A SEMI-AUTOMATIC METHOD FOR MULTIPLE SCLEROSIS LESION SEGMENTATION ON DUAL-ECHO MAGNETIC RESONANCE IMAGES: APPLICATION IN A MULTICENTER CONTEXT

L. Storelli[1], E. Pagani[1], M.A. Rocca[1,2], M.A. Horsfield[3], A. Gallo[4,5], A. Bisecco[4,5], M. Battaglini[6], N. De Stefano[6], H. Vrenken[7], D.L. Thomas[8], L. Mancini[8], S. Ropele[9], C. Enzinger[9, 10], P. Preziosa[1,2], M. Filippi[1,2]

[1]Neuroimaging Research Unit, and [2]Department of Neurology, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, Italy; [3]Xinapse Systems, Colchester CO6 3BW, UK; [4]MRI Center "SUN-FISM," Second University of Naples and Institute of Diagnosis and Care "Hermitage-Capodimonte," Naples, Italy; [5]I Division of Neurology, Department of Medical, Surgical, Neurological, Metabolic and Aging Sciences, Second University of Naples, Naples, Italy; [6]Department of Neurological and Behavioral Sciences, University of Siena, Italy; [7]Department of Radiology and Nuclear Medicine, MS Centre Amsterdam, VU Medical Centre, Amsterdam, Netherlands; [8]Neuroradiological Academic Unit, UCL Institute of Neurology, London, UK; [9]Department of Neurology, Medical University of Graz, Austria; [10]Clinical Division of Neuroradiology, Vascular and Interventional Radiology; Department of Radiology, Medical University of Graz, Austria.

Text word count: 4500; **Figure count**: 3.

Supplementary material.

Correspondence should be addressed to: Prof. Massimo Filippi, Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Via Olgettina, 60, 20132 Milan, Italy. Telephone number: #39-02-2643-3033; Fax number: #39-02-2643-3031; E-mail address: filippi.massimo@hsr.it.

Abstract

Background and Purpose. A multicenter validation of a proposed semi-automatic method for hyperintense MS lesion segmentation on dual-echo MRI is presented.

Materials and Methods. The classification technique used by the method is based on a region growing approach starting from manual lesion identification by an expert observer, with a final segmentation refinement step. The method was validated in a cohort of 52 relapsing-remitting MS patients, with dual-echo images acquired in 6 different European centers.

Results. A mathematical expression was found that made the optimization of the method independent of the need of a training dataset. The automatic segmentation was in good agreement with the manual segmentation (dice similarity coefficient = 0.62 and root mean square error = 2 ml). Assessment of the segmentation errors showed no significant differences in algorithm performance between the different MR scanner manufacturers (p value > 0.05).

Conclusion. The method proved to be robust, and no center-specific training of the algorithm was required, giving the possibility for the application in a clinical setting. Adoption of the method should lead to improved reliability and lower operator time required for image analysis in research and clinical trials in MS.

Key Words: Neuroimaging, MRI post-processing, Multiple sclerosis, Lesion segmentation.

ABBREVIATIONS: DE = dual-echo; PD = proton density; SD = standard deviation; EDSS = Expanded Disability Status scale; DSC = dice similarity coefficient; RMSE = root mean square error; FPF = false positive fraction; FNF = false negative fraction; TPF = true positive fraction.

Introduction

Assessment of the disease burden on MR images from MS patients, both for research and for clinical trials, requires quantification of the volume of hyperintense lesions on T_2-weighted images.[1] However, lesion segmentation remains challenging, and the required accuracy and reproducibility are difficult to achieve. Ideally, segmentation should be automated, or require the minimum of operator input in order to minimize the operator time required, and reduce bias;[2-4] however, manual segmentation is still the gold standard.

Although several methods for fully-automated MS lesions segmentation have been published, their performance is difficult to compare. This is because they are usually validated without a common framework[5] and, even if within the same framework such as the MS lesion segmentation challenge presented at the MICCAI 2008,[6] the validation is done using a small dataset of cases and does not include a DE proton-density (PD)/T_2-weighted images dataset. Moreover, the majority of the methods are optimized and tested on FLAIR MR images that benefit from CSF signal suppression and better contrast between focal lesions and the surrounding tissue[7-9] in comparison with the more established techniques that use DE sequences. Large datasets of DE MR images are available from past studies and their acquisition is still common both for research and for clinical trials, so that there is still the need to develop methods for lesion segmentation on these data.[10]

We have previously proposed a semi-automated method for MS lesion segmentation on DE MR images based on a region growing approach, that results in a considerable reduction in the amount of time required for lesion segmentation compared with manual segmentation and shows good agreement with the ground truth.[11]

Most large MRI studies of MS involve multiple scanning centers with different scanner manufacturers.[12] While all centers would use a common scanning protocol with pulse sequence parameters restricted within certain ranges, there are inevitable differences in image contrast due to hardware and software differences. The aims of the current work were to analyze the training

procedure required by the algorithm and to validate the lesion segmentation method proposed in a multicenter context. The method was validated by comparing the lesion segmentations obtained using the proposed method with manual segmentations, across different MR scanner manufacturers.

Methods

Background. The method was presented at the BrainLes MICCAI workshop 2015[11] and validated for a single acquisition center on 20 patients. Refer to the supplementary material for the methodological framework of the lesion segmentation technique.

MRI Acquisition. The dataset consisted of 52 MS patients, part of a project on imaging correlates of cognitive impairment in MS, acquired in six European centers which are part of the MAGNIMS consortium (Amsterdam, Graz, London, Milan, Naples and Siena) using 3.0 Tesla MRI scanners from a range of manufacturers (2 scanners from Philips Medical Systems; 2 scanners from General Electric Medical Systems; 2 scanners from Siemens Medical). To be included, patients had to be aged between 20 and 65 years, have a diagnosis of relapsing-remitting (RR) MS,[13] no relapse or corticosteroids treatment within the month prior to scanning and no history of psychiatric conditions (see Supplementary Table 1). Only MRI sequences without visually relevant artifacts were selected for the current analysis.

