Fully automatic detection of deep white matter T1 hypointense lesions in multiple sclerosis

2013 Phys. Med. Biol. 58 8323

(http://iopscience.iop.org/0031-9155/58/23/8323)

IOP Publishing Physics in Medicine and Biology

Phys. Med. Biol. **58** (2013) 8323–8337 doi:10.1088/0031-9155/58/23/8323

Fully automatic detection of deep white matter T1 hypointense lesions in multiple sclerosis

Lothar Spies[1,4], Anja Tewes[1], Per Suppa[1], Roland Opfer[1], Ralph Buchert[2], Gerhard Winkler[3,5] and Alaleh Raji[3,5]

[1] jung diagnostics GmbH, Hamburg, Germany
[2] Department of Nuclear Medicine, Charité, Berlin, Germany
[3] Neurozentrum Hamburg, Hamburg, Germany

E-mail: lothar.spies@jung-diagnostics.de

Received 17 May 2013, in final form 28 August 2013
Published 11 November 2013
Online at stacks.iop.org/PMB/58/8323

Abstract

A novel method is presented for fully automatic detection of candidate white matter (WM) T1 hypointense lesions in three-dimensional high-resolution T1-weighted magnetic resonance (MR) images. By definition, T1 hypointense lesions have similar intensity as gray matter (GM) and thus appear darker than surrounding normal WM in T1-weighted images. The novel method uses a standard classification algorithm to partition T1-weighted images into GM, WM and cerebrospinal fluid (CSF). As a consequence, T1 hypointense lesions are assigned an increased GM probability by the standard classification algorithm. The GM component image of a patient is then tested voxel-by-voxel against GM component images of a normative database of healthy individuals. Clusters ($\geqslant$0.1 ml) of significantly increased GM density within a predefined mask of deep WM are defined as lesions. The performance of the algorithm was assessed on voxel level by a simulation study. A maximum dice similarity coefficient of 60% was found for a typical T1 lesion pattern with contrasts ranging from WM to cortical GM, indicating substantial agreement between ground truth and automatic detection. Retrospective application to 10 patients with multiple sclerosis demonstrated that 93 out of 96 T1 hypointense lesions were detected. On average 3.6 false positive T1 hypointense lesions per patient were found. The novel method is promising to support the detection of hypointense lesions in T1-weighted images which warrants further evaluation in larger patient samples.

(Some figures may appear in colour only in the online journal)

[4] Author to whom any correspondence should be addressed.
[5] Equally contributing as senior authors.

Introduction

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system, which also has a neurodegenerative component. Magnetic resonance (MR) imaging of the brain plays a pivotal role in the diagnosis of MS in addition to clinical assessment. MR-based diagnostic criteria have been established and resulted in earlier diagnosis and treatment (Polman *et al* 2011, McDonald *et al* 2001). Hallmark of the disease are lesions due to inflammatory relapses in the brain and spinal cord that disseminate in time and space. Two imaging biomarkers (diagnostic markers) are frequently used to characterize MS lesion patterns and changes thereof: the number of lesions seen in T2-weighted images and the number of gadolinium enhancing lesions in T1-weighted images. T2-weighted fluid-attenuated inversion recovery (FLAIR) images are often preferred over conventional T2-weighted images, because they provide better lesion contrast (Filippi *et al* 1996, Woo *et al* 2006).

Quantitative assessment of MR imaging is useful not only in the management of MS patients in clinical routine, but also in the context of MS therapy trials. MR-based markers have been used in trials to assess potential neuroprotective effects of the drug (Barkhof *et al* 2009, Wolinsky *et al* 2013, Radue *et al* 2012). Detection and characterization of T1 hypointense lesions are particularly useful for this purpose, as these lesions are correlates of the loss of myelin and axons (van Walderveen *et al* 1998). There is accumulating evidence that T1 hypointense lesion accrual is associated with worsening in clinical disability (Giorgio *et al* 2013).

Manual detection and characterization (including volumetry and categorization of the localization) of T1 hypointense lesions are time consuming and prone to significant intra- and inter-reader variability. Therefore, automatic procedures might be useful. In clinical trials, the elimination of intra- and inter-reader variability is expected to improve the statistical power which would result in reduced sample size required to demonstrate therapy effects.

However, whereas there is a large body of literature on automated detection and volumetry of lesions in T2-weighted images (an overview of automated algorithms is given in table 1 in Schmidt *et al* 2012), there are only very few studies which addressed the automatic detection of hypointense lesions in T1-weighted images. Adams *et al* (1999) measured T1 hypointense lesion volumes using a region growing algorithm based on the manual annotation of seed points. Molyneux *et al* (2000) deployed a semi-automated contour technique. These concepts were further advanced by Datta *et al* (2006) who presented a technique which requires only minimal operator intervention.

The relative lack of methods for the automatic analysis of T1-weighted images compared to T2-weighted images might be related to the fact that a T1 hypointense lesion by definition is associated with a hyperintense lesion in T2 (Sahraian *et al* 2010), for the detection and volumetry of which there are automatic methods available. However, there is no exact one-to-one correspondence between hyperintense voxels in T2-weighted images and hypointense voxels in T1-weighted images: the hypointense lesion in T1 may be significantly smaller than the corresponding hyperintense lesion in T2 (Neema *et al* 2007). Thus, fully automatic methods for the independent detection and characterization of T1 hypointense lesions might be useful, in particular since improving accuracy of T1 lesion volumetry seems to enhance correlations with clinical measures (Tam *et al* 2011, Giorgio *et al* 2013).

The aim of this study, therefore, is to propose a novel algorithm for automatic detection and quantification of T1 hypointense lesions. The method takes advantage of the fact that hypointense white matter (WM) lesions in T1 present with intensity similar to normal gray matter (GM). The first step of the method is to partition native, high resolution T1-weighted images into GM, WM and cerebrospinal fluid (CSF). T1 hypointense lesions are assigned an

increased probability of being GM. This is the rationale for the second step in which the GM probability map of the patient is tested voxel-by-voxel against GM maps of healthy subjects. Clusters with significantly increased GM density within a predefined region-of-interest (ROI) for WM are 'candidates' for T1 hypointense lesions. In the third step, the contour of each candidate lesion is transferred to the coregistered T2 or FLAIR image of the same patient in order to check for the presence of a corresponding T2 hyperintense lesion. Currently, this is done by a human rater. In the future, the automatic detection of T1 hypointense lesions will be combined with the automatic detection of T2 hyperintense lesions.

