Neuroinform
DOI 10.1007/s12021-016-9301-1

ORIGINAL ARTICLE

Validation of White-Matter Lesion Change Detection Methods on a Novel Publicly Available MRI Image Database

Žiga Lesjak [1] · Franjo Pernuš [1] · Boštjan Likar [1,2] · Žiga Špiclin [1,2]

Abstract Changes of white-matter lesions (WMLs) are good predictors of the progression of neurodegenerative diseases like multiple sclerosis (MS). Based on longitudinal magnetic resonance (MR) imaging the changes can be monitored, while the need for their accurate and reliable quantification led to the development of several automated MR image analysis methods. However, an objective comparison of the methods is difficult, because publicly unavailable validation datasets with ground truth and different sets of performance metrics were used. In this study, we acquired longitudinal MR datasets of 20 MS patients, in which brain regions were extracted, spatially aligned and intensity normalized. Two expert raters then delineated and jointly revised the WML changes on subtracted baseline and follow-up MR images to obtain ground truth WML segmentations. The main contribution of this paper is an objective, quantitative and systematic evaluation of two unsupervised and one supervised intensity based change detection method on the publicly available datasets with ground truth segmentations, using common pre- and post-processing steps and common evaluation metrics. Besides, different combinations of the two main steps of the studied change detection methods, i.e. dissimilarity map construction and its segmentation, were tested to identify the best performing combination.

Keywords Multiple sclerosis · Lesion · Magnetic resonance · Image segmentation · Change detection · Quantitative evaluation · Validation dataset

Introduction

Serial or longitudinal imaging of the brain is performed routinely on patients with certain cerebrovascular and neurodegenerative diseases, for instance, in multiple sclerosis, small vessel disease, Alzheimer's and other dementias. The evolution of these diseases has been strongly correlated to changes of brain structures (Ramirez et al. 2014; Rocca et al. 2013; Susanto et al. 2015), which often appear ahead of clinical symptoms (Lebrun et al. 2008; Risacher et al. 2009). Structural changes can manifest only locally, affecting specific brain structures or locations, and resulting for instance in white matter lesions (WMLs), or may have a gross effect on the whole brain, resulting in atrophy. Here, we focus on multiple sclerosis (MS) and the detection of associated local changes of WML, since they were established as good predictors of MS disease progression and long term patient disability (Patti et al. 2015; Popescu et al. 2013). For detection of MS WML changes (Rocca et al. 2013), magnetic resonance (MR) tomographic imaging is by far the most sensitive imaging technique. Clinically relevant time intervals for observing WML changes range from several months to up to 2 years. Because WML changes may be very subtle (small changes in volume and/or MR intensities), their detection in longitudinally acquired MR images requires highly sensitive image analysis techniques (Vrenken et al. 2013; Wei et al. 2004; Lladó et al. 2012).

Detection and quantification of WML changes is nowadays mostly performed by comparing corresponding manually delineated WMLs in the baseline and follow-up MR images.

✉ Žiga Lesjak
ziga.lesjak@fe.uni-lj.si

[1] University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia

[2] Computer Vision Systems, Sensum, Tehnološki Park 21, 1000 Ljubljana, Slovenia

Published online: 20 May 2016

When characterizing changes of WMLs, a single brain MR imaging session may consist of several MR modalities, e.g. T1-, T2-weighted (T1w, T2w), diffusion weighted, proton-density weighted (PDw) and fluid-attenuated inversion recovery (FLAIR), which may all need to be jointly observed to reliably detect and delineate the WMLs (Vrenken et al. 2013). However, manual delineation slice-by-slice across multiple three-dimensional (3D) MR images is tedious, time consuming, and most of all subjective. Because manual delineations generally suffer from high inter- and intra-rater variabilities, subtle WML changes cannot be reliably identified and accurately quantified. To avoid the aforementioned shortcomings of the manual method and to improve the accuracy, reliability and reproducibility of change detection in longitudinal MR images, several automated methods have been developed.

Automated WML Change Detection

The problem of automated WML change detection can be addressed by three different strategies (Lladó et al. 2012): 1) longitudinal volumetric analysis, 2) deformable image registration and 3) longitudinal analysis of MR intensity (Patriarche and Erickson 2004). Automated longitudinal volumetric analysis independently delineates (segments) the WMLs in the baseline and follow-up MR images (García-Lorenzo et al. 2013; Llado et al. 2012) and thus mimics the manual method. The variability of results obtained by this method is often too high to consistently accurately segment small WMLs, let alone their changes. Deformable image registration aligns the baseline and follow-up MR images and then extracts changes from deformation fields. This type of methods have the potential to measure enlarging and shrinking WMLs, while their potential in case of newly appearing or disappearing WMLs is not that clear (Rey et al. 2002; Studholme et al. 2006). Moreover, the deformations must be physically constrained, which is generally difficult to model in case of diffuse-appearing structures like WMLs. The longitudinal intensity analysis employs a more simple rigid or affine registration, followed by change detection based on the analysis of intensity values at corresponding sites of the baseline and follow-up MR images (Moraal et al. 2009). As MR intensity changes are an important feature of longitudinal changes of WMLs, the methods based on longitudinal intensity analysis or, shortly, the intensity based methods seem to be the most promising of the three strategies for detecting enlarging, shrinking, newly appearing and disappearing WMLs. In the following we thus focus on this class of methods.

Intensity based methods generally consist of four main steps (Fig. 1), i.e. 1) preprocessing of MR images, 2) construction of a dissimilarity map (DM), 3) segmentation of DM to obtain change mask and 4) postprocessing of the change mask resulting in WML changes. Most authors of intensity based methods use a rather similar preprocessing sequence, which comprises established procedures for brain extraction, white-matter (WM) masking, MR intensity normalization and spatial image co-registration. These preprocessing tools were rigorously evaluated in several recent studies, e.g. (Diez et al. 2013; Roura et al. 2014; Shinohara et al. 2014). Postprocessing aims to eliminate false positive WML changes by relying on expert knowledge of WML morphology, appearance and location, obtained by rule-based methods (Battaglini et al. 2014) or size and intensity based filtering (Ganiler et al. 2014). Postprocessing is quite similar across different methods for WML change detection (Lladó et al. 2012).

The intensity-based change detection methods mainly differ in the way the DM is computed and segmented. The DM computations are either unsupervised or supervised, whereas the former use solely the baseline and follow-up MR images, while the latter also require a set of training images with accurately segmented WML changes. A class of unsupervised methods computes the DM by simply subtracting the co-registered baseline and follow-up MR images (Battaglini et al. 2014; Duan et al. 2008; Ganiler et al. 2014; Moraal et al. 2009, 2010a, b). Battaglini et al. (2014) segment the DM by a low threshold so as to obtain an overestimated mask of candidate lesions within a subject-specific WM mask. By

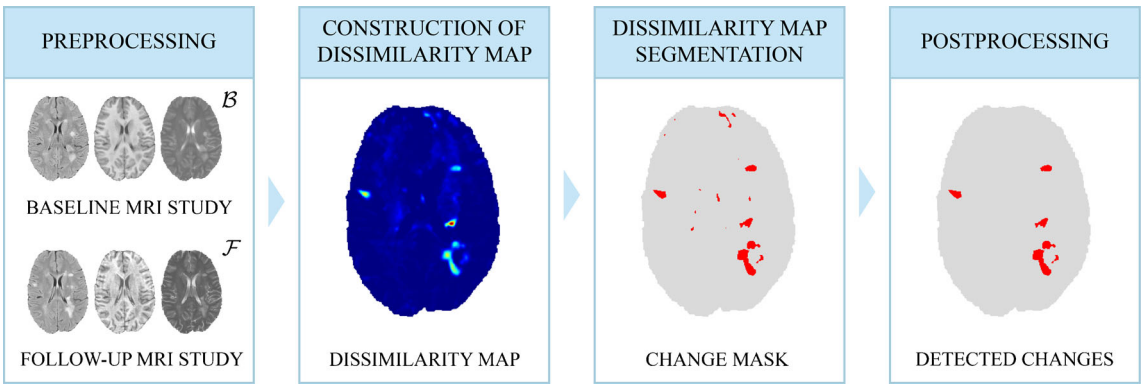

Fig. 1 The four main steps of WMLs change detection

rule-based postprocessing they finally reduce the false positive hyperintense clusters of voxels based on their extent, shape, and intensity. Ganiler et al. (2014) use a threshold of mean plus five standard deviations of the DM values to segment the DM and keep only those segmented structures, which are larger than three voxels. Direct point-by-point image subtraction may result in a DM that is highly susceptible to image noise, imperfect intensity normalization and registration errors and thus may not provide an objective change detection criterion. To overcome some of these drawbacks, a statistical change detection based on Generalized Likelihood Ratio (GLR) (Bosc et al. 2003), a voxel-wise test statistic that represents the DM, was proposed. The GLR considers the intensity distributions within local patches of baseline and follow-up images and may be easily extended to multimodal change detection (Bosc et al. 2003; Simoes and Slump 2011). Simoes and Slump 2011 automatically determined a threshold of the GLR by minimizing the offset of the angular histogram of change vectors in the space spanned by T1w and T2w modalities. Nika et al. (2014) used patches of baseline and follow-up images for adaptive dictionary learning. To assess the (dis)similarity between the images, each patch in the follow-up image was expressed as a linear combination of patches from the dictionary. To extract the best features for segmentation and for higher efficiency, the obtained linear coefficients were projected to a lower-dimensional subspace by principal component analysis and the DM was formed as the L1 norm of subspace features. The performance of patch-based methods on real MR images is rather unclear because they were only evaluated qualitatively and/or on synthetic MR images.

Supervised intensity-based approaches determine the best features from the annotated training datasets. Sweeney et al. (2013) used a Logistic Regression Model (LRM) based on multi-modality images of baseline and follow-up MRs, the difference images and the time between studies to estimate a DM. Elliott et al. (2013) used the baseline and one or more follow-up MR images to perform a joint temporally consistent Bayesian segmentation of brain tissues. The obtained tissue class probabilities, baseline, follow-up and difference images were then applied in a random decision forest classifier to get a change probability map, which represented the DM. The changes were obtained by a user-defined threshold of the DM. Because the supervised methods are trained on images acquired by a specific MR machine, parameters and modalities, their performance is likely to deteriorate substantially if applied to images acquired under conditions different to those in the training dataset.

Validation Issues

The authors of the aforementioned methods reported rather good change detection results. However, an objective comparison of their methods is difficult, because the authors used different validation datasets, with differently obtained ground truth, and even a different set of performance metrics (Lladó et al. 2012). Furthermore, a fair cross-comparison is also difficult due to pre- and post-processing differences, which are not the core of the methods. In image segmentation, the ground truth is generally obtained manually. For WML change segmentation, synthetic MR images from the BrainWeb (Cocosco et al. 1997) simulator can be used. However, the problem is that BrainWeb provides only one brain template that contains MS lesions. Although synthetic datasets are often used to verify a change detection method, they are not appropriate for an objective and reliable validation. Within a recent WML segmentation challenge, a dataset of 20 MS patients, each with 3-5 MR studies and lesions independently segmented by two raters on baseline and follow-up MR images, was made available (Pham n.d.). Since these WML segmentations were created on a per study basis where the inter-rater variability is generally large, they are not the best for validating WMLs change detection methods. A better approach would be to consider a consensus across segmentations of WML changes provided by several experienced neuroradiologists (Styner et al. 2008). To the best of our knowledge, validation datasets with such a ground truth are not yet publicly available.