The research protocol was approved by the local ethics review boards of participating centers, and all subjects gave written informed consent.

A similar MRI acquisition protocol was used for all patients: DE TSE; TR=ranging from 4000 to 5380 ms, TE_1=ranging from 10 to 23 ms, TE_2=ranging from 90 to 102 ms, echo train length=ranging from 5 to 11, 44 contiguous, 3-mm thick axial slices, parallel to the AC-PC plane, with a matrix size=256 x 256, recFOV=75% and a FOV=250 x 250 mm^2.
The characteristics of MR hardware and number of patients acquired at each center are summarized in Supplementary Table 2.

Analysis of the training procedure. The use of different scanners could cause hardware-dependent differences in image quality. In this study, we assumed that patients scanned on different scanners from the same manufacturer using the same RF coils and MRI protocol would have comparable image quality and could therefore be grouped together for the analysis.

Manual identification of lesions was used to initialize the algorithm, while manual segmentation was used for the training and validation of the proposed method. Both tasks were performed using software for medical image analysis Jim Version 6 (Xinapse Systems, Colchester, UK). Manual identification and segmentation of lesions was performed by an experienced rater with 7 years' experience in MS lesion segmentation. In the case of doubts in lesion identification, a senior rater was consulted.

For image standardization (*step 1*), a group of 12 patients (two from each center) with a low lesion load was selected. A high lesion load was avoided since a high number of hyperintense lesions could significantly alter the shape of the image intensity histograms and affect the estimation of the standard parameters. For the computation of the standard parameters, using scans from healthy subjects would be preferable, but these are not always available in a clinical environment.

Since the method required a training step, the selection of a reliable set representative of the entire dataset, in terms of lesion load and sample size for each MR manufacturer, was investigated. Patients were grouped by scanner manufacturer. A threshold function (*step 3*) was calculated for each group and steadily decreasing the number of MS patients included in the training set. First, all patients were included, and then at each step three patients were removed from each group. The choice of which patients to remove was made by attempting to maintain a balanced lesion load (i.e., a variation within ±10 %) across the three different MR manufacturers. This analysis was performed in order to assess the relationship between the sample size and the threshold function for each MR manufacturer, to lead to a proper selection of the training set for this method. A straight line was fitted to the seed intensity values plotted against the optimal threshold values, obtaining the

threshold function for the initial region growing. The linear relationship between the normalized seed intensity and the optimal threshold values was empirically obtained.

To evaluate the sensitivity of the segmentation results to the slope (m) of the threshold function, 17 simulated threshold functions were generated to initialize the region growing. These functions consisted of a straight line passing from a common point (described in detail in the results) and with a slope varying from 0.1 to 0.9 in steps of 0.05 (a wider range of values than that founded in the training). The lesion segmentation was performed without the refinement step, to evaluate only the effect of a different slope on the results.

The optimal threshold function was selected from the simulated ones by maximizing the Dice similarity coefficient (DSC) between the manually and automatically outlined lesions (as described in the paragraph below).

Moreover, we investigated whether the 2 parameters identifying the optimal training straight line could be estimated directly from the image to be segmented, thus avoiding the need of a training procedure implying the acquisition of an extra group of patients and the manual lesion segmentation.

Since it was found that the training based on manual segmentation could be avoided (see results), the entire dataset could be used as test dataset and the optimized procedure was applied to the whole group of MS patients.

Statistical analysis

The root mean square error (RMSE) in lesion volume for the proposed method relative to the manual segmentation was computed. The RMSE values, grouped by scanner manufacturer, were compared to evaluate any performance differences between MR manufacturers. It was assumed that the observations from the three manufacturer groups were independent of each other. The Wilcoxon-Mann-Whitney test was used to test for differences in errors between the groups. This is a non-parametric test of the null hypothesis that two independent samples come from the same

population, against an alternative hypothesis. The test was performed pairwise between the three groups: test 1 was performed between the lesion segmentation errors on the images acquired on Siemens scanners compared to Philips scanners; test 2 was between Siemens scanners and General Electrics scanners; test 3 was between General Electric scanners and Philips scanners. The segmentations produced by the proposed method were compared to manual segmentations performed by an expert physician, using the DSC. DSC values range from zero to one, where zero corresponds to no overlap between the two segmentations, and one corresponds a perfect overlap. The false positive fraction (FPF), false negative fraction (FNF) and true positive fraction (TPF) were computed for each lesion to indicate the percentage of voxels correctly or incorrectly classified as lesion by the method. The "ground truth" for assessing the true and false positive rates was the binary lesion mask obtained after manual segmentation, comparing individual lesions pixel-by-pixel between the manual and automatic mask.

Results

The threshold functions (plots of threshold value against seed intensity) showed a similar trend with decreasing of the number of patients included in the training set: as the sample size decreased the fitted lines maintained a similar slope and approximately they pass through a similar point (Figure 1). The seed intensity at this 'common point' was found to be the intensity of the GM peak on the standardized histogram. This is due to the fact that the image standardization process fixed the GM peak for all the PD-w images to the same intensity value. Thus, this value as seed point would produce similar thresholds during the training, and after the fitting operation on the training set these points were interpolated producing a single 'common point' between the functions. Furthermore, on the y-axis this point represents the intensity variation on the GM standard intensity distribution that discriminates the lesion intensity values, which mostly overlap with GM intensity values, from the surrounding tissue (WM).

The effect of a different slope of the threshold functions on the segmentation results was evaluated. The slope of the threshold function was varied between 0.1 and 0.9, and higher DSC scores were found (DSC>0.6) at higher values of slope (m>0.7), although this improvement was not significant.

From those findings, the thresholds used in initial seed growing were expressed as:

$$T = m * (Iseed_i - I_{GM}) + \sigma_{GM}; \quad [1]$$

where m was fixed to 0.9, T is the threshold for the region growing, $Iseed_i$ is the seed intensity value for lesion i; I_{GM} and σ_{GM} were respectively the intensity of the GM peak and the standard deviation of the GM distribution on the standard histogram. Equation [1] was used to compute the threshold function, and then the method was performed without training on manual segmentation.

