Magnetic Resonance Imaging 62 (2019) 59–69

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.elsevier.com/locate/mri

Original Contribution

Pathology-preserving intensity standardization framework for multi-institutional FLAIR MRI datasets☆

Brittany Reiche[a], A.R. Moody[b], April Khademi[c,*]

[a] School of Engineering, University of Guelph, Guelph, Canada
[b] Department of Medical Imaging, University of Toronto, Toronto, Canada
[c] Image Analysis in Medicine Laboratory, Ryerson University, Toronto, Canada

ARTICLE INFO

Keywords:
Brain
Fluid-attenuated inversion recovery
Intensity standardization
White matter lesions
Alzheimer's disease
Vascular disease
Segmentation

ABSTRACT

Fluid-Attenuated Inversion Recovery (FLAIR) MRI are used by physicians to analyze white matter lesions (WML) of the brain, which are related to neurodegenerative diseases such as dementia and vascular disease. To study the causes and progression of these diseases, multi-centre (MC) studies are conducted, with images acquired and analyzed from multiple institutions. Due to differences in acquisition software and hardware, there is variability in image properties, which creates challenges for automated algorithms. This work explores this variability, known as the MC effect, by analyzing nearly 5000 MC FLAIR volumes and proposes an intensity standardization framework to normalize intensity non-standardness in FLAIR MRI, while ensuring the appearance of WML. Results show that original image characteristics varied significantly between scanner vendors and centres, and that this variability was reduced with standardization. To further highlight the utility of intensity standardization, a threshold-based brain extraction algorithm is implemented and compared with a classifier-based approach. A competitive Dice Similarity Coefficient of 81% was achieved on 183 volumes, demonstrating that optimized pre-processing can effectively reduce the variability in MC studies, allowing for simplified algorithms to be applied on large datasets robustly.

1. Introduction

The economic burden for all neurological disease in Canada is estimated at $60 billion per year, or about 38% of the total burden presented by disease [1]. To reduce mortality rates and long-term disability, as well as to alleviate economic burden, the pathology of neurological disease must be understood so that interventions can be optimized. To this end, researchers have begun to investigate magnetic resonance images (MRI) of the brain to look for precursors and biomarkers of disease. One feature that has been identified on MRI are white matter lesions (WML), which are thought to be expressions of vessel disease [2], and are associated with ischemic stroke [3], Alzheimer's Disease (AD) [2] and demyelinating diseases such as multiple sclerosis.

To better understand the relationships between these diseases and WML, images from large patient cohorts must be analyzed. Accurate and quantitative calculation of WML volume, as well as other measurements, can be used to model disease progression, correlate with outcome/survival, and identify new risk factors [4–11]. Unfortunately, the visual analysis of medical images is subjective, error prone, and inefficient, which ultimately affects diagnostic accuracy and the ability to conduct large research studies [12,13]. Automated analysis techniques are a better alternative as they perform calculations in an objective, efficient, and reproducible manner.

Fluid-Attenuated Inversion Recovery (FLAIR) MRI has been gaining momentum in its use for diagnosis and treatment of neurodegenerative disease, and is becoming increasing important sequence for this task [14–18]. FLAIR is advantageous over T2-weighted sequences because the normally high signal of cerebral spinal fluid (CSF) is nulled, which allows the high signal of WML to be better visualized [14]. Although FLAIR is growing in popularity, only a few methods exist for its independent analysis, and novel algorithms that can operate on large patient databases are needed. However, designing algorithms for multi-centre (MC) databases is challenging due to the variability in data

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

* Corresponding author.
E-mail address: akhademi@ryerson.ca (A. Khademi).

https://doi.org/10.1016/j.mri.2019.05.001
Received 24 August 2018; Received in revised form 1 May 2019; Accepted 1 May 2019

caused by differing scanner vendor hardware, software and acquisition protocols used to create the images. Even if the same patient is scanned at the same centre, using the same scanner and acquisition parameters, it is possible that it would result in different intensity values and ranges for anatomy and pathology across the images. Collectively, we define this variability as the "multicentre effect (MCE)", which ultimately creates differences in pixel intensities, image contrast, and noise across datasets. Small changes in intensity values can have a large negative impact on the reliability of automated results, and therefore, the MCE does not permit for algorithms to be applied on all MC databases consistently. Many existing algorithms tend to be designed for a specific database (i.e. images acquired using identical protocols), and are not known to generalize well to other datasets due to the MCE and variability [18–22].

Intensity non-standardness is a major source of variability in MRI. A subject can have two scans of the same anatomy, with the same scanner and protocol, and will yield different pixel intensities for the same anatomies in the resultant images [23]. Such an effect is prevalent in MC data, resulting in the same tissues being expressed by different intensity values, which can severely affect automated algorithms [24]. In the analysis of longitudinal data, which is critical to modelling neurological disease progression, there is a need to accurately compare images from different time points. Without standardization, many of these differences may simply be variability created at the time of acquisition. Some works have attempted to standardize the FLAIR intensity scale [25]; however, whether the appearance of WML are altered or maintained during transformation of the intensity scale is unknown. Similarly, approaches for T1- and T2-weighted images [23] are not easily applied, as changes to the intensity scale can suppress the appearance of WML, reducing the clinical utility of the images.

For these reasons, this work is focused on the design of an intensity standardization framework for FLAIR MRI, which makes way for large-scale neurological studies of MC data. This framework normalizes variability in databases of FLAIR MRI acquired by different imaging centres and scanner vendors. This approach is designed solely for FLAIR MRI, which eliminates dependence on other sequences, such as T1 and T2, which is common in many FLAIR analysis methods. Standardization ensures that longitudinal images are normalized, allowing for the robust and consistent analysis of each time point. Additionally, algorithms do not need to be modified for different cases within the same database. This is one of the first works that provides an analysis and framework that can manage intensity variability in MC FLAIR data, using only the FLAIR modality. The methodology in the paper is the subject of a pending PCT patent application [26].

The framework is validated using nearly 5000 FLAIR MRI volumes (approximately 200,000 image slices) acquired at over 60 centres from subjects with vascular disease and various stages of dementia, making it one of the largest studies of its kind. To further highlight the utility of standardization, the performance of a simple, thresholding-based brain extraction tool is validated on 183 volumes, acquired from 31 centres, demonstrating that optimized pre-processing can effectively reduce the variability in MC studies, allowing for simplified algorithms to be applied on large datasets robustly.

2. Methods and materials

In this section, first, the standardization framework is described, which is used to normalize differences created by acquisition noise, bias field, and intensity non-standardness, resulting in images with similar intensity distributions across the datasets. The major components that are used to standardize the images are shown in Fig. 1. This is followed by a section detailing the validation of these techniques. Each pixel within an image is defined as $I(x,y,z)$ and the atlas (a template of the brain - more details can be found in Section 4.2) is defined as $A(x,y,z)$, where x, y are the spatial coordinates in each image, and z represents the slice number in each volume. Fig. 1 shows a flowchart of the proposed framework.

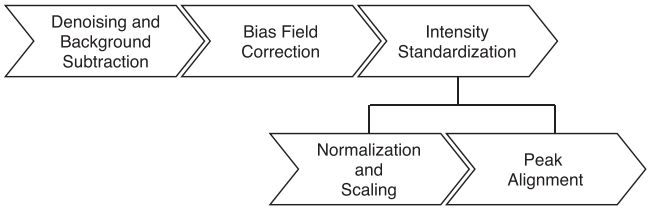


Fig. 1. The standardization framework.