The novel method was inspired by de Boer *et al* (2009), who used tissue misclassification in a T1-weighted image to aid the segmentation of T2 hyperintense lesions in a corresponding FLAIR image. Patterns of misclassifications in T1-weighted images have not yet been used for the characterization of lesions in T1 itself. Other groups considered T1 hypointense lesions as a source of error in tissue segmentation and solved this problem by avoiding explicit segmentation (Stamatakis and Tyler 2005) or by introducing a lesion class in addition to the standard tissue classes of GM, WM and CSF (Seghier *et al* 2008, Shiee *et al* 2010).

The novel method was evaluated by computer simulations and by application to a small sample of relapsing-remitting and secondary progressive MS patients.

Methods

Lesion detection algorithm

In T1-weighted brain MR images, MS lesions appear darker than surrounding normal WM. Since standard classification algorithms are optimized for normal brain anatomy, misclassification of dark appearing WM (DAWM) as GM is frequently observed (Sanfilipo *et al* 2005). Hence the basic idea of the novel lesion detection algorithm is to utilize the misclassification of DAWM for the localization of lesions. A voxelwise statistical test then evaluates whether the GM composition of DAWM voxels is significantly different from that of corresponding voxels of a normative population of healthy subjects.

The novel lesion detection algorithm comprises four major processing steps (see the flowchart in figure 1). First, T1-weighted images are classified into standard tissue classes, i.e. GM, WM and CSF. Second, component images are deformed to match a stereotactic atlas. Third, a voxelwise statistical test is applied, which detects significantly increased GM densities in a single MS patient. Finally, connected regions larger than 0.1 ml are identified and defined as 'candidate' T1 hypointense lesions. For the first three processing steps voxel-based morphometry (VBM) (Ashburner and Friston 2000) as implemented in SPM8 (release April 2009; Wellcome Trust Centre for Neuroimaging, London, UK) was deployed, which was adapted to single subject cases.

Input data are three-dimensional (3D) high-resolution T1-weighted MR images of a single subject, for whom candidate T1 hypointense lesions are to be automatically detected, and 3D high-resolution T1-weighted MR images of subjects of a normative database as reference data for the voxelwise statistical test. A basic prerequisite is that input data were acquired with the same scanner and image acquisition protocol.

Segmentation. The standard segmentation engine of VBM was used, which is the unified segmentation algorithm from Ashburner and Friston (2005). The performance depends on prior tissue probability maps (TPMs) for GM, WM and CSF, which assist the segmentation process by providing prior information on tissue composition of a normal brain. We deployed freely available TPMs which were generated from a population of 662 healthy elderly subjects

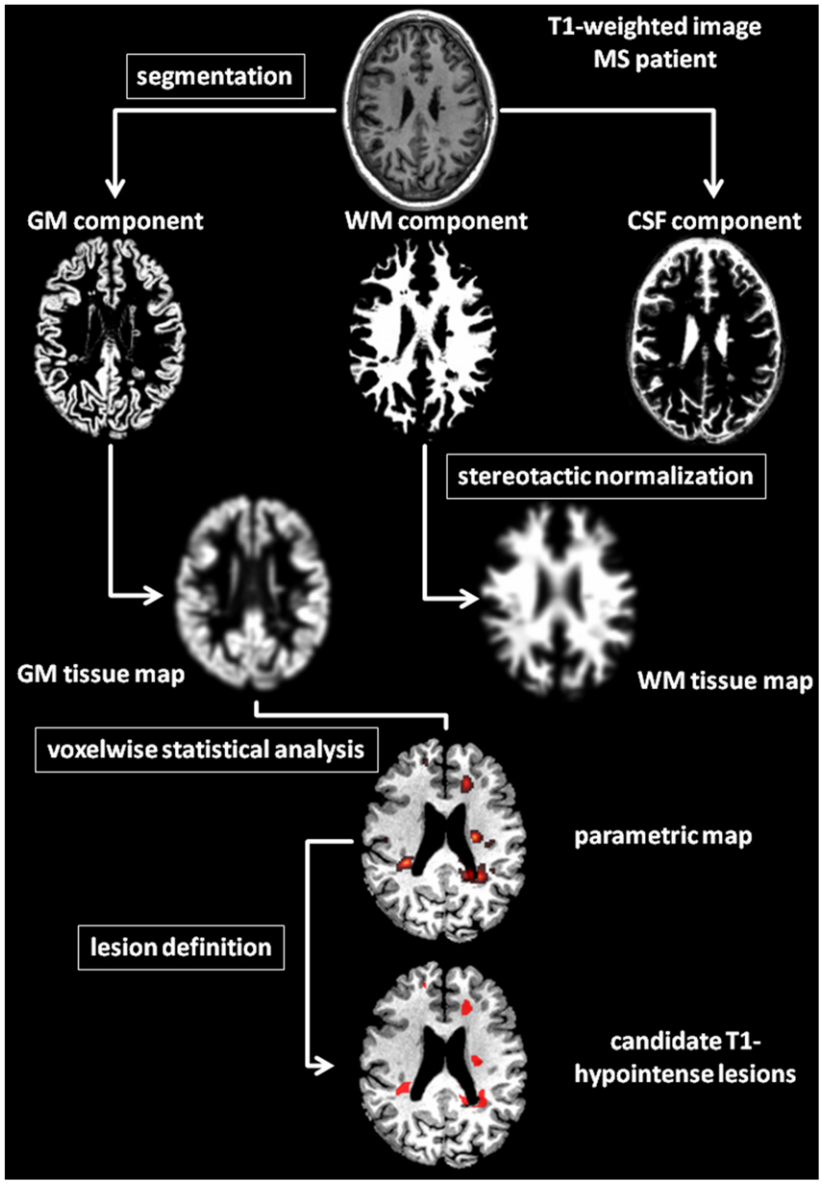

Figure 1. Flowchart of the lesion detection algorithm. It comprises four main processing steps which are segmentation of T1-weighted MR input images, stereotactic normalization, voxelwise statistical analysis and lesion definition. The processing steps are described in the text in more detail.

aged between 63 and 75 years (Lemaitre *et al* 2005). TPMs feature an isotropic resolution of 1 mm. Prior to segmentation, images were rigidly coregistered to a T1 template defined in the same space and with the same dimensions as the TPMs, which was demonstrated to improve the segmentation accuracy of the unified segmentation algorithm (Klauschen *et al* 2009).

