NeuroImage: Clinical 3 (2013) 462–469

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs)☆

Martijn D. Steenwijk [a,*], Petra J.W. Pouwels [b], Marita Daams [a,c], Jan Willem van Dalen [d], Matthan W.A. Caan [e], Edo Richard [d], Frederik Barkhof [a], Hugo Vrenken [a,b]

[a] Department of Radiology and Nuclear Medicine, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands
[b] Department of Physics and Medical Technology, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands
[c] Department of Anatomy and Neurosciences, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands
[d] Department of Neurology, Academic Medical Centre Amsterdam, The Netherlands
[e] Department of Radiology, Academic Medical Centre Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 20 September 2013
Received in revised form 1 October 2013
Accepted 3 October 2013
Available online 14 October 2013

Keywords:
Segmentation
White matter lesions
Multiple sclerosis
Cerebrovascular disease
MRI

ABSTRACT

Introduction: The segmentation and volumetric quantification of white matter (WM) lesions play an important role in monitoring and studying neurological diseases such as multiple sclerosis (MS) or cerebrovascular disease. This is often interactively done using 2D magnetic resonance images. Recent developments in acquisition techniques allow for 3D imaging with much thinner sections, but the large number of images per subject makes manual lesion outlining infeasible. This warrants the need for a reliable automated approach. Here we aimed to improve *k* nearest neighbor (*k*NN) classification of WM lesions by optimizing intensity normalization and using spatial tissue type priors (TTPs).

Methods: The *k*NN-TTP method used *k*NN classification with 3.0 T 3DFLAIR and 3DT1 intensities as well as MNI-normalized spatial coordinates as features. Additionally, TTPs were computed by nonlinear registration of data from healthy controls. Intensity features were normalized using variance scaling, robust range normalization or histogram matching. The algorithm was then trained and evaluated using a leave-one-out experiment among 20 patients with MS against a reference segmentation that was created completely manually. The performance of each normalization method was evaluated both with and without TTPs in the feature set. Volumetric agreement was evaluated using intra-class coefficient (ICC), and voxelwise spatial agreement was evaluated using Dice similarity index (SI). Finally, the robustness of the method across different scanners and patient populations was evaluated using an independent sample of elderly subjects with hypertension.

Results: The intensity normalization method had a large influence on the segmentation performance, with average SI values ranging from 0.66 to 0.72 when no TTPs were used. Independent of the normalization method, the inclusion of TTPs as features increased performance particularly by reducing the lesion detection error. Best performance was achieved using variance scaled intensity features and including TTPs in the feature set: this yielded ICC = 0.93 and average SI = 0.75 ± 0.08. Validation of the method in an independent sample of elderly subjects with hypertension, yielded even higher ICC = 0.96 and SI = 0.84 ± 0.14.

Conclusion: Adding TTPs increases the performance of *k*NN based MS lesion segmentation methods. Best performance was achieved using variance scaling for intensity normalization and including TTPs in the feature set, showing excellent agreement with the reference segmentations across a wide range of lesion severity, irrespective of the scanner used or the pathological substrate of the lesions.

1. Introduction

Focal white matter (WM) pathology in the brain has been associated with various disorders, including multiple sclerosis (MS), cerebrovascular disease and dementia. Magnetic resonance imaging (MRI) plays a key role in diagnosing, monitoring and studying these diseases (Polman et al., 2011; Provenzano et al., 2013). Perhaps one of the most important contributions of MRI is that it can be used to visualize lesions in the WM. Treatment effects are studied in clinical trials by counting these lesions

☆ This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author at: VU University Medical Center, Department of Radiology and Nuclear Medicine, PO Box 7057, 1007 MB Amsterdam, The Netherlands. Tel.: +31 20 444 4596; fax: +31 20 444 0397.
E-mail address: m.steenwijk@vumc.nl (M.D. Steenwijk).

2213-1582/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.nicl.2013.10.003

and quantifying their volumes through lesion segmentation, and epidemiological studies are performed to understand how the lesions affect the brain (Kappos et al., 2007; Mortamais et al., 2013).

Quantification of white matter lesions (WMLs) is traditionally performed by visual rating or manual outlining on 2D proton density (PD) weighted, T2-weighted, or fluid attenuated inversion recovery (FLAIR) images with slice thicknesses of 3 mm or more (Fazekas et al., 1987; Olsson et al., 2013; Schoonheim et al., 2012). Recent advances in acquisition techniques enable 3D imaging with much better spatial resolution, typically around 1 mm isotropic. The much larger number of images per subject makes manual outlining of lesions infeasible, and warrants the need for reliable automated lesion segmentation techniques.

A number of automated WML segmentation techniques have been described (Mortazavi et al., 2012). Based on the performance reported in literature and the explicit use of a priori information, we selected the *k*-nearest neighbor (*k*NN) method described by Anbeek et al. (2004) as a starting point for our method. *k*NN classification is a supervised pattern recognition technique, which performs segmentation by comparing new data to a collection of labeled examples in a training set. For each new voxel to be classified, the algorithm computes the probability of the voxel being a lesion, by determining the fraction of *k* nearest neighbors that were labeled as a lesion in the feature space of the training set. Previous studies showed that *k*NN classification provides good WML segmentation results when both signal intensities and spatial coordinates are used as features (Anbeek et al., 2004, 2005).

Here, we sought to improve on the method by Anbeek et al., first, by adding GM, WM and CSF tissue type priors (TTPs) derived from healthy controls to allow the inclusion of anatomical information and reduce the number of false positive voxels. The use of such tissue type information has been shown to improve WML segmentation in previous studies (Schmidt et al., 2012). Second, we optimize the method of signal intensity normalization by comparing different normalization strategies. We trained and evaluated the method in patients with MS and elderly subjects with hypertension using manually developed reference segmentations, constructed by expert raters who perform these segmentations routinely.

The aim of the present study was to quantify the effect of adding TTPs and optimizing intensity normalization on the performance of *k*NN WML classification. This was done by measuring the segmentation performance (i.e. spatial correspondence with the manual reference segmentation) of *k*NN-TTP with various intensity normalization methods, using a leave-one-out approach in a sample of MS patients. Finally, the robustness of the method across different scanners and patient populations was studied by applying it in an independent sample of elderly subjects with hypertension.