2.1. Standardization framework

2.1.1. Denoising and background subtraction

To remove high frequency acquisition noise, a median filter was implemented in the spatial domain, as it is simple, and does not significantly modify the image while performing denoising [27]. The filter is defined as:

$$g(x,y) = median\{I(x,y), (x,y) \in w\}, \quad (1)$$

where $g(x,y)$ is the resultant image and w is the kernel window.

For background subtraction, the upper and lower 2% of the histogram were cropped to remove spurious noise and provide robust intensity limits; this process is also known as a Percentile Contrast Stretch. A K-Means classifier (k = 2) was applied to segment the denoised image into foreground and background. K-Means was used as it is able to discriminate between low intensity background pixels and higher intensity tissue pixels without having to implement hard thresholds, making it more robust to MC variability. This was calculated using the 3D volume. The background mask was then used to zero out all non-tissue pixels, thus suppressing background noise.

2.1.2. Bias field correction

Using the foreground mask, bias field correction was performed in a way similar to [28], where the mode of the mask was used to fill in the background of the image to reduce edge effects in later steps. The image was divided by a low-pass filtered (LPF) version of itself, which represents the low-frequency bias field artifact. The resultant volume $b(x,y,z)$ is defined as:

$$b(x,y,z) = k \cdot \frac{g(x,y,z)}{LPF(g(x,y,z))}, \quad (2)$$

where LPF means low-pass filter and k is the non-zero mode of the image. The resultant volume is then multiplied by the foreground mask to suppress border effect. This calculation is performed on the 3D volume, and removes low frequency variations in intensities within the same tissue class. Parameters for denoising and bias field correction were optimized previously [37].

2.1.3. Intensity standardization

For effective intensity standardization, the histograms of all images should be similar, so that the tissue classes from every image occupy the same intensity intervals. To accomplish this, a novel method inspired by [29] was developed to reflect the histogram characteristics of neurological FLAIR MRI. While [29] scales image histograms to have similar means and standard deviations between reference and target images, our approach also scales the histograms, but adjusts the standard deviation in a way that does not affect the appearance of WML. The goal of this approach is to maximally align the intensities of the gray matter and white matter tissue classes in all histograms within a database, while preserving the appearance of WML. A FLAIR template atlas [30] is used as the gold standard to which all other images were matched.

This algorithm performs intensity standardization on unimodal histograms in stages, as outlined in Fig. 1. First, the image histograms

are normalized and scaled so that they have similar magnitudes for comparison. Peak alignment is then performed by applying a global, linear re-scaling of the entire histogram with a factor determined by the difference between the mode of an atlas and a test image. In contrast to other works that use piece-wise linear transformations between landmarks [23], this method allows the bounds of the histogram to expand or contract without restriction and does not require explicit or implicit tissue segmentation first. This is especially important for maintaining the appearance of WML, as other methods can wash out their appearance or reduce contrast between WML and brain tissue.

2.1.3.1. Normalization and scaling. The normalized histograms were calculated, so that the magnitudes of the counts would be similar for comparison:

$$h(n) = \frac{v_i}{L \cdot M \cdot N}, \tag{3}$$

where $h(n)$ is the histogram (h_I for the image, and h_A for the atlas), n is the number of intensity bins, $L \cdot M \cdot N$ are the respective image dimensions, and v_i is the number of pixels in the image with intensity i. This calculation does not actually affect the intensities in the images; rather, it simply normalizes the histogram frequencies for analysis as a percent of the image, as images of size 256 × 256 have significantly less pixels than those of 560 × 560 images.

Due to background suppression, bins at low intensities of the histogram tend to be empty. To utilize the full range, the first non-zero intensity with a non-zero number of counts is selected, and will be referred to as τ_I. This intensity is subtracted from the entire volume:

$$c(x,y,z) = b(x,y,z) - \tau_I, \tag{4}$$

where $c(x,y,z)$ is the resultant image, and the intensity range of non-background pixels now start at one.

2.1.3.2. Alignment of GM/WM peaks. In FLAIR images, the gray matter (GM) and white matter (WM) tissue classes appear as a single peak in the histogram, and this landmark will be referred to as the GM/WM peak. To find the location of this peak in the atlas histogram, h_A, the maximum count is found:

$$A_{GM/WM} = n\text{argmax}\, h_A(n), \tag{5}$$

where $A_{GM/WM}$ is the intensity that corresponds with the GM/WM peak in the atlas. The same landmark is detected in the image as well, and is denoted as $I_{GM/WM}$. The ratio between the peaks is defined as:

$$\alpha = \frac{A_{GM/WM}}{I_{GM/WM}}, \tag{6}$$

where α is a multiplicative factor that is used to align the image GM/WM peak with that of the atlas, as in:

$$d(x,y,z) = \alpha \cdot c(x,y,z), \tag{7}$$

yielding the resultant image, $d(x,y,z)$. Intensity bins at the upper end of the histograms tend to be sparsely populated (some bins would contain no counts) after this alignment. To fully utilize the intensity scale, all bins with counts of zero were deleted from this histogram; this resulted in the decrementation of the upper pixel values in the image.

2.2. Validation methods

To ensure that the framework is effectively suppressing variability, several validation metrics were calculated. This section is split into the validation of intensity standardization, followed by the brain extraction algorithm that is applied on MC, standardized images, which highlights the utility of the standardization method.

2.2.1. Histogram comparison

To measure the similarity between images, the histogram of each image is calculated, and the Kullback-Leibler (KL) divergence between all image histograms and the mean histogram of all images before and after standardization is computed as:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx \tag{8}$$

where P and Q represent the histograms of two different images. The KL divergence was calculated between each dataset and the mean histogram of all images across all datasets in order to quantify the similarity of intensity distributions before and after standardization.

An independent samples t-test was conducted to analyze significant changes to the KL divergence following standardization, as improvement in this metric implies better alignment of intensity intervals, yielding more consistency between tissue intensities in different images.

2.2.2. Pathology preservation

To measure whether contrast was maintained between WML and the surrounding tissues, a WML segmentation method [31] was applied to both original and standardized images. This measurement is important, as a reduction in this contrast could reduce the ability of an algorithm to discriminate between WML and GM/WM classes.

With the WML masks, the Contrast Improvement Ratio (CIR) was calculated [32,33] as a percentage:

$$CIR = \frac{\sum_{(x,y)\in\mathbb{R}} |c(x,y) - \tilde{c}(x,y)|^2}{\sum_{(x,y)\in\mathbb{R}} c(x,y)} \times 100\%, \tag{9}$$

where c and $\tilde{c}$ represent the local contrast before and after standardization, respectively. Local contrast was computed as:

$$c(x,y) = \left|\frac{\mu_O - \mu_N}{\mu_O + \mu_N}\right|, \tag{10}$$

where μ_O is the mean value of the lesion, and μ_N is the mean value of the surrounding tissues. Each lesion was dilated with a disk size of fifteen, and the original lesion mask was subtracted; therefore, no lesions are included in this mask. This yields the neighbourhood region of the GM/WM around the WML. The lesion mask itself is used as the centre region. An increase in local contrast and CIR demonstrates that the boundaries of the WML are maintained, yielding edges that can allow forthe robust discrimination between WML and GM/WM tissue classes in further analysis. t-tests were used to compare local contrast measurements before and after standardization to verify significant improvement following standardization.

2.2.3. Brain extraction Performance

As a proof-of-concept, brain extraction was performed using a thresholding-based approach. As the images have a standardized intensity scale, the intensity boundaries of the brain should be similar across all images, making thresholding a viable option for segmentation.