We used the default settings of the unified segmentation engine. A mixture of Gaussians to model intensity distributions of GM, WM, CSF and 1–GM–WM–CSF was deployed. The number of Gaussians was 2, 2, 2 and 4, respectively. Further parameters were: 25 mm for

the cut-off of 3D discrete cosine transform basis functions for spatial warping, very light regularization (0.0001), and 60 mm width for the Gaussian smoothness of the intensity bias field.

The unified segmentation algorithm produces GM, WM and CSF component images with dimensions of the original image (native space). Voxel values of a component image range from 0 to 1.

Stereotactic normalization. GM and WM component images were simultaneously normalized using a high-dimensional elastic registration technique (DARTEL) (Ashburner 2007). The DARTEL registration produces flow fields. Flow fields are vector fields describing the transformation of an image from native space to a stereotactic atlas defined in the Montreal Neurological Institute (MNI) space. Flow fields have an isotropic spatial resolution of 2 mm. They were applied to GM component images yielding normalized GM component images. Normalized GM component images are called GM tissue maps in what follows. Volume is preserved (<0.2 per mille volume difference) since the determinant of the Jacobian of the transformation field is locally applied (modulation). Modulated GM tissue maps given in MNI space for each subject were obtained as a result. Modulated GM tissue maps feature an isotropic voxel grid with a grid size of 2 mm. The total voxel volume is thus 0.008 ml.

Voxelwise statistical analysis. Prior to testing, modulated GM tissue maps of the single subject and subjects of the normative database were spatially smoothed by an isotropic Gaussian filter. This ensures that voxel values across the normative database are normally distributed, which is a prerequisite for applying the *t*-test (Salmond *et al* 2002a).

Age and total intracranial volume (TIV), which is the total of GM, WM and CSF volumes, have a significant impact on the GM volume in a normal brain (Pell *et al* 2008). As a consequence GM dependence on age and TIV has to be controlled. A bilinear relationship was used to model the effect of TIV and age in each voxel across all normative subjects. TIV and age were considered as independent variables. For each voxel the bilinear model was fitted to the modulated and smoothed GM tissue maps of the normative database. Voxel values of the single subject, $\overline{\mathrm{gm}}$, were then corrected to mean age, $\langle\mathrm{age}\rangle$, and mean TIV, $\langle\mathrm{TIV}\rangle$, of the normative database using bilinear modeling: $\mathrm{gm} = \overline{\mathrm{gm}} + a\cdot(\mathrm{age} - \langle\mathrm{age}\rangle) + b\cdot(\mathrm{TIV} - \langle\mathrm{TIV}\rangle)$. Here, a and b are the voxelwise regression coefficients. In the equation above, age and TIV are parameters of the single subject.

For voxelwise statistical analysis we deployed the general linear modeling framework adapted to single case studies. Such an adaptation was suggested by Muhlau *et al* (2009). Their approach, which we followed in this paper, uses an independent two-sample *t*-test. Both samples are assumed to have the same distribution with estimated means and variances given by the distribution of the normative database. Hence, the voxelwise two-sample *t*-test simplifies to

$$t = \frac{\langle\mathrm{gm}\rangle - \mathrm{gm}}{\sqrt{s^2/N + s^2}}. \tag{1}$$

Here, $\langle\mathrm{gm}\rangle$ denotes the voxelwise GM volume averaged over all modulated and smoothed GM tissue maps of the normative database comprising N healthy subjects. s^2 is the estimated variance of the respective normative database. gm represents the corrected GM voxel volume of the single subject as described above. Since GM enhancements of the single subject versus subjects of the normative database are of interest, contrasts $t < 0$ (one-sided) are considered.

Statistical analysis was restricted to regions dominated by WM and thus defined by those voxels which contained a minimum of 85% WM in an average normal subject (represented

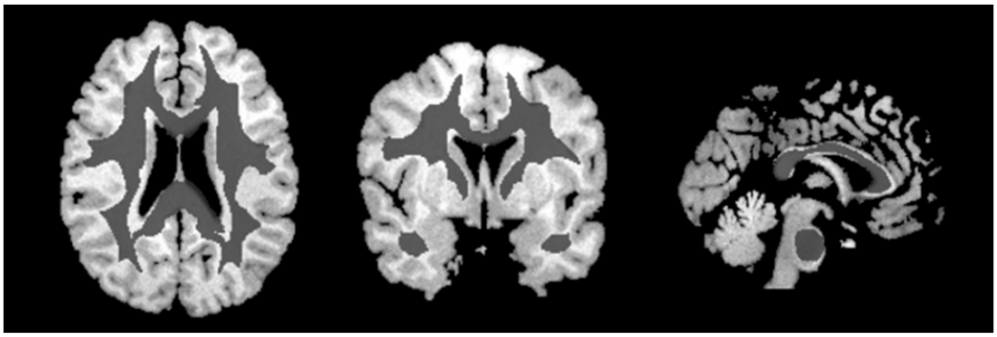

Figure 2. WM mask (dark gray) shown on sample slices in axial, sagittal and coronal view covers regions with more than 85% WM content on average used as a representation for deep WM regions of the brain. Additionally voxels of the corpus callosum were included.

by the WM TPM). Additionally, we included all voxels which belong to the corpus callosum as defined by the wfu pick atlas (Maldjian *et al* 2003). A respective binary mask (figure 2), denoted as WM mask, was derived, which renders one for a voxel belonging to this voxel set and zero outside. Anatomically it can be interpreted as a representation for deep WM regions of the brain.

The test produces parametric *t*-maps with the same dimension and voxel volume as the modulated and spatially smoothed GM tissue maps. No corrections for multiple comparisons were applied because we wanted the test to be most sensitive.

Lesion definition. Lesion masks were derived from parametric *t*-maps. Clusters of voxels with negative *t*-values lower than threshold *t*-values determined by significance levels (*p*-values) of interest were considered as lesions if cluster voxels were interconnected by at least one edge. For automatic lesion labeling and counting, a method developed by Thurfjell *et al* (1992) was applied. All voxels belonging to a lesion define a lesion mask. The corresponding lesion volume is the total of all voxel volumes of the lesion mask. Lesion volumes below a threshold of 0.1 ml (13 voxels) were discarded. In this way, for each parametric *t*-map and each *p*-value a lesion map of 'candidate' T1 hypointense lesions was produced.