2. Materials and methods

2.1. Subjects

We primarily investigated MR images of patients with clinically-definite MS and healthy controls who were part of a larger cohort. The validation sample consisting of elderly subjects with hypertension will be described in the section '*Validation in an independent cohort of elderly subjects with hypertension*' below.

The institutional ethics review board approved the study and all subjects gave written informed consent prior to participation. From a larger study cohort, we selected a subset of 20 patients with MS showing a wide variety of pathology in terms of lesion burden. Their ages varied between 29 and 67 years (mean age: 52.5 ± 7.7 years), and 13 of them were women. Disease severity was measured on the day of scanning using the expanded disability status scale (EDSS) (Kurtzke, 1983). The median EDSS score was 4, ranging between 2.5 and 8.0. From the same cohort we randomly selected the MR images of 16 healthy controls (mean age: 51.7 ± 5.8, 8 of them were women) for use as an atlas in the TTP creation step of the segmentation method (see details below).

2.2. MR imaging

MR imaging was performed on a 3.0 T whole body scanner (GE Signa HDxt, Milwaukee, WI, USA) using an eight-channel phased-array head coil. The protocol contained among others two 3D sequences: a fat-saturated 3DFLAIR (TR: 8000 ms, TE: 125 ms, TI: 2350 ms, 250 × 250 mm^2 field of view (FOV), 132 sagittal slices of 1.2 mm thickness, 0.98 × 0.98 mm^2 in-plane resolution) for lesion detection, and a 3DT1 weighted fast spoiled gradient echo (FSPGR) sequence (TR 7.8 ms, TE 3 ms, FA 12°, 240 × 240 mm^2 FOV, 176 sagittal slices of 1 mm thickness, 0.94 × 0.94 mm^2 in-plane resolution) for anatomical information.

2.3. Manual reference segmentation

A reference WML segmentation was constructed manually using the 3DFLAIR and 3DT1 images. Before constructing the reference segmentation, the 3DT1 image of each subject was rigidly registered to its respective 3DFLAIR image using FLIRT which is part of the FMRIB Software Library (FSL 5.0.2) (Jenkinson and Smith, 2001). Subsequently, both 3DT1 and 3DFLAIR images were orthogonally reformatted to the axial plane, which resulted in 256 slices with a thickness of 0.94 mm for each dataset.

The axially reformatted images were then used to identify and outline the WMLs. Lesion identification was performed by three raters in consensus (two PhD-students with two years of experience each and an experienced neuroradiologist) using the 3DFLAIR images, while the raters were allowed to view the corresponding co-registered 3DT1 image. Lesions were only identified if they were larger than 3 voxels in-plane and visible on at least two consecutive slices. In the next step, two trained technicians manually outlined the identified lesions on the 3DFLAIR using MIPAV (http://mipav.cit.nih.gov). Each technician was randomly assigned to 10 of the 20 patients, and outlined the identified WMLs on each slice. The 20 reference segmentations thus produced were used to train and evaluate the automatic lesion segmentation algorithm.

To assess interobserver reliability of the manual segmentations, each technician also outlined six randomly selected consecutive slices of each subject assigned to the other technician. Furthermore, both technicians outlined twenty consecutive slices of one of the subjects for a second time during the project, to obtain information about intraobserver reliability.

2.4. Automatic white matter lesion segmentation

*k*NN classification compares new data with a collection of examples (i.e. the training set) in a feature space. In this feature space, each voxel is characterized by 3DFLAIR intensity, 3DT1 intensity, MNI-normalized spatial coordinates and tissue type probability. Based on the manual reference segmentations, the voxels in the training set are labeled as being lesion or not. The algorithm classifies a new voxel based on the labels of its neighbors in feature space. The full algorithm consists of five stages, namely image preprocessing, feature extraction, feature normalization, classification and post-processing. These different stages are discussed in the following sections.

2.5. Image preprocessing

First, non-brain tissue was removed from the co-registered 3DT1 image using the FSL brain extraction tool (BET) (Smith, 2002), using standardized parameters for brain extraction, including bias field correction and robust brain center estimation as recommended by Popescu et al. (2012). The resulting brain mask was also applied to the

3DFLAIR image. Finally, radio frequency (RF) field inhomogeneity correction was performed on both images using the N3 algorithm (Sled, 1997).

2.6. Feature extraction

The features used for *k*NN classification in the current study were: 3DFLAIR and 3DT1 signal intensity, MNI-normalized spatial coordinates *x*, *y* and *z*, and tissue type probabilities pCSF, pGM, and pWM (see Fig. 1).

The normalized spatial coordinates *x*, *y*, and *z* were derived by linear registration of the 3DT1 image to MNI space using FLIRT. By applying the inverse transformation, the voxelwise corresponding MNI coordinates were subsequently warped back to subject-space. This resulted in *x*, *y*, and *z* features comparable between subjects.

The TTPs were obtained using a procedure commonly referred to as *multi-atlas segmentation* as follows (Aljabar et al., 2009). For the 3DT1 images of the 16 healthy control subjects, voxelwise "hard" segmentations of CSF, GM and WM were generated using FSL-FAST (Zhang et al., 2001). Then the 3DT1 image of each healthy control was non-linearly registered to the 3DT1 image of the subject of interest using Elastix, which involved an affine and B-spline transformation, both using mutual information as cost-function, gradient descent optimizers, a four-stage pyramidal approach and a final control point resolution of 2.5 mm (Klein et al., 2010). The resulting transformations were applied to the voxelwise CSF, GM and WM segmentations using nearest neighbor interpolation. Then for each voxel the probability of being CSF, WM or GM was estimated by computing the frequency of the respective tissue class in the registered segmentations (Aljabar et al., 2009; De Boer et al., 2009).

2.7. Feature normalization

As different features have different ranges, the features should be normalized to obtain meaningful distances in feature space for selecting the *k* "nearest" neighbors. A common way of feature normalization is variance scaling, which subtracts the within-subject mean feature value from each voxel's feature value and divides the result by the within-subject standard deviation, resulting in zero mean and unit variance in the normalized feature set. This approach however, may be sensitive to differences in feature distribution, such as signal intensity distribution differences between patients with different lesion loads. We therefore also investigated the effect of two other feature normalization strategies which might be less sensitive to differences in feature distribution between subjects, namely robust range normalization (De Boer et al., 2009) and histogram matching (Lao et al., 2008; Younis et al., 2008). Robust range normalization linearly scales a feature such that the 4th percentile of the histogram is matched to value 0 and the 96th percentile is matched to value 1. Histogram matching finds, for each new patient, the linear transformation that maximizes the overlap between the normalized histogram of the transformed feature and the normalized histogram of a reference histogram. This reference histogram is selected by finding the 'most typical' histogram among the subjects, and scaling this between zero and one using robust range normalization. The histogram overlap was maximized using Genetic Algorithms, as described in Younis et al. (2008).