Contribution of the Paper

The aim of this paper is to objectively validate and compare several intensity based methods for detecting WML changes using longitudinal MR image datasets with accurate consensus-based ground truth. We focused on the methods proposed by Ganiler et al. (2014), Simoes and Slump (2011) and Sweeney et al. (2013), since they seem to be able to capture all types of WML changes (i.e. newly appearing and disappearing, enlarging and shrinking WMLs) and since several researchers (Ganiler et al. 2014; Rousseau et al. 2007; Seo and Milanfar 2009; Simoes and Slump 2011) have already reported a rather good performance of these methods. Three methods, two unsupervised (Ganiler et al. 2014; Simoes and Slump 2011) and one supervised (Sweeney et al. 2013), were studied according to the main steps of a change detection method, i.e. DM formation and segmentation. Performance of each of these steps was individually validated to get a better insight into the tested methods, while different combinations of steps were validated with the aim to maximize the overall performance of change detection. The methods were tested on a longitudinal database of 20 MS patients, on which ground truth was created by two expert raters who segmented WML changes on preprocessed and subtracted baseline and follow-up T1w, T2w and FLAIR MR images. The consensus segmentations were jointly refined by the two raters until they agreed on what was considered the ground truth segmentation.

Among the tested methods the subtraction-based DM computed on the FLAIR modality and automated confidence level thresholding (Ganiler et al. 2014) provided most accurate change detection in terms of median Dice similarity coefficient (0.48) and was considered reliable as it had the highest and the most consistent detection success rate (>75 %) across different volumes of lesion changes.

Methods and Material

The WML change detection process can be divided into four steps (Fig. 1): 1) preprocessing, in which all images are spatially aligned into a common coordinate frame and MR intensities are corrected for non-uniformities and normalized. Since WML changes are constrained to WM, the whole brain and the WM mask are also extracted in this step. 2) DM is computed as a function of co-located intensities of the preprocessed baseline and follow-up images. The fuzzy value of a DM voxel corresponds to the probability of change. 3) segmentation of DM, which results in a change mask, indicating the detected changes. 4) postprocessing of the change mask to reduce false positives due to imperfect steps 1–3. Table 1 lists the characteristics of the three validated intensity based change detection methods (Ganiler et al. 2014; Simoes and Slump 2011; Sweeney et al. 2013 (SuBLIME, RRID:SCR_014409)) according to the four steps described above. The methods mainly differ in the DM creation and segmentation (steps 2 and 3, respectively).

Throughout this paper, the symbols $\mathcal{B}$ and $\mathcal{F}$ will refer to sets of preprocessed baseline and follow-up MR images, respectively, where $\mathcal{B}=\{\mathcal{B}_{\mathcal{M}}(\mathbf{x})\}$ and $\mathcal{F}=\{\mathcal{F}_{\mathcal{M}}(\mathbf{x})\}$ and $\mathcal{M}$ is a set of MR modalities, e.g. $\mathcal{M}=\{\text{T1w, T2w, FLAIR, }\ldots\}$.

Preprocessing

The preprocessing steps of the three validated methods presented in Table 1 are quite similar but not equal. Based on the preprocessing used by the three methods, we have developed a preprocessing pipeline (Fig. 2), which was applied to all three methods. In this way, preprocessing did not bias a method. The input to the preprocessing step are the raw baseline and follow-up T1w, T2w, and FLAIR images, while the output are intensity inhomogeneity corrected and intensity normalized T1w, T2w, and FLAIR images, and brain and WM masks that are all registered and transformed into a common reference frame.

In the preprocessing pipeline, brain masks are first extracted from the baseline $\mathcal{B}$ and follow-up $\mathcal{F}$ T1w MR images using BET 2 (Smith 2002). After brain extraction, the N4 bias correction (Tustison et al. 2010) and Gaussian mixture model based Atropos segmentation (Avants et al. 2011) are iteratively executed on the masked T1w images until there are no changes to the bias field and the segmentation. The result are bias corrected masked T1w $\mathcal{B}$ and $\mathcal{F}$ images and segmentations of these images into normal appearing brain structures (NABS), comprising WM, gray matter (GM) and cerebrospinal fluid (CSF). Next, the intra-study T1w, T2w and FLAIR images are registered. It is expected that the intra-study registration will perform better if, besides the T1w, the T2w and FLAIR images are also masked and bias corrected. Therefore, the registration is performed in two-stages. First, the bias corrected masked T1w $\mathcal{B}$ and $\mathcal{F}$ images are registered to their corresponding raw T2w and FLAIR images using the affine transformation and mutual information maximization (Avants et al. 2014; Maes et al. 1997). The obtained transformations are used to align the T1w brain mask to T2w and FLAIR images, which are then bias corrected by the N4. Second, after all images are bias corrected they are again affinely registered by using FLAIR as a reference image and the affine transformations from the first step to initialize the registration. Next, inter-study registration is performed between the $\mathcal{B}$ and $\mathcal{F}$ masked FLAIR images using the affine transformation and normalized correlation maximization (Avants et al. 2014). After registration, all images are resampled into a new reference frame defined "half-way" between $\mathcal{B}$ and $\mathcal{F}$ FLAIR images so as to harmonize the impact of interpolation artifacts across $\mathcal{B}$ and $\mathcal{F}$ images. Namely, given the affine transformation T from $\mathcal{B}$ to $\mathcal{F}$ FLAIR image, the reference frame is defined by transformations $T_{1/2}$ and $T_{-1/2}$ with respect to the $\mathcal{B}$ and $\mathcal{F}$ FLAIR images, respectively, such that $T=T_{1/2}\cdot T_{-1/2}^{-1}$. The obtained intra-study $\mathcal{B}$ and $\mathcal{F}$ and respective $T_{1/2}$ and $T_{-1/2}$ transformations are composed so as to align all the images in $\mathcal{B}$ and $\mathcal{F}$ into a common reference frame. Image registration and resampling of transformed images, N4 bias field correction and Atropos segmentation were all performed using the ANTs toolbox (Avants et al. n.d.).

Since only changes within the WM are of our interest, the $\mathcal{B}$ and $\mathcal{F}$ WM masks $\Omega_{\mathcal{B}}$ and $\Omega_{\mathcal{F}}$ are transformed into the common reference space. Their union defines the domain Ω where WML changes are searched for. Because the CSF and GM are excluded from further analysis, false change detections due to imperfect registration at structures' borders, partial volume artifacts and signal overshoots are reduced.

Let $\mathbf{x}_i\in\Omega$, $i=1,\ldots,N$ represent lexicographically ordered spatial coordinates, where N is the number of voxels within Ω. The mean μ and standard deviation σ of WM intensities within Ω are used to normalize each MR modality $\mathcal{M}$ in $\mathcal{B}$ and $\mathcal{F}$ as in (Sweeney et al. 2013). E.g., for modalities in $\mathcal{B}$:

Table 1 Overview of the three validated methods according to four steps of change detection

Method	Preprocessing	Construction of dissimilarity map	Segmentation of dissimilarity map	Postprocessing
Ganiler et al. (2014)	• Brain extraction • Bias field correction • Atlas and model-based WM segmentation • Histogram based intensity normalization • Rigid image registration	Subtraction of intensity images (STI)	Confidence level thresholding (CLT)	Size and intensity based filtering
Simoes and Slump (2011)	• Affine image registration • Brain extraction • Bias field correction • Histogram based intensity normalization	Generalized likelihood ratio (GLR)	Change vector angular histogram thresholding (CVAHT)	/
Sweeney et al. (2013)	• Isotropic image resampling • Rigid image registration • Brain extraction • Normal-appearing model-based WM segmentation • Intensity normalization based on normal-appearing WM	Logistic regression model (LRM)	Manual threshold	/

$$\mathcal{B}'_{,\mathcal{M}}(\mathbf{x}) = \frac{\mathcal{B}_{,\mathcal{M}}(\mathbf{x}) - \mu(\mathcal{B}_{,\mathcal{M}}(\mathbf{x_i}); \mathbf{x_i} \in \Omega)}{\sigma(\mathcal{B}_{,\mathcal{M}}(\mathbf{x_i}); \mathbf{x_i} \in \Omega)}, \quad (1)$$

where $\mathcal{B}'_{,\mathcal{M}}(\mathbf{x})$ is the intensity-normalized image. Analogously, the normalization is carried out on modalities in $\mathcal{F}$. Note that the normalization applies to all voxels $\mathbf{x}$ in the image. For notational brevity we assign $\mathcal{B}_{,\mathcal{M}}(\mathbf{x}) \leftarrow \mathcal{B}'_{,\mathcal{M}}(\mathbf{x})$ and $\mathcal{F}_{,\mathcal{M}}(\mathbf{x}) \leftarrow \mathcal{F}'_{,\mathcal{M}}(\mathbf{x})$ and use only the intensity normalized images in the following steps.

Construction of Dissimilarity Map

The preprocessed $\mathcal{B}$ and $\mathcal{F}$ MR image sets are used to create the DMs by the STI (Ganiler et al. 2014), GLR (Simoes and Slump 2011) and LRM (Sweeney et al. 2013) methods, which are based on subtraction of intensity images, generalized likelihood ratio and logistic regression model, respectively. In the following, we briefly describe these methods.

Subtraction of Intensity Images (STI)

A simple way to construct a DM is to subtract co-located intensities in the baseline $\mathcal{B}_{,\mathcal{M}}$ and follow-up $\mathcal{F}_{,\mathcal{M}}$ MR images of the same modality $\mathcal{M}$ (Ganiler et al. 2014):

$$\mathrm{DM}_{\mathrm{STI}}(\mathbf{x_i}) = \mathcal{F}_{,\mathcal{M}}(\mathbf{x_i}) - \mathcal{B}_{,\mathcal{M}}(\mathbf{x_i}); i \in \Omega, \quad (2)$$

where $\|\cdot\|$ is the L2 norm of voxel-wise intensity differences. A higher value of $\mathrm{DM}_{\mathrm{STI}}(\mathbf{x})$ corresponds to a higher likelihood of change at voxel $\mathbf{x}$. Computation of $\mathrm{DM}_{\mathrm{STI}}(\mathbf{x})$ may employ multiple modalities or MR sequences, e.g. T1-, T2-weighted and/or FLAIR, such that $\mathcal{F}_{,\mathcal{M}}(\mathbf{x})$ and $\mathcal{B}_{,\mathcal{M}}(\mathbf{x})$ represent vectors of corresponding intensity values. Because the best result of WML change detection are obtained by subtracting only the FLAIR images, we used only FLAIR to compute $\mathrm{DM}_{\mathrm{STI}}$.