Comparison of data between the different scanner manufacturers (Siemens *vs* Philips, Siemens *vs* GE and GE *vs* Philips) showed that there was no evidence that lesion segmentation errors came from different distributions. The mean values of segmentation errors for each MR manufacturer were: $RMSE_{GE}$ = 1.99 ml, $RMSE_{PHILIPS}$ = 1.59 ml, $RMSE_{SIEMENS}$ = 1.86 ml. The statistical test performed between the groups revealed no differences of segmentation performance between manufacturers: p_{test1}=0.65, p_{test2}=0.44 and p_{test3}=0.30.

The validation metrics were extracted for each lesion load of each patient, considering each lesion as a connected region in 3-D space for the computation of its total volume. In Figure 2, the metrics evaluated for each patient over all lesions are graphically reported. The following were obtained after averaging the metrics over all patients: DSC = 0.62; RMSE = 2 ml; TPF = 0.76; FPF = 0.36; FNF = 0.22.

An example lesion segmentation result is shown in Figure 3.

Discussion

Since manual segmentation is time-consuming and is subject to inter- and intra- observer variability, automatic segmentation of MS lesions is an active research field with many proposals

presented in the last years.[5] The method validated in this study has several advantages. First, it works on DE MR images. Most of the proposed methods segment lesions on FLAIR sequences, that benefit from suppression of the CSF signal and better contrast between focal lesions and the surrounding background.[7-9, 14-16] However, large amounts of data have been and are currently being acquired for research and clinical trials using DE PD/T_2w images. Thus, with use of the proposed method, it should be possible to rapidly analyze these large sets of images. Second, despite the limitation of the manual identification of lesions by an expert physician, this initialization ensures the correct identification of all lesions and avoids the problem of the identification of entire false positive lesions (since only possible misclassification of lesion pixels can occur). This is a common challenge for fully automatic lesion segmentation methods, which tend to be sensitive to the image quality.[4] In the method proposed, we avoid this issue by maintaining manual identification of lesions and automating the segmentation task that is the most time-consuming operation. Some automatic lesion segmentation tools with available code (LST, SLS and Lesion-TOADS) expect as input FLAIR images. As a consequence, a comparison with our method would be unfair. Moreover, the majority of the proposed methods have been validated on a restricted number of cases and within single centers or simulated MRI acquisitions.[5, 17, 18] Also a validation of the method on data provided by the MICCAI Grand Challenge workshop 2008 would be unfeasible due to the absence of a DE sequence in the dataset.[6, 19] In this study, a validation of the method against manual segmentation in a multicenter context was presented, proving that the method was robust to scanner differences and its performance was not MR software and hardware-dependent.

During an initial assessment of the size of the training set needed, it was found that the threshold functions extracted for the initial region growing algorithm were not noticeably affected by including smaller numbers of subjects, and that there were no significant differences between the thresholds functions computed from each scanner manufacturer group. Moreover, using the simulated threshold functions, it emerged that once their intersection point was found, changes to the slope introduce only a small non-significant improvement at higher values; thus, the most

important feature of the threshold function was the crossing point of the lines, which was a result of the standardization process.

These results allowed us to find an expression for the threshold function used in the initial region growing part of the algorithm, thus avoiding the training step using manual segmentation. Since the segmentation results improved when using a higher slope of the threshold function, m=0.9 was selected to allow the use of higher thresholds and a less restricted region growing segmentation. This is because of the stop condition on the threshold value (see Eq. [1] of the supplementary material): a higher threshold implies a higher difference between the seed point and the *i-th* pixel intensity value that stops the region growing, so a larger range of intensities classifiable as lesion (less restricted segmentation). This was made possible because we included an edge detection step in the segmentation that acts as a barrier to stop the region growing even if a too high threshold is used. Because of noise or artifacts on the images, the two stop conditions were used in combination for a good result. However, using a high slope for the threshold function might generate a bias between lesions with higher and lower intensity values relative to the crossing point of the straight line: that is, with a high slope, lower intensity lesions would have lower threshold values, causing a more restricted region growing while the opposite would be observed for higher intensity lesions. This bias was avoided by applying a threshold refinement step, in which a more robust threshold is computed to restart the region growing, thus correcting too restricted segmentation due to lower threshold values. Hence, using Eq. [1] to find the threshold function, we avoided the training step using manual segmentation, making the applicability of the proposed technique easier in clinical settings. Regarding the possible bias between different lesion loads, from Figure 2 (lower-right graph) it seemed that the difference between automatic and manual lesion load becomes larger with increasing lesion load. This could be explained by the fact that a high lesion load could be due to many small lesions or a few but very large lesions. In the first case, a difference of a few pixels between the automatic and manual segmented lesion (that is visually undetectable), summed up for all lesions, could result in a relevant difference in the quantification of lesion load between the two

methods. In the second case, a difference of more pixels, for example at lesion border (again visually undetectable), could result in a relevant difference in lesion load quantification between manual and automatic segmentation.

The stability and robustness of the method was assessed when working on data from different scanner manufacturers. The initial step in image analysis standardizes the intensity values between the PD-w MRI scans, allowing the use of fixed intensity parameters. The method was not significantly affected by possible hardware or software dependent differences between MRI scanners.

Lesion segmentation performed using the new method showed good agreement with the ground truth (DSC = 0.62 and TPF = 0.76). The difference between the lesion load estimated using the proposed method and with manual segmentation gave a mean error of 19% (RMSE = 2 ml), with low misclassification of lesion voxels (FNF = 0.22 and FPF = 0.36).

The evidence of the benefit for the operator time required to segment lesions was demonstrated in our previous work.[11] In the current study, the significant reduction in time for the segmentation task was confirmed. For the lesion loads we considered, the average time for manual lesion segmentation of a single MRI scan was about 50 minutes for the segmentation task only, while for the new method the average time for the same task was about 55 seconds, a reduction in time of about 98.2%.