Since we expected non-intensity feature distributions to be relatively constant, all non-intensity features were always scaled using variance scaling.

2.8. Classification

The probability that a new voxel is a lesion was defined as the fraction of the *k* nearest examples that were labeled as being a lesion in the training set. This can be converted to a binary segmentation by applying a threshold *p* to the probability map. Based on values used in the literature (Anbeek et al., 2008), *k* was set to 40 in the current study. Using a leave-one-out procedure, a probability map was computed for each patient. Subsequently, the optimal threshold was determined by applying different thresholds $p = 0.05, 0.10, ..., 0.95$ to each probability map, and calculating the SI of the resulting binary segmentations with the manual reference segmentation. The threshold *p* resulting in the highest average SI across the 20 datasets was selected as the optimal threshold.

2.9. Post-processing

The binary segmentation sometimes contained small false positive regions, which are often too small to be considered as a true lesion. To remove these small false positive regions, we applied a simple post-processing step which removes all lesions with a volume smaller than a threshold *C*. From the binary probability maps obtained using the optimal threshold *p* in the leave-one-out procedure, the optimal *C* was

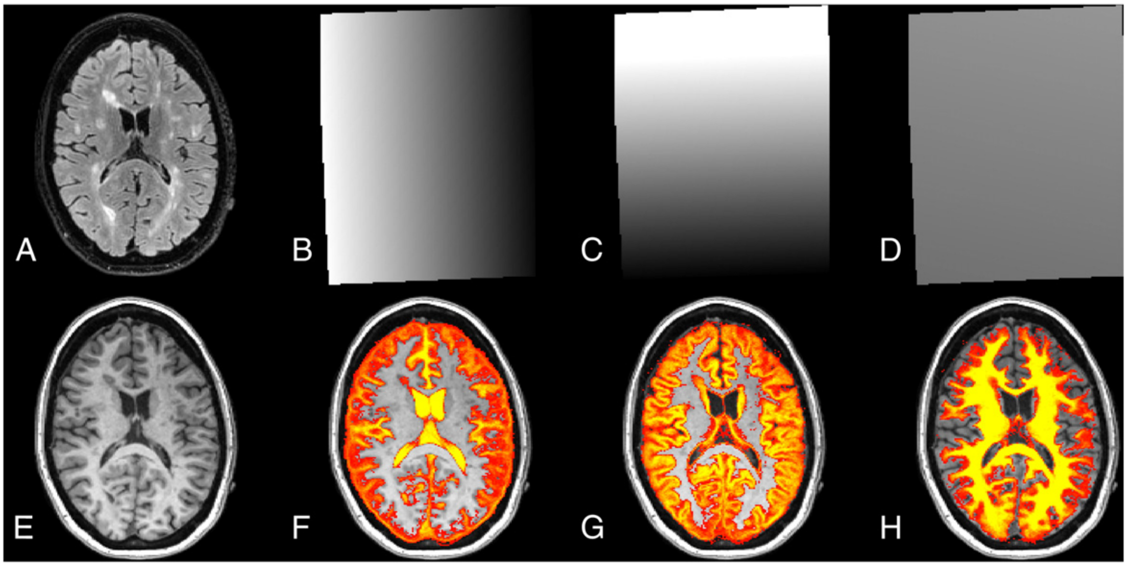

Fig. 1. Features used for the *k*NN classification: 3DFLAIR intensity (A), MNI-normalized spatial coordinate *x* (B), spatial coordinate *y* (C), spatial coordinate *z* (D), 3DT1 intensity (E), pCSF (F), pGM (G), and pWM (H).

selected by applying different minimum lesion volumes, and selecting the threshold *C* which results on average in the highest overlap with the manual reference segmentation.

2.10. Evaluation metrics

We tested the performance of six different configurations by altering the normalization procedure for the intensity features (i.e., variance scaling, robust range normalization and histogram matching), and either including TTPs in the feature set or omitting them (see Table 1).

Each configuration was evaluated using both volumetric and spatial correspondence measures. Volumetric correspondence between the automatic segmentation and the manual reference segmentations was measured using the intraclass correlation coefficient (ICC; two-way mixed model with absolute agreement definition) for the total lesion volume (Koch, 1982). Spatial correspondence at voxel level was evaluated using Dice's similarity index (SI) (Dice, 1945) and sensitivity, respectively defined as $SI = 2 \times TP/(2 * TP + FP + FN)$, and $sensitivity = TP/(TP + FN)$, where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives. Since SI is affected by lesion burden (Admiraal-Behloul et al., 2005), we also computed the lesion volume independent similarity index $SI_{estimate}$, and the outline error rate (OER) (Wack et al., 2012). As a logical extension to OER, we also computed detection error rate (DER), defined as $DER = DE / MTA$, where DE is detection error and MTA is mean total area such as described in Wack et al. (2012). In the leave-one-out approach, SI was regarded as the primary outcome measure.

2.11. Validation in an independent cohort of elderly subjects with hypertension

In order to evaluate the robustness of the optimal configuration across different scanners and patient populations, we finally applied the previously described training procedure, parameter selection and cross-validation to an independent dataset consisting of 20 high resolution MR images, selected from a larger cohort of elderly subjects with hypertension. In order to include a wide variety of vascular WML severity, subjects were selected based on the severity of WMLs. Age varied from 74 to 81 years (mean ± SD: 77.1 ± 7.0), 11 were women, and mean blood pressure varied from 113 to 188 mm Hg systolic (mean ± SD: 142.0 ± 17.0) and 66 to 90 mm Hg diastolic (mean ± SD: 79.0 ± 7.0).