Generalized Likelihood Ratio (GLR)

The GLR (Simoes and Slump 2011) was computed from the baseline and follow-up T1w and FLAIR MR images. The GLR assumes that the intensities are normally distributed, which, in general, holds for intensities within the WM mask. To compute the GLR at each voxel location $\mathbf{x}_i \in \Omega$ within the WM mask Ω, a window of sidelength W is centered at $\mathbf{x}_i$ in the $\mathcal{B}$ and $\mathcal{F}$ images and the dissimilarity value is then computed as:

$$\begin{aligned}\mathrm{DM}_{\mathrm{GLR}}(\mathbf{x}_i) = -\frac{1}{2}\sum_{j \in W(\mathbf{x}_i)} \Big[&\left(B_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_B\right)^{\mathrm{T}} \boldsymbol{C}_v^{-1} \left(B_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_B\right) \\ &+ \left(F_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_F\right)^{\mathrm{T}} \boldsymbol{C}_v^{-1} \left(F_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_F\right) \\ &- \left(B_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_0\right)^{\mathrm{T}} \boldsymbol{C}_v^{-1} \left(B_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_0\right) \\ &- \left(F_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_0\right)^{\mathrm{T}} \boldsymbol{C}_v^{-1} \left(F_{,\mathcal{M}}(\mathbf{x}_j) - \hat{\mu}_0\right) \Big],\end{aligned} \quad (3)$$

where $\boldsymbol{C}_v$ is the noise covariance matrix, while $\hat{\mu}_B$ and $\hat{\mu}_F$ represent the mean intensities of WM in the $\mathcal{B}$

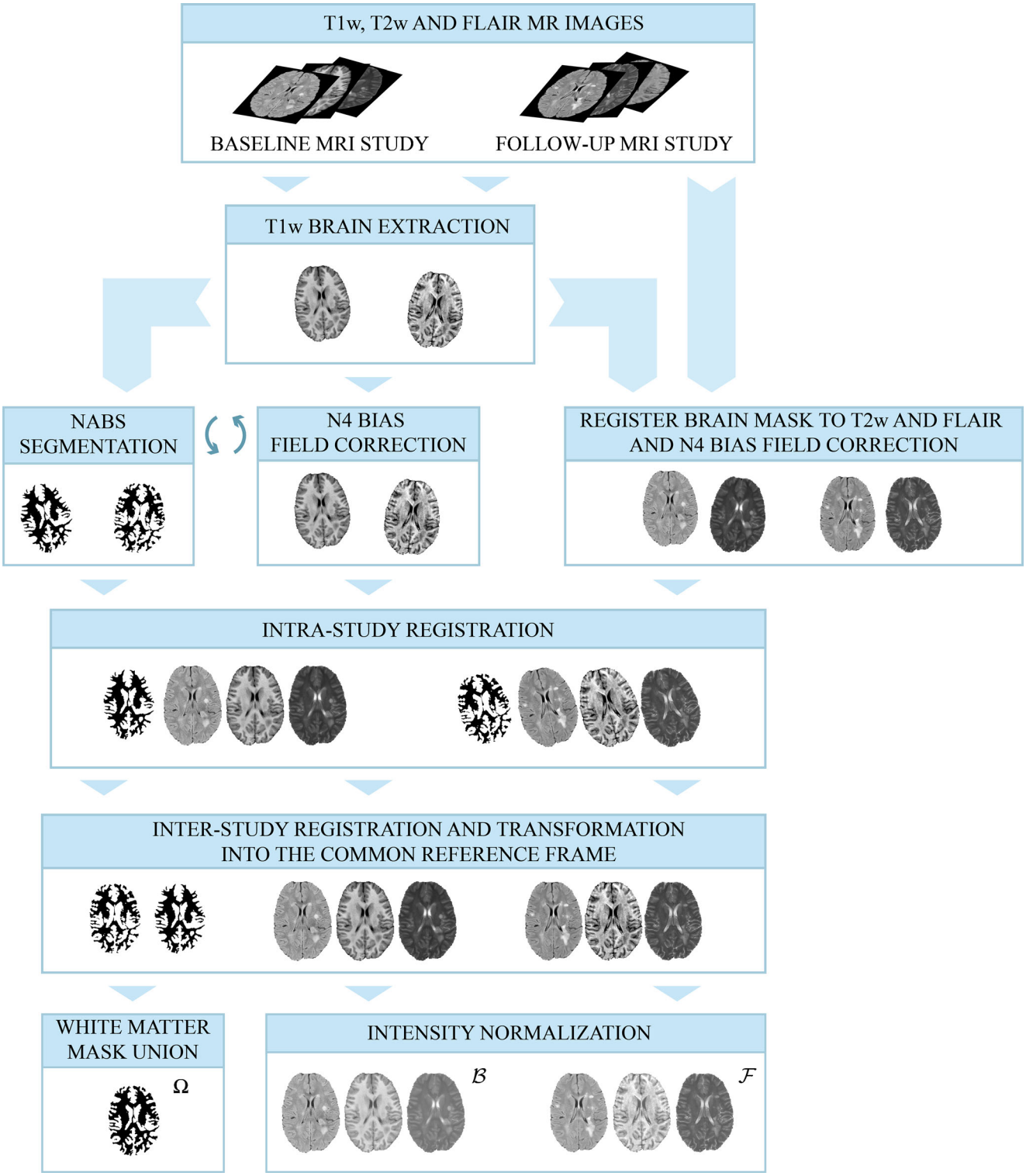

Fig. 2 Preprocessing of baseline and follow-up MR images. First, a brain mask is extracted using BET 2 (FSL, RRID:SCR_002823), followed by N4 bias field correction and Atropos-based NABS segmentation (ANTS - Advanced Normalization ToolS, RRID:SCR_004757). Next, intra- and inter-study registration is performed on the bias field corrected images and, then, the images and NABS segmentations are transformed into a common reference frame. The registered NABS segmentations are used to 1) define the mask for change detection (Ω) as a WM union of the baseline and follow-up WM masks and 2) to perform WM intensity normalization on the registered images

and $\mathcal{F}$ images, respectively, and $\hat{\mu}_0 = (\hat{\mu}_B + \hat{\mu}_F)/2$ (Simoes and Slump 2011). Equation (3) yields a DM, in which higher values correspond to a higher likelihood of change.

Logistic Regression Model (LRM)

The DM is obtained by feeding a trained LRM (Sweeney et al. 2013) with multiple MR image modalities, pairwise intra- modality difference images and time difference between studies (Δt) to estimate voxel-level probabilities of lesion change:

$$\begin{aligned} \mathrm{DM_{LRM}} = \; & \beta_0 + \beta_1 \Delta t + \beta_2 . \mathcal{B}_{FLAIR} \\ & + \beta_3 (\mathcal{F}_{FLAIR} - \mathcal{B}_{FLAIR}) + \beta_4 (\mathcal{F}_{FLAIR} - \mathcal{B}_{FLAIR})\, \Delta t \\ & + \beta_5 . \mathcal{B}_{T1} + \beta_7 (\mathcal{F}_{T1} - \mathcal{B}_{T1}) \Delta t \; + \beta_8 . \mathcal{B}_{T2} \\ & + \beta_9 (\mathcal{F}_{T2} - \mathcal{B}_{T2}) + \beta_{10} (\mathcal{F}_{T2} - \mathcal{B}_{T2}) \Delta t, \end{aligned} \tag{4}$$

where $\beta_0 \ldots \beta_{10}$ are trained coefficients corresponding to the T1w, T2w, and FLAIR MR modalities. Again, higher values of $\mathrm{DM_{LRM}}$ correspond to a higher likelihood of change.

Dissimilarity Map Segmentation

Each of the three tested methods originally employed a different DM segmentation approach (Table 1), which resulted in a tentative change mask. The LRM based DM, which is a probability map with a range from 0 to 1, was manually thresholded in (Sweeney et al. 2013). A threshold could also be automatically computed from the DM. The DM segmentation method in (Ganiler et al. 2014) computes a threshold based on nonparametric statistical testing of an empirical probability density function (PDF), which is computed from the DM. A predefined confidence level α determines the significant changes (outliers) of the empirical PDF. Based on the nature of DM, the outliers are selected either from two or one tail of the empirical PDF. The disadvantage of this so-called confidence level thresholding (CLT) method is that the value of α is related to the number of voxels representing significant changes. Therefore, it is difficult to select a value of α that will optimally segment the DM of patients with different volumes of changes, as these are generally not known in advance. Therefore, for optimal change detection, the value of α needs to be adjusted on a case by case basis.

The segmentation method in (Simoes and Slump 2011) computes the threshold by analyzing the angular histogram of a change vector (CV), defined as $CV_{\mathcal{M}} = \mathcal{F}_{\mathcal{M}} - \mathcal{B}_{\mathcal{M}}$. The angular histogram of change vectors is computed from angles:

$$\angle CV = \arctan\left(\frac{CV_{FLAIR}}{CV_{T1w}}\right). \tag{5}$$

At a low threshold, the angular histogram will generally appear noisy and will contain a certain offset (Fig. 3). By increasing the DM threshold, less significant changes are excluded, and the histogram's offset decreases towards 0. The threshold of DM may be selected by increasing a tentative DM threshold from a low to a high value until the histogram's offset drops below a certain predefined small value ε. This segmentation will be referred to as Change Vector Angular Histogram Thresholding (CVAHT).

Since both the CLT and CVAHT segmentations are based on thresholding, a third segmentation based on the optimal DM threshold, found by maximizing Dice similarity coefficient (DSC) between the computed and the reference change masks on each dataset using exhaustive search, was included for comparative performance evaluation.

Postprocessing

The change mask often contains many false positives resulting from various artifacts such as partial volume, imperfect intensity normalization and/or registration or spurious high

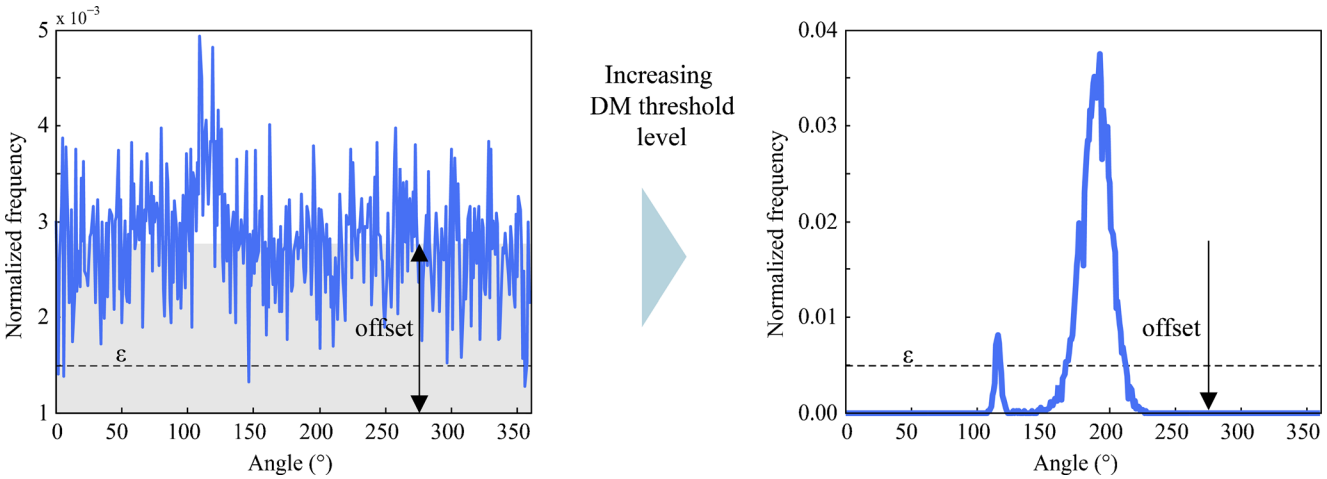

Fig. 3 DM thresholding by the CVAHT method: by increasing the DM threshold the angular histogram's offset progressively decreases towards zero. For numerical reasons, the DM threshold is chosen when the histogram's offset falls below some small value ε

intensity values due to image noise. These false positives are mostly isolated voxels or small clusters of voxels and can thus be efficiently removed by size based filtering of the change mask (Ganiler et al. 2014). The size based filtering is performed by connected component analysis that isolates regions of connected voxels. If the volume of a region is less than some threshold ϑ, the region is removed from the change mask.