In cases where lesions have intensity similar to that of CSF, the method gives segmentations that extend beyond the real boundary of the lesions. This happened in very few cases in this study, and was mainly for periventricular lesions. It may be possible to improve this in future by introducing further information about lesions, perhaps using other MR tissue contrast such as co-registered T1-w images. This improvement could also be useful for a more certain lesion boundary delineation in case of diffuse lesions in patients with high lesion load. Notably, the method did not encounter difficulties in segmenting subcortical/cortical lesions. This is due to the edge detection step using the high pass filter: the border of subcortical/cortical lesions were well-defined with respect to the

surrounding tissue, differently from what happened to periventricular lesions, that had intensity values similar to the CSF on DE scans.

The method implemented is based on a 2-D region growing approach since it started from initial seed points positioned in 2-D. The choice of 2-D implementation was due to the fact that images were not acquired using 3-D MR sequences, therefore resolution along z-axis (slice-thickness) is lower than the axial one. The adaptation of the method to 3-D approach could be a future extension when 3-D MR sequences are available, to reduce the interaction time of the expert. Similarly, the applicability of the method on different images (e.g, pre- and post-contrast T1-weighted sequences) would require some modifications and retraining of the method for the new contrasts.

The algorithm relies on manual identification of lesions which must be performed by an expert operator, while the most time consuming task, i.e. outlining each lesion, is fully automated. However, it would obviously be preferable to avoid all manual intervention to remove any operator dependence. In future it may be possible to fully automate T_2-hyperintense lesion segmentation by using other MRI contrasts such as FLAIR or Double Inversion Recovery sequences.[20] Finally, the reproducibility of the method should be evaluated in longitudinal studies.

Conclusions

In this study, we evaluated the performance and stability of a semi-automatic method for MS lesion segmentation using DE data acquired from different centers with different scanners, compared with manual segmentation by an expert physician. The method proved to be robust and stable when working on data from different scanner manufacturers. It emerged also that no center-specific training of the algorithm was required, making the method suitable for direct use on a wide range of images. Adoption of the method should lead to improved reliability and lower operator time required for image analysis in research and clinical trials in MS.

Acknowledgments

This study was partially supported by Fondazione Italiana Sclerosi Multiple (FISM2013/S/1). DLT is supported by the UCL Leonard Wolfson Experimental Neurology Centre (PR/ylr/18575).

References

1. Filippi M, Rocca MA, De Stefano N, et al. Magnetic resonance techniques in multiple sclerosis: the present and the future. *Archives of neurology* 2011;68:1514-1520

2. Johnston B, Atkins MS, Mackiewich B, et al. Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. *IEEE transactions on medical imaging* 1996;15:154-169

3. Sajja BR, Datta S, He R, et al. Unified approach for multiple sclerosis lesion segmentation on brain MRI. *Annals of biomedical engineering* 2006;34:142-151

4. Van Leemput K, Maes F, Vandermeulen D, et al. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE transactions on medical imaging* 2001;20:677-688

5. Garcia-Lorenzo D, Francis S, Narayanan S, et al. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical image analysis* 2013;17:1-18

6. Styner M, Lee J, Chin B, et al. 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. *MIDAS Journal* 2008;MICCAI - New York

7. Garcia-Lorenzo D, Prima S, Arnold DL, et al. Trimmed-likelihood estimation for focal lesions and tissue segmentation in multisequence MRI for multiple sclerosis. *IEEE transactions on medical imaging* 2011;30:1455-1467

8. Khayati R, Vafadust M, Towhidkhah F, et al. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and markov random field model. *Comput Biol Med* 2008;38:379-390

9. Souplet J, Lebrun C, Ayache N, et al. An Automatic Segmentation of T2-FLAIR Multiple Sclerosis Lesions. In: Springer-Verlag, ed. *MICCAI Grand Challenge Workshop: Multiple Sclerosis Lesion Segmentation Challenge*; 2008:1-11

10. Erbayat Altay E. EF, S.E. Jones, C. Hara-Cleaver, J-C. Lee and R.A. Rudick. Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurol* 2013;70:338-344

11. Storelli L, Pagani E, Rocca MA, et al. A Semi-automatic Method for Segmentation of Multiple Sclerosis Lesions on Dual-Echo Magnetic Resonance Images. In: Springer-Verlag, ed. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*: Springer International Publishing; 2015:80-90

12. Schnack HG, van Haren NE, Hulshoff Pol HE, et al. Reliability of brain volumes from multicenter MRI acquisition: a calibration study. *Human brain mapping* 2004;22:312-320

13. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to theMcDonald criteria. *Ann Neurol* 2011;69:292-302

14. Subbanna N, Precup D, Arnold D, et al. IMaGe: Iterative Multilevel Probabilistic Graphical Model for Detection and Segmentation of Multiple Sclerosis Lesions in Brain MRI. *Inf Process Med Imaging* 2015;24:514-526

15. Jain S, Sima DM, Ribbens A, et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *Neuroimage Clin* 2015;8:367-375

16. Mechrez R, Goldberger J, Greenspan H. Patch-Based Segmentation with Spatial Consistency: Application to MS Lesions in Brain MRI. *Int J Biomed Imaging* 2016;2016:7952541

17. Freifeld O, Greenspan H, Goldberg J. Multiple sclerosis lesion detection using constrained GMM and curve evolution. *Int J of Biomed Imaging* 2009;2009:13

18. Galimzianova A, Pernus F, Likar B, et al. Stratified mixture modeling for segmentation of white-matter lesions in brain MR images. *Neuroimage* 2016;124:1031-1043

19. Strumia M, Schmidt F, Anastasopoulos C, et al. White Matter MS-Lesion Segmentation Using a Geometric Brain Model. *IEEE transactions on medical imaging* 2016