By analyzing the histogram of the atlas template, bounds of [200 400] were selected to represent the boundaries of the GM/WM class. These thresholds were applied to the images, yielding a binary mask of the brain. However, as the intensities of WML lay outside of this range, they are often not included in this mask, and we must rely on post-processing to include them again.

A post-processing scheme that employs mathematical morphology was then applied to remove artifacts from the segmentations, as intensities found in WML are often also present in non-brain tissues, such as the skull [32]. First, the rough segmentation masks are eroded by a disk with a kernel size of 3, and all remaining small objects (i.e. non-brain tissues) are removed. The resulting mask is then dilated by a disk with a kernel size of 3% of the smallest image dimension (i.e. an image of size 256 × 256 would have a kernel of 8). Next, and holes within the mask are filled, and the mask is eroded with a disk with a kernel size of

3. Both erosion kernels were not dependent on image size, as only a small erosion was required to “clean up” the segmentations. This process yielded a binary mask encapsulating the brain tissue. This segmentation approach could not be applied to images prior to standardization, as the threshold boundaries would vary between images, and would not yield intelligent segmentations across large datasets.

Results from the proposed thresholding-based brain extraction was were compared to an existing Random Forest classifier-based approach [37]. The classifier uses intensity, position, and texture-based features to compute a brain segmentation. The classifier was trained using a portion of the dataset for which groundtruth masks were available, and tested on a holdout set.

Segmentation accuracy was objectively assessed using multiple metrics. To measure the amount of intersection between a segmented object and the groundtruth, the Dice Similarity Coefficient (*DSC*) [34] was calculated:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}, \tag{11}$$

where A and B are the binary masks of the brain for the groundtruth reference and automatic segmentation, respectively. The Hausdorff Distance (*HD*) was also calculated, which is a measure of maximum surface-to-surface distance [35]. It is calculated as the sum of distances between boundary points of the automatic segmentation and their closest neighbours in the groundtruth mask. In contrast to the *DSC*, this metric penalizes cases in which two overlapping objects still have different boundaries.

In addition to these metrics, classification accuracy was further quantified using sensitivity (*sens*), also known as Overlap Fraction, and is a measure of the true positive (*TP*) rate:

$$sens = \frac{TP}{TP + FN}, \tag{12}$$

where *FN* are false negatives. In addition, the specificity (*spec*) was calculated as a measure of the true negative (*TN*) rate:

$$spec = \frac{TN}{TN + FP}, \tag{13}$$

where *FP* are false positives. Extra Fraction [36] was also calculated, which is a measure of the false positive rate:

$$EF = \frac{FP}{TP + FN}. \tag{14}$$

In an ideal automatic segmentation, the *DSC*, specificity, and sensitivity measures should be close to one, while *HD* and *EF* should be close to zero.

2.3. Experimental data

2.3.1. Atlas template

A FLAIR template atlas [30], acquired on a Siemens 3 T scanner, with 0.5 × 0.5 × 1.0 mm resolution was used for standardization, created from images of 336 subjects with WML. The acquisition parameters are TR/TI/TE = 5000/1800/353 ms, with a flip angle of 180°.

2.3.2. Standardization Validation Datasets

A summary of data used to validate the standardization framework can be found in Table 1, and their respective demographic information can be found in Table 2. The Canadian Atherosclerosis Imaging Network (CAIN) database contains data from 236 subjects with cerebrovascular risk factors and a varying numbers of follow-up scans, yielding a total of 700 volumes. Images were acquired on scanners from three vendors (GE, Siemens, and Philips), from nine institutions, with variable acquisition parameters. The ischemic disease Sunnybrook (SB) database, which contains images acquired at 1.5 T, was also used to quantitatively validate standardization. The Alzheimer's Disease Neuroimage Initiative (ADNI) dataset was also used, as it contains longitudinal imaging data from 889 subjects, acquired at 58 imaging centres, resulting in a total of 4264 image volumes for analysis (ADNI-2 cohort). This dataset contains subjects within the following classifications: Normal, Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), Subjective Memory Concerns (SMC), and AD. This large and diverse dataset is an ideal representation of the MC centre problem that we are trying to solve, and will be a good indicator of the robustness of the framework.

Table 1
A summary of data used to validate the image standardization framework.

	CAIN	SB	ADNI
Disease	Vascular disease	Vascular disease	Alzheimer's disease
Total subjects	236	27	889
Total image volumes	700	27	4264
Total image slices	35,000	945	150,000
Centres	9	1	58
Scanner vendors	GE, Philips, Siemens	GE	GE, Philips, Siemens
Magnetic field strength (T)	3	1.5	3
TR (ms)	8000–11,000	8000	6000–11,900
TE (ms)	117–150	128	90–193
TI (ms)	2200–2800	2000	2000–2800
Pixel spacing (mm)	0.4286–1	0.5	0.7813–1.0156
Slice thickness (mm)	3	6	5

Table 2
Summary of demographic factors in the CAIN and ADNI databases.

		Women/men (no.)	Age [range]
CAIN	Vascular disease	95/141	73.3 ± 8.20, [50 94]
SB	Vascular disease	17/10	63.1 ± 24.4, [24 90]
ADNI	Normal	105/101	74.5 ± 6.75, [56.3 94.7]
	SMC	51/33	72.1 ± 5.56, [59.8 90.2]
	EMCI	132/167	71.4 ± 7.4, [55.2 88.7]
	LMCI	77/96	72.8 ± 7.8, [55.1 91.5]
	AD	51/76	74.7 ± 8.1, [55.7 90.4]

2.3.3. Brain extraction validation Dataset

A summary of the data used to validate brain extraction can be found in Table 3. For this task, 135 subjects from the CAIN dataset had binary masks of the brain available (evenly sampled from each centre and scanner vendor). Subjects from the ADNI database were also selected to validate brain segmentation. Twenty-one subjects were selected from different centres, with 7 subjects selected randomly for each scanner vendor. Subjects were also selected from all disease classifications in the ADNI database. The progression of the disease can affect the prevalence of WML, as well as the morphological characteristics of the brain, making this a diverse dataset that will be ideal for validating the robustness of the framework. All subjects (27 volumes) from the Sunnybrook database had brain masks available for validation. This yielded a total of 183 volumes for brain extraction validation. Groundtruth masks were generated using the Pathcore Sedeen Viewer[2] by the authors of this work and the protocol specified that the masks include all GM/WM structures. All groundtruth masks were generated prior to the development of this work, so algorithm performance did not affect how masks were drawn, and therefore, observers were blind to results.

As mentioned, the proposed thresholding-based brain extraction is also compared to a classifier-based approach [37]. The classifier was trained using a portion of the CAIN dataset for which groundtruth masks were available (108 volumes), and tested on a holdout CAIN set

[2] http://www.pathcore.ca/sedeen/

Table 3
A summary of data used to validate brain extraction via thresholding.

	CAIN	ADNI	SB
Total image volumes	135	21	27
Centres	9	21	1
Scanner vendors	GE, Philips, Siemens	GE, Philips, Siemens	GE
Magnetic field strength	3 T	3 T	1.5 T
TR (ms)	8000–11,000	650–11,900	8000
TE (ms)	117–150	20–193	128
TI (ms)	2200–2800	200–2800	2000
Pixel spacing (mm)	0.4286–1	0.7813–1.0156	0.5
Slice thickness (mm)	3	5	5

(27 volumes), as well as the ADNI set (21 volumes).

3. Results

This section details the experimental results and validation of the proposed standardization framework. The sections are split into intensity standardization validation and brain extraction validation. The major goals of this work is to analyze multicentre variability and the effects of MRI scanner vendor on FLAIR images, to demonstrate that effective standardization can simplify improve consistency in the intensity scale in multi-centre datasets while maintaining the appearance of WML and that such tools improve segmentation accuracy and reliability. Fig. 2 highlights MC dataset variability by summarizing measurements of mean signal intensity for each of the 700 volumes from the CAIN database as a function of scanner vendor. As shown by the standardized images, the mean intensities are similar across the nearly 700 volumes. Fig. 3 shows image histograms from the CAIN, SB, and ADNI datasets, from each scanner vendor, before and after standardization, and Fig. 4 shows sample standardized images. These figures demonstrate that within a given database, there is a range of intra- and inter-scanner variability; but also that effective standardization can reduce this variability, increasing the ease of which these images can be analyzed. The volume histograms are much more aligned in the standardized data, indicating that intensity ranges of tissues are being mapped to the same ranges. This is visually confirmed by inspecting the resultant images in Fig. 4, as the brain tissue across datasets have similar intensities after standardization.

3.1. Standardization validation

Standardization validation was conducted on the full CAIN (700 volumes), SB (27 volumes), and ADNI (4264 volumes) datasets. To demonstrate that the intensity standardization method normalizes image histograms and aligns the intensities of similar tissues to the same range, the KL divergence of histograms from all datasets (CAIN, ADNI, SB) were compared with the mean histogram of all images, both before and after standardization. This metric quantifies the similarity of the images with each other, where improvement in this metric following standardization indicates an increased similarity in image characteristics. Table 4 summarizes the results of this test for each scanner, and shows significant improvement for all scanner vendors; this effect is obvious when analyzing Fig. 3, as standardization clearly aligns the histograms within each scanner vendor, but also between vendors and datasets. Fig. 5 shows the histograms of the images with theworst and best KL divergence metrics, respectively. As can be seen, even the “worst” cases show close alignment, highlighting the robustness and consistency of this approach.

Fig. 6 shows the average histograms from each centre in the CAIN and ADNI studies (a total of 67 centres), before and after standardization. In Fig. 6(a), it can be seen that there is significant variability between the histograms of images acquired at different centres. In particular, it can be seen that the histograms associated with the ADNI study have relatively small peaks, indicating low contrast between tissues in the images; the locations of these peaks are also variable between centres, highlighting the lack of a standard intensity scale. However, Fig. 6(b) shows the histograms following standardization – the peaks of the histograms are now aligned, and the peaks corresponding to the GM/WM class are more prominent in all images. These results clearly demonstrate the utility of standardization: not only are the intensities associated with different tissues now aligned between images, but image properties, such as contrast, are normalized as well.

To ensure that the appearance of WML were not modified or altered in a negative way, contrast analysis was performed using areas surrounding WML. Based on inclusion criteria of a diameter greater than 3 mm, 3126 volumes acquired at 3 T, and all 27 volumes at 1.5 T were used for analysis. Average local contrast measurements for all scanners are shown in Table 4. It was found that there was significant improvement in local contrast for all images. This resulted in a mean CIR of 1.2% ± 1.52, and an improvement of contrast in 99% of images, with a mean improvement of 11%. Therefore, the appearance of WML were maintained in the standardized version.

3.2. Segmentation results

Fig. 7 shows sample results of brain extraction using the same treshold across the datasets. Table 5 contains the quantitative validation metrics for both the tresholding-based and classifier-based brain extraction tools. As shown, brain extraction via thresholding was successful, irrespective of pathology, scanner vendor, and acquisition parameters. This approach achieved a DSC of 81.8 ± 6.5 for 183 image volumes across different datasets. These results are competitive with those achieved using a classifier-based approach [37], implying that these segmentations are accurate, and therefore viable for further analysis. It should be noted that the classifier yielded an increased *DSC* for the CAIN dataset; this is because it was trained to generalize to WML, whereas the thresholding-based approach initially excludes WML, and works to “regain” them using mathematical morphology. The ADNI *DSC* is similar for both approaches; this is likely because the classifier was trained using only CAIN data, and did not generalize as well to the ADNI images due to the differences in pathology. For both

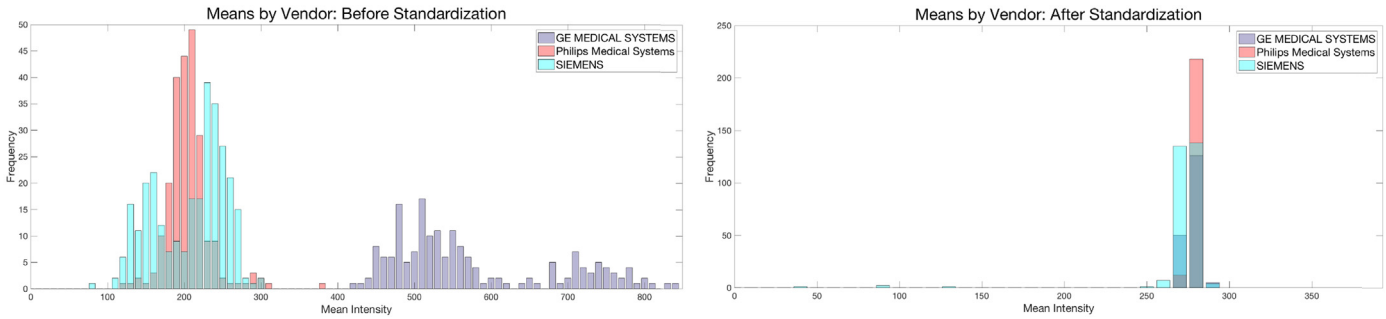
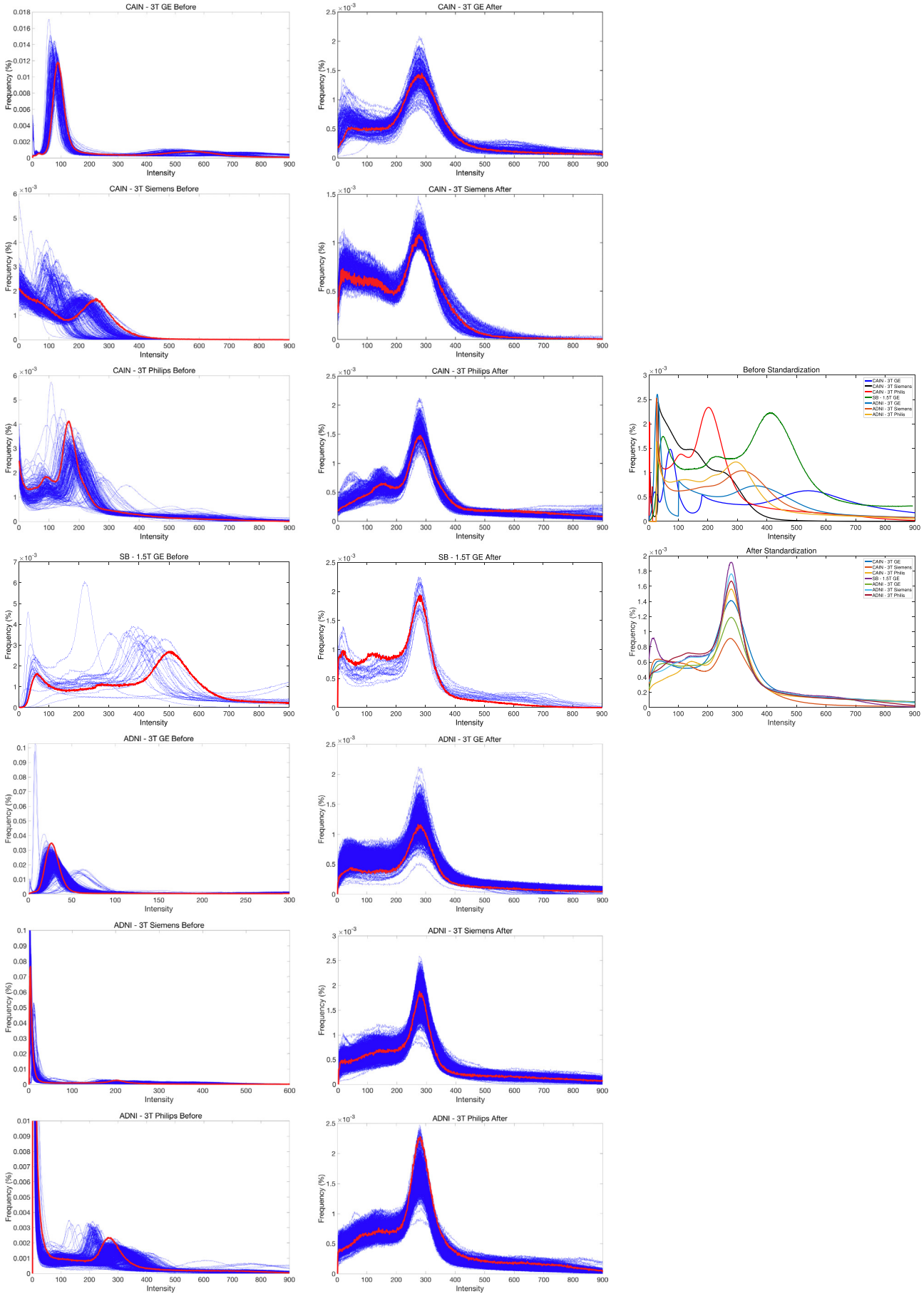


Fig. 2. Mean scanner signal between three different vendors in the CAIN dataset, before and after intensity standardization. Best viewed in colour.



(caption on next page)

Fig. 3. Scanner histograms before and after standardization for each dataset. Histograms of sample images in Fig. 4 are shown in red. Best viewed in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

datasets, the classifier yielded a marked decrease in false positives (Extra Fraction metric), which can likely be attributed to the fact that it does not rely on mathematical morphology (erosion and dilation) to compute the segmentation. To summarize, the classifier-based approach yielded better results, but provided a baseline for comparison of the proposed threshold-based approach. As can be seen, the very simple method proposed here yielded results that were close to the baseline. This demonstrates that thorough pre-processing of datasets can have a substantially beneficial effect on further processing and analysis.

In addition, it should be noted that brain segmentation via thresholding only takes a few seconds to compute; in [22], a convolutional neural network was implemented using Graphics Processing Units (GPUs), and computation took an average of 40 to 60 s. In a clinical setting, where GPU computation may not be available, these same calculations could take up to 12 times longer on conventional CPUs [38].

4. Discussion

The multi-centre (MC) effect describes the variability in image properties created by differences in scanning software and hardware between institutions. A major source of variability is intensity non-standardness, which severely impedes the ability of algorithms to robustly quantify disease in MC datasets. In this work, a novel intensity standardization framework for FLAIR MRI, which maintains the appearance of WML pathology, is presented and validated. Intensity standardization is performed in two stages. First, the volume histogram of each subject is pre-processed, which includes percentile trimming and shifting to ensure the histogram begins at zero. Second, the volume histogram is matched to a FLAIR template's volume histogram by peak

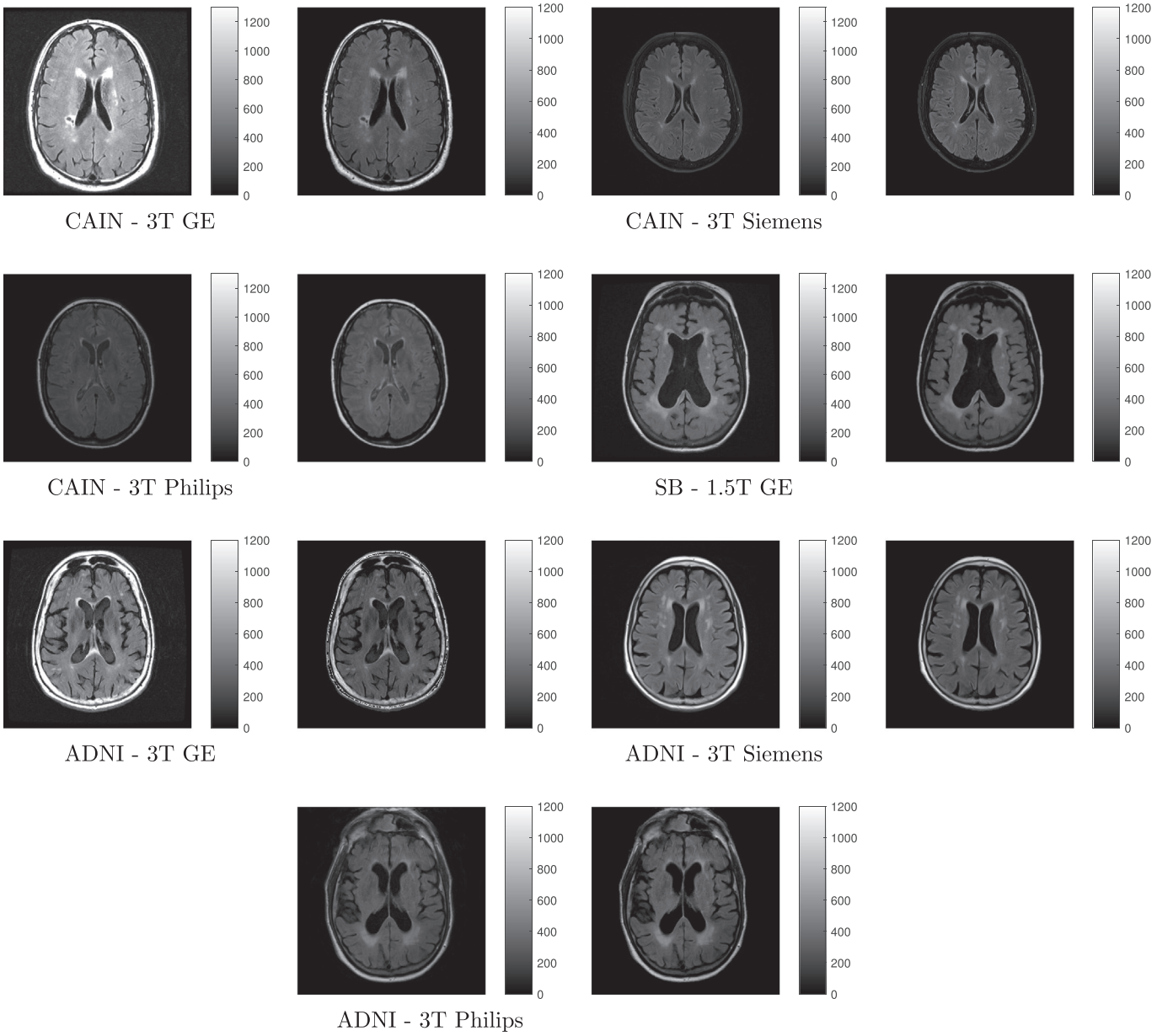


Fig. 4. Original images and results from standardization steps. Each original image in shown, followed by its standardized version to the right of the original image.

Table 4
Summary of Intensity Standardization Metrics for *T*-Tests: KL Divergence and Local Contrast. * indicates that results are at significance.

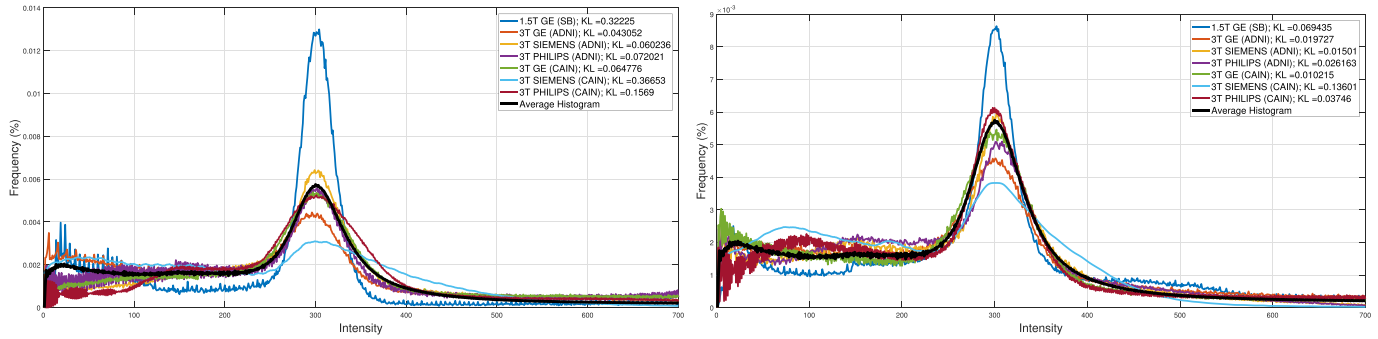
	Before	After
KL divergence		
1.5 T GE	1.482 ± 0.109	0.142 ± 0.010*
CAIN - 3 T GE	1.70 ± 0.416	0.037 ± 0.014*
CAIN - 3 T Philips	0.373 ± 0.086	0.097 ± 0.028*
CAIN - 3 T Siemens	0.553 ± 0.100	0.170 ± 0.039*
ADNI - 3 T GE	0.699 ± 0.408	0.039 ± 0.017*
ADNI - 3 T Philips	0.802 ± 0.592	0.042 ± 0.015*
ADNI - 3 T Siemens	1.202 ± 0.506	0.028 ± 0.012*
Overall	1.013 ± 1.635	0.094 ± 0.057*
Local contrast		
1.5 T GE	0.3093 ± 0.0617	0.3282 ± 0.0415*
CAIN - 3 T GE	0.378 ± 0.128	0.446 ± 0.118*
CAIN - 3 T Philips	0.376 ± 0.146	0.423 ± 0.148*
CAIN - 3 T Siemens	0.399 ± 0.171	0.449 ± 0.01*
ADNI - 3 T GE	0.272 ± 0.139	0.317 ± 0.132*
ADNI - 3 T Philips	0.228 ± 0.107	0.247 ± 0.082*
ADNI - 3 T Siemens	0.273 ± 0.104	0.280 ± 0.082*
Overall	0.3348 ± 0.058	0.339 ± 0.091*

detection and linear scaling. Validation was completed on over 5000 FLAIR MRI image volumes collected from over sixty imaging centres of patients with vascular disease and various stages of dementia, making it one of the largest studies of its kind. Several metrics were computed to quantify the performance of intensity standardization, such as the KL divergence of the image histograms before and after standardization with each other; *t*-test results showed that standardization improved KL divergence significantly over the all datasets, and results were presented as a function of dataset, scanner manufacturer (GE, Siemens, Philips), and magnetic field strength (1.5 T and 3 T). Visual histogram analysis of these centres also shows the alignment of histograms as a function of scanner vendor over different diseases.

Another novelty of this work is that the method preserves the appearance of pathology, and this was validated using a contrast-based feature, which shows that local contrast over the lesion boundaries was maintained or improved. This is an extremely important innovation: WML analysis schemes cannot compute accurate segmentations on images in which pathology has been altered by pre-processing. Therefore, this method should increase the robustness and accuracy of WML segmentation algorithms. This is contrasted to three other commonly used intensity standardization algorithms [23,29,39]. It was found that these methods were not as robust to MC data, as they did not account for WML, which were often blended into the surrounding GM/WM. In contrast to the previously cited work that uses piece-wise linear transformations to align histogram landmarks [23], the proposed method allows the histogram bounds to change without restriction, which is essential for maintaining the appearance of pathology and WML. Although the proposed approach does not *directly* account for variations in pathology (i.e. hydrocephalus, previous strokes and infarcts), the ability of the histogram bounds to change without restriction allows the relationships between pixel intensities to be retained regardless of lesion load, ensuring that the appearance of pathology is maintained while still standardizing the intensity scale. This also applies to the analysis of dirty-appearing white matter (DAWM), which is defined as regions of the brain with an intermediate intensity between those of WML and normal-appearing white matter [40], which is beginning to gain substantial attention in the research community. The authors hypothesize that the standardization of brain pixel intensities will also align the intensity range of DAWM, which may allow for future works to analyze this phenomenon more objectively.

Because the standardization framework could robustly suppress image variability and transform the intensity distributions of each image into the same space, a threshold-based method of brain extraction could be implemented that is based on the exact same thresholds for 183 volumes from 31 centres of patients with vascular disease and dementia. This approach achieved an average DSC of 81%, which approaches the range of accuracy produced by more complex approaches [37]. The relatively competitive accuracy achieved by this threshold-based technique highlights the true power of standardization: complex models are no longer required and a simple threshold can be identified that corresponds to the same anatomies in all images, from all scanners and centres. The main shortcoming of the threshold-based brain extraction technique was that it was not robust to WML located near the brain boundary, as thresholding created holes at these locations that cannot be filled with mathematical morphology. A classifier-based approach was shown to improve on this limitation. However, brain extraction via thresholding was proposed simply as a proof-of-concept, and demonstrates that thorough pre-processing of images can yield more consistent and robust results in subsequent analysis.

This work applied a standardization framework to three multi-centre datasets, and demonstrated that image standardization can significantly reduce variability in a MC database, regardless of scanner vendor, acquisition parameters and type of disease. Some works have previously proposed standardization approaches for MRI [23,25]; however, these approaches have not validated whether the appearance of WML are maintained. What differentiates this work from previous work in this regard is that the proposed standardization approach allows the intensity scale to be unbounded, which only imposes a lower bound on WML intensities. This allows for the standardization of the intensity scale without affecting the appearance of WML –the intensity distribution of WML is not assumed, and linear scaling allows for standardization without changing the appearance of WML. A sensitivity analysis of the effect of lesion load on standardization would be beneficial, and the authors hope to include it in future work that also demonstrates that standardization yields an increased performance of



(a) Worst Cases (b) Best Cases

Fig. 5. Histograms of images with worst and best KL divergence measurements following standardization, per scanner. This demonstrates that even "worst" alignment calculated via the KL divergence yields good overlap of similar tissue classes. Best viewed in colour.

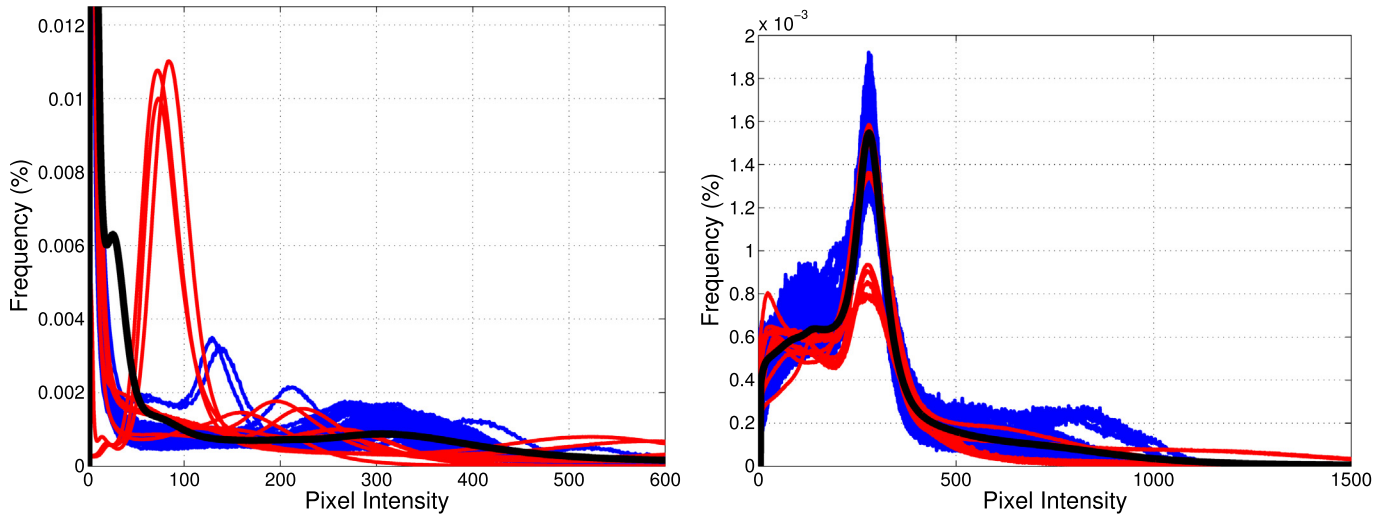


Fig. 6. Average histograms of each centre in the CAIN and ADNI datasets, before and after standardization, respectively. Red lines represent the 9 centres from CAIN, the blue lines are the 58 centres from ADNI, and the black line is the average histogram from both studies. Best viewed in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

WML segmentation approaches.

Although this investigation was conducted solely on FLAIR MRI, the authors believe that this concept can be applied to all MRI datasets, as standardization of the intensity scale can simplify algorithms for further analysis, while increasing accuracy and robustness. This framework permits for the consistent and robust processing of large populations of subjects with longitudinal data. With this information, it may be possible to correlate measurements derived from FLAIR MRI with patient outcomes in order to gain valuable insight into neurodegenerative disease pathology.

5. Conclusion

This work demonstrates that image standardization can significantly reduce variability in a MC database, regardless of scanner vendor and acquisition parameters. Pre-processing images via standardization allows for the application of simpler segmentation and quantification algorithms, which may be robustly applied to large datasets. As the intensity scale is standardized, automatic segmentation of pathology should result in more precise and accurate measurements when compared to non-standardized images.

As shown through the validation studies in this work (applied to more cases than other leading-edge approaches), the method robustly standardizes the intensity scale of FLAIR MRI regardless of lesion load and disease type, which allows large-scale studies to be efficiently conducted with simplified models, on a scale that would be too time-consuming with manual processing. As shown by the local contrast results, and investigated further in the discussion, the intensity standardization algorithm presented here ensures that the appearance of WML are maintained, which is an issue that has not yet been addressed in the literature and is critical for WML quantification. A threshold-based brain extraction method is also presented based on the intensity-standardized images; the achieved segmentation accuracy further demonstrates the utility of the proposed normalization algorithm.

This framework represents one of the first FLAIR MRI standardization frameworks that uses only the FLAIR sequence, and that focuses on the preservation of WML. We eliminate common challenges of processing FLAIR MRI with this method, which includes theneed to co-register FLAIR images with T1- and T2-weighted images, which increases computational time and errors associated with registration. In addition, standardization should make the analysis of longitudinal data more consistent and accurate across different time points.

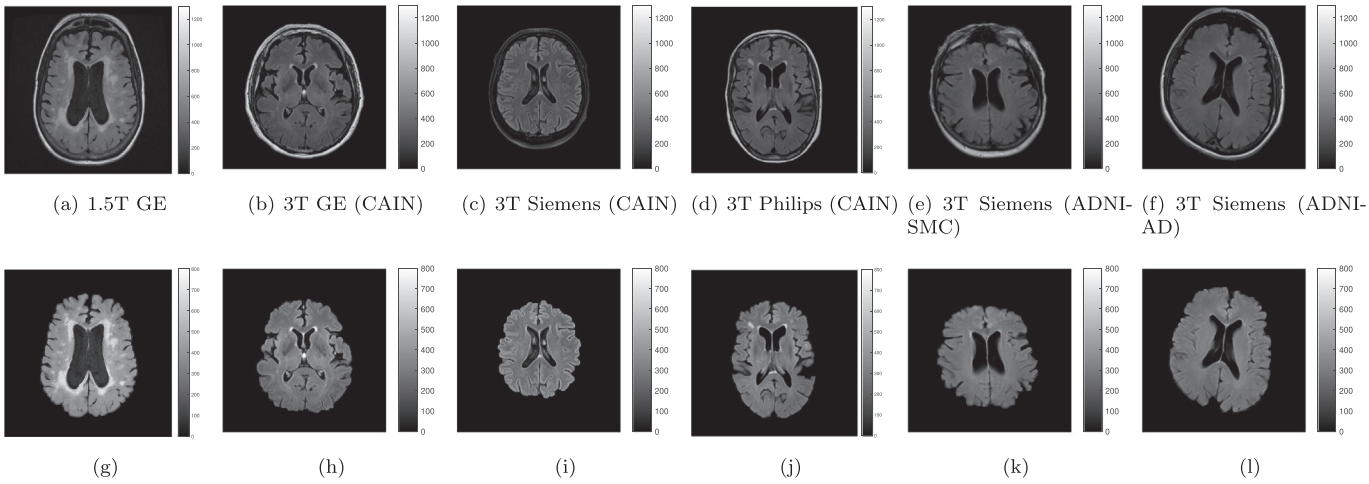


Fig. 7. Images of varying lesion loads before and after thresholding-based brain extraction. (a)–(f) are original images, (g)–(l) are segmented images.

Table 5
Comparison of segmentation performance of two methods across MC and multi-disease datasets.

Method	Dataset	DSC	HD	Sensitivity	Specificity	Extra fraction
Thresholding	CAIN	81.4 ± 10.9	2.2 ± 2.6	94.9 ± 7.2	96.7 ± 3.4	19.0 ± 21.5
	ADNI	86.1 ± 5.1	3.10 ± 1.84	97.8 ± 1.3	97.5 ± 1.36	14.0 ± 7.03
	SB	78.1 ± 3.6	0.51 ± 0.09	99.9 ± 0.004	90.0 ± 1.98	28.3 ± 5.8
Classifier [37]	CAIN	91 ± 1.52	1.11 ± 0.8	96.9 ± 2.1	98.4 ± 0.62	6.55 ± 3.1
	ADNI	86.2 ± 5.4	3.71 ± 2.83	95.5 ± 3.1	97.5 ± 1.6	11.1 ± 6.46

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through the NSERC Discovery grant.

The Canadian Atherosclerosis Imaging Network (CAIN) was established through funding from a Canadian Institutes of Health Research Team Grant for Clinical Research Initiatives (CIHR- CRI 88057). Funding for the communication and imaging infrastructure for CAIN was received through a grant from the Canada Foundation for Innovation (CFI CAIN 20099), with matching funds provided by the governments of Alberta, Ontario, and Quebec.

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

References

[1] Brain Canada Foundation, Annual Report 2015, Technical Report, Brain Canada Foundation, Montreal, Quebec, 2015. URL: http://www.braincanada.ca/files/BrainCanada_2015_ENG_WEB.pdf.

[2] Oppedal K, Eftestl T, Engan K, Beyer MK, Aarsland D. Classifying dementia using local binary patterns from different regions in magnetic resonance images. International Journal of Biomedical Imaging 2015;2015:1–14. https://doi.org/10.1155/2015/572567.

[3] Brant-Zawadzki M, Atkinson D, Detrick M, Bradley WG, Scidmore G. Fluid-attenuated inversion recovery (FLAIR) for assessment of cerebral infarction: initial clinical experience in 50 patients. Stroke 1996;27:1187–91.

[4] Schmidt P, Gaser C, Arsic M, Buck D, Frschler A, Berthele A, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. NeuroImage 2012;59:3774–83.

[5] Dufouil C, De Kersaint-Gilly A, Besancon V, Levy C, Auffrey E, Brunnereau L, et al. Longitudinal study of blood pressure and white matter hyperintensities: the EVA MRI cohort. Neurology 2001;56:921–6.

[6] van Dijk EJ, Breteler MM, Schmidt R, Berger K, Nilsson L-G, Oudkerk M, et al. The association between blood pressure, hypertension, and cerebral white matter lesions: cardiovascular determinants of dementia study. Hypertension 2004;44:625–30.

[7] Ferguson SC, Blane A, Perros P, McCrimmon RJ, Best JJ, Wardlaw J, et al. Cognitive ability and brain structure in type 1 diabetes relation to microangiopathy and preceding severe hypoglycemia. Diabetes 2003;52:149–56.

[8] Gons RAR, van Norden AGW, de Laat KF, van Oudheusden LJB, van Uden IWM, Zwiers MP, et al. Cigarette smoking is associated with reduced microstructural integrity of cerebral white matter. Brain 2011;134:2116–24.

[9] Staals J, Makin SD, Doubal FN, Dennis MS, Wardlaw JM. Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden. Neurology 2014;83:1228–34.

[10] Wardlaw JM, Allerhand M, Doubal FN, Hernandez MV, Morris Z, Gow AJ, et al. Vascular risk factors, large-artery atheroma, and brain white matter hyperintensities. Neurology 2014;82:1331–8.

[11] Debette S, Seshadri S, Beiser A, Au R, Himali JJ, Palumbo C, et al. Midlife vascular risk factor exposure accelerates structural brain aging and cognitive decline. Neurology 2011;77:461–8.

[12] Bilello M, Suri N, Krejza J, Woo JH, Bagley LJ, Mamourian AC, et al. An approach to comparing accuracies of two Flair MR sequences in the detection of multiple sclerosis lesions in the brain in the absence of gold standard. Acad Radiol 2010;17:686–95.

[13] Wilke M, de Haan B, Juenger H, Karnath H-O. Manual, semi-automated, and automated delineation of chronic brain lesions: a comparison of methods. NeuroImage 2011;56:2038–46.

[14] Malloy P, Correia S, Stebbins G, Laidlaw D. Neuroimaging of white matter in aging and dementia. Clin Neuropsychol 2007;21:73–109.

[15] Altaf N, Morgan PS, Moody A, MacSweeney ST, Gladman JR, Auer DP. Brain white matter hyperintensities are associated with carotid intraplaque hemorrhage. Radiology 2008;248:202–9.

[16] Altaf N, Daniels L, Morgan P, Lowe J, Gladman J, MacSweeney S, et al. Cerebral white matter hyperintense lesions are associated with unstable carotid plaques. Eur J Vasc Endovasc Surg 2006;31:8–13.

[17] de Groot M, Verhaaren BF, de Boer R, Klein S, Hofman A, van der Lugt A, et al. Changes in normal-appearing white matter precede development of white matter lesions. Stroke 2013;44:1037.

[18] Wardlaw JM, Valds Hernndez MC, MuozManiega S. What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. J Am Heart Assoc 2015;4:001140.

[19] Garcia-Lorenzo D, Francis S, Narayanan S, Arnold DL, Collins DL. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med Image Anal 2013;17:1–18.

[20] Fennema-Notestine C, Ozyurt IB, Clark CP, Morris S, Bischoff-Grethe A, Bondi MW, et al. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Hum Brain Mapp 2006;27:99–113.

[21] Iglesias JE, Liu Cheng-Yi, Thompson PM, Tu Zhuowen. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans Med Imaging 2011;30:1617–34.

[22] Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, et al. Deep MRI brain extraction: a 3d convolutional neural network for skull stripping. NeuroImage 2016;129:460–9.

[23] Nyul LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med 1999;42:1072–81.

[24] Palumbo D, Yee B, O'Dea P, Leedy S, Viswanath S, Madabhushi A. Interplay between bias field correction, intensity standardization, and noise filtering for t2-weighted mri. Engineering in Medicine and Biology Society, EMBC, IEEE. 2011. p. 5080–3.

[25] Jager F, Hornegger J. Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging. IEEE Trans Med Imaging 2009;28:137–50http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4601458https://doi.org/10.1109/TMI.2008.2004429.

[26] Khademi A, Reiche B, Arezza G. Method and System for Standardized Processing of MR Images. 2018. PCT Patent Application No. PCT/CA2018/051606, Filed: Dec. 14, 2018.

[27] Isa IS, Sulaiman SN, Mustapha M, Darus S. Evaluating denoising performances of fundamental filters for T2-weighted MRI images. Procedia Computer Science 2015;60:760–8.

[28] Zhong Y, Utriainen D, Wang Y, Kang Y, Haacke EM. Automated white matter hyperintensity detection in multiple sclerosis using 3d T2 FLAIR. International Journal of Biomedical Imaging 2014;2014:1–7.

[29] Reinhard E, Ashikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Comput Graph Appl 2001;21:34–41.

[30] Winkler A, Kochunov P, Glahn D. FLAIR templates Available at https://brainder.org/download/flair/; 2019.

[31] Khademi A, Venetsanopoulos A, Moody AR. Robust white matter lesion

segmentation in FLAIR MRI. IEEE Transactions on Biomedical Engineering 2012;59:860–71.
[32] Khademi A, Venetsanopoulos A, Moody AR. Automatic contrast enhancement of white matter lesions in FLAIR MRI. International Symposium on Biomedical Imaging (ISBI). IEEE; 2009. p. 322–5.
[33] Wyatt C, Wang Y-P, Freedman M, Loew M, Wang Y. Chapter 7. Biomedical Information Technology. Elsevier; 2007. p. 165–9.
[34] Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004;11:178–89.
[35] Beauchemin M, Thomson KP, Edwards G. On the Hausdorff distance used for the evaluation of segmentation results. Canadian Journal of Remote Sensing 1998;24:3–8.
[36] Stokking R, Vincken KL, Viergever MA. Automatic Morphology-Based Brain Segmentation (MBRASE) from MRI-T1 data. NeuroImage 2000;12:726–38.
[37] Reiche B, Moody AR, Khademi A. Effect of image standardization on flair mri for brain extraction. Signal, Image and Video Processing 2015;9:11–6.
[38] Jia Y. Learning Semantic Image Representations at a Large Scale Ph.D. thesis Berkeley: EECS Department, University of California; 2014. URL http://www2.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-93.html.
[39] Rolland J, Vo V, Abbey C. Fast algorithms for histogram-matching: application to texture synthesis. Journal of Electronic Imaging 2000;9:39–45.
[40] Filippi M, Rocca M. Dirty-appearing white matter: a disregarded entity in multiple sclerosis. Am J Neuroradiol 2010;31:390–1.