Validation Database and Ground Truth

The validation database contained baseline and follow-up MR images of 20 MS patients. Patient demographic and treatment data is summarized in Table 2. The images were acquired on a 1.5 T Philips MRI machine at the University Medical Centre Ljubljana (UMCL). All 20 subjects have given written informed consent at the time of enrollment for imaging. The authors, who have obtained approval from the UMCL to use the data, confirm that the data was anonymized. Each patient's MR dataset contained a 2D T1-weighted (spin echo sequence, repetition time (TR) = 600 ms, echo time (TE) = 15 ms, flip angle (FA) = 90°, sampling of 0.9 × 0.9 × 3 mm with no inter-slice gap resulting in a 256 × 256 × 45 lattice), a 2D T2-weighted (spin echo sequence, TR = 4500 ms, TE = 100 ms, FA = 90°, sampling of 0.45 × 0.45 × 3 mm with no inter-slice gap resulting in a 512 × 512 × 45 lattice) and a 2D FLAIR image (TR = 11,000, TE = 140, TI = 2800, FA = 90, sampling of 0.9 × 0.9 × 3 mm with no inter-slice gap resulting in a 256 × 256 × 49 lattice). The median time between the baseline and follow-up studies was 311 days, ranging from 81 to 723 days with the interquartile range (IQR) of 223 days (Fig. 4). Some examples of database images are shown in Fig. 5.

For evaluation purposes, the reference or ground truth of changes was created by two expert raters. Initially, each rater independently segmented lesion changes in all 20 patient image datasets. Segmentation was performed on preprocessed and subtracted baseline and follow-up FLAIR images. To facilitate segmentation the raters could observe in side-by-side view the subtracted FLAIR image as well as baseline and follow-up FLAIR, T2w and T1w images and use manual and local semi-automated segmentation tools to segment the lesion changes. The raters focused on hypo- and hyper-intense regions of the subtracted FLAIR image, taking into consideration both the change in intensities and shapes of WMLs when deciding if hypo- or hyper-intense changes are due to MRI artifacts and image misregistration. The raters then jointly revised and merged their individual delineations to obtain a consensus, which was used as the ground truth segmentation of the changes. According to the ground truth, the median volume of lesion changes (LCs) per patient was 6.2 cm^3 (IQR: 6.5 cm^3). The distribution of LC volumes across all datasets is shown in Fig. 4. Individual LCs were stratified according to their volume into five categories listed in Table 3, similarly as in (Elliott et al. 2013; Ganiler et al. 2014).

Table 2 Patient demographic and treatment data

Gender	Age	MS diagnosis	Therapy
14 female 6 male	19 to 50 years Median: 34 years	1 secondary progressive 14 relapse remitting 5 unspecified	2 no therapy 6 Gilenya 2 Copaxon 2 Tysabri 1 Tecfidera 1 Aubagio 6 unspecified

Experiments and Results

Validation of change detection focused on evaluating the two main steps of change detection, namely DM creation and segmentation. Therefore, combinations of three methods for DM creation (STI, GLR and LRM) and three for DM segmentation (CLT, CVAHT and optimal threshold) were tested on 20 MR image datasets with a common preprocessing and postprocessing as shown in Fig. 6.

Dissimilarity Map Construction and Evaluation

Using the preprocessed baseline and follow-up T1w, T2w, and FLAIR images, DM_{STI}, DM_{GLR}, and DM_{LRM} dissimilarity maps were constructed according to Eqs. (2), (3) and (4), respectively. The DM_{STI} was computed from the $\mathscr{B}$ and $\mathscr{F}$ FLAIR images, since FLAIR exhibits the best contrast between WM and the WMLs. The DM_{GLR} was computed from the T1w and FLAIR images while DM_{LRM} employed the T1w, T2w and FLAIR images. To compute the DM_{LRM}, the coefficients β_i provided by the original authors in [28] were used. Figure 7 shows the DM_{STI}, DM_{GLR} and DM_{LRM} computed on five $\mathscr{B}$ and $\mathscr{F}$ MR images from the validation dataset.

In order to determine the capability of the three different DMs to express WML changes, a Receiver Operating Characteristic (ROC) curve was first computed based on modifying α of CLT from 0 to 1, which is equivalent to changing the DM threshold from a high to a low value (Fig. 8, *left*). In the range of 0-0.1 of false positive rate (FPR), the ROC curves indicate substantially lower sensitivity to changes (i.e. lower true positive rate or TPR) of the LRM compared to STI and GLR DMs.

The change detection capability of each DM was then measured by calculating the area under the curve (AUC). Figure 8 (*right*) shows the obtained AUCs for the three tested DMs. The STI and GLR performed best with median AUCs of 0.94 (IQR: 0.03) and 0.93 (IQR: 0.03), respectively, while the

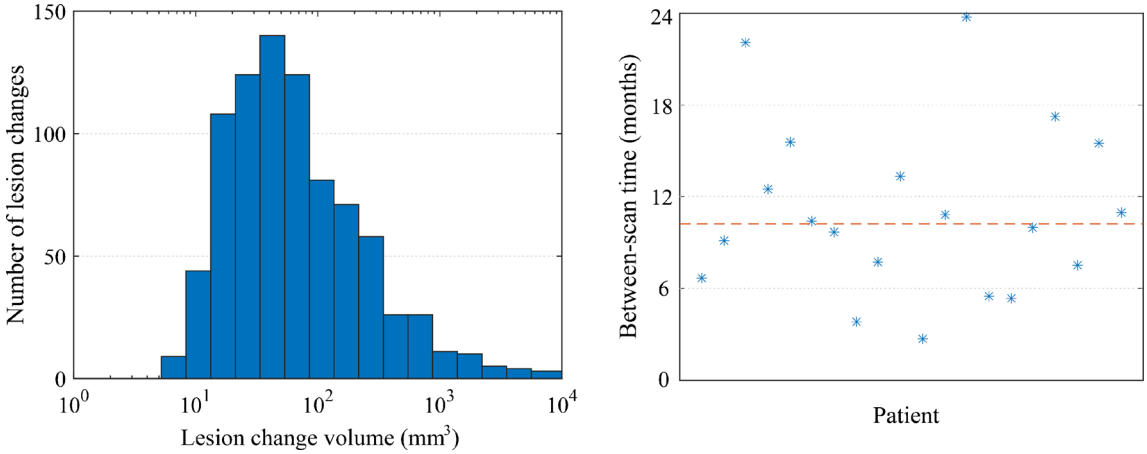

Fig. 4 Distribution of lesion change volumes across the 20 MR image datasets (*left*) and scatter plot of time difference between baseline and follow-up study for each of the patients (*right*)

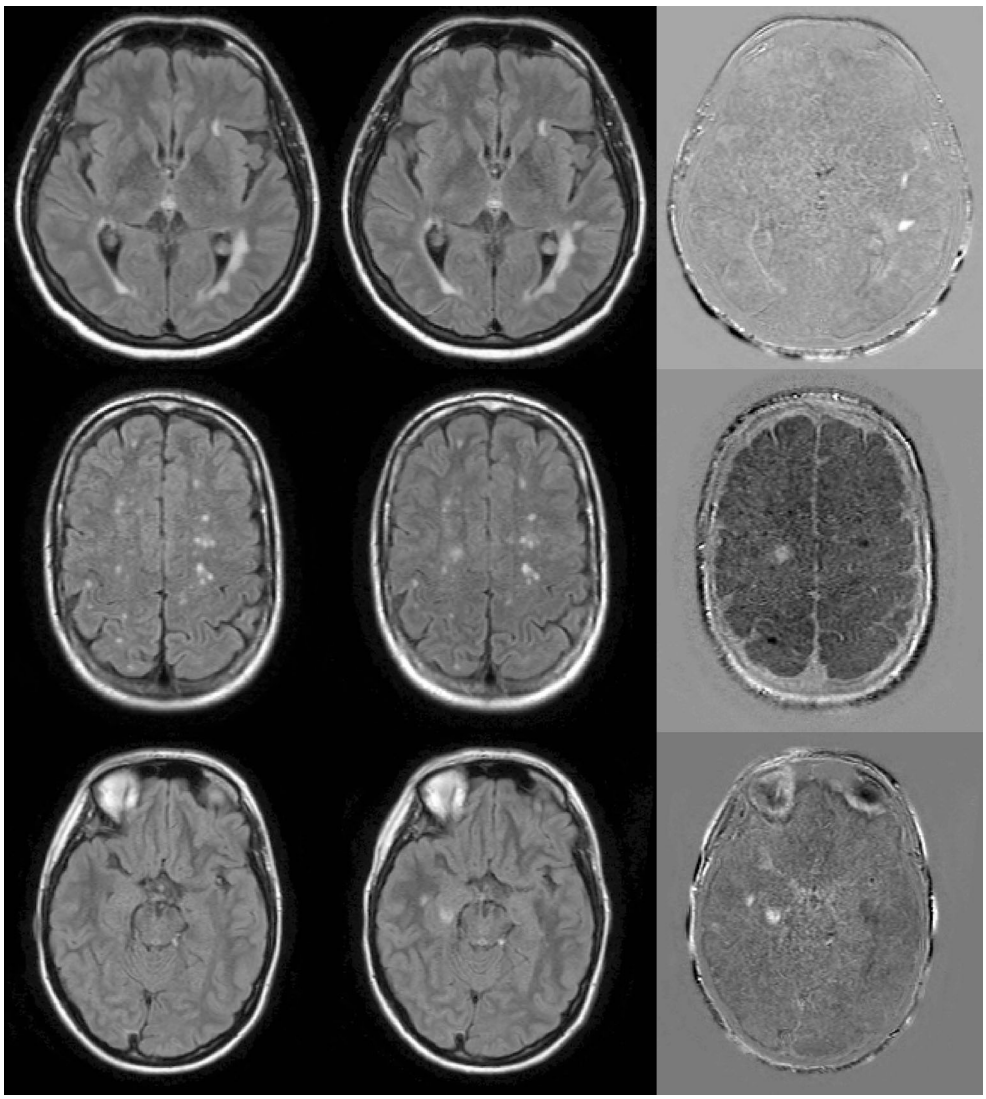

Fig. 5 Axial slices of FLAIR MRI images from three different MS patients: baseline image (1st column), follow-up (2nd column) and their difference (3rd column)

Table 3 Categorization of lesion changes (LCs) according to their volume

Category	LC voxel count	LC volume (mm^3)	Number of LCs	Number of patients
Very small	up to 10	up to 24	161	20
Small	11–20	24–48	201	20
Medium	21–50	48–120	205	20
Large	51–100	120–240	110	18
Very large	101 and larger	240 and larger	126	18

LRM resulted in a lower median AUC (0.78) and had substantially higher variance (IQR: 0.18).