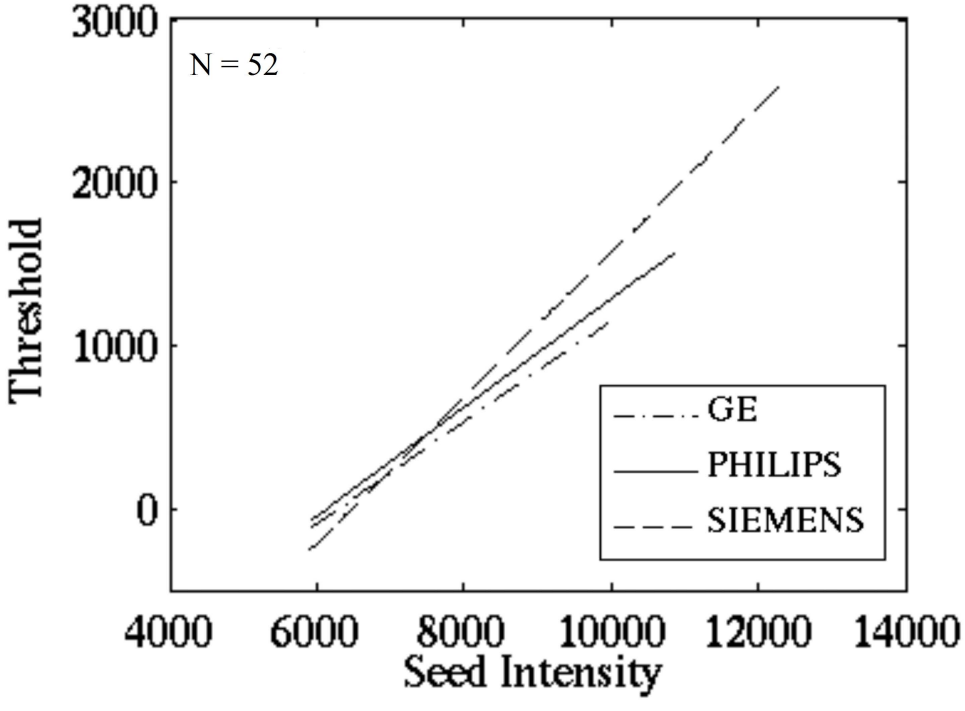
20. Veronese E, Calabrese M, Favaretto A, et al. Automatic Segmentation of Gray Matter Multiple Sclerosis Lesions on DIR Images. *XIII Mediterranean Conference on Medical and Biological Engineering and Computing*; 2013:241-244

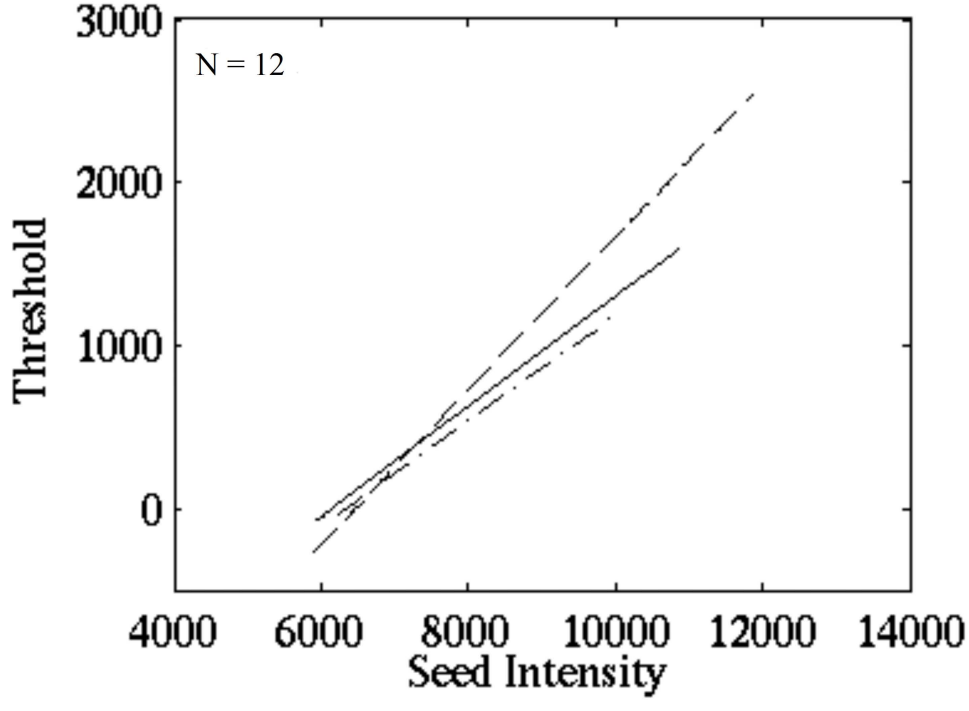
Figure legends

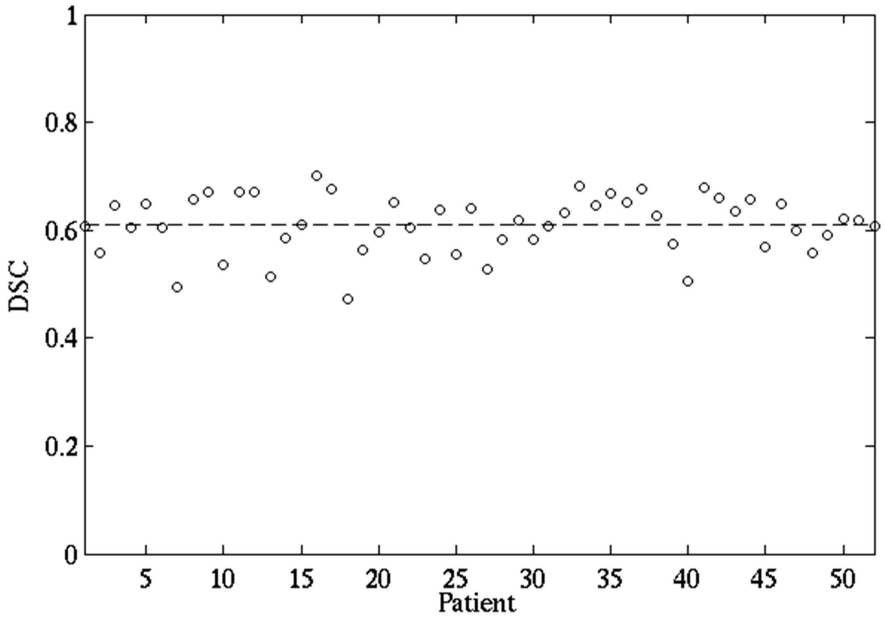
Figure 1. Threshold functions obtained after the training step for each different scanner manufacturers at the decreasing of the training set sample sizes (i.e., number of subjects included) as indicated. It is possible to observe that with decreasing sample size, the linear regression functions did not modify their trends.