MR imaging of this dataset was performed on a 3.0 T Intera whole body scanner (Philips Medical Systems, Best, The Netherlands) using a phased-array SENSE-eight-channel head coil. The protocol contained among others a 3DFLAIR sequence (TR: 4800 ms, TE: 355 ms, TI: 1650 ms, 250×250 mm^2 field of view (FOV), 160 saggital slices of 1.12 mm thickness, interpolated to 0.56 mm thick (overcontiguous) slices during reconstruction, 1.1×1.1 mm^2 in-plane resolution) for lesion detection, and a sagittal MPRAGE (magnetization prepared rapid acquisition gradient echo) sequence (TR: 6.6 ms, TE: 3.1 ms, FA: 9°, 270×270 mm^2 FOV, 170 sagittal slices of 1.2 mm thickness, 1.1×1.1 mm^2 in-plane resolution) for anatomical information.

Table 1
The different configurations.

Configuration	Description
Variance scaling	Variance scaling 3DFLAIR, 3DT1, *x*, *y*, *z*
Robust range normalization	Robust range normalization of 3DFLAIR and 3DT1 Variance scaling of *x*, *y* and *z*
Histogram matching	Histogram matching of 3DFLAIR and 3DT1 Variance scaling of *x*, *y* and *z*
Variance scaling + tissue type priors	Variance scaling of 3DFLAIR, 3DT1, *x*, *y*, *z*, pCSF, pGM, and pWM
Robust range normalization + tissue type priors	Robust range normalization of 3DFLAIR and 3DT1 Variance scaling of *x*, *y*, *z*, pCSF, pGM, and pWM
Histogram matching + tissue type priors	Histogram matching of 3DFLAIR and 3DT1 Variance scaling of *x*, *y*, *z*, pCSF, pGM, and pWM

In the reference segmentation, segmentations of the vascular WMLs were constructed using the 3DFLAIR images as follows. First, RF field inhomogeneity correction was performed using the N3 algorithm (Sled, 1997) implemented in 3D Slicer software (version 4.0, www.slicer.org). Subsequently, the images were orthogonally reformatted to the axial plane and WMLs were labeled by a single, trained rater. Afterwards, voxelwise thresholding was applied to the labeled areas to only include voxels with an intensity higher than the cortex at the level of the insula. Such a thresholding approach is well known in aging studies, as it allows for a much more consistent definition of lesion boundaries, which are often not clear in vascular WM lesions (Olsson et al., 2013).

3. Results

3.1. Reliability of manual reference segmentation

The manual reference segmentation showed a very good intra-observer agreement at the voxel level, with SI between the first and the second segmentation of 0.93 for the first technician, and 0.92 for the second technician. Inter-observer agreement was also very good, both concerning volumes, with ICC = 0.96, as well as at the voxel level, with an average SI of 0.84 ± 0.04 across all 120 slices on which lesions were outlined by both technicians. Mean and SD lesion volume in the final manual reference segmentation was 16.33 ± 11.49 mL with a median of 13.92 mL and volumes per patient ranging from 1.88 to 50.95 mL, quite typical for the range of lesion volumes in established MS patients.

3.2. Quantitative analysis of WML segmentation configurations

Table 2 lists the SI, sensitivity, $SI_{estimate}$, DER, OER and ICC for the configurations that were tested. Fig. 2 displays the average similarity index as a function of the binary threshold *p* for the different configurations.

In terms of volumetric correspondence, the configurations including TTPs within the feature set resulted in higher ICCs compared to the configurations without TTPs. The highest ICC was achieved using variance scaling with TTPs (ICC = 0.92). Robust range normalization without TTPs resulted in the lowest ICC (ICC = 0.80), indicating that intensity normalization and TTPs have a strong effect on volumetric correspondence.

The combination of variance scaling and inclusion of TTPs also led to maximum performance in terms of spatial correspondence (SI = 0.74 ± 0.09). In general, again better spatial performance was measured using the configurations where TTPs were added as features, although the addition of TTPs only had a marginal effect in the case of histogram matching, and a large effect in the case of variance scaling. Similar to SI, $SI_{estimate}$ showed the best performance when variance scaling + TTPs was used ($SI_{estimate}$ = 0.73 ± 0.05) and a lower performance when no TTPs were used.

Sensitivity was overall reasonable, with histogram matching + TTPs giving the best results (sensitivity = 0.73 ± 0.13). DER and OER showed that particularly a reduced detection error is responsible for the increased SI when including TTPs in the feature set. While outline error is relatively constant throughout the different configurations, the average detection error reduces from 0.21 in the worst case of variance scaling without TTPs to 0.09 in the configuration of variance scaling with TTPs.

Based on these results we selected variance scaling with TTPs as the optimal configuration.

Table 2
Evaluation of different configurations in MS patients.

Method	p	SI	Sensitivity	$SI_{estimate}$	DER	OER	ICC
Variance scaling	0.40	0.66 ± 0.12	0.63 ± 0.12	0.64 ± 0.11	0.21 ± 0.18	0.47 ± 0.12	0.84
Robust normalization	0.40	0.66 ± 0.12	0.62 ± 0.13	0.65 ± 0.09	0.19 ± 0.16	0.50 ± 0.15	0.80
Histogram matching	0.35	0.72 ± 0.09	0.72 ± 0.14	0.70 ± 0.07	0.11 ± 0.08	0.47 ± 0.13	0.90
Variance scaling + tissue type priors	0.40	0.74 ± 0.09	0.72 ± 0.11	0.73 ± 0.05	0.09 ± 0.08	0.44 ± 0.11	0.92
Robust range normalization + tissue type priors	0.35	0.72 ± 0.09	0.71 ± 0.11	0.72 ± 0.05	0.09 ± 0.08	0.46 ± 0.11	0.91
Histogram matching + tissue type priors	0.35	0.72 ± 0.09	0.73 ± 0.13	0.72 ± 0.05	0.09 ± .070	0.46 ± 0.13	0.91

p: optimal threshold for configuration; SI: Dice's similarity index; DER: detection error ratio; OER: outline error ratio; ICC: intra-class coefficient. All spatial correspondence metrics are listed (mean ± SD).

3.3. Post-processing and detailed analysis of the optimal configuration: variance scaling with TTPs

Post-processing was applied to the binary segmentation of the optimal configuration to reduce the number of small false positive regions. Variation in the size threshold C (integer values between 1 and 10 voxels) only caused small variations in performance. The highest mean SI was obtained after removing lesions smaller than 5 voxels, increasing the average SI from 0.74 ± 0.09 at $p = 0.40$ (no post-processing) to 0.75 ± 0.08 at $p = 0.35$. Volumetric correspondence in terms of ICC also increased, from 0.92 before to 0.93 after post-processing, respectively. Post-processing reduced both outline and detection error.

An example segmentation of a patient with average lesion load is shown in Fig. 3. To obtain more insight in the performance characteristics of the optimal configuration, the spatial correspondence metrics are listed in Table 3 for patients with low, intermediate, and high lesion loads. This shows that SI increases with lesion burden: datasets with lesion volume ≤5 mL have an average SI of 0.65, while datasets with lesion volume ≥15 mL have an average SI of 0.81. Similar behavior was seen for mean $SI_{estimate}$ which increases from 0.64 (<5 mL) to 0.77 (>15 mL). Furthermore, DER decreases strongly when lesion burden is lower. Although less pronounced, a similar relationship was seen for OER.

3.4. Validation in an independent cohort of elderly subjects with hypertension

Mean and SD lesion volume in the dataset of elderly subjects with hypertension was 8.21 ± 8.02 mL with a median of 5.36 mL and volumes per patient ranging from 0.57 to 31.20 mL. We first performed segmentation of the elderly subjects by using the MS reference segmentations (based on data acquired using a different scanner) as training set, the ‘variance scaling + TTPs’ configuration, and the previously derived optimal $p = 0.35$ and $C = 5$. As expected, this yielded suboptimal results: volumetric ICC = 0.60, average SI = 0.50 ± 0.24, sensitivity = 0.87 ± 0.06, $SI_{estimate}$ = 0.49 ± 0.17, DER = 0.53 ± 0.43 and OER = 0.45 ± 0.12. Retraining was then performed using the elderly reference segmentations and ‘variance scaling + TTPs’ configuration. Cross-validation measured maximal segmentation performance at $p = 0.5$ and $C = 2$, with volumetric ICC = 0.96, average SI = 0.84 ± 0.10, sensitivity = 0.86 ± 0.14, $SI_{estimate}$ = 0.83 ± 0.05, DER = 0.07 ± 0.06 and OER = 0.25 ± 0.16, thus showing substantial improvement in all measures except sensitivity, which on average stayed the same. The retrained method in the elderly tended to show, as the MS patients did, a lower segmentation performance when subjects had a low lesion volume, compared to elderly subjects with a high lesion volume (see Table 4). Here it should be noted that 9 of the elderly subjects had a lesion volume of <5 mL, 4 subjects had a lesion volume of 5–10 mL, 2 subjects had a lesion volume of 10–15 mL and only 3 subjects had a lesion volume of >15 mL.

4. Discussion

An automated WML segmentation algorithm on 3DFLAIR and 3DT1 images was presented and validated using manual reference segmentations. The optimal method used variance scaling, tissue type priors, 3DFLAIR intensities, 3DT1 intensities, and MNI-normalized spatial coordinates as features, and achieved very good voxelwise agreement with the reference segmentation. The results were further improved by applying a post-processing step which removed regions too small to be classified as a lesion from the segmentation.

The results of our study show that adding TTPs improves the results of kNN lesion segmentation considerably. This is in line with results of other studies showing increased performance when using tissue type information in the segmentation procedure (Schmidt et al., 2012). Adding TTPs improved lesion segmentation particularly by reducing the average detection error, while average outline error was fairly

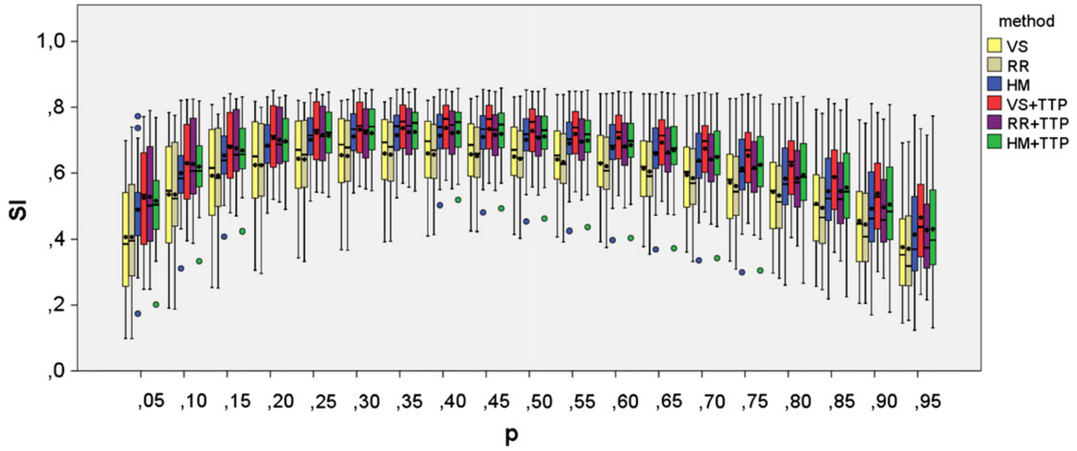

Fig. 2. Segmentation performance for different configurations in the MS patients. Boxplots showing for different configurations the distribution of the similarity indices across the 20 MS datasets as a function of threshold p. VS: variance scaling; RR: robust range normalization; HM: histogram matching; TTP: tissue type priors.

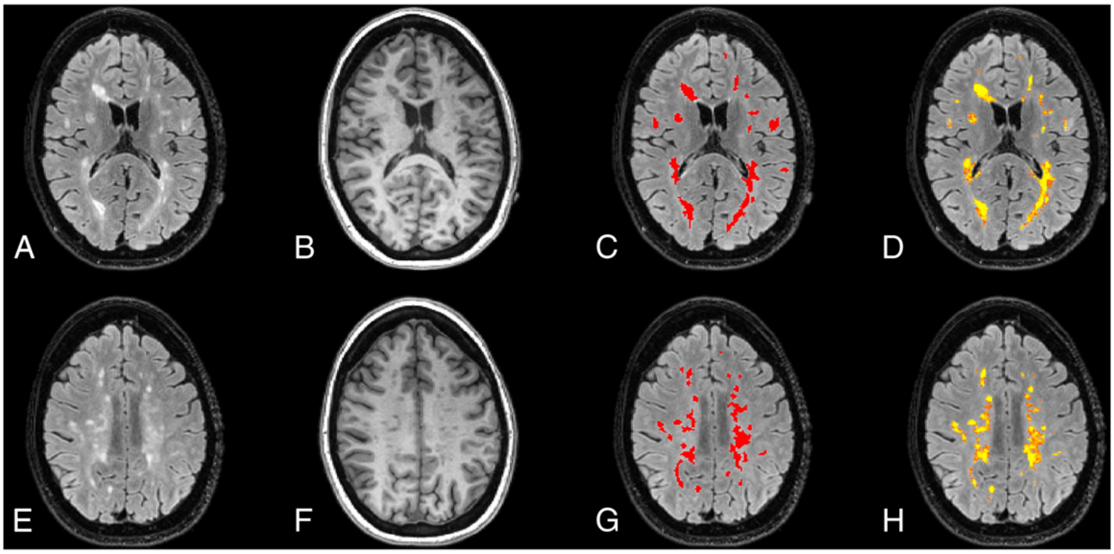

Fig. 3. Two slices showing the result of the automatic segmentation in a 39 year old relapsing–remitting MS patient (EDSS 2.5). 3DFLAIR (A, E), 3DT1 (B, F), manual reference segmentation (C, G), and thresholded probability map (red-yellow: $p = [0.35–1.0]$; D, H).

constant. Furthermore the results confirm the large influence of the choice of feature normalization on segmentation performance, emphasizing that feature normalization is an important aspect to consider in the design of a supervised lesion segmentation algorithm. Additional post-processing (i.e. removal of regions too small to be considered as a lesion) showed only a minor improvement in segmentation performance. Visually however, after post-processing, segmentation results were considerably smoother compared to without post-processing.

The final algorithm is fully automatic, and segmented a single dataset on a standard eight-core machine on average in about 23 min, of which 19 min was needed for nonlinear registration of the healthy controls to the dataset of interest and 4 min for the actual segmentation and post-processing.

For applicability of automated WML segmentation procedures in clinical studies, both good spatial and volumetric correspondence are critical. First, it is important to find the correct regions, but second, it is also important to outline them as accurately as possible, since lesion volumes are often used as outcome parameters or explanatory variables (Kappos et al., 2007; Mikol et al., 2008; Polman et al., 2006; Schoonheim et al., 2012) and lesion masks are increasingly used to perform lesion filling for obtaining accurate brain atrophy measurements (Battaglini et al., 2012; Chard et al., 2010). Using our final method, the volumetric correspondence reached ICC values up to 0.93, which we regard as excellent agreement. Furthermore, using TTPs, spatial correspondence measured by SI was higher than 0.7, which is regarded as excellent as well (Anbeek et al., 2004; Bartko, 1991). The final method also showed the lowest SI variance, indicating that *k*NN segmentation with TTPs delivers robust performance across our 20 datasets chosen to reflect the heterogeneity typically observed in MS populations, which is important for its applicability in clinical trials.

Table 3
Detailed evaluation of 'variance scaling + tissue type priors' configuration including post-processing in MS patients.

	N	SI	Sensivity	$SI_{estimate}$	DER	OER
<5 mL	3	0.65 ± 0.04 (0.60–0.68)	0.65 ± 0.08 (0.57–0.73)	0.64 ± 0.08 (0.56–0.70)	0.19 ± 0.06 (0.10–0.27)	0.50 ± 0.06 (0.43–0.56)
5–10 mL	4	0.72 ± 0.08 (0.61–0.78)	0.71 ± 0.13 (0.54–0.82)	0.73 ± 0.02 (0.71–0.75)	0.08 ± 0.06 (0.04–0.16)	0.47 ± 0.11 (0.39–0.63)
10–15 mL	5	0.73 ± 0.07 (0.63–0.80)	0.72 ± 0.10 (0.57–0.83)	0.76 ± 0.01 (0.75–0.76)	0.07 ± 0.03 (0.03–0.10)	0.48 ± 0.11 (0.37–0.63)
>15 mL	8	0.81 ± 0.05 (0.69–0.86)	0.79 ± 0.09 (0.68–0.94)	0.77 ± 0.01 (0.76–0.78)	0.04 ± 0.02 (0.01–0.08)	0.34 ± 0.09 (0.25–0.53)
Total	20	0.75 ± 0.08 (0.60–0.86)	0.74 ± 0.10 (0.54–0.94)	0.74 ± 0.05 (0.56–0.78)	0.08 ± 0.07 (0.01–0.27)	0.43 ± 0.11 (0.25–0.63)

N: number of subjects per group; SI: Dice's similarity index; DER: detection error rate; OER: outline error rate mean ± SD (minimum–maximum).

Validation using an independent dataset, obtained on a different scanner, involving vascular WMLs in elderly hypertensive subjects, again yielded very good voxelwise performance, demonstrating the robustness of the *k*NN-TTP segmentation method irrespective of the scanner used or the pathological substrate of the WMLs.

Many methods for WML segmentation have been published (Admiraal-Behloul et al., 2005; Akselrod-Ballin et al., 2009; Anbeek et al., 2005; Damangir et al., 2012; De Boer et al., 2009; Geremia et al., 2011; Khayati et al., 2008; Schmidt et al., 2012; Shiee et al., 2010; Van Leemput et al., 2001). Comparison of different methods however, should be done with care, since the measured performance is highly dependent on the dataset and the reference segmentation being used for evaluation. Factors known to influence segmentation performance include the pulse-sequence being used (i.e., sequence type, 2D versus 3D) (Anbeek et al., 2005), the way the reference segmentation was constructed (i.e., manual or semi-automatic), the heterogeneity of pathology in the sample (i.e., easier to achieve high performance in a homogeneous dataset), and overall lesion burden (i.e., higher lesion load generally leads to better spatial segmentation performance) (Wack et al., 2012). This is illustrated by the better performance in the validation dataset compared to the dataset consisting of patients with MS: the vascular pathology in the validation dataset is more homogeneous and the construction of the reference segmentation involved a semi-automatic segmentation step, which might partially explain the higher segmentation performance in this sample. Taking these considerations into account, and given that the use of such a semi-automatic procedure is defendable since the described approach is common in aging studies (Olsson et al., 2013), our method performs very well.

Comparing our method to others, some studies reported poorer performance in terms of SI (Akselrod-Ballin et al., 2009; De Boer et al., 2009; Shiee et al., 2010; Van Leemput et al., 2001), whereas others reported comparable or higher performance (Admiraal-Behloul et al., 2005; Anbeek et al., 2004; Khayati et al., 2008; Schmidt et al., 2012). One study reporting high performance is the original study by Anbeek et al. which was the first to use *k*NN to classify WMLs (average SI = 0.80). In that study, WMLs of 20 patients with vascular disease were segmented using spatial coordinates, and 2D T1, IR, PD, T2, and FLAIR-intensities as features. The method used in that study was very

Table 4
Similarity index versus lesion load in various studies.

	Total	<5 mL	5–10 mL	10–15mL	>15mL
Current study ‘variance scaling + tissue type priors’, MS patients	0.75	0.65	0.72	0.73	0.81
Current study ‘variance scaling + tissue type priors’, elderly subjects with hypertension	0.84	0.78	0.92	0.79	0.91
Schmidt et al. (2012)	0.75	0.67	0.76	0.82	0.85
Khayati et al. (2008a)1	0.75	0.73	0.75		0.81
Sajja et al. (2006)	0.78	0.67		0.84	
Admiraal–Behloul et al. (2005)	0.75	0.70	0.75		0.82
Anbeek et al. (2004b)2	0.80	0.50	0.75		0.85

[a] Different definition of lesion load: (LV < 4 mL), moderate (4 mL < LV < 18 mL), large (LV > 18 mL).

[b] Definition of lesion load based on diameter of largest diffuse white matter lesion and location of periventricular white matter lesions.

similar to our ‘variance + no TTPs’ configuration which resulted in much lower performance in our MS sample ($\mathrm{SI} = 0.66$). This difference can possibly be explained by the different sequences used and different pathologies addressed in both studies, and it illustrates the difficulty of comparing performance using different reference datasets.

As expected, SI was lower in subjects with lower lesion burden. It has also been reported by others that small errors have a relatively larger effect on a smaller reference (Admiraal-Behloul et al., 2005; Anbeek et al., 2004; Khayati et al., 2008; Sajja et al., 2006; Schmidt et al., 2012). Table 4 compares the SI for different lesion loads of our study with other studies and shows that our method performs equally well, despite the use of 3D sequences and manual reference segmentation, across the full range of lesion loads.

A limitation of our method is that the algorithm requires new training when applied to data originating from other scanners or other acquisition protocols. This is necessary since 3DFLAIR and 3DT1 signal characteristics may differ among MR scanners and pulse sequences, and is illustrated by the much better performance after retraining in the sample with elderly subjects. Secondly, the outlining of the manual MS reference segmentations was performed by two technicians who, while highly trained and performing MS lesion outlining on 2D images on a daily basis, were not used to working with the high-resolution 3D images used in the current study. Therefore, to optimize their performance with these new images, we provided limited additional training prior to the study. The resulting manual segmentation was of high quality, as evidenced by the high reproducibility, both between sessions of the same technician and between the two technicians (inter-observer $\mathrm{SI} = 0.84$). Furthermore, our manual MS reference segmentations were based on a single consensus scoring to determine which regions were MS WMLs. This could have led to artificial higher inter- and intra-observer agreements since detection errors could not occur when outlining the lesions. Finally, we did not optimize the value of k in the present work, but selected a value of 40 based on the literature. To rule out that other values of k would have resulted in large performance differences, we performed a post-hoc analysis in which the training and evaluation of the optimal configuration for the dataset with MS patients was repeated for different values of k, namely $k = 20$, 80 and 160. Here, it should be noted that classification takes longer when larger values of k are used, since more nearest neighbors have to be found. The results of this post-hoc analysis ($k = 20$: $p = 0.35$, $C = 6$, $\mathrm{SI} = 0.74 \pm 0.08$; $k = 40$: $p = 0.35$, $C = 5$, $\mathrm{SI} = 0.75 \pm 0.08$ (previously reported); $k = 80$: $p = 0.30$, $C = 6$, $\mathrm{SI} = 0.75 \pm 0.08$; and $k = 160$: $p = 0.30$, $C = 8$, $\mathrm{SI} = 0.75 \pm 0.08$) confirmed that k in the current range is suitable for this type of segmentation problems.

In conclusion, we improved kNN classification for the segmentation of WMLs by adding TTPs and showed that intensity normalization has a strong impact on segmentation performance. The optimal configuration showed excellent agreement in terms of volumetric and spatial measures with fully manual 3D reference segmentations across a wide range of WML severity, irrespective of the scanner used or the pathological substrate of the WML.

Acknowledgments

This work was supported by the Dutch MS Research Foundation through a program grant to the VUmc MS Center Amsterdam (grant number 09-358d). The MRI scans of the validation cohort were obtained with support from ‘Internationale Stichting Alzheimer Onderzoek’ (grant number 10507). The authors would like to thank the Image Analysis Center (IAC) Amsterdam for contributing to the development of the manual reference segmentations of the patients with MS.

References

Admiraal-Behloul, F., van den Heuvel, D.M.J., Olofsen, H., van Osch, M.J.P., van der Grond, J., van Buchem, M.A., Reiber, J.H.C., 2005. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. Neuroimage 28, 607–617.

Akselrod-Ballin, A., Galun, M., Gomori, J.M., Filippi, M., Valsasina, P., Basri, R., Brandt, A., 2009. Automatic segmentation and classification of multiple sclerosis in multichannel MRI. IEEE Trans. Biomed. Eng. 56, 2461–2469.

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46, 726–738.

Anbeek, P., Vincken, K.L., van Osch, M.J.P., Bisschops, R.H.C., van der Grond, J., 2004. Probabilistic segmentation of white matter lesions in MR imaging. Neuroimage 21, 1037–1044.

Anbeek, P., Vincken, K.L., van Bochove, G.S., van Osch, M.J.P., van der Grond, J., 2005. Probabilistic segmentation of brain tissue in MR imaging. Neuroimage 27, 795–804.

Anbeek, P., Vincken, K., Viergever, M., 2008. Automated MS-lesion segmentation by k-nearest neighbor classification. Midas J. 1–8.

Bartko, J.J., 1991. Measurement and reliability: statistical thinking considerations. Schizophr. Bull. 17, 483–489.

Battaglini, M., Jenkinson, M., De Stefano, N., 2012. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Hum. Brain Mapp. 33, 2062–2071.

Chard, D.T., Jackson, J.S., Miller, D.H., Wheeler-Kingshott, C.A.M., 2010. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. J. Magn. Reson. Imaging 32, 223–228.

Damangir, S., Manzouri, A., Oppedal, K., Carlsson, S., Firbank, M.J., Sonnesyn, H., Tysnes, O., O'Brien, J.T., Beyer, M.K., Westman, E., Aarsland, D., Wahlund, L., Spulber, G., 2012. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. J. Neurol. Sci. 322, 211–216.

De Boer, R., Vrooman, H.A., van der Lijn, F., Vernooij, M.W., Ikram, M.A., van der Lugt, A., Breteler, M.M.B., Niessen, W.J., 2009. White matter lesion extension to automatic brain tissue segmentation on MRI. Neuroimage 45, 1151–1161.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Fazekas, F., Chawluk, J.B., Alavi, A., Hurtig, H.I., Zimmerman, R.A., 1987. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. AJ. Am. J. Roentgenol. 149, 351–356.

Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57, 378–390.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Kappos, L., Freedman, M.S., Polman, C.H., Edan, G., Hartung, H.P., Miller, D.H., Montalbán, X., Barkhof, F., Radü, E.-W., Bauer, L., Dahms, S., Lanius, V., Pohl, C., Sandbrink, R., 2007. Effect of early versus delayed interferon beta-1b treatment on disability after a first clinical event suggestive of multiple sclerosis: a 3-year follow-up analysis of the BENEFIT study. Lancet 370, 389–397.

Khayati, R., Vafadust, M., Towhidkhah, F., Nabavi, M., 2008. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model. Comput. Biol. Med. 38, 379–390.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29, 196–205.

Koch, Gary G., 1982. Intraclass correlation coefficient. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences. John Wiley & Sons, New York, New York, USA, pp. 213–217.

Kurtzke, J.F., 1983. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33, 1444–1452.

Lao, Z., Shen, D., Liu, D., Jawad, A.F., Melhem, E.R., Launer, L.J., Bryan, R.N., Davatzikos, C., 2008. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. Acad. Radiol. 15, 300–313.

Mikol, D.D., Barkhof, F., Chang, P., Coyle, P.K., Jeffery, D.R., Schwid, S.R., Stubinski, B., Uitdehaag, B.M.J., 2008. Comparison of subcutaneous interferon beta-1a with glatiramer acetate in patients with relapsing multiple sclerosis (the REbif vs Glatiramer Acetate in Relapsing MS Disease [REGARD] study): a multicentre, randomised, parallel, open-label trial. Lancet Neurol. 7, 903–914.

Mortamais, M., Reynes, C., Brickman, A.M., Provenzano, F.A., Muraskin, J., Portet, F., Berr, C., Touchon, J., Bonafé, A., le Bars, E., Maller, J.J., Meslin, C., Sabatier, R., Ritchie, K., Artero, S., 2013. Spatial distribution of cerebral white matter lesions predicts progression to mild cognitive impairment and dementia. PLoS One 8, e56972.

Mortazavi, D., Kouzani, A.Z., Soltanian-Zadeh, H., 2012. Segmentation of multiple sclerosis lesions in MR images: a review. Neuroradiology 54, 299–320.

Olsson, E., Klasson, N., Berge, J., Eckerström, C., Edman, A., Malmgren, H., Wallin, A., 2013. White matter lesion assessment in patients with cognitive impairment and healthy controls: reliability comparisons between visual rating, a manual, and an automatic volumetrical MRI method—the Gothenburg MCI study. J. Aging Res. 2013, 198471.

Polman, C.H., O'Connor, P.W., Havrdova, E., Hutchinson, M., Kappos, L., Miller, D.H., Phillips, J.T., Lublin, F.D., Giovannoni, G., Wajgt, A., Toal, M., Lynn, F., Panzara, M.A., Sandrock, A.W., 2006. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. N. Engl. J. Med. 354, 899–910.

Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F.D., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A.J., Waubant, E., Weinshenker, B., Wolinsky, J.S., 2011. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann. Neurol. 69, 292–302.

Popescu, V., Battaglini, M., Hoogstrate, W.S., Verfaillie, S.C.J., Sluimer, I.C., van Schijndel, R.A., van Dijk, B.W., Cover, K.S., Knol, D.L., Jenkinson, M., Barkhof, F., de Stefano, N., Vrenken, H., 2012. Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. Neuroimage 61, 1484–1494.

Provenzano, F.A., Muraskin, J., Tosto, G., Narkhede, A., Wasserman, B.T., Griffith, E.Y., Guzman, V.A., Meier, I.B., Zimmerman, M.E., Brickman, A.M., 2013. White matter hyperintensities and cerebral amyloidosis: necessary and sufficient for clinical expression of Alzheimer disease? JAMA Neurol. 70, 455–461.

Sajja, B.R., Datta, S., He, R., Mehta, M., Gupta, R.K., Wolinsky, J.S., Narayana, P.A., 2006. Unified approach for multiple sclerosis lesion segmentation on brain MRI. Ann. Biomed. Eng. 34, 142–151.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Mühlau, M., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. Neuroimage 59, 3774–3783.

Schoonheim, M.M., Popescu, V., Rueda Lopes, F.C., Wiebenga, O.T., Vrenken, H., Douw, L., Polman, C.H., Geurts, J.J.G., Barkhof, F., 2012. Subcortical atrophy and cognition: sex effects in multiple sclerosis. Neurology 79, 1754–1761.

Shiee, N., Bazin, P.L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage 49, 1524–1535.

Sled, J.G., 1997. A Non-parametric Method for Automatic Correction of Intensity Non-uniformity in MRI Data.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20, 677–688.

Wack, D.S., Dwyer, M.G., Bergsland, N., Di Perri, C., Ranza, L., Hussein, S., Ramasamy, D., Poloni, G., Zivadinov, R., 2012. Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates. BMC Med. Imaging 12, 17.

Younis, A., Ibrahim, M., Kabuka, M., John, N., 2008. An artificial immune-activated neural network applied to brain 3D MRI segmentation. J. Digit. Imaging 21 (Suppl. 1), S69–S88.

Zhang, Y., Brady, M., Smith, S.M., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.