Dissimilarity Map Segmentation

The CLT and CVAHT segmentations of all DMs were tested and the detected WML changes were evaluated by comparing their segmentations to the reference changes using the DSC. To evaluate the results obtained by the CLT segmentation, we performed a leave-one-out training/validation to determine the optimal confidence level α with respect to DSC. For CVAHT, the optimal offset threshold level (ε) for the given validation datasets was determined similarly as for CLT, using leave-one-out training. Figure 9 shows box-whisker plots of DSCs obtained by the CLT, CVAHT, and optimal thresholding on the STI, GLR, and LRM dissimilarity maps across all 20 datasets, while Table 4 reports the median DSCs and corresponding IQRs. Based to the optimal DM threshold, the GLR achieved the highest median DSC of 0.57 (IQR: 0.17), followed by STI with median DSC of 0.54 (IQR: 0.13) and LRM with median DSC of 0.43 (IQR: 0.33). Among the tested DMs, the GLR had the highest and most consistent DSC regardless of the DM segmentation approach. However, compared to the optimal threshold, the CLT and CVAHT segmentations performed significantly worse (Wilcoxon signed rank test, $p<0.05$) on all three DMs.

In order to quantify the overlap of the number of detected and reference changes with respect to the LC volume, a Regional DSC (RDSC) was computed. To find corresponding volume related changes, a connected components analysis was performed on both the reference change mask and on the computed change mask. Two connected components, one in the computed change mask and the other in the reference change mask, had to overlap in at least one voxel in order to be considered corresponding. The number of true positives (TP) was then determined as the number of overlapping voxels between the corresponding connected components. Voxels belonging to the connected component in reference change mask that did not overlap with the component in the computed change mask were considered false negatives (FN) and, vice versa, voxels in the connected component in the computed change mask that did not overlap with the connected component in the reference change mask were considered false positives (FP). The obtained numbers of TPs, FNs and FPs are illustrated in Fig. 6 and were used to compute the RDSC, similarly to DSC. Besides the RDSC, the detection success rate was computed, such that a particular lesion was considered to be successfully detected if the obtained LC overlapped with the ground truth LC in at least 5 % of voxels or at least 1 voxel for the group of very small changes.

Combined box-whisker plots and graphs of RDSC and detection success rate, respectively, are shown in Fig. 10 with respect to the LC volume. To determine the optimal combination of DM creation (STI, GLR or LRM) and segmentation (CLT, CVAHT or Optimal threshold) for each of the five LC groups, all combinations were evaluated. The results indicated that the success rate of LC detection, in general, increased with the volume of LCs, which was especially apparent for the GLR, whose success rate was significantly higher (Wilcoxon signed rank test, $p<0.05$) for medium and very large LCs as compared to small and very small LCs. The median RDSC of successfully detected lesions ranged from 0.29 to 0.71, where larger LCs generally had higher values than small LCs. To indicate significant differences between the DMs and between CLT or CVAHT and the Optimal DM thresholding, the Wilcoxon rank sum test was performed at a significance level of 0.05. The significances are indicated in Fig. 10.

Effect of Postprocessing

The goal of size based filtering is to decrease the FPs. The parameter ϑ of the size based filter was set to 7.29 mm^3, which corresponded to 3 voxels [21]. To demonstrate the effect on the median DSC and median FPR, Table 4 reports corresponding values before and after the postprocessing. The DSC of change segmentations was consistently increased by the postprocessing, overall by 3.9 % (IQR: 13.9 %). The overall number of FPs decreased by 26.8 % (IQR: 44.3 %) with the application of postprocessing. In STI and LRM DMs combined with any of the three segmentations reduced the FPR by at least 17 % and at most 67 %, while the reduction of FPR was negligible (<5 %) on GLR as GLR seems to effectively

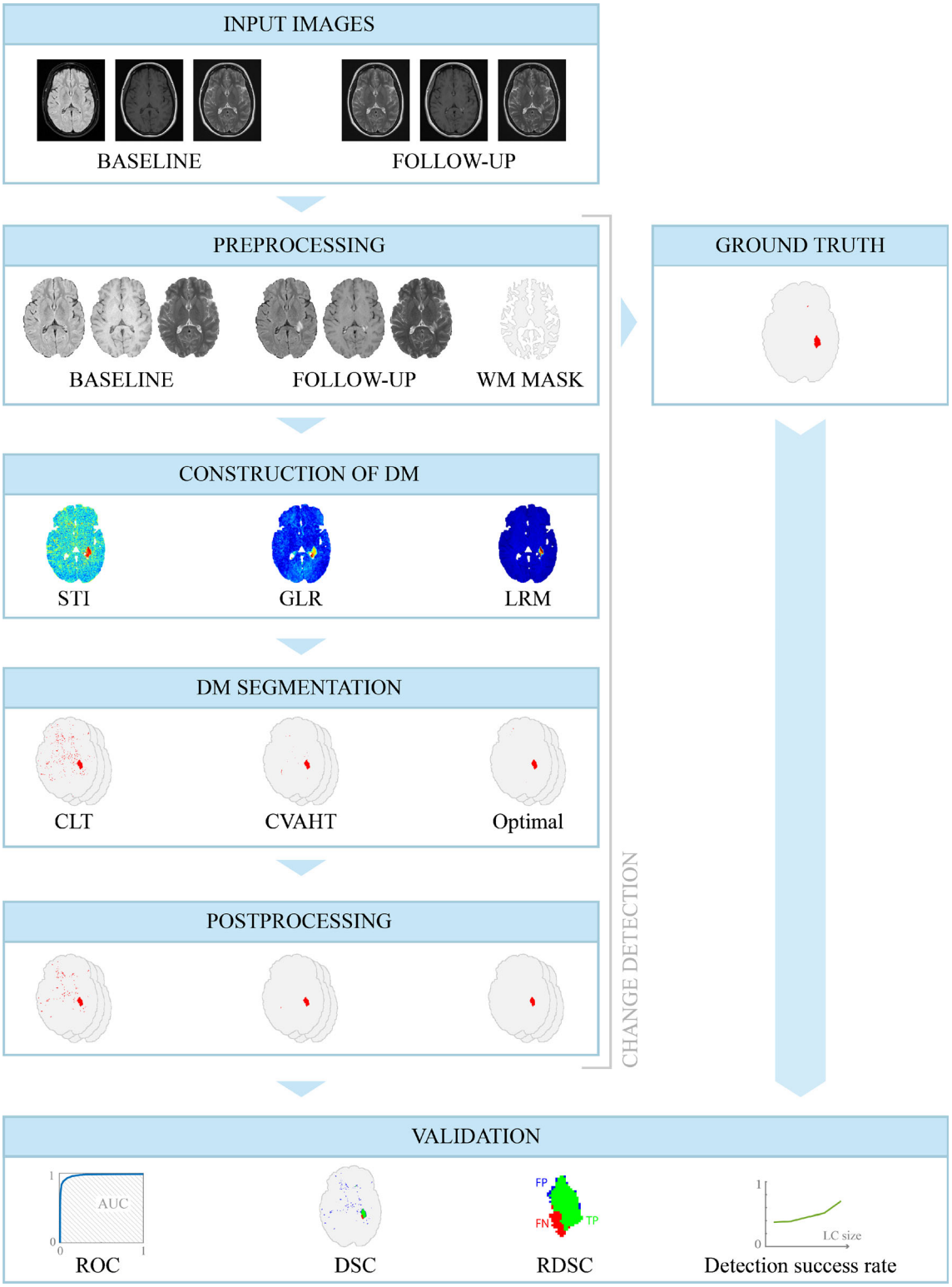

Fig. 6 Change detection validation pipeline. Baseline and follow-up input images are preprocessed (Fig. 2) and used to create the ground truth segmentation. They are also used to construct three different DMs: STI, GLR and LRM. Each DM is segmented with three different methods (CLT, CVAHT and optimal thresholding) resulting in 9 different segmentations of lesion changes. Postprocessing involving size-based filtering is performed the segmentations to reduce false positives, then, the obtained segmentations were validated against the ground truth by computing metrics like ROC analysis, voxel-wise Dice similarity coefficient (DSC), regional DSC and detection success rate

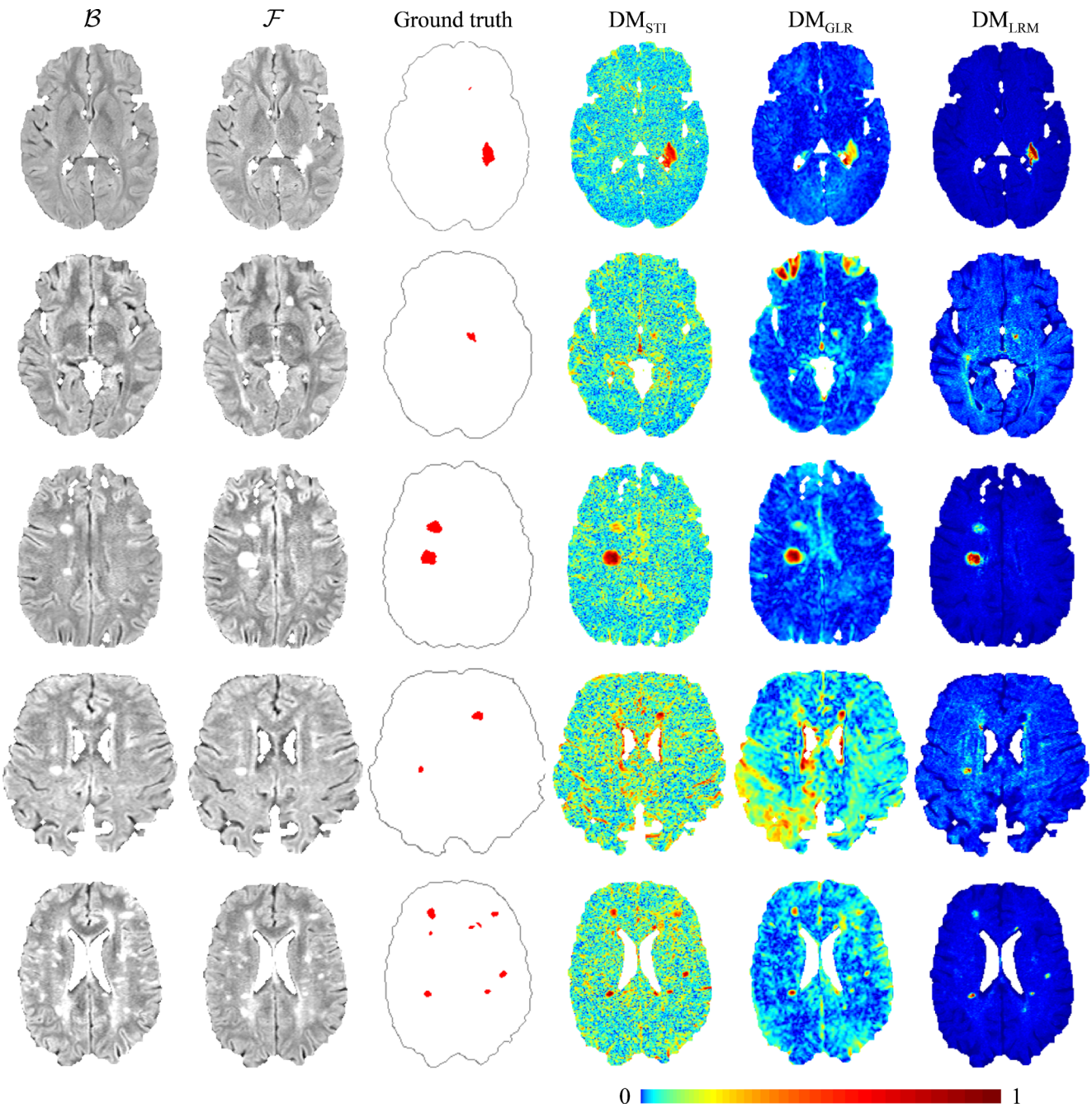

Fig. 7 Axial cross-sections of corresponding baseline (*1st column*) and follow-up (*2nd column*) FLAIR images of seven patients and the consensus ground truth segmentation of changes (*3rd column*). The right three columns show the corresponding dissimilarity maps DM_{STI}, DM_{GLR} and DM_{LRM}

suppress small volume changes through local window-based smoothing (cf. Eq. (3)).

Discussion

In this paper, we objectively evaluated three intensity based methods and their variants for the detection of lesion changes on longitudinal MR image datasets of 20 MS patients. On each dataset, accurate manual reference delineations of changes between the baseline and follow-up MR images were created by the consensus of two expert raters. The goal of the present research was to provide an objective evaluation of the main steps (computation of DM, segmentation of DM and postprocessing) common to intensity-based change detection methods. Namely, the tested methods mainly differ in the way the DM is computed and segmented. To the best of our knowledge these methods have so far been tested either on public

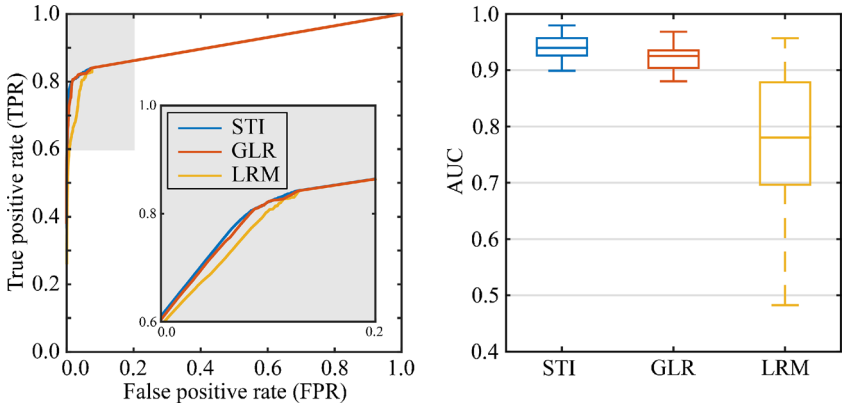

Fig. 8 ROC curves (*left*) and AUC (*right*) for the three DMs based on CLT segmentation across 20 MR datasets

datasets, comprising a rather limited number of synthetic MR images (e.g. BrainWeb) (Cocosco et al. 1997) or on a specific MR image dataset available only to the authors. Based on the reported results in the literature it is, therefore, impossible to objectively and reliably compare the performance of these methods. Hence, one of our main contributions is an objective, quantitative and systematic evaluation of the methods performed on 20 image datasets, using common pre- and post-processing steps and common evaluation metrics. Furthermore, various combinations of steps for DM computation and segmentation were tested to identify the best performing combination.

In order to systematically evaluate and compare the performance of the methods, the MR images were first preprocessed by a common preprocessing pipeline, which consisted of brain extraction, intensity inhomogeneity correction, spatial image co-registration, intensity normalization and white-matter masking (Fig. 2). A similar preprocessing pipeline was recently used by Ganiler et al. (2014) but the preprocessing pipeline used herein differs from Ganiler's in two important aspects. First, Ganiler et al. (2014) employed a histogram matching based intensity normalization (Shah et al. 2011), which requires learning and might, in cases with large discrepancy of pathology, incorrectly match the intensity levels (Shinohara et al. 2014). Instead, we normalized the intensities of each MR modality using the estimated mean and standard deviation of the WM intensity distribution as in (Shinohara et al. 2014; Sweeney et al. 2013). Second, to avoid atlas based WM segmentation, which requires nonlinear registration and is thus generally less robust, we used the Atropos method (Avants et al. 2011) for WM segmentation. The performance of the preprocessing pipeline was assessed only qualitatively. The main observations were that the preprocessing pipeline performs very reliably if the baseline and follow-up images are acquired on the same MR machine and if the same acquisition protocols are used.

When images are acquired on different scanners, which is a common situation in clinical practice, the proposed preprocessing pipeline would likely need to be improved. We managed to obtain a longitudinal dataset of three cases with baseline and follow-up MR study acquired on two scanners of different vendors. Assessment of the performance of our preprocessing pipeline on these three cases showed that, because of a high discrepancy between the intensity levels of the same tissue on the baseline and follow-up MR images, the intensity normalization step may need to be improved. Histogram matching based normalization method for multi-scanner data developed by Shah et al. (2011) proved insufficient as it may badly scale the lesion intensity (Shinohara et al. 2014). In general, there are rare cases of a large initial misregistration between the baseline and follow-up MR images where the preprocessing may fail, however, this can be easily detected and corrected by a coarse manual registration prior to running the preprocessing. The degree of influence of the

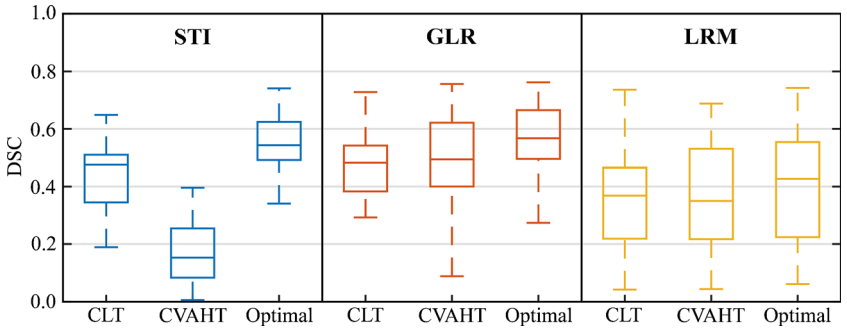

Fig. 9 Performance of three DMs and three segmentations in terms of DSC. The CLT and CVAHT segmentations performed significantly worse (Wilcoxon signed rank test, $p < 0.05$) compared to optimal thresholding on all three DMs

Table 4 Effect of postprocessing on the various instances of white matter lesion segmentation

Metric	Postprocessing	Segmentation↓/ Dissimilarity→	STI	GLR	LRM
Median DSC (IQR)	Before	CLT	0.48 (0.16)	0.48 (0.16)	0.37 (0.25)
		CVAHT	0.15 (0.17)	0.49 (0.22)	0.35 (0.31)
		Optimal threshold	0.54 (0.13)	0.57 (0.17)	0.43 (0.33)
	After	CLT	0.52 (0.24)	0.51 (0.20)	0.38 (0.31)
		CVAHT	0.19 (0.20)	0.49 (0.22)	0.36 (0.34)
		Optimal threshold	0.58 (0.13)	0.57 (0.17)	0.45 (0.34)
Median FPR $\times 10^{-2}$ ($IQR \times 10^{-2}$)	Before	CLT	0.31 (0.34)	0.24 (0.26)	0.25 (0.21)
		CVAHT	10.26 (11.71)	0.09 (0.08)	0.23 (0.24)
		Optimal threshold	0.24 (0.40)	0.25 (0.32)	0.25 (0.44)
	After	CLT	0.09 (0.09)	0.23 (0.24)	0.15 (0.12)
		CVAHT	8.59 (11.91)	0.09 (0.08)	0.15 (0.14)
		Optimal threshold	0.08 (0.19)	0.24 (0.303)	0.12 (0.32)

Results are shown as the median DSC (IQR) and the median FPR (IQR) before and after postprocessing

preprocessing pipeline on the measurement error of lesion change volume should be well understood before any image-based method that measures lesion changes can be applied for monitoring disease progress in the clinic. In the proposed preprocessing pipeline, we have employed the best methods for brain extraction, bias field correction, intensity normalization and image registration, which are all publicly available and which were extensively validated in the context of brain image analysis (Klein et al. 2009, Diez et al. 2013, Roura et al. 2014; Shinohara et al. 2014). Therefore, rather than focusing on the impact of the methods employed for preprocessing, this paper focused mainly on previously inadequately validated methods to detect and segment lesion changes between two images, a rigorous approach to assess their performance and on creation of a clinical dataset with reference change delineations.