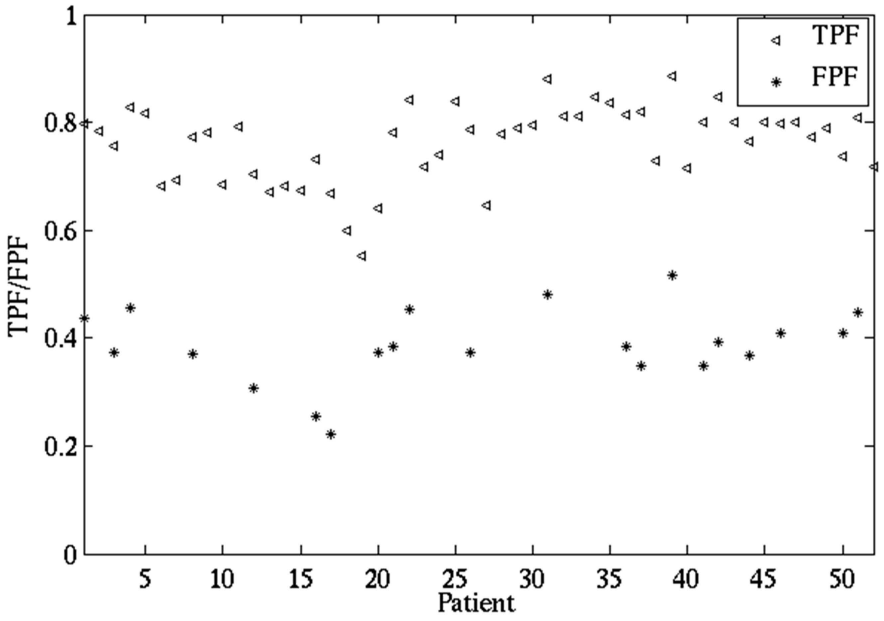
Figure 2. DSC values (top left), mean TPF/ FPF values (top right) and mean FNF values (bottom left) are shown for each patient. In the bottom right, a scatter plot to compare manual lesion load against automatic lesion load is shown. The dashed line is the line of identity.

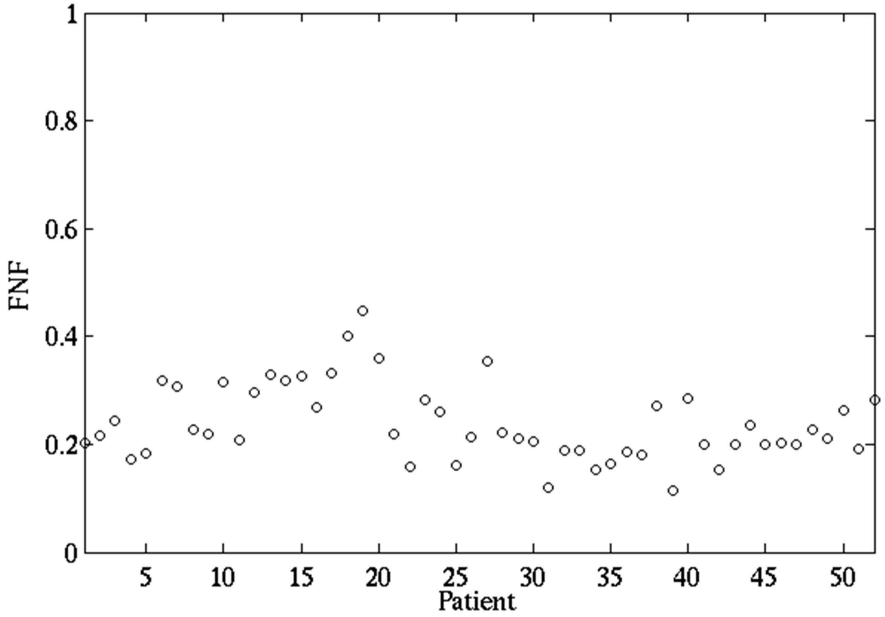
Figure 3. Example lesion segmentations for three patients (in the three rows) from three different scanners by the proposed method (in red) compared to the expert operator segmentation (in blue). The corresponding T2-w images are shown in the right column.

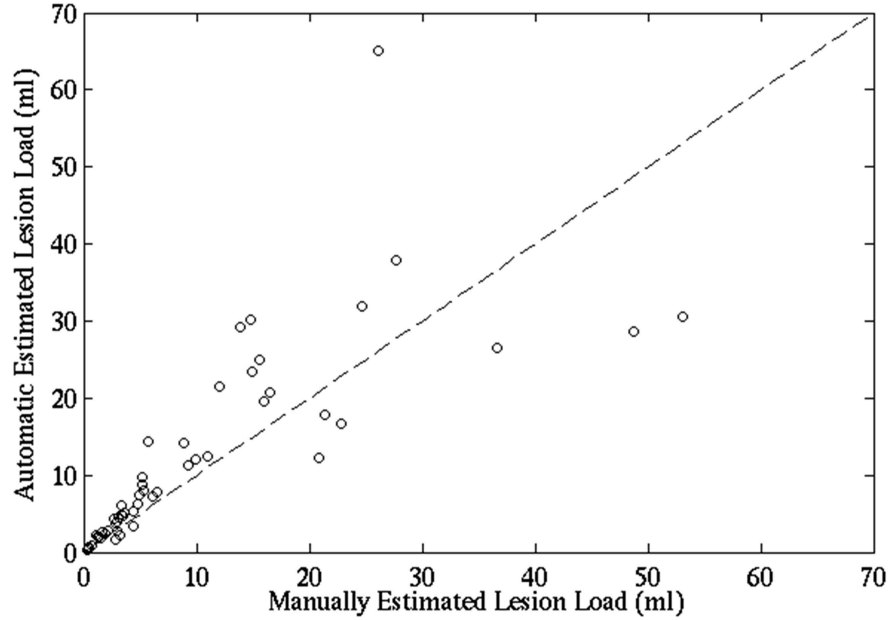

N = 52
3000
2000
1000
0
Threshold
4000
6000
8000
10000
12000
14000
Seed Intensity
GE
PHILIPS
SIEMENS

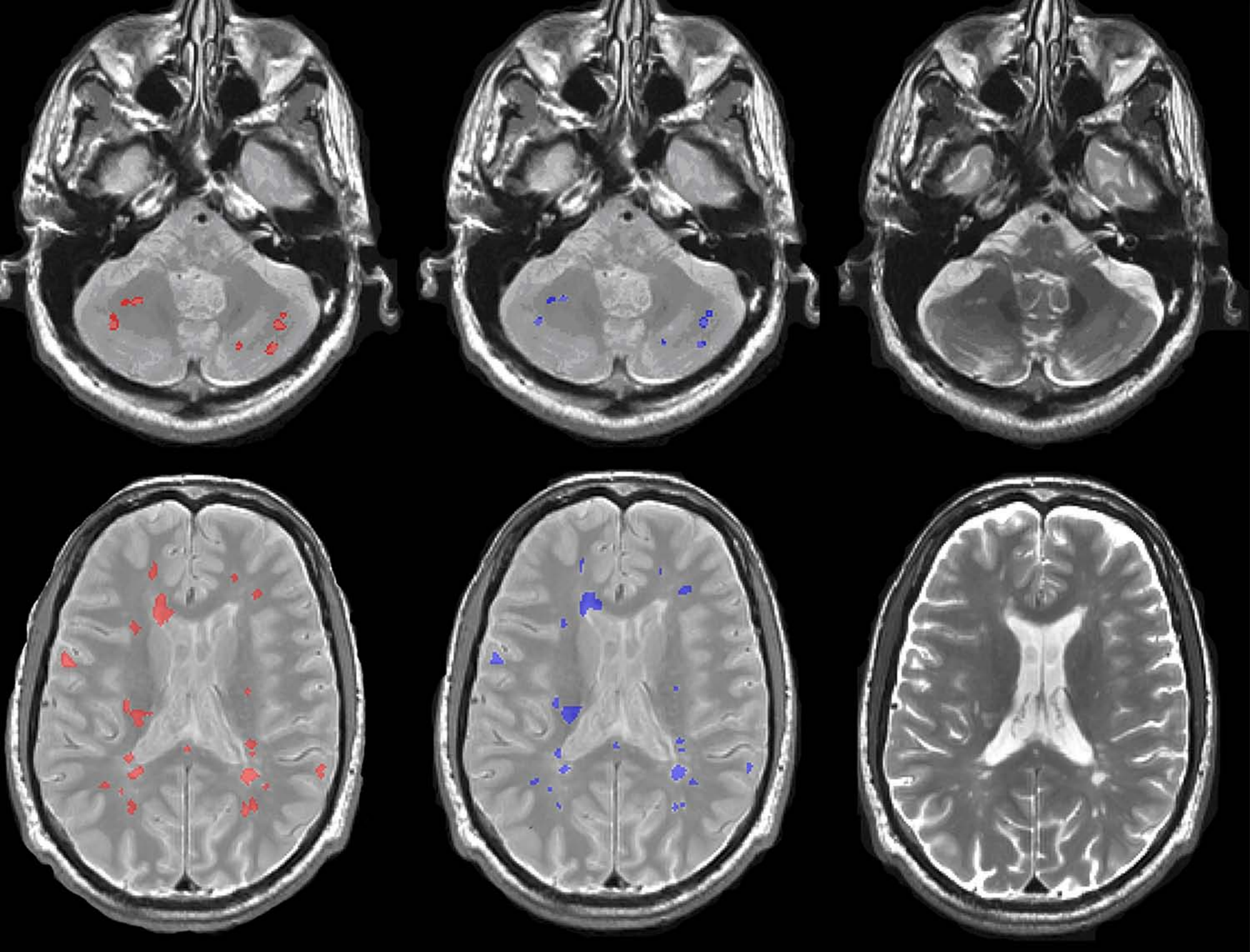

N = 12
3000
2000
1000
0
Threshold
4000
6000
8000
10000
12000
14000
Seed Intensity


DSC
Patient


TPF
FPF
TPF/FPF
Patient


FNF
Patient


Automatic Estimated Lesion Load (ml)
Manually Estimated Lesion Load (ml)



Supplementary material

The methodological structure of the lesion segmentation technique is described below.

1 – Image standardization. One difficulty with non-quantitative MRI techniques is that the image intensities are arbitrary, even within the same protocol, for the same scanner and the same subject.[1] This issue becomes important in multicenter studies where different MRI scanners are used and when a segmentation technique is applied that requires an intensity threshold value. In such cases, standardization of image intensities is vital to correct for the arbitrary intensity scaling for different acquisitions. The standardization procedure used here transforms each intensity value of the original PD-weighted image into a new unique intensity value on the standardized image.[2] To do this, a linear transformation is performed between the intensity values of the two images, such that the histogram of the transformed image has the GM peak position and the first and the last percentiles (1% and 98%) projected into standardized values.[2] After the standardization process, the histograms have comparable intensities and a fixed intensity value for the highest intensity mode.

For the computation of the standard values, an initialization step must be performed only once for a given MRI protocol on a cohort of patients, in which the three intensity parameters (GM peak, and the first and last percentiles) are estimated from each histogram of the training set, averaged and then used to calculate the linear transformation required to generate the standardized image for the protocol.

2 – Region growing algorithm. The core of the algorithm is the pixel-based region growing approach. This clustering method examines neighboring pixels of initial "seed points" and determines whether the pixel neighbors should be added to the region according to similarity constraints.[3] The process is iterated until a similarity condition is violated. The main constraint used for the growth of the segmented region is the intensity similarity, based on a threshold that varies according to a relationship determined during a training process.

3 – Training. Region growing is used on a training dataset of lesions that were previously manually identified in 2-D on the PD-weighted images using a marker point and outlined by an

expert physician. Starting from the marker in each lesion on each 2-D slice, the segmentation continues to the adjacent pixels in the slice and stops before exceeding manual segmentation lesion borders, used as reference result. The absolute *differences* between the seed intensity value and the intensity values inside each segmented region which stops region growing during the training are collected as optimal thresholds. From the threshold values, a straight line is fitted to define a function that associates seed point intensities with a threshold value. As the lesion gets brighter, the difference in intensity between it and the surrounding normal tissue (which should have constant intensity) becomes larger.

4 – Segmentation. The rater identifies each lesion in 2-D using a marker inside the lesion (not necessarily in the center). Then the segmentation starts from each seed point marker and continues using 4 pixels connectivity until the threshold value extracted through the training function is reached. The condition on intensity similarity is combined with the detection of lesion edges to stop the segmentation. An image with enhanced lesion edges is obtained by high-pass filtering a half-way contrast image, obtained by averaging of the non-standardized PD-w and the non-standardized T_2-w MRI scans. 1st echo and 2nd echo scans are used for the edge detection step to take advantage of the contrast in both sequences; high frequency components (edges) are amplified by high-pass unsharp mask filtering the half-way contrast image.[4] The filtered image is then subtracted from the original image, to obtain an image (If_s) in which lesion edges are at zero-crossing points between negative (inside the lesion) and positive (outside the lesion) intensity values. The subtraction would yield the desired values inside and outside the lesion boundary just because of the difference between a high frequency enhanced image and the same but no filtered one. As a consequence, no normalization was needed on the subtraction image. Thus, the intensity values are combined with the edge information to obtain a unique stop condition for the region growing segmentation:

$$Stop\ Condition = (|I_S - I_i| > T) \cap If s_i > 0; \quad [1]$$

where I_s is the intensity of the seed point and I_i is the intensity of an adjacent pixel to be classified on the standardized PD-w image; T is the threshold value (extracted from the training function; different for each lesion). If this condition is violated for all the adjacent pixels, the region growing stops for that lesion.

It was selected for running the segmentation algorithm to use the standardized PD-w image instead of the half-way contrast image because of an easier pre-processing/training, since it was done only on a MRI sequence; while the advantage of both PD/T_2-w sequences was integrated in the method including the lesion edge detection in the stop condition.

5 – Threshold refinement step. After initial segmentation, the intensity distribution of each lesion is employed to estimate a more robust intensity threshold for the region growing. The percentiles of the standardized PD-w intensity distribution of the lesions are used to compute a new threshold to restart the region growing from the results of the first segmentation. The percentile of the intensity distribution used as a new threshold is selected according to the size of the lesion, which gives the sample size of the distribution. The lower the sample size (until 3 pixels), the higher the percentile selected (linearly from 5^{th} until 20^{th} percentile) in order to avoid the inclusion of outliers. The feasibility of the percentiles selected was empirically assessed.

The lesion segmentation method was implemented in Matlab®.

Supplementary Table 1. Main demographic, clinical and conventional MRI characteristics of patients enrolled in this study at six European centers. Center A: Amsterdam, B: Naples, C: Graz, D: London, E: Milan, F: Siena.

Center	**A**	**B**	**C**	**D**	**E**	**F**
Men/Women	4/4	3/8	5/1	3/4	3/7	1/9
Mean age (SD) (years)	45.2 (6.7)	38.8 (7.8)	37.2 (10.2)	39.4 (9.8)	37.4 (8.3)	39.8 (6.2)
Median EDSS (range)	3.5 (2.0-4.0)	1.5 (1.0-6.0)	2.5 (0-4.0)	2.0 (1.0-4.0)	1.5 (1.5-4.0)	1.5 (1.0-4.0)
Mean disease duration (SD) (years)	7.4 (4.5)	12.4 (7.9)	8.9 (7.3)	4.4 (2.5)	7.4 (9.1)	8.1 (4.4)
Mean T2 lesion load (SD) [ml]	11.65 (10.72)	7.28 (5.90)	21.20 (21.14)	13.51 (17.30)	8.63 (7.75)	5.16 (5.68)
Mean # of lesions	66	65	90	91	57	73

Abbreviations: SD, standard deviation; EDSS, Expanded Disability Status scale.

Supplementary Table 2. MRI acquisition centers with scanner manufacturer, coil, and the number of patients included at each center. Center A: Amsterdam, B: Naples, C: Graz, D: London, E: Milan, F: Siena.

Center	Manufacturer/Model	Coil	# of patients
A	General Electric Medical Systems – Signa HDxt	8-channel High Resolution Brain coil	8
B			11
C	Siemens MAGNETOM Tim Trio (Syngo MR B15)	32-channel head coil (12 head anterior + 20 head poterior)	6
D			7
E	Philips Medical System - Achieva	SENSE Head coil 8-elements	10
F			10

References

1. Nyul LG, Udupa JK. New variants of a method of MRI scale normalization. *IEEE transactions on medical imaging* 2000;19:142-150

2. Nyul LG, Udupa JK. On standardizing the MR image intensity scale. *Magnetic resonance in medicine* 1999;42:1072-1081

3. Krishna RK, Kamdi S. Image segmentation and region growing algorithm. . *Int j comput technol electron eng* 2012;2:103-107

4. Luft T, Colditz C, Deussen O. Image enhancement by unsharp masking. *ACM TOG* 2006;20:1206-1213