Original Research

Quantitative Follow-up of Patients With Multiple Sclerosis Using MRI: Reproducibility

Charles R.G. Guttmann,* MD[1] Ron Kikinis, MD,[1] Mark C. Anderson, MS,[1] Marianna Jakab, MS,[1] Simon K. Warfield, PhD,[1] Ron J. Killiany, PhD,[1] Howard L. Weiner, MD,[2] and Ferenc A. Jolesz, MD[1]

The reproducibility of an automated method for estimating the volume of white matter abnormalities on brain magnetic resonance (MR) images of multiple sclerosis (MS) patients was evaluated. Twenty MS patients underwent MR imaging twice within 30 minutes. Measurement variability is introduced mainly by MRI acquisition and image registration procedures, which demonstrate significantly worse reproducibility than the image segmentation. The correction of partial volume artifacts is essential for sensitive measurements of overall lesion burden. The average lesion volume difference (bias) between two MR exams of the same MS patient ($N = 20$) was 0.05 cm^3, with a 95% confidence interval between −0.17 and +0.28 cm^3, suggesting that the proposed measurement system is suitable for clinical follow-up trials, even in relatively small patient cohorts. The limits of agreement for lesion volume were between −1.3 and +1.5 cm^3, implying that in individual patients changes in lesion load need to be at least this large to be detected reliably. This automated method for estimating lesion burden is a reliable tool for the evaluation of MS progression and exacerbation in patient cohorts and potentially also in individual patients. J. Magn. Reson. Imaging 1999; 9:509–518. © **1999 Wiley-Liss, Inc.**

Index terms: sclerosis, multiple; magnetic resonance (MR), image processing; magnetic resonance (MR), volume measurement; magnetic resonance (MR), treatment planning; brain, volume; brain, white matter

A COMPREHENSIVE METHOD for measuring lesion volumes on serial brain MRI studies of multiple sclerosis (MS) patients is presented in this same issue (1). Here we present a systematic evaluation of the reproducibility and consistency of the entire procedure and of its individual components.

[1]Department of Radiology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115.

[2]Department of Neurology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115.

Contract grant sponsor: National Institutes of Health; contract grant numbers: N01-NS-0–2397, IP41RR13218-01, and P01 CA67165; contract grant sponsor: National Multiple Sclerosis Society; contract grant number: RG 2318-A-1; contract grant sponsor: Pharmacia & Upjohn; contract grant sponsor: Swiss National Science Foundation.

*Address reprint requests to: C.R.G.G., Department of Radiology, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, MA 02115. E-mail: guttmann@bwh.harvard.edu

Presented in part at the 1997 ISMRM meeting.

Received August 18, 1997; Accepted October 7, 1998.

Assessment of the accuracy of volumetric measurements of MS lesions on MRI is impeded by the difficulty of having a reference measurement of the "ground-truth" from independent measurements, such as the identification of lesions on corresponding histological sections of the imaged brains. Therefore, at the present stage, the major requirement for quantitative assessment of lesion load progression in MS is measurement reproducibility. No phantom is able to reproduce satisfactorily the complexity of MS lesions and white matter in composition, relaxation characteristics, shape, and spatial distribution. We therefore elected to perform MRI repeatedly on each of 20 MS patients within very short time intervals, during which lesion volumes were not expected to change. Statistical analysis of measurements obtained by applying various combinations of the image processing chain was performed to assess strengths and limitations of this approach to lesion burden estimation.

MATERIALS AND METHODS

Patient Selection and Positioning in the MR Scanner

Twenty patients suffering from clinically definite MS who were enrolled in a randomized double-blind placebo-controlled therapeutic trial of roquinimex (quinoline-3-carboxamide; Linomide, Pharmacia & Upjohn, Kalamazoo, MI) were selected for this evaluation (2). Brain MR images of each patient were acquired twice within 30 minutes. Between the two imaging sessions, the patients exited and re-entered the scanner room. At each time the patient's head was positioned in a neutral position (subjectively comfortable for the patient) the same way this had been performed during the course of a 1-year follow-up study of 46 untreated patients as well as during the ongoing MRI monitoring of the therapeutic trial (2,3). Two radiology technologists alternated in positioning the patient, to minimize bias. Foam pads and tape were the sole devices used to stabilize the head within the standard head-holder in the quadrature head coil.

For one of the patients in this study, a 36-year-old woman, we analyzed 27 MRI studies that had been acquired over a 3-year period. This patient had com-

pleted a monthly follow-up study over 1 year before enrolling in the ongoing randomized double-blind therapeutic trial. Twelve retrospective time-points were available for this patient from this study. Nine further MRI studies of this patient were acquired in the course of the ongoing drug trial. Six additional data sets were acquired from this patient within a 2 hour interval by two radiology technologists who alternated in positioning this patient in the scanner, three times each, following the procedure described above. Salient events in this patient's clinical history were recorded and are described in the results section.

MR Image Acquisition

MR images were acquired on a 1.5 T system (Signa, GE Medical Systems, Milwaukee, Wisconsin) using product pulse sequences and options. At each time-point a T1-weighted spin-echo sagittal localizer with 600 ms / 19 ms / 1 (repetition time / echo time / excitations) was followed by two interleaved dual-echo spin-echo (variable-echo multiplanar) sequences covering the brain with contiguous 3 mm slices from the foramen magnum to the upper convexity. The pulsing parameters for the latter sequence were 3000 ms / 30 ms; 80 ms / 0.5 (repetition time / echo times / excitations), and the nominal in-plane voxel size was 0.94 mm × 0.94 mm [24 cm field of view (FOV) with a 256 x 192 acquisition matrix]. Variable receiver bandwidth was used for the first (10.67 kHz) and second (4.27 kHz) echos. Standard flow compensation was achieved using first-order gradient moment nulling. Spins were saturated in an 8-cm-thick slab inferior to the imaging volume in the slice-select direction. No per-study shimming was performed. Scan duration was kept at 11 minutes and 36 seconds by using the half-Fourier technique.

Image Processing

The elements of the image analysis pipeline described in detail in the accompanying paper (1) were applied to the dual-echo MR images in various combinations, which are specifically noted in the results section. The individual elements systematically evaluated in this work were:

- Interactive segmentation of the intracranial cavity (ICC) (4).
- Anisotropic diffusion filter (5)
- Partial volume correction (1)
- Patient position
- Automated image registration (6)
- Fully automatic image segmentation into four tissue classes: gray matter, white matter, lesions, and cerebrospinal fluid (CSF) (7)

Outcome Measures and Statistical Analysis

The reproducibility of the various image processing strategies was calculated using statistical measures of variability of the measured lesion volumes. Lesion volumes were transformed from pixels to cm^3 by applying the equation:

$$\text{lesion volume} = \text{number of pixels} \times (\text{FOV} / \text{matrix size})^2 \times \text{slice thickness} = 0.00264 \times \text{number of pixels}$$

Interrater, intrarater, and method-related agreements were assessed by applying the statistical approach described by Bland and Altman (8). In essence, the agreement of two sets of measurements of a sample of a target population is assessed by relating the measurement difference in each subject to the average of the two measurements. The bias is defined as the average of the differences over the sample. Limits of agreement are the bias ±2 standard deviations (SD) of the differences. The 95% confidence intervals were computed as described in Bland and Altman (8)

Coefficients of variability were calculated as the ratio between standard deviation and mean (expressed as percent of the mean) to ensure proper comparison of our results with the work reported and discussed by Jackson et al (9).

Analysis of variance (ANOVA) and the Pearson's correlation coefficient were calculated using commercially available software routines (10).

Semi-quantitative evaluation of artifacts due to bulk movements of the head during MRI was performed by one neuroradiologist: one image at the level of the lateral ventricles was thresholded, to visualize the typical artifact patterns in the phase-encoding direction. The degree of motion artifact was graded into three classes (none / moderate / severe).

RESULTS

Interactive Segmentation of the Intracranial Cavity

Two operators independently identified the ICC in the first and second data sets, for each of the 20 patients ($N = 40$). Both operators repeated this operation 2 or more months later on both data sets from each patient to assess intrarater variability. Fully automatic tissue segmentation was repeated on each of the exams by masking the data with each ICC mask, obtained by the two independent operators on two different occasions. Averages, SDs, and ranges are summarized for total ICC volume and for total lesion volume in Table 1. Lesion volumes were compared after correcting for partial volume artifacts between brain or bone and CSF using the sequence of morphometric operators as described in the companion paper (1).

The average intrarater coefficient of variability of total ICC volume over the 40 MR scans was 0.55% and 0.47% for operators A and B, respectively. The average intrarater coefficient of variability of total lesion volume over the 40 MR scans was 1.87% and 0.93% for operators A and B, respectively. The mean measurement for each operator was used to compute interrater coefficients of variability for each of the 40 MR exams. The average interrater coefficient of variability for total ICC volume was 0.002%. The average interrater coefficient of variability for total lesion volume was 0.03%.

Table 1
Reproducibility of Volumetric Measurements of the Intracranial Cavity and Total Lesion Volume

	ICC volume		Total lesion volume	
	First MRI	Second MRI	First MRI	Second MRI
Operator A (first measurement) ($n = 20$)	1438.4 ± 135.7 (1228.2–1673.0)	1430.2 ± 134.3 (1233.6–1689.6)	8.53 ± 4.28 (3.02–20.00)	8.58 ± 4.26 (3.23–19.24)
Operator A (second measurement) ($n = 20$)	1436.9 ± 129.0 (1241.4–1649.7)	1430.2 ± 129.9 (1227.1–1645.4)	8.44 ± 4.36 (3.06–20.11)	8.54 ± 4.28 (3.17–19.23)
Operator B (first measurement) ($n = 20$)	1430.5 ± 134.0 (1219.8–1662.4)	1432.2 ± 132.2 (1229.9–1664.0)	8.51 ± 4.28 (2.95–20.00)	8.64 ± 4.32 (3.23–19.41)
Operator B (second measurement) ($n = 20$)	1435.8 ± 131.8 (1236.7–1665.4)	1436.0 ± 128.7 (1241.5–1650.3)	8.57 ± 4.30 (2.92–20.16)	8.68 ± 4.28 (3.28–19.41)

Data are in cm^3, average ± SD, with range in parentheses.

Three-way ANOVA compared the variability of ICC and total lesion volumes within the patient sample with variability due to operator interaction and repeated imaging procedures. The three factors evaluated in this statistical model were the MRI procedure (repeated scan), image analysis by different operators (interrater variability), and image analysis repeated by the same operator (intrarater variability). Eight measurements for each of the 20 patients were included for this purpose: each of the two measurements performed by each of two operators on each of two MRI exams from each patient. The two outcome volumes, the ICC volume and total lesion volume, were evaluated separately. None of the three factors nor any of the interaction terms between them demonstrated a significant effect on the variability of the outcome measures. This indicates that the variance due to operator interaction and repeated MRI measurement was insignificant compared with the variance of lesion volumes within the MS patient population, exemplified by these 20 subjects. F-ratios were smaller than 0.05 in all cases, and P values ranged between 0.8417 and 0.9996.

Anisotropic Diffusion Filter

Fully automatic tissue segmentation within the ICC (interactively segmented by one operator) was performed on the original as well as on the anisotropically filtered dual-echo MR images. The anisotropic diffusion filter (5) was applied by consistently setting the filtering parameter κ to 7 and the number of iterations to 3. Partial volume correction was applied for the brain/CSF and bone/CSF interfaces, but no image registration was performed.

The measurements of total lesion volume calculated from filtered and unfiltered images revealed a bias (average difference between lesion volumes calculated with and without filtering) of -0.1 cm^3; limits of agreement (bias ± 2 SD) were -0.6 and 0.4 cm^3, respectively (Fig. 1A). With the exception of four outliers, none of which showed abnormal levels of motion artifact, the difference in filtered and unfiltered lesion volume was less than 3% of the average between the two measurements. The maximum discrepancy was 1.5 cm^3 or 9.3% of the mean between the two measurements for that patient.

One set of images was repeatedly filtered, each time using a different filtering strength by varying the value of κ to illustrate the influence of "noise" reduction on lesion quantitation (Fig. 1B). Lesion segmentation was applied to each filtered set of images as well as directly to the unfiltered data. Total lesion volume was calculated before and after applying brain/CSF and bone/CSF partial volume correction (Fig. 1B). Total lesion volume measurements appeared to be largely independent of the filter strength for $\kappa \leq 17$, when correcting for partial volume effects. Without the correction for partial volume artifacts, lesion volume estimates became smaller as stronger anisotropic diffusion filtering was applied to the original MR images.

Partial Volume Correction

On average over the 40 MRI exams, the partial volume correction step at the boundaries between brain parenchyma (or skull bone) and CSF reduced the estimate of total lesion volume by 84.8% (± 4.3% SD) of the lesion volume estimated without this partial volume correction step. Before partial volume correction, the average difference in total lesion volumes between the two studies of each patient was -0.5 cm^3 and the limits of agreement (mean ± 2 SD) were -6.6 and 5.6 cm^3. After elimination of "false-positive" lesions through the partial volume correction, the corresponding average difference and the limits of agreement were 0.1 cm^3 and $-1.3/1.5$ cm^3, respectively. In percentage of the mean measurement, the bias for uncorrected lesion volumes was -0.5% (1.4% after correction) and the limits of agreement were -9.9% and 9.0% (-12.5% and 15.4% after correction). The average unsigned difference (average of the absolute value of the difference) between two studies in the same patient was 2.0 cm^3 (range: 0.03–10.7 cm^3; SD: 2.4 cm^3) before partial volume correction and 0.5 cm^3 (range: 0.04–1.7 cm^3; SD: 0.5 cm^3) after partial volume correction. As a percent of the mean between each pair of studies the average unsigned differences were 3.5% (range: 0.03–11.8%; SD: 3.1%) before partial volume correction and 5.8% (range: 0.04–11.6%; SD: 3.9%) after partial volume correction. The Pearson's correlation coefficient between the interstudy differences calculated with or without partial voluming correction was $r = 0.55$.

The estimated lesion volumes used to evaluate measurement agreement between two independently obtained sets of MR images were the averages of four

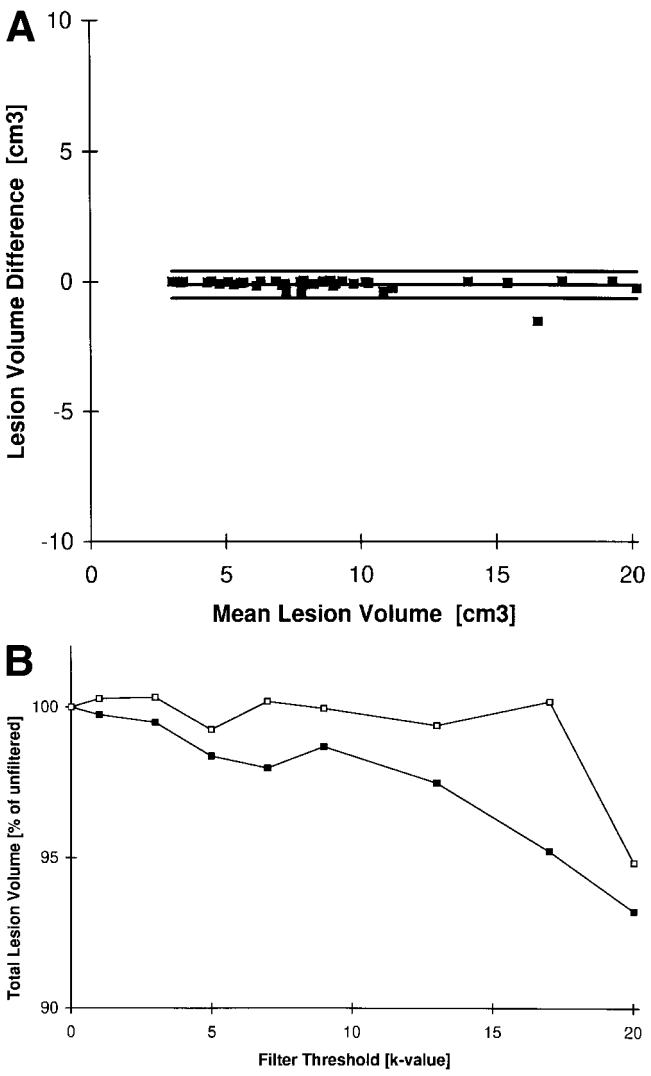

Figure 1. Noise reduction using anisotropic diffusion filtering: effects on the automated quantitation of lesion volumes. **A:** Automated segmentation of lesion volumes in 40 MRI studies (2 for each of 20 MS patients) was performed with and without applying the anisotropic diffusion filter ($\kappa = 7$, 3 iterations) to the gray-scale PDw and T2w images. The differences between the two measurements are plotted against their average to illustrate the agreement between them. The bias (mean difference) and the limits of agreement (mean ± 2 SD) are shown by the continuous lines. Top line = mean +2 SD, middle line = mean, bottom line = mean –2 SD. **B:** Repeated automated segmentation of total lesion volume from one set of dual-echo MR brain images of an MS patient. Each measurement was obtained with a different value for the filtering parameter κ (three iterations for all measurements). The normalized total lesion volumes were calculated before (filled boxes) and after (open boxes) applying the brain/CSF and bone/CSF partial voluming (PV) correction step. Total lesion volumes are normalized to the respective values for the unfiltered images.

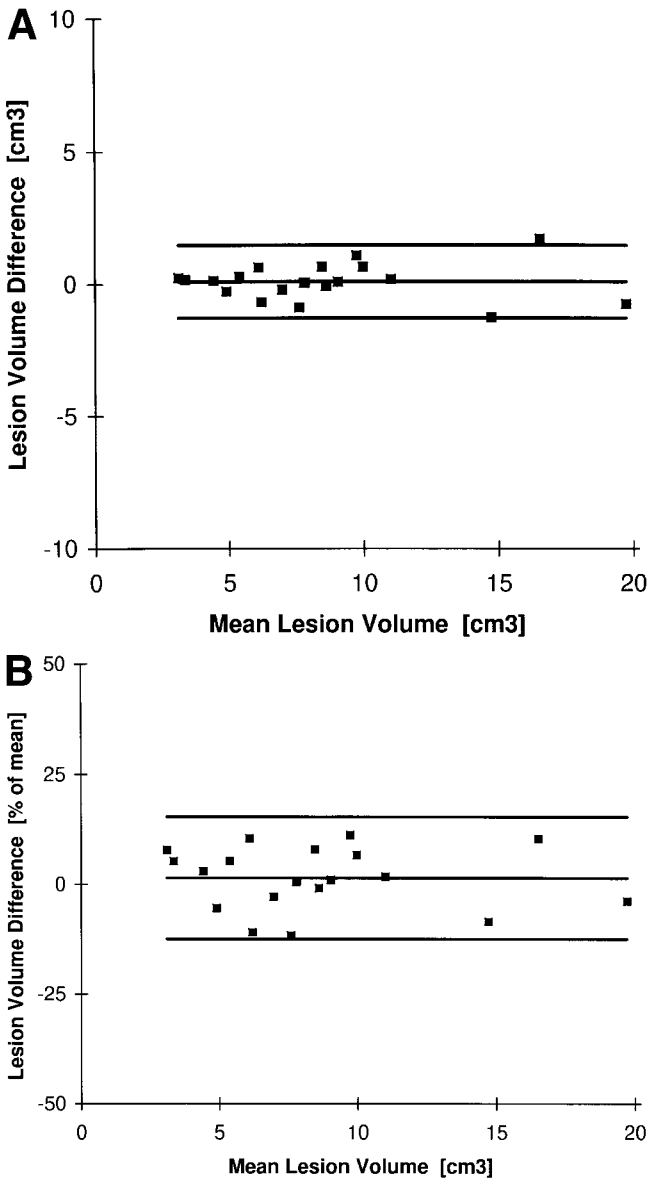

Figure 2. Patient repositioning and repeated MR imaging: effects on the automated quantitation of lesion volumes. Automated lesion segmentation was performed on two MR exams each for 20 patients. Anisotropic diffusion threshold $\kappa = 7$, brain/CSF and bone/CSF partial voluming (PV) correction step was applied, and no image registration was performed. **A:** The differences between the two lesion volume measurements for each patient are plotted against their means. The measurement difference appears to increase with larger lesion load. **B:** The differences are expressed as percent of the mean of the two measurements. The percent change appears to be uniform over the whole range of lesion loads. Top lines = mean +2 SD, middle lines = mean, bottom lines = mean –2 SD.

values (two from each of two operators), to minimize the influence of intra- and interoperator differences.

Patient Position

Three-way ANOVA demonstrated no significant variance of total lesion volume between the first and second MR exams in the 20 patients examined. The pairwise differences in lesion volume estimates are summarized in the preceding section and are illustrated in Fig. 2.

The bias was 0.1 cm^3 and the limits of agreement were -1.3 and 1.5 cm^3 (the 95% confidence interval for the bias is -0.12–0.32 cm^3). Expressed as percentage of the mean measurement, the bias was 1.4% with limits of agreement of -12.5% and 15.4%.The maximum mea-

surement difference was 1.7 cm^3 (average ± SD: 0.5 ± 0.5 cm^3) or 11.6% (average ± SD: 5.8 ± 3.9%; median: 5.4%) of the average (estimated) lesion volume in that patient. The measurement difference appeared to grow with larger lesion load (Fig. 2).

A neuroradiologist (F.A.J.), who was blinded to the results of the computerized volumetric evaluation, assessed in a semi-quantitative fashion the extent of motion artifacts in each set of images. Five of the 20 patients showed some degree of motion artifact in the second set of images (possibly due to the lengthiness of the whole experiment). Only one of these patients demonstrated some motion artifact also on the first data set, although the degree of motion increased in the second set of images also in this patient. Only two patients demonstrated strong motion artifact in the second set of images compared with their first set. One of these two patients was the one with the biggest difference in automated lesion volume measurement: the scan with the motion artifacts yielded 1.7 cm^3 (11.6%) more lesion volume than the previous scan.

Automated Image Registration

The effect of data registration on lesion volume measurements was evaluated by using each patient's second set of gray-scale images in filtered ($\kappa = 7$) form as a starting point. The 20 data sets were repositioned and resampled ("resliced") to match automatically the three-dimensional (3D) surface of their respective ICCs to the reference ICC (patients' first MR image sets) (6). Following this step, automatic segmentation and partial volume correction were performed in this order. Lesion volumes were compared with the corresponding measurements from the identical analysis sequence with the sole omission of the registration step (Fig. 3).

On average, measurements of total lesion volumes after image reformatting were 1.34 cm^3 smaller than the corresponding values for the original images. The measurement difference appeared to be smaller for larger lesion loads: a mean of 1.7 cm^3 for estimated lesion loads <10 cm^3 compared with 0.3 cm^3 for estimated lesion loads ≥ 10 cm^3 . This difference was significant with $P = 0.0004$ (one-tailed Student's *t*-test). The significant effect of image registration on the assessment of small lesion loads is evident in Fig. 3B.

Consistency of the Overall Procedure

In the first part of our analysis, to isolate contributions of each step in the analysis, we averaged lesion volumes obtained by two operators at two different times. In real-life application of the tools presented here, lesion volumes are likely to come from measurements obtained by using one operator for each time point. This second part of the analysis tested the overall reproducibility of patient positioning, imaging, and image processing by considering only one measurement per set of MR images.

For each time point, a value was picked randomly from the four values available to simulate the randomness of operators analyzing studies in a long-term follow-up of MS patients. No registration was performed

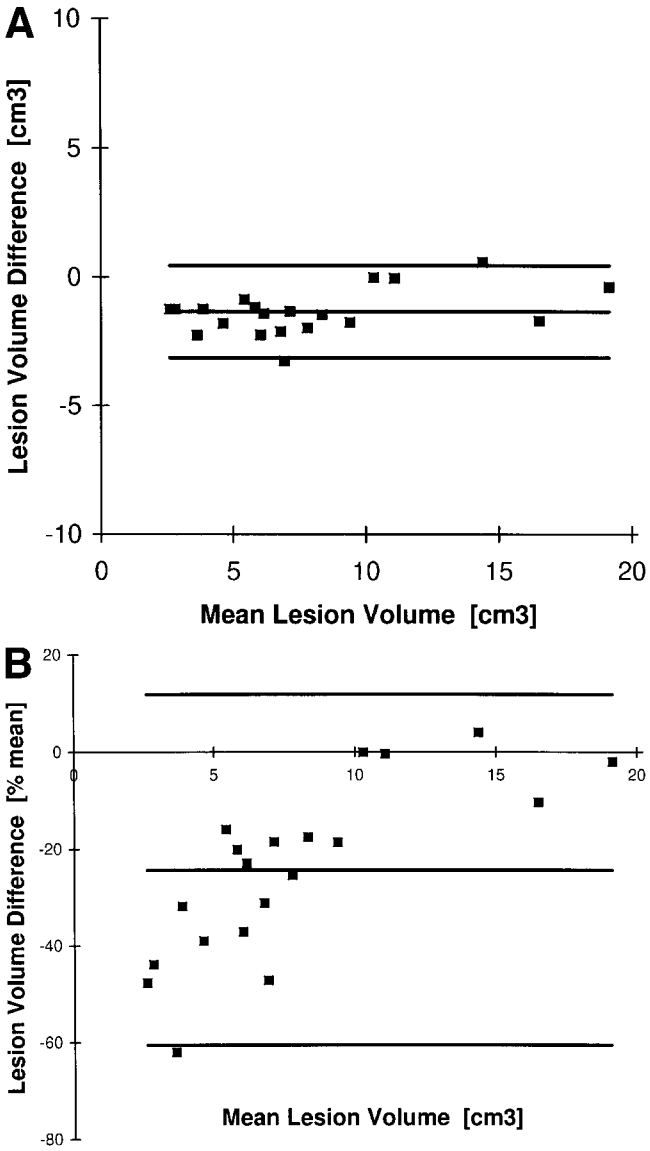

Figure 3. Image registration and lesion volume measurement. The images from the second MR exam of each patient were resampled to match those of the patient's first MR exam (registration). Lesion volumes were segmented from the second exam before and after the registration step. **A:** Agreement between lesion volumes measured on the second MR exam with and without registration ($n = 20$). The volumes obtained by image segmentation after registration are subtracted from those obtained without image repositioning. Anisotropic diffusion threshold $\kappa = 7$; brain/CSF and bone/CSF partial voluming (PV) correction step was applied. **B:** As in A but lesion volume differences are expressed as percent of the mean of the two measurements (with and without registration) for each MR exam. Top (thicker) lines = mean +2 SD, middle lines = mean, bottom lines = mean –2 SD.

to avoid the associated bias. Image registration is not required when the goal is to measure total lesion volume and tracking of individual lesions is not the aim. Calculated this way, the average difference (bias) between each patient's two measurements was 0.05 cm^3, and the limits of agreement were -1.35 and 1.46 cm^3 (expressed as percentage of the mean, these values were 0.86%,

−14.24%, and 15.95%, respectively). The 95% confidence interval of the bias was between −0.17 and 0.28 cm^3 (or −1.56–3.27% of the mean estimate of total lesion volume).

A patient who was imaged 27 times in the course of 3 years, including six repeated measurements within 2 hours, exemplifies our findings (Fig. 4). These results adequately summarize the strengths and limitations of the current implementation of our lesion burden estimation method. The patient was a 36-year-old woman with relapsing-remitting progressive (secondary progressive) MS. The patient had been clinically stable and attack free during the 18 months prior to enrollment. The first 12 MR studies (monthly studies between day 0 and day 350) revealed little fluctuation in total lesion volume. However, 10 months later on day 651 a new measurement demonstrated what appeared to be an inconsistent, if limited, decline in lesion volume of approximately 3 cm^3. This measurement was confirmed 1 month later (day 681), before an apparent rise in lesion volume yet another month later (day 707). We fortuitously performed six consecutive measures on day 709, only 2 days after this apparent activity, and found measurements consistent with that measured 1 month earlier, on day 681. The outlier was the only measurement that had been performed on a different (but technically identical) machine (1.5 T GE Signa MR imager) in our service (Fig. 4).

In Fig. 4B the lesion volumes are normalized by the measurement of total white matter (including abnormal white matter, ie, lesions). The artifactual fluctuations described above appear to be largely corrected, while the other, presumably disease-related, fluctuations are maintained.

No clinically defined attack was documented in the course of the initial 1-year follow-up, although internuclear ophthalmoplegia was detected 1 month after the clinical enrollment examination, at the first timepoint in the MR study (day 0), at which a relatively high lesion volume was also detected. No clinical attack was noted in the period between the natural course study and enrollment in the subsequent drug study (day 350

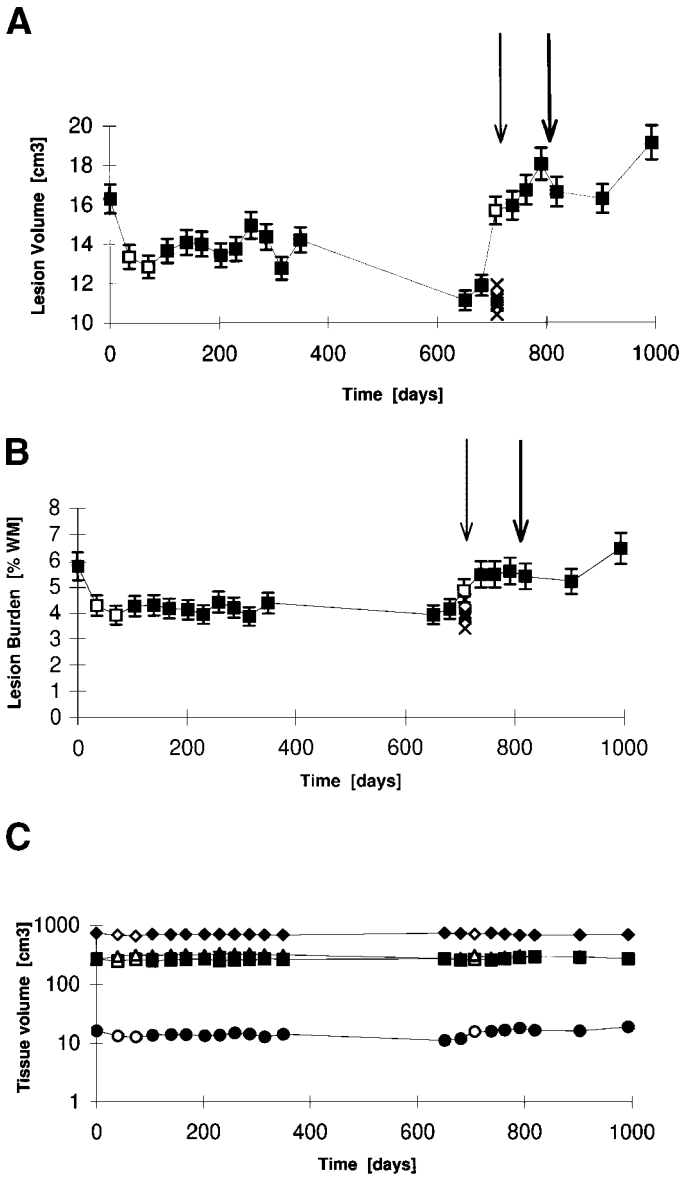

Figure 4. Serial lesion volume measurements in a relapsing remitting progressive (secondary progressive) MS patient. A 36-year-old woman was followed over 3 years at up to monthly intervals. MR imaging and lesion volume quantitation were performed 27 times. Anisotropic diffusion threshold $\kappa = 7$; brain/CSF and bone/CSF partial voluming (PV) correction step was applied; no image registration was performed. During follow-up only one attack (exacerbation), characterized by weakness in the left lower extremity, was clinically diagnosed (thick arrow). Approximately 2 months prior to this attack, the patient reported flu-like symptoms, including fever, fatigue, headache and gastrointestinal symptoms (thin arrow). The exacerbation was treated for 8 days (from day 809 to day 813 and from day 825 to day 827) with a daily oral dose of 8 mg dexamethasone (Decadron, Merck, West Point, PA). Open symbols indicate MR exams performed on a different imager of the same type as that used for the filled symbols. Crosses indicate the six lesion volume measurements performed on this patient within 2 hours. These values were used to calculate the coefficient of variation (CV = standard deviation/average) that was applied to each measurement (error bars = ±CV × measurement). **A:** Lesion volumes. **B:** Lesion load normalized for total white matter: lesion burden = 100 x lesions/(white matter + lesions). **C:** Brain tissue volumes. Diamonds = gray matter, squares = cerebrospinal fluid (csf), triangles = white matter, circles = lesions.

to day 651), although elements of disease progression were evident [the patient started using a cane and her expanded disability status scale (EDSS) score worsened from 3.5 to 6]. During the Linomide trial (day 651 through day 993), the patient received a placebo and experienced a clinically well-documented exacerbation that was treated with corticosteroids (see Fig. 4 for details).

A statistically significant growth in lesion burden occurred approximately at the time the patient reported flu-like symptoms and appeared to peak at the time of the documented exacerbation, some 2 months later on day 791 (Fig. 4A,B). Regression in lesion volume was observed during corticosteroid treatment (Fig. 4A,B). In Fig. 4C we present the time course for the other tissue compartments that were measured: white matter, gray matter, and cerebrospinal fluid. A detailed analysis of measurement reproducibility for these compartments transcends the scope of this work and is presented elsewhere (11).

DISCUSSION

The usefulness of lesion burden measurements as surrogate markers of disease activity and disability hinges on the biological significance of such a measure as well as on its reproducibility. Several studies, including a preliminary analysis of a longitudinal MR study of the natural course of MS with the methods presented in this paper, indicate that changes in MR-based lesion load measurements correlate with changes in clinical outcome measures, such as the EDSS and the number or frequency of attacks (12,13). Given that the measured volume changes appear to be relevant for estimating disease progression, the sensitivity of such measurements needs to be established.

We set out to assess the reproducibility of our measurements of lesion volume. Because of the inherent difficulty in determining these measures by independent methods, such as histopathologic techniques, we were unable to assess the specificity of our method, which is a function of the complex interplay among such diverse factors as tissue composition, MR physics, and image processing. The measured lesion volumes are therefore not to be considered accurate in the histological sense. It is the changes in these measurements that are expected to yield information about disease progression in groups of subjects and eventually in individual patients. Our study design reflects this hypothesis and rests on the assumption that no disease-related change occurred in the 30 minutes between each patient's two MRI sessions.

Interactive Segmentation of the Intracranial Cavity

Identification of the ICC is the only step in our image analysis procedure that required operator interaction, however limited. Our results here and elsewhere (4) emphasize the high reproducibility in determining total volume of the ICC. However, this measurement has little significance in the follow-up of most brain pathologies and is mostly useful as a normalizing factor, similar to more traditional measures of head size (14). We directed our attention to this step's influence on the variability of total lesion volume, which is, typically, the measure of primary interest in determining MS progression. We conclude from this study that computer-assisted definition of the ICC by a human operator had negligible influence on the measurement of lesion volume.

Anisotropic Diffusion Filter

Image filtering using the concept of anisotropic diffusion was proposed by Perona and Malik (15) and implemented for the treatment of MR images by Gerig et al (5), with the goal of reducing the noise level while maintaining the boundaries of significant structures in the images. Since then several groups have applied this pre-processing step (9,16–18). To our knowledge, only Mitchell et al (16) have systematically evaluated its usefulness for the measurement of lesion volumes in MS. These authors evaluated the influence of anisotropic filtering on the results of manual contouring of five individual lesions seen on MR images of a single MS patient. They concluded that the filtering significantly reduced intra- and interrater variability in images acquired at 1.5 T with an MR protocol (dual-echo spin echo with TR/TE1,TE2 2800 ms/30 ms, 80 ms, variable bandwidth, 5 mm slice thickness) and filtering parameters ($\kappa = 6$ and 3 iterations) very similar to those used in our study.

In contrast, we evaluated the effect of anisotropic filtering on total lesion volumes estimated by automatic tissue segmentation (Fig. 1). The results of our evaluation suggest that filtering of the images reduces the number of pixels classified as lesions by the automatic tissue segmentation. Lesion volumes steadily diminish with growing κ-value (Fig. 1B). When correcting for brain/CSF and bone/CSF partial voluming, however, this dependency of lesion volume from filter strength disappears for the range of κ-values that would be considered of potential use. The anisotropic diffusion filter and the partial volume correction thus appear to have similar targets (ie, partially volumed pixels at the interface between CSF and brain parenchyma or bone). In conclusion, the anisotropic diffusion filter did not provide any advantage in the context of this automated segmentation of total lesion volume and is not needed for this purpose.

Partial Volume Correction

Similar post-processing approaches to correct for misclassified pixels (false-positive lesions) using a priori anatomical knowledge have been previously proposed and successfully applied also by Zijdenbos et al (18) and Kapouleas et al (19). Correcting for artifacts due to partial voluming between brain parenchyma (or bone) and CSF reduced our estimate of total brain lesions by over 80% by accounting for what are with very high likelihood false-positive MS abnormalities. This reduction in false-positive lesion classifications is comparable to that recently found by Kamber et al (20) by repositioning patient brain MR images to a 3D brain tissue probability model in Talairach's stereotactic space. Warf-

ield et al (21) have used an elastic matching approach to create an explicit anatomical model for each patient, thereby also largely avoiding this type of error. In contrast to both these methods, however, our method does not assume an explicit anatomical model of an average human brain and requires far less computation.

Assuming that the rate of change in MS lesion load is independent of initial lesion volume, we are most interested in the precision of absolute differences. As evidenced by the approximately fourfold improvement in the limits of agreement, the lesion volume estimate after partial volume correction is more precise (in absolute terms) and therefore can be expected to be more sensitive to real changes in lesion load. Our poor correlation between lesion volume differences calculated with and without partial volume correction probably indicates that the false-positive lesions confound the results by adding largely uncorrelated noise to the measurement of total lesion burden. We therefore believe that this step greatly enhances the specificity and sensitivity of our method to real changes in lesion burden and that such correction will be needed by any lesion segmentation algorithm relying exclusively on multispectral information from conventional long-TR dual-echo spin-echo sequences.

Patient Position and Reproducibility of Data Acquisition

During image acquisition several variables can limit the reproducibility of the evaluation of lesion load. Among them are the intrinsic reproducibility of MR measurements, patient position with respect to the reference frame of the imager (interscan displacement), and patient motion during the MR examination (intrascan motion).

We analyzed the effect of interscan variations in patient position by simulating conditions of clinical trials: we asked each patient to leave the scanner room between their serial exams and required different radiological technologists to alternate in positioning the patient. We found a 5.4% median difference (in percent of the average of the two volume estimates) between lesion volumes derived from each patient's scan pair. This measure can be compared with the 9.9% median reported by Gawne-Cain et al (22) in a similar analysis for the worst case repositioning in each of their five MS patients.

However, our finding of a very low lesion volume bias relative to patient position (1.4% of the mean volume) indicates that artifactual changes in lesion volume related to random patient positioning tend to cancel out over a patient cohort (regression to the mean), while this is not expected to be the case for lesion accumulation due to disease progression. This bias with its 95% confidence interval should ensure that our method detects the recently reported median yearly changes in MS lesion volume of between 5% and 10%, even without accurate patient repositioning (23). Nevertheless, accurate patient repositioning should only improve matters and we have recently developed a non-invasive stereotactic method for this purpose (24).

Our method appears to be fairly resistant to intrascan patient motion. When our experienced neuroradiologist judged such motion to be extreme, however, the error in our automated lesion volume estimation appeared significant and could mimic a real change in lesion burden. This finding points to the need to establish quality standards for the inclusion of MRI data in clinical studies that take into account the maximum degree of acceptable motion artifact in MR images.

This reliability study was not designed to evaluate systematically the effects of long-term fluctuations in the response function of MR imagers on lesion burden estimates. These fluctuations are due to the instability of the main magnetic field B_0 as well as to other hardware-related variations. We have purposely evidenced such effects in the 3-year follow-up of one of our patients. Figure 4 illustrates that apparently significant changes in lesion load can be ascribed to hardware changes, although the evidence we present here is anecdotal. To address this issue more specifically a separate study would be required, eg, one in which patients would be scanned on several imagers in the same day. We thought it important, especially for the correct interpretation of multicenter clinical trials involving MRI measurements, to emphasize this aspect, in spite of its limited evaluation in the present study. Use of internal normalization of lesion volumes to the estimated white matter volume may be a way to identify spurious lesion volume fluctuations. Further evaluation of this promising approach is needed.

Automated Image Registration

Correcting for imperfect patient repositioning within the MR imager is only necessary when following events, in this case focal tissue abnormalities (lesions), at specific anatomic locations. This procedure is not required if the sole concern is follow-up of total lesion volume within the brain as the whole brain is covered by the MR protocol we used. Our study addressed the effect of image registration, and in particular the interpolation step necessary for image reformatting, on the quantitation of lesion volume.

Our data indicate that the image resampling procedure used for image registration (ie, trilinear interpolation of the gray scale values in the original images), significantly corrupted the measure of lesion volume, especially when the lesion burden was small (Fig. 3). The consistent reduction in measured lesion volume on interpolated images was probably due to synthetic partial voluming of the small lesions with the surrounding tissue. This blurring of the lesions decreases the sensitivity in their detection with the automatic segmenter. This measurement error was not uniformly distributed over the range of estimated lesion volumes and appeared to be especially noteworthy for cases with relatively small lesion loads. This type of error should be evaluated in lesion burden segmentation approaches that apply image resampling before lesion segmentation (20,25).

Overall Performance

The reproducibility characteristics reported here appear to be comparable to or better than those of previously proposed segmentation techniques (9,17,25–32). To compare our work with a detailed reproducibility analysis of supervised segmentation of MS lesions using multispectral cluster analysis, we calculated coefficients of variability according to that work's specifications (9). Our values were approximately 20 times lower. Our average intrarater coefficient of variation was 1.4% vs. their 20.3%; our interrater coefficient of variation was 1.06% vs. their 21.2%. Similarly poor values were reported for comparable methods by Herman (28) and Wang et al (17).

Filippi et al (31) compared a semiautomated threshold-based method (SAT) with an arbitrary lesion scoring system (ASS) and a manual lesion tracing method (MTM). After analyzing intra- and interrater reproducibility, they concluded that the SAT technique had the best reproducibility characteristics (31). When calculating the agreement index between two volumes, as these authors proposed

$$\text{agreement index} = 1 - |\text{volume 1} - \text{volume 2}| / (\text{volