R. A., Buckle, G. J., Gauthier, S. A., Stankiewicz, J., Meier, D., Egorova, S., Arora, A., Guss, Z. D., Glanz, B., Khoury, S. J., Guttmann, C. R. G., and Weiner, H. L. "Predicting Clinical Progression in Multiple Sclerosis With the Magnetic Resonance Disease Severity Scale." Archives of Neurology, 65(11):1449–1453 (2008).
URL https://doi.org/10.1001/archneur.65.11.1449

Cannon, T. D., Sun, F., McEwen, S. J., Papademetris, X., He, G., van Erp, T. G., Jacobson, A., Bearden, C. E., Walker, E., Hu, X., et al. "Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis." Human brain mapping, 35(5):2424–2434 (2014).

Dworkin, J., Sati, P., Solomon, A., Pham, D., Watts, R., Martin, M., Ontaneda, D., Schindler, M., Reich, D., and Shinohara, R. "Automated Integration of Multimodal MRI for the Probabilistic Detection of the Central Vein Sign in White Matter Lesions." American Journal of Neuroradiology, 39(10):1806–1813 (2018).

Filippi, M., Wolinsky, J. S., Comi, G., Group, C. S., et al. "Effects of oral glatiramer acetate on clinical and MRI-monitored disease activity in patients with relapsing multiple sclerosis: a multicentre, double-blind, randomised, placebo-controlled study." The Lancet Neurology, 5(3):213–220 (2006).

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. "Harmonization of cortical thickness measurements across scanners and sites." NeuroImage, 167:104–120 (2018).

Fortin, J.-P., Parker, D., Tunc, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., et al. "Harmonization of multi-site diffusion tensor imaging data." Neuroimage, 161:149–170 (2017).

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

Ghassemi, R., Brown, R., Narayanan, S., Banwell, B., Nakamura, K., and Arnold, D. L. "Normalization of White Matter Intensity on T1-Weighted Images of Patients with Acquired Central Nervous System Demyelination." Journal of Neuroimaging, 25(2):184–190 (2015).

Hauser, S. L., Bar-Or, A., Comi, G., Giovannoni, G., Hartung, H.-P., Hemmer, B., Lublin, F., Montalban, X., Rammohan, K. W., Selmaj, K., Traboulsee, A., Wolinsky, J. S., Arnold, D. L., Klingelschmitt, G., Masterman, D., Fontoura, P., Belachew, S., Chin, P., Mairon, N., Garren, H., and Kappos, L. "Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis." New England Journal of Medicine, 376(3):221–234 (2017). PMID: 28002679.

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

Kappos, L., Antel, J., Comi, G., Montalban, X., O'Connor, P., Polman, C. H., Haas, T., Korn, A. A., Karlsson, G., and Radue, E. W. "Oral Fingolimod (FTY720) for Relapsing Multiple Sclerosis." New England Journal of Medicine, 355(11):1124–1140 (2006). PMID: 16971719.

Keshavan, A., Paul, F., Beyer, M. K., Zhu, A. H., Papinutto, N., Shinohara, R. T., Stern, W., Amann, M., Bakshi, R., Bischof, A., et al. "Power estimation for non-standardized multisite studies." NeuroImage, 134:281–294 (2016).

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. "Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI)." Alzheimer's & Dementia, 1(1):55–66 (2005).

Nyúl, L. G., Udupa, J. K., et al. "On standardizing the MR image intensity scale." image, 1081 (1999).

Oh, J., Bakshi, R., Calabresi, P. A., Crainiceanu, C., Henry, R. G., Nair, G., Papinutto, N., Constable, R. T., Reich, D. S., Pelletier, D., et al. "The NAIMS cooperative pilot project: Design, implementation and future directions." Multiple Sclerosis Journal, 24(13):1770–1772 (2018).

Papinutto, N., Bakshi, R., Bischof, A., Calabresi, P. A., Caverzasi, E., Constable, R. T., Datta, E., Kirkish, G., Nair, G., Oh, J., et al. "Gradient nonlinearity effects on upper cervical spinal cord area measurement from 3D T1-weighted brain MRI acquisitions." Magnetic resonance in medicine, 79(3):1595–1601 (2018).

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. "Mapping reliability in multicenter MRI: Voxel-based morphometry and cortical thickness." Human brain mapping, 31(12):1967–1982 (2010).

Schnack, H. G., van Haren, N. E., Hulshoff Pol, H. E., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T., Huttunen, M., Murray, R., and Kahn, R. S. "Reliability of brain volumes from multicenter MRI acquisition: a calibration study." Human brain mapping, 22(4):312–320 (2004).

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. "Multisite reliability and repeatability of an advanced brain MRI protocol." Journal of Magnetic Resonance Imaging (2019).

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

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

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. "Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis." American Journal of Neuroradiology, 38(8):1501–1509 (2017).

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. "Statistical normalization techniques for magnetic resonance imaging." NeuroImage: Clinical, 6:9–19 (2014).

Shinohara, R T and Muschelli, J. WhiteStripe: White Matter Normalization for Magnetic Resonance Images using
(2018). R package version 2.3.1.
URL http://CRAN.R-project.org/package=WhiteStripe

Smith, S. M. "Fast robust automated brain extraction." Human brain mapping, 17(3):143–155 (2002).

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. "Registration of Functional Data Using Fisher-Rao Metric." arXiv preprint arXiv, 1103.3817 (2011).

Sweeney, E., Shinohara, R., Shea, C., Reich, D., and Crainiceanu, C. "Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI." American Journal of Neuroradiology, 34(1):68–73 (2013a).

Sweeney, E. M., Shinohara, R. T., Shiee, N., Mateen, F. J., Chudgar, A. A., Cuzzocreo, J. L., Calabresi, P. A., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. "OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI." NeuroImage: clinical, 2:402–413 (2013b).

Tucker, J. D. fdasrvf: Elastic Functional Data Analysis (2017). R package version 1.8.1.
URL http://CRAN.R-project.org/package=fdasrvf

Tucker, J. D., Wu, W., and Srivastava, A. "Generative models for functional data using phase and amplitude separation." Computational Statistics and Data Analysis, 61:50–66 (2013).

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. "N4ITK: improved N3 bias correction." IEEE transactions on medical imaging, 29(6):1310–1320 (2010).

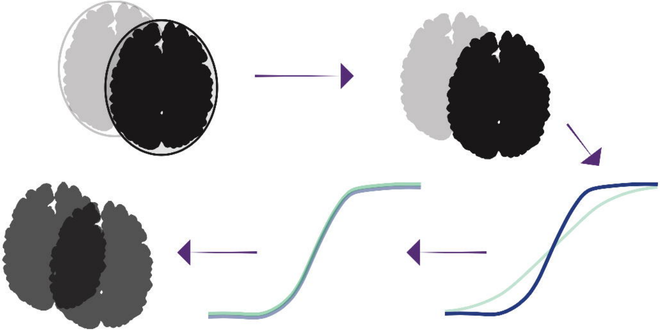
Valcarcel, A. m., Linn, K. A., Vandekar, S. N., Satterthwaite, T. D., Muschelli, J., Calabresi, P. A., Pham, D. L., Martin, M. L., and Shinohara, R. T. "MIMoSA: An Automated Method for Intermodal Segmentation Analysis of Multiple Sclerosis Brain Lesions." Journal of Neuroimaging (2018).

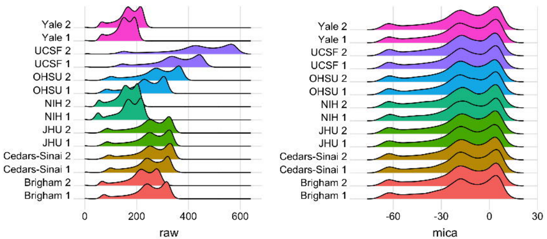
Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. "The WU-Minn human connectome project: an overview." Neuroimage, 80:62–79 (2013).

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. "Multi-atlas segmentation with joint label fusion." IEEE transactions on pattern analysis and machine intelligence, 35(3):611–623 (2013).

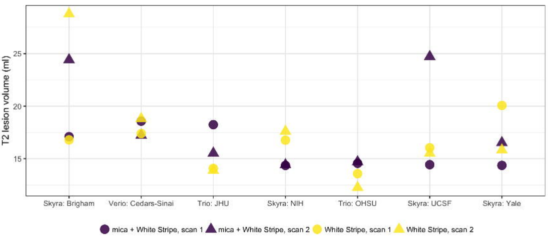
Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. "Registration for Exponential Family Functional Data." Biometrics (2018).

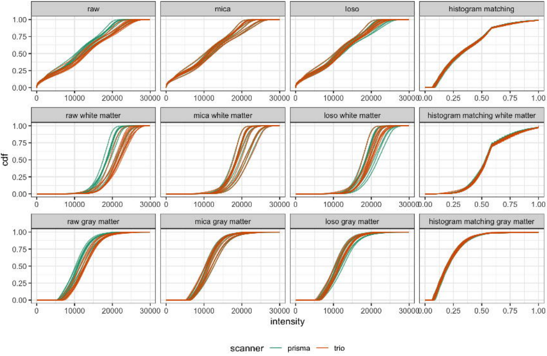
Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., and Sheline, Y. I. "Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data." Human brain mapping, 39(11):4213–4227 (2018).

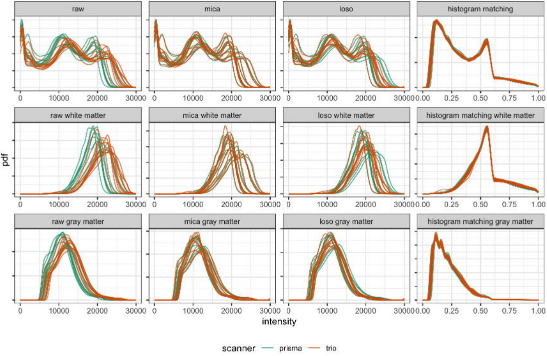


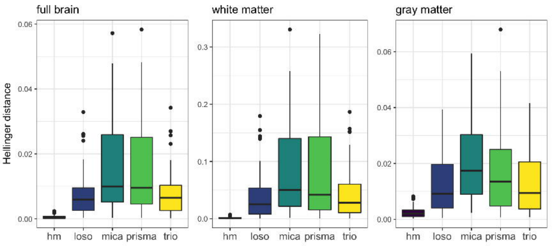


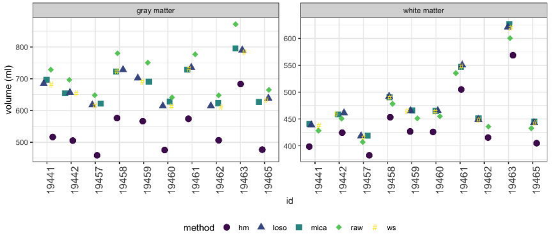
Yale 2
Yale 1
UCSF 2
UCSF 1
OHSU 2
OHSU 1
NIH 2
NIH 1
JHU 2
JHU 1
Cedars-Sinai 2
Cedars-Sinai 1
Brigham 2
Brigham 1
0
200
400
600
raw
Yale 2
Yale 1
UCSF 2
UCSF 1
OHSU 2
OHSU 1
NIH 2
NIH 1
JHU 2
JHU 1
Cedars-Sinai 2
Cedars-Sinai 1
Brigham 2
Brigham 1
-60
-30
0
30
mica


T2 lesion volume (ml)
25
20
15
Skyra: Brigham
Verio: Cedars-Sinai
Trio: JHU
Skyra: NIH
Trio: OHSU
Skyra: UCSF
Skyra: Yale
● mica + White Stripe, scan 1
▲ mica + White Stripe, scan 2
● White Stripe, scan 1
▲ White Stripe, scan 2


raw
mica
loso
histogram matching
raw white matter
mica white matter
loso white matter
histogram matching white matter
raw gray matter
mica gray matter
loso gray matter
histogram matching gray matter
cdf
1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
0.00
0.25
0.50
0.75
1.00
intensity
scanner
prisma
trio


raw
mica
loso
histogram matching
raw white matter
mica white matter
loso white matter
histogram matching white matter
raw gray matter
mica gray matter
loso gray matter
histogram matching gray matter
pdf
intensity
0
10000
20000
30000
0.00
0.25
0.50
0.75
1.00
scanner
prisma
trio


full brain
white matter
gray matter
Hellinger distance
0.00
0.02
0.04
0.06
0.0
0.1
0.2
0.3
0.00
0.02
0.04
0.06
hm
loso
mica
prisma
trio
hm
loso
mica
prisma
trio
hm
loso
mica
prisma
trio


gray matter
white matter
volume (ml)
500
600
700
800
400
450
500
550
600
19441
19442
19457
19458
19459
19460
19461
19462
19463
19465
id
method
● hm
▲ loso
■ mica
◆ raw
# ws