Evaluation

Simulation data. The Brainweb MR simulator provides customizable high-resolution MR images of normal brains and brains with MS lesions for T1, T2 and PD contrast (Collins *et al* 1998). It has been frequently used in the literature and meanwhile it is considered a standard for testing methods for MR-based lesion detection in MS patients. In this study, we used simulated T1-weighted images generated with the default settings of the simulation engine (T1-weighted spoiled FLASH pulse sequences with a repetition time of 18 ms, echo time of 10 ms, flip angle of 30°, 1 mm isotropic voxel grid and matrix size of $181 \times 217 \times 181$).

Eleven images of a healthy brain, denoted as a non-lesion database, and 11 images with a typical MS lesion load (using the simulator's 'severe' lesion model), denoted as a lesion database, were simulated with random noise and non-uniformity level with a mean noise level of 3% (relative to the brightest tissue) and a mean image non-uniformity of 20%, respectively. We suppose that images with a 3% noise level and 20% non-uniformity on average are representative for patient images of current scanners.

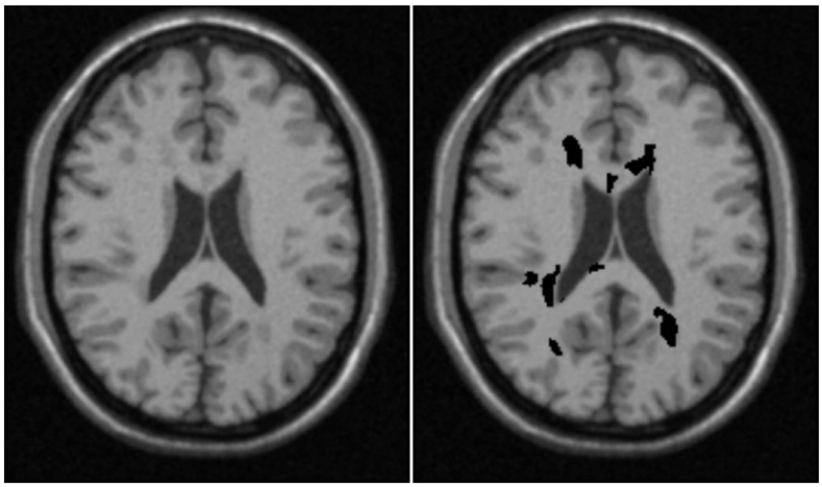

Figure 3. Axial view of a simulated MS brain with T1 lesions stemming from the Brainweb MR simulator using the severe lesion model (left). Right shows the same slice but with the lesion ROI image superimposed (black). The lesion ROI image was generated by thresholding the LM ground truth image at 0.2. Slices are displayed in radiological orientation (left to right).

The simulator offers ground truth information. For a normal brain, ground truth information is provided as GM, WM, CSF ground truth images. Ground truth images feature a voxel grid with identical size and dimensions as simulation images. For a brain with MS pathology, an additional class is provided, which is distinct from GM, WM and CSF and denoted as a lesion matter (LM) ground truth image. All ground truth images feature voxel values from 0 to 1 denoting the probability of the respective tissue for that voxel.

Patient cohort and normative databases. The patient cohort included 10 MS patients (7 females, 3 males; age 31–63 yr, median 44 yr). Six patients had been diagnosed with relapsing-remitting MS and four patients with secondary progressive MS.

A normative database comprises 28 healthy individuals with age ranging from 22 to 80 years without history of current neurological or psychiatric disorder. Median age is 43 years. The health status of each eligible subject was confirmed by a clinical examination.

Normal subjects and MS patients were all acquired using a 3T MR scanner (GE Signa HDxt) and deploying the same imaging sequence with 3D high-resolution T1-weighted magnetization prepared rapid gradient echo (MPRAGE) acquisition. The following protocol settings were used: TR = 6.79 ms, TE = 1.93, TI = 450 ms. Flip angle was 8°. No contrast agent was administered prior to MPRAGE scanning. Slice thickness is 1 mm and pixels are 1 mm in size for both directions.

For all MS patients, FLAIR images were additionally acquired. The protocol settings were TR = 8000 ms, TE = 120 ms, TI = 2247 ms and flip angle = 90°. The slice thickness was 2.6 mm with a 2.6 mm gap and the isotropic in-plane voxel size is 0.51 mm.

Effect of misclassification. The effect of misclassification was studied in synthetically generated lesions deploying the severe lesion model of the Brainweb MR simulator: a lesion ROI image was defined by thresholding LM ground truth image voxels at 0.2 (figure 3). Voxels with a value greater than 0.2 were set to 1 and define the lesion ROI. All other voxels were set to zero. The lesion ROI volume was calculated to be 19.0 ml. All 22 simulated images

(with and without lesions) were segmented into GM, WM and CSF component images using the unified segmentation algorithm. GM, WM and CSF volumes within the lesion ROI were estimated by summing up all voxels of the GM, WM and CSF component images belonging to the lesion ROI normalized to the volume of the lesion ROI. Mean GM, WM and CSF volumes were calculated.

Parameter optimization. We were interested in assessing the performance of the algorithm for a wider spectrum of T1 hypointense lesions ranging from gray to almost normal appearing WM. Since the Brainweb MR simulator does not offer a model which allows tuning of lesion severity, we devised our own simulation model.

For this simulation, all 28 healthy subjects of the intra-scanner normative database were segmented into GM and WM component images. GM and WM component images were then simultaneously registered to the stereotactic MNI space using DARTEL and modulated to warrant volume preservation. The lesion ROI image as defined above (figure 3) was deployed and was transformed into MNI space. No modulation was applied to ensure that the lesion ROI remained binary after transformation.

Since the segmentation algorithm identifies more GM in DAWM areas as normally expected, an artificial lesion in normal tissue, gm_β, can be modeled by adding extra GM tissue to normal GM tissue according to the following equation:

$$\mathrm{gm}_\beta = \mathrm{gm}_0 + \beta \cdot \mathrm{wm}_0, \tag{2}$$

where gm_0 and wm_0 are modulated GM and WM tissue maps of a normal healthy subject. Outside of the lesion ROI the parameter β is always zero, i.e. $\mathrm{gm}_\beta = \mathrm{gm}_0$. The parameter β, termed as 'lesion effect size', models the impact of the lesion on WM contrast (lesion severity) and is varied from 0 to 1 in increments of 0.1. A β value of zero models the situation that no lesion tissue is present or the lesion appears as normal WM. A value of 1 signifies that the WM lesion tissue is misclassified as GM. Since CSF content in WM is relatively small and not altered in DAWM lesions, CSF was disregarded in the above model for the sake of simplicity. Composite GM tissue maps were spatially smoothed with the same Gaussian filter and then submitted to the voxelwise statistical test (equation (1)).

To obtain a robust estimation of the effect, this procedure was iterated for all subjects of the normative database as follows: select one subject, apply the artificial lesion with varying strengths and test against the remaining original subjects of the database (leave-one-out). For each subject and each parameter setting (lesion effect size, Gaussian filter size and p-value), a parametric map was generated and lesions were defined as described above. Gaussian filter sizes with a full width at half maximum (FWHM) of 4 and 8 mm and p-values of 0.05 and 0.005 constituted the parameter range of interest.

For validation on voxel level, lesion maps were analyzed as follows: WM voxels were classified as true positive (TP) if they belonged to a lesion ROI and to a 'candidate' T1 hypointense lesion. False positive (FP) voxels were voxels outside of the lesion ROI, which belonged to a 'candidate' T1 hypointense lesion. Analogously, true negative (TN) voxels were voxels outside of the lesion ROI which did not match a 'candidate' T1 hypointense lesion and false negative (FN) voxels were voxels within the lesion ROI which did not match a 'candidate' T1 hypointense lesion.

The dice coefficient (DC) (Dice 1945) was chosen as a primary performance measure to characterize correspondence between the lesion map and ground truth image. DC is a very common similarity measure to evaluate and compare the performance of lesion segmentation algorithms. Based on voxel classification, the DC is given by $\mathrm{DC} = \frac{2 \cdot \mathrm{num(TP)}}{2 \cdot \mathrm{num(TP)} + \mathrm{num(FP)} + \mathrm{num(FN)}}$

Table 1. WM, GM and CSF volumes, denoted as vol (wm) , vol (gm), vol (csf), within the lesion ROI applied to normal and brain with MS lesions (MS brain) from the Brainweb simulation database.

	vol (wm) (%)	vol (gm) (%)	vol (csf) (%)
MS brain	73.8	24.9	1.3
Normal	93.1	5.6	1.3

(Zijdenbos *et al* 1994). Here, num (TP) is the number of all TP voxels of a lesion map. Numbers num (FP) and num (FN) are defined analogously.

Automatic lesion detection in MS patients. Since no widely accepted and publicly available ground truth for detection of T1 hypointense lesions exists, automatic lesion detection was evaluated against visual rating by two independent experts (AR and GW). For visual rating rigidly coregistered FLAIR and MPRAGE images were used. First, lesions were confirmed as 'true' T1 hypointense lesions by the two raters if the criteria of Sahraian *et al* (2010) were met, which are T1 hypointensity and partial concordance with T2 hyperintensity. MPRAGE images of the study cohort were supplied to the automatic lesion detection algorithm which produced lesion maps with candidate T1 hypointense lesions. Raters then evaluated whether correspondence between true T1 hypointense lesions and candidate T1 hypointense lesions existed. Correspondence of true T1 lesions and candidate hypointense lesions was established if they partly overlapped as judged by visual inspection by the two experts. Correspondence in size and volume was not assessed. Both confirmation of lesion and definition of correspondence were decided in consensus amongst the two experts.

A candidate T1 hypointense lesion with a correspondence to a true T1 hypointense lesion was labeled as a TP lesion. Analogously, an FP lesion is a candidate T1 hypointense lesion without correspondence to a true T1 hypointense lesion. Detection performance was specified in terms of the TP fraction (TPF) and the mean number of FP lesions per patient. TPF was defined as the total number of TP lesions divided by the total number of true T1 hypointense lesions.

Results

Effect of misclassification

Mean GM, WM, and CSF volumes for the region defined by the lesion ROI applied to simulation images of the non-lesion and lesion database are presented in table 1. We found a 19.3% shift in volume from WM to GM. CSF was not affected. Taking advantage of equation (2), the lesion effect size is given by $\frac{\mathrm{vol}(\mathrm{gm}_\beta)-\mathrm{vol}(\mathrm{gm}_0)}{\mathrm{vol}(\mathrm{wm}_0)}$. Here, $\mathrm{vol}(\mathrm{wm}_0)$ and $\mathrm{vol}(\mathrm{gm}_0)$ are the WM and GM volume of a lesion ROI applied to an image of the non-lesion database. $\mathrm{vol}(\mathrm{gm}_\beta)$ is the GM volume of a lesion ROI applied to an image of the lesion database. Values were taken from table 1 and yielded an effect size of 0.2 for the severe T1 lesion model of the Brainweb MR simulator (see figure 3).

Parameter optimization

A DC was calculated for each parameter set (p-value, FWHM and β) and each subject of the normative database (28 in total) as described above. DCs were averaged over all subjects of the normative database.

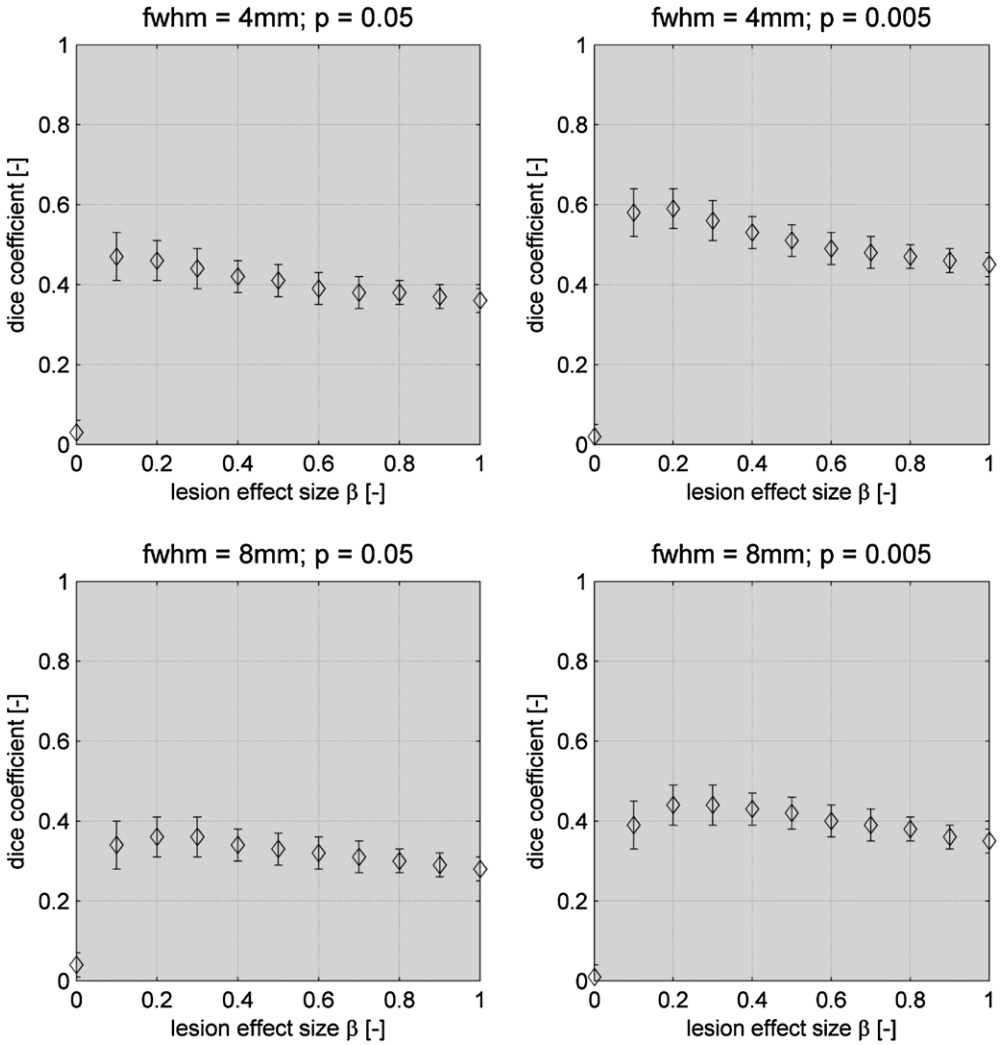

Figure 4. DCs as a function of the lesion effect size β ranging from 0 to 1. Results are presented for $p = 0.05$ and $p = 0.005$ and two Gaussian filter sizes with FWHM of 4 and 8 mm.

Mean DC values with errorbars given by the standard deviations are reported as a function of the effect size in figure 4. Plots were generated for p-values 0.05 and 0.005 and Gaussian filter sizes with FWHM of 4 and 8 mm. DC values reach a plateau or a local maximum at effect sizes around 0.2 and then drop off for higher effect sizes.

For a p-value of 0.005 and an FWHM of 4 mm, a maximum DC of 60% was obtained for $\beta = 0.2$. This value is the maximum for all parameter configurations and all β values. For $p = 0.005$ and FWHM $= 4$ mm, the minimum value of 45% is reached at $\beta = 1$, meaning that the detection performance is less for lesions which appear almost as dark as cortical GM. For $\beta = 0$, DCs are finite due to the presence of FP voxels. Standard deviations are always smaller than 5%. The parameter setting, $p = 0.005$ and FWHM $= 4$ mm, shows the best performance along the entire range of β values.

For a representative β value of 0.2, a summary of the model predictions are given in table 2.

Table 2. Simulation model predictions for mean DC at a lesion effect size, β, of 0.2 tabulated for p-values of 0.05 and 0.005 and FWHMs of 4 and 8 mm. The DC was defined as DC $= 2 \cdot \text{num(TP)}/2 \cdot \text{num(TP)} + \text{num(FP)} + \text{num(FN)}$, where num(TP), num(FP) and num(FN) are the numbers of all TP, FP and FN voxels of a lesion map.

FWHM (mm)	p	DC (%)
4	0.005	60
8	0.005	43
4	0.05	44
8	0.05	38

Automatic lesion detection in MS patients

The parameter setting, which demonstrated the best performance in the simulation experiment ($p = 0.005$ and FWHM $= 4$ mm), was applied to MS patient data.

Readers identified and confirmed 96 T1 hypointense lesions in deep WM and 7 lesions within the surrounding WM for the total of 10 MS patients of the study cohort.

The novel algorithm detected a total of 129 candidate T1 hypointense lesions in deep WM. Ninety-three out of 129 candidate T1 hypointense lesions matched true T1 hypointense lesions. The TPF was thus 97%. 36 candidate T1 hypointense lesions were FP lesions resulting in 3.6 FP lesions per patient on average.

Analyzing FP lesions in more detail, we found that 6 out of 36 FP lesions could be matched to lesions seen in MPRAGE data but not in FLAIR images. Nine FP lesions could not be matched to either of the two MR imaging modalities. 21 FP lesions had only counterparts in FLAIR images but not in MPRAGE. In these cases, the algorithm seems to be more sensitive than visual inspection by experienced readers. If we relaxed the definition of T1 hypointense lesions, accepting that a T1 hypointense lesion is either identified in MPRAGE or in FLAIR, then the mean number of FP lesions per patient would improve to 0.9. The TPF would reach 98%.

For illustration, figure 5 depicts automatically detected candidate T1 hypointense lesions for a representative MS patient of the study cohort. Lesions are presented as contours on coregistered MPRAGE and FLAIR images, for which non-parenchymal matter was removed.

Discussion

TPMs play an important role in the segmentation process, which is an essential part of the detection algorithm. Since TPMs were built from elderly subjects aged between 63 and 75 years (Cyceron TPMs), they misrepresent the population of MS patients which is typically younger on average. To evaluate the age mismatch, default TPMs of the SPM8 package (SPM TPMs) were used for comparison. SPM TPMs were derived from 452 T1-weighted scans of normal young adult brains and feature a 2 mm isotropic resolution. Detailed age information was not published for this cohort. We found an improved CSF segmentation performance in normative data for Cyceron TPMs as opposed to SPM TPMs. The inclusion of T2 information to improve the definition of CSF as realized in the construction of Cyceron TPMs has a significant impact. This impact seems to outweigh the age mismatch. We hypothesize that a set of TPMs matching the age of a typical MS population and which incorporates T2 information is best suited to minimize segmentation failures. Since no T2-weighted images were available for subjects of the normative database, we could not build dedicated TPMs for this study.

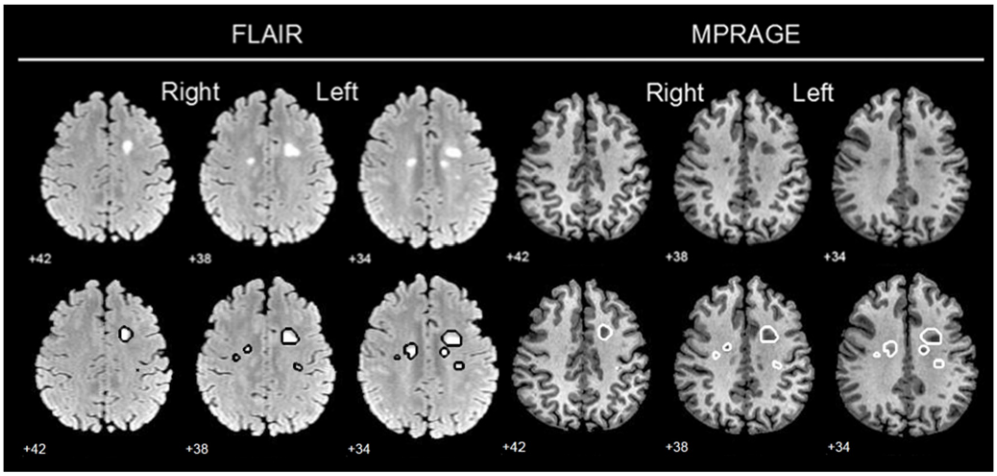

Figure 5. Fully automatically detected T1 hypointense lesions for a typical MS patient. Axial views of FLAIR (left) and corresponding MPRAGE (right) images in MNI space are presented. White contours delineate regions on MPRAGE images (right) with finite t scores for $p < 0.005$ and a minimum volume of 0.1 ml. The same contours (black) are overlaid on the FLAIR image. Views are shown in radiological convention.

A lesion effect size of about 0.2 was derived from simulation data using the severe T1 lesion model of the Brainweb MR simulator. To investigate this further we calculated effect sizes for T1 hypointense lesions in real MS patient data featuring a similar contrast as the lesions in the Brainweb data. We found lesion effect sizes of the same order ranging from 0.18 to 0.23. As a consequence, a value of 0.2 seemed to be a good representation for this type of MS lesions in T1-weighted images. However, the simulation has limitations since it does not cover lesions which have a contrast similar to CSF. This would require a simulation model which includes CSF. Such an extension is beyond the scope of this work and subject of further studies. Moreover, segmentation using the unified segmentation engine is mainly based on a frequency distribution of image intensities. This warrants a direct proportionality between lesion contrast in WM and WM misclassification, i.e. the darker the contrast of lesion tissue in WM the greater the proportion of GM. This is countered, however, by the fact that the unified segmentation engine utilizes TPMs to correct for obvious misclassifications. The probability of a voxel to contain a specific tissue is not only determined by the image intensity but also by the location of the voxel in the brain as defined by the TPMs. Hence it is not straightforward to conclude how the algorithm performs in the presence of darker WM lesions, i.e. with lesion effect sizes greater than 0.2. Further work and more data are needed to clarify this point. Despite these indeterminacies another group recently reported the usefulness of the unified segmentation algorithm for the detection of increased iron content in MPRAGE images which correlated with GM and WM misclassifications in the basalganglia (Goto *et al* 2012).

Although increasing β values produce a darker contrast, which suggests a greater DC, indeed we observe that the DC slightly decreases with increasing lesion contrast. This is because the number of TP voxels steeply rises with increasing β values and saturates for $\beta > 0.2$. Beyond this value only the number of FP voxels gets greater due to increasing differences between the GM content of voxels inside and outside of the lesion ROI. Consequently, more GM content is smeared out by the Gaussian filter. This eventually leads to a significant increase in FP voxels while the number of TP voxels remains almost constant, thus compromising detection performance.

For each voxel of the WM mask, distributions were generated comprising voxel values of smoothed GM tissue maps for all subjects of the normative database. The Shapiro–Wilk formalism was applied on voxel-basis to test for normality. The majority of voxels failed to meet the requirements for normality if the associated maps were smoothed with a Gaussian filter of FWHM = 4 mm. Many papers advise to use high filter sizes in order to assure that the condition of normal distributed residuals is met for statistical testing, e.g. Muhlau *et al* (2009). But Salmond *et al* (2002b) delivered an argument that testing for an increase in voxel values is robust and does not require normality as a prerequisite. Implicitly the validity of their argument is confirmed by our results.

TIV and age are used as covariates for all results presented in this paper. However, we also tested both covariates independently of each other. We found that age does not have a significant impact on the statistical results. Consequently, the fact that the exact age of the simulated brains is not extractable from the literature is uncritical. GM volume suggests that the age of the brain is between 20 and 30 years. Therefore we used a dummy age of 25 years where needed.

No widely accepted standard for the interpretation of DC coefficients exists. Some authors consider DC of 70% as 'excellent' (Anbeek *et al* 2004). Others interpret levels of 40%, 60% and 80% as 'moderate', 'substantial' and 'almost perfect' (Landis and Koch 1977). We approximated the DC to be near 60%, which can hence be interpreted as substantial agreement. Datta *et al* (2006) validated their semi-automated approach on 14 subjects. They found DCs varying between 40% and 90%. A DC of 60% falls well in between their range of values.

Conclusion

A novel voxel-based algorithm for fully automatic detection of candidate T1 hypointense lesions was presented. It exploits the effect that standard gray and white matter classification schemes systematically misclassify voxels, which normally belong to white matter but appear darker than the surrounding white matter tissue. Changes in white matter contrast are often caused by disease processes which alter tissue properties.

A simulation model was devised to characterize the performance of the novel algorithm. We found a maximum DC of 60% for T1 hypointense lesions, which are darker than normal appearing white matter and brighter than cortical gray matter.

Visual inspection of MS patient images demonstrated that 97% of the automatically detected candidate T1 hypointense lesions in deep white matter could be matched to 'true' T1 hypointense lesions as verified in MPRAGE and FLAIR images. 3.6 false positive T1 hypointense lesions per patient were found on average.

We conclude that the proposed algorithm shows promise to become a robust tool for fully automatic T1 hypointense lesion detection in deep white matter regions. This warrants testing in larger retrospective and prospective clinical studies.

Acknowledgments

LS, AT, RO and PS are employees of jung diagnostics GmbH. RB, AR and GW have nothing to disclose.

References

Adams H P *et al* 1999 Hypointense and hyperintense lesions on magnetic resonance imaging in secondary-progressive MS patients *Eur. Neurol.* **42** 52–63

Anbeek P, Vincken K L, van Osch M J P, Bisschops R H C and van der Grond J 2004 Automatic segmentation of different-sized white matter lesions by voxel probability estimation *Med. Image Anal.* **8** 205–15

Ashburner J 2007 A fast diffeomorphic image registration algorithm *NeuroImage* **38** 95–113

Ashburner J and Friston K J 2000 Voxel-based morphometry—the methods *NeuroImage* **11** 805–21

Ashburner J and Friston K J 2005 Unified segmentation *NeuroImage* **26** 839–51

Barkhof F, Calabresi P A, Miller D H and Reingold S C 2009 Imaging outcomes for neuroprotection and repair in multiple sclerosis trials *Nature Rev. Neurol.* **5** 256–66

Collins D L, Zijdenbos A P, Kollokian V, Sled J G, Kabani N J, Holmes C J and Evans A C 1998 Design and construction of a realistic digital brain phantom *IEEE Trans. Med. Imaging* **17** 463–8

Datta S, Sajja B R, He R, Wolinsky J S, Gupta R K and Narayana P A 2006 Segmentation and quantification of black holes in multiple sclerosis *NeuroImage* **29** 467–74

de Boer R *et al* 2009 White matter lesion extension to automatic brain tissue segmentation on MRI *NeuroImage* **45** 1151–61

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

Filippi M *et al* 1996 Quantitative assessment of MRI lesion load in multiple sclerosis *Brain* **119** 1349–55

Giorgio A *et al* 2013 Relevance of hypointense brain MRI lesions for long-term worsening of clinical disability in relapsing multiple sclerosis *Mult. Scler.* at press

Goto M *et al* 2012 Association between iron content and gray matter missegmentation with voxel-based morphometry in basal ganglia *J. Magn. Reson. Imaging* at press

Klauschen F, Goldman A, Barra V, Meyer-Lindenberg A and Lundervold A 2009 Evaluation of automated brain MR image segmentation and volumetry methods *Hum. Brain Mapp.* **30** 1310–27

Landis J R and Koch G G 1977 An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers *Biometrics* **33** 363–74

Lemaitre H, Crivello F, Grassiot B, Alperovitch A, Tzourio C and Mazoyer B 2005 Age- and sex-related effects on the neuroanatomy of healthy elderly *NeuroImage* **26** 900–11

Maldjian J A, Laurienti P J, Kraft R A and Burdette J H 2003 An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of FMRI data sets *NeuroImage* **19** 1233–9

McDonald W I *et al* 2001 Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis *Ann. Neurol.* **50** 121–7

Molyneux P D *et al* 2000 The precision of T1 hypointense lesion volume quantification in multiple sclerosis treatment trials: a multicenter study *Mult. Scler.* **6** 237–40 (PMID: 10962544)

Muhlau M *et al* 2009 Voxel-based morphometry in individual patients: a pilot study in early Huntington disease *Am. J. Neuroradiol.* **30** 539–43

Neema M, Stankiewicz J, Arora A, Guss Z D and Bakshi R 2007 MRI in multiple sclerosis: what's inside the toolbox? *Neurotherapeutics* **4** 602–17

Pell G S, Briellmann R S, Chan C H P, Pardoe H, Abbott D F and Jackson G D 2008 Selection of the control group for VBM analysis: influence of covariates, matching and sample size *NeuroImage* **41** 1324–35

Polman C H *et al* 2011 Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria *Ann. Neurol.* **69** 292–302

Radue E-W *et al* 2012 Impact of fingolimod therapy on magnetic resonance imaging outcomes in patients with multiple sclerosis *Arch. Neurol.* **69** 1259–69

Sahraian M A, Radue E-W, Haller S and Kappos L 2010 Black holes in multiple sclerosis: definition, evolution, and clinical correlations *Acta Neurol. Scand.* **122** 1–8

Salmond C H, Ashburner J, Vargha-Khadem F, Connelly A, Gadian D G and Friston K J 2002a Distributional assumptions in voxel-based morphometry *NeuroImage* **17** 1027–30

Salmond C H, Ashburner J, Vargha-Khadem F, Connelly A, Gadian D G and Friston K J 2002b The precision of anatomical normalization in the medial temporal lobe using spatial basis functions *NeuroImage* **17** 507–12

Sanfilipo M P, Benedict R H B, Sharma J, Weinstock-Guttman B and Bakshi R 2005 The relationship between whole brain volume and disability in multiple sclerosis: a comparison of normalized gray versus white matter with misclassification correction *NeuroImage* **26** 1068–77

Schmidt P *et al* 2012 An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis *NeuroImage* **59** 3774–83

Seghier M L, Ramlackhansingh A, Crinion J, Leff A P and Price C J 2008 Lesion identification using unified segmentation-normalisation models and fuzzy clustering *NeuroImage* **41** 1253–66

Shiee N, Bazin P-L, Ozturk A, Reich D S, Calabresi P A and Pham D L 2010 A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions *NeuroImage* **49** 1524–35

Stamatakis E A and Tyler L K 2005 Identifying lesions on structural brain images—validation of the method and application to neuropsychological patients *Brain Lang.* **94** 167–77

Tam R C, Traboulsee A, Riddehough A, Sheikhzadeh F and Li D K B 2011 The impact of intensity variations in T1-hypointense lesions on clinical correlations in multiple sclerosis *Mult. Scler.* **17** 949–57

Thurfjell L, Bengtsson E and Nordin B 1992 A new three-dimensional connected components labeling algorithm with simultaneous object feature extraction capability *CVGIP, Graph. Models Image Process.* **54** 357–64

van Walderveen M A *et al* 1998 Histopathologic correlate of hypointense lesions on T1-weighted spin-echo MRI in multiple sclerosis *Neurology* **50** 1282–8

Wolinsky J S *et al* 2013 Magnetic resonance imaging outcomes from a phase III trial of teriflunomide *Mult. Scler.* at press

Woo J H, Henry L P, Krejza J and Melhem E R 2006 Detection of simulated multiple sclerosis lesions on T2-weighted and FLAIR images of the brain: observer performance *Radiology* **241** 206–12

Zijdenbos A P, Dawant B M, Margolin R A and Palmer A C 1994 Morphometric analysis of white matter lesions in MR images: method and validation *IEEE Trans. Med. Imaging* **13** 716–24