After the preprocessing, a dissimilarity map is computed and segmented to obtain a change mask. The DM's capacity to capture changes and the segmentation results were evaluated individually using ROC analysis and quantitative metrics such as the AUC, DSC and RDSC. Besides, different combinations of the DM computation and segmentation methods were evaluated using the same quantitative metrics. Comparative evaluation of three DM variants (Fig. 8) showed a high median AUC (>0.93) for the unsupervised STI and GLR based DMs, but a lower median AUC (0.78) for supervised LRM based DM. The STI is easy to extend from one to two or more MR modalities, however, the best results were obtained by using FLAIR only. The reason is that while near the ventricles the detection of lesion changes from T2w is adversely affected by high-intensity CSF signals, the T1w indicates only a subtype of WMLs corresponding to chronic tissue injury or severe inflammatory edema (Ge 2006). These phenomena introduce undesired variations into the values of DM, which renders it more difficult for segmentation. The GLR based DM computation involves estimation of the WM noise covariance matrix from MR intensities within the WM mask. Possible errors in WM segmentation could therefore directly impact the performance of the GLR method. Although the lowest AUC score was obtained for the LRM method, the resulting DM seems very resilient to image noise and FP intensity differences near edges of the WM mask (Fig. 7). The LRM based DM was executed by using the regression coefficients provided by the authors (Sweeney et al. 2013). For this reason care was taken to normalize the MR image intensities as described in (Sweeney et al. 2013). By retraining the regression coefficients on the current MR image dataset, the results would most likely improve, however, this would

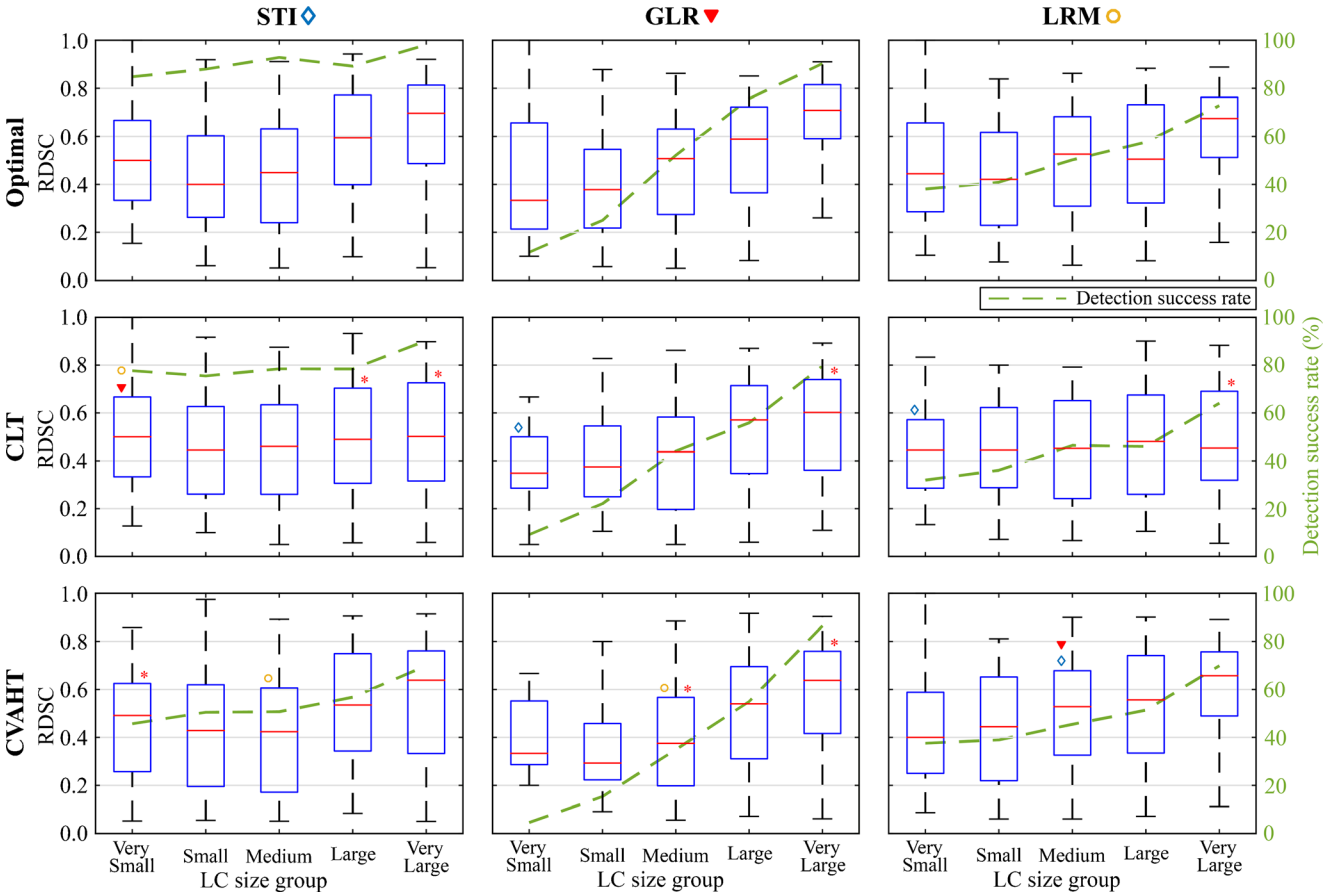

Fig. 10 Performance of change detection in terms of RDSC and detection success rate for all combinations of three DMs and three DM segmentations. A *star* (*) indicates significantly lower RDSC (Wilcoxon signed rank test, $p < 0.05$) in comparison to optimal thresholding on the same DM. A *diamond* (◊), *triangle* (▼) or *circle* (○) indicate significant difference (Wilcoxon signed rank test, $p < 0.05$) between STI, GLR or LRM, respectively, for the same DM segmentation method

certainly bias the evaluation in favor of the LRM based DM. Hence, we feel it is of greater interest to use the provided regression coefficients trained on one dataset and evaluate the method performances on another MR image dataset.

Two automated methods for DM segmentation were evaluated, namely confidence level thresholding (CLT) and change vector angular histogram thresholding (CVAHT). For comparison purposes, the optimal threshold was determined individually for each dataset by maximizing the DSC. Both CLT and CVAHT performed worse than optimal thresholding in terms of DSC (Table 4), which was also verified by the Wilcoxon rank sum test that indicated statistically significant ($p < 0.05$) difference compared to the DSC of segmentation based on the optimal threshold. This result suggests that, since CLT and CVAHT segmentations are relatively simple approaches, the use of more advanced or the development of novel methods for DM segmentation could substantially improve the accuracy of change detection.

The potential of change detection methods was also analyzed with respect to the volume of LC (Fig. 10). Among the tested methods the combination of STI based DM and any of the three segmentations resulted in the highest and most stable detection success rates with respect to the different volumes of LCs. On the other hand, the use of GLR based DM rendered the change detection highly sensitive to the volume of LCs, since less than 50 % of LCs of very small, small and medium volumes were successfully detected. Besides, the RDSC of the GLR varied most with respect to different LC volumes. This result also suggests that the high DSC obtained for the GLR (Table 4) is mainly due good performance on large LCs.

For very small LCs the RDSC of STI in combination with the CLT segmentation was significantly higher compared to the GLR and LRM. However, compared to the optimal threshold, the CLT segmentation applied to any of the DMs resulted in significantly lower RDSC for very large LCs as well as for large LCs when using STI. The reason is that the value of confidence level α corresponds to the expected volume of LCs, thus a fixed value of α inevitably leads to suboptimal change segmentation performance on real datasets, in which the volume of LCs is generally quite varying. Overall, the combination of STI dissimilarity map construction and CLT

segmentation provides fairly accurate change detection in terms of DSC (0.48) and RDSC (>0.45) and seems quite reliable as it has the highest and the most consistent detection success rate (>75 %) across different volumes of LCs. However, it is yet to be determined if such a performance is sufficient for LC measuments for clinical use.

From a clinical point of view, the expected total volume of hyperintense T2 lesion change observed RRMS patients in a span of 1 year is 0.25 ± 0.5 ml (Giorgio et al. 2014). Based on our MR studies with a $1 \times 1 \times 3$ mm 2D FLAIR acquisition, which adheres to current clinical guidelines for MR imaging of MS patients (Rovira et al. 2015), this amounts to total lesion change volume of 83 ± 167 voxels. Clearly, using a 1 mm isotropic 3D FLAIR acquisition would remedy this problem to some extent, however, due to extended scanning time and possible motion artifacts this is not always feasible in clinical practice. Hence, robust detection of the very small lesions (up to 10 voxels) could prove important in achieving a reliable measurement for clinical use.

Change detection accuracy can be further improved by applying a postprocessing step which reduces the false positives. Size based filtering discards any regions in the change mask that are smaller than 3 voxels. In the present study, postprocessing increased the median DSC by 3.9 % (IQR: 13.9 %) and reduced the median FPR by 26.8 % (IQR: 44.3 %). We have tested two additional postprocessing methods described in (Ganiler et al. 2014), however, they did not improve the final results on our datasets. The first of the two methods aims to remove false positives near the WM-GM tissue interface, which are mainly caused by misregistration. It seems that due to good registration and precise WM masking in the preprocessing step, this had no impact on the final results. The second postprocessing method compares the intensity in the region of LC against the intensities of neighboring voxels, excluding the region if the intensity difference is below some threshold. However, applying this postprocessing did not improve the results. Besides, this method seems to perform well only in cases of newly appearing lesions, while it may discard true positives around an existing lesion that only changed in shape, volume or intensity.

Alternative strategy to remove false positives at characteristic locations like the WM-GM tissue interface or the medial longitudinal fissure could be by masking based on co-registered atlas. This approach would, however, require a good nonlinear registration of the atlas. To differentiate changes of existing lesions from false positives one strategy is to perform temporal lesion change shape modeling (Goldberg-Zimring et al. 2003) and then exclude only those detected changes that are not well captured by the model. In a similar way, machine learning methods that recently gained momentum could be applied for this purpose (Wang et al. 2011). The downside of these approaches is that they require highly accurate lesion segmentation and a large number of training datasets.

All of the tested methods assume, in some part, that there definitely exist changes between the baseline and follow-up MR images. Because of this assumption, the methods generally return a huge amount of false positives if applied to cases with very little or no changes. In order to fully automate the change detection, a method should either be capable to deal with such cases or be followed by postprocessing which is able to better detect and eliminate false positives. Simple postprocessing methods have limited success, but recent machine learning algorithms based on high-level features derived from appearance, shape and location of change regions in the change mask probably have a much higher potential (Wang et al. 2011).

Accurate and reliable detection of structural changes from longitudinal MR brain images remains a challenging task, since it requires careful tuning of all steps involved in the change detection process according to the quality of input MR images. Surprisingly, the results obtained in this study were not nearly as good as the ones reported by the authors (Ganiler et al. 2014; Simoes and Slump 2011; Sweeney et al. 2013). This suggests that the performances of the evaluated change detection methods might either be very data dependent or dependent on the accuracy of ground truth segmentation. It is important to note that for the purpose of this study the MR machine and acquisition protocols were the same across all longitudinal MR datasets. This is often not the case in clinical practice, but nevertheless it is expected that a change detection method will perform consistently in terms of accuracy and reliability on datasets acquired across different MR machines and acquisition parameters. However, to objectively evaluate a method a repository containing multiple publicly available longitudinal MR datasets acquired on various scanners with accurate reference change delineations is required. As a step towards this goal and to enable other researchers to reproduce the results in this study, we intend to further expand and publicly disseminate our longitudinal MR image datasets on our website http://lit.fe.uni-lj.si/tools. We consider this work to be the first, but important step in the direction of creating an ecosystem of publicly available resources for validation and comparison of longitudinal change detection methods.

Information Sharing Statement

The magnetic resonance (MR) images for this study were acquired on a 1.5 T Philips MR machine at the University Medical Centre Ljubljana (UMCL). All 20 subjects have given written informed consent at the time of enrollment for imaging. The authors, who have obtained approval from the UMCL to use the data, confirm that the data was anonymized. The MR datasets and ground truth segmentations of white-matter lesion changes will be made publicly available in raw raster format on our website http://lit.fe.uni-lj.si/tools.

Acknowledgments This research was supported by the Ministry of Education, Science and Sport, Slovenia, under grants J2-5473, L2-5472, and J7-6781. The authors would also like to acknowledge A.K. and M.L. from the University Medical Centre Ljubljana for creating the reference segmentations.

References

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., & Gee, J. C. (2011). An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics, 9*(4), 381–400. doi:10.1007/s12021-011-9109-y.

Avants, B. B., Tustison, N. J., Stauffer, M., Song, G., Wu, B., & Gee, J. C. (2014). The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, *8*. doi:10.3389/fninf.2014.00044.

Avants, B. B., Tustison, N. J., & Johnson, H. J. (n.d.). Advanced normalization tools (ANTs). http://stnava.github.io/ANTs/. Accessed 16 Mar 2016.

Battaglini, M., Rossi, F., Grove, R. A., Stromillo, M. L., Whitcher, B., Matthews, P. M., & De Stefano, N. (2014). Automated identification of brain new lesions in multiple sclerosis using subtraction images. *Journal of Magnetic Resonance Imaging, 39*(6), 1543–1549. doi:10.1002/jmri.24293.

Bosc, M., Heitz, F., Armspach, J. P., Namer, I., Gounot, D., & Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage, 20*(2), 643–656. doi:10.1016/S1053-8119(03)00406-3.

Cocosco, C. A., Kollokian, V., Kwan, R. K.-S., Pike, G. B., & Evans, A. C. (1997). BrainWeb: online interface to a 3D MRI simulated brain database. *NeuroImage, 5*, 425.

Diez, Y., Oliver, A., Cabezas, M., Valverde, S., Martí, R., Vilanova, J. C., et al. (2013). Intensity based methods for brain MRI longitudinal registration. a study on multiple sclerosis patients. *Neuroinformatics, 12*(3), 365–379. doi:10.1007/s12021-013-9216-z.

Duan, Y., Hildenbrand, P. G., Sampat, M. P., Tate, D. F., Csapo, I., Moraal, B., et al. (2008). Segmentation of subtraction images for the measurement of lesion change in multiple sclerosis. *AJNR. American Journal of Neuroradiology, 29*(2), 340–346. doi:10.3174/ajnr.A0795.

Elliott, C., Arnold, D. L., Collins, D. L., & Arbel, T. (2013). Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Transactions on Medical Imaging, 32*(8), 1490–1503. doi:10.1109/TMI.2013.2258403.

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology, 56*(5), 363–374. doi:10.1007/s00234-014-1343-1.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., & Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical Image Analysis, 17*(1), 1–18. doi:10.1016/j.media.2012.09.004.

Ge, Y. (2006). Multiple sclerosis: the role of MR imaging. *American Journal of Neuroradiology, 27*(6), 1165–1176.

Giorgio, A., Stromillo, M. L., Bartolozzi, M. L., Rossi, F., Battaglini, M., De Leucio, A., … & Amato, M. P. (2014). Relevance of hypointense brain MRI lesions for long-term worsening of clinical disability in relapsing multiple sclerosis. *Multiple Sclerosis Journal*, *20*(2), 214–219.

Goldberg-Zimring, D., Achiron, A., Guttmann, C. R., & Azhari, H. (2003). Three-dimensional analysis of the geometry of individual multiple sclerosis lesions: detection of shape changes over time using spherical harmonics. *Journal of Magnetic Resonance Imaging, 18*(3), 291–301.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M. C., … & Song, J. H. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, *46*(3), 786–802.

Lebrun, C., Bensa, C., Debouverie, M., De Seze, J., Wiertlievski, S., Brochet, B., et al. (2008). Unexpected multiple sclerosis: follow-up of 30 patients with magnetic resonance imaging and clinical conversion profile. *Journal of Neurology, Neurosurgery, and Psychiatry, 79*(2), 195–198. doi:10.1136/jnnp.2006.108274.

Llado, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J. C., Quiles, A., et al. (2012). Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. *Information Sciences, 186*(1), 164–185. doi:10.1016/j.ins.2011.10.011.

Lladó, X., Ganiler, O., Oliver, A., Martí, R., Freixenet, J., Valls, L., et al. (2012). Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology, 54*(8), 787–807. doi:10.1007/s00234-011-0992-6.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging, 16*(2), 187–198. doi:10.1109/42.563664.

Moraal, B., Meier, D. S., Poppe, P. A., Geurts, J. J. G., Vrenken, H., Jonker, W. M. A., et al. (2009). Subtraction MR images in a multiple sclerosis multicenter clinical trial setting. *Radiology, 250*(2), 506–514. doi:10.1148/radiol.2501080480.

Moraal, B., van den Elskamp, I. J., Knol, D. L., Uitdehaag, B. M. J., Geurts, J. J. G., Vrenken, H., et al. (2010a). Long-interval T2-weighted subtraction magnetic resonance imaging: a powerful new outcome measure in multiple sclerosis trials. *Annals of Neurology, 67*(5), 667–675. doi:10.1002/ana.21958.

Moraal, B., Wattjes, M. P., Geurts, J. J. G., Knol, D. L., van Schijndel, R. A., Pouwels, P. J. W., et al. (2010b). Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. *Radiology, 255*(1), 154–163. doi:10.1148/radiol.09090814.

Nika, V., Babyn, P., & Zhu, H. (2014). EigenBlock algorithm for change detection – an application of adaptive dictionary learning techniques. *Journal of Computational Science, 5*(3), 527–535. doi:10.1016/j.jocs.2013.10.008.

Patriarche, J., & Erickson, B. (2004). A review of the automated detection of change in serial imaging studies of the brain. *Journal of Digital Imaging, 17*(3), 158–174. doi:10.1007/s10278-004-1010-x.

Patti, F., De Stefano, M., Lavorgna, L., Messina, S., Chisari, C. G., Ippolito, D., et al. (2015). Lesion load may predict long-term cognitive dysfunction in multiple sclerosis patients. *PloS One, 10*(3), e0120754. doi:10.1371/journal.pone.0120754.

Pham, D. (n.d.). Longitudinal Multiple Sclerosis Lesion Segmentation Challenge | ISBI 2015. http://biomedicalimaging.org/2015/3d-segmentation-of-neurites-in-em-images/. Accessed 10 Apr 2015.

Popescu, V., Agosta, F., Hulst, H. E., Sluimer, I. C., Knol, D. L., Sormani, M. P., et al. (2013). Brain atrophy and lesion load predict long term disability in multiple sclerosis. *Journal of Neurology, Neurosurgery, and Psychiatry, 84*(10), 1082–1091. doi:10.1136/jnnp-2012-304094.

Ramirez, J., McNeely, A. A., Scott, C. J., Stuss, D. T., & Black, S. E. (2014). Subcortical hyperintensity volumetrics in Alzheimer's disease and normal elderly in the Sunnybrook Dementia Study: correlations with atrophy, executive function, mental processing speed, and verbal memory. *Alzheimer's Research & Therapy, 6*(4), 49. doi:10.1186/alzrt279.

Rey, D., Subsol, G., Delingette, H., & Ayache, N. (2002). Automatic detection and segmentation of evolving processes in 3D medical images: application to multiple sclerosis. *Medical Image Analysis, 6*(2), 163–179. doi:10.1016/S1361-8415(02)00056-7.

Risacher, S. L., Saykin, A. J., West, J. D., Shen, L., Firpi, H. A., McDonald, B. C., & Alzheimer's Disease Neuroimaging Initiative (ADNI). (2009). Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Current Alzheimer Research, 6*(4), 347–361.

Rocca, M. A., Anzalone, N., Falini, A., & Filippi, M. (2013). Contribution of magnetic resonance imaging to the diagnosis and monitoring of multiple sclerosis. *La Radiologia Medica, 118*(2), 251–264. doi:10.1007/s11547-012-0811-3.

Roura, E., Oliver, A., Cabezas, M., Vilanova, J. C., Rovira, À., Ramió-Torrentà, L., & Lladó, X. (2014). MARGA: multispectral adaptive region growing algorithm for brain extraction on axial MRI. *Computer Methods and Programs in Biomedicine, 113*(2), 655–673. doi:10.1016/j.cmpb.2013.11.015.

Rousseau, F., Faisan, S., Heitz, F., Armspach, J.-P., Chevalier, Y., & Blanc, F., et al. (2007). An A Contrario Approach for Change Detection in 3D Multimodal Images: Application to Multiple Sclerosis in MRI. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007* (pp. 2069–2072). Presented at the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007. doi:10.1109/IEMBS.2007.4352728.

Rovira, À., Wattjes, M. P., Tintoré, M., Tur, C., Yousry, T. A., Sormani, M. P., … & Barkhof, F. (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis – clinical implementation in the diagnostic process. *Nature Reviews Neurology.*

Seo, H. J., & Milanfar, P. (2009). A non-parametric approach to automatic change detection in MRI images of the brain. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009. ISBI '09* (pp. 245–248). Presented at the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. doi:10.1109/ISBI.2009.5193029.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., & Arbel, T. (2011). Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis, 15*(2), 267–282. doi:10.1016/j.media.2010.12.003.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage Clinical, 6*, 9–19. doi:10.1016/j.nicl.2014.08.008.

Simoes, R., & Slump, C. (2011). Change detection and classification in brain MR images using change vector analysis. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society,EMBC* (pp. 7803–7807). Presented at the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society,EMBC. doi:10.1109/IEMBS.2011.6091923.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping, 17*(3), 143–155. doi:10.1002/hbm.10062.

Studholme, C., Drapaca, C., Iordanova, B., & Cardenas, V. (2006). Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change. *IEEE Transactions on Medical Imaging, 25*(5), 626–639. doi:10.1109/TMI.2006.872745.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., et al. (2008). 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. *MIDAS Journal*, 1–6. Accessed 9 Mar 2015.

Susanto, T. A. K., Pua, E. P. K., & Zhou, J. (2015). Cognition, brain atrophy, and cerebrospinal fluid biomarkers changes from preclinical to dementia stage of Alzheimer's disease and the influence of apolipoprotein e. *Journal of Alzheimer's Disease, 45*(1), 253–268. doi:10.3233/JAD-142451.

Sweeney, E. M., Shinohara, R. T., Shea, C. D., Reich, D. S., & Crainiceanu, C. M. (2013). Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *American Journal of Neuroradiology, 34*(1), 68–73. doi:10.3174/ajnr.A3172.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging, 29*(6), 1310–1320. doi:10.1109/TMI.2010.2046908.

Vrenken, H., Jenkinson, M., Horsfield, M. A., Battaglini, M., van Schijndel, R. A., Rostrup, E., et al. (2013). Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis. *Journal of Neurology, 260*(10), 2458–2471. doi:10.1007/s00415-012-6762-5.

Wang, H., Das, S. R., Suh, J. W., Altinay, M., Pluta, J., Craige, C., … & Alzheimer's Disease Neuroimaging Initiative (2011). A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, *55*(3), 968–985.

Wei, X., Guttmann, C. R. G., Warfield, S. K., Eliasziw, M., & Mitchell, J. R. (2004). Has your patient's multiple sclerosis lesion burden or brain atrophy actually changed? *Multiple Sclerosis (Houndmills, Basingstoke, England), 10*(4), 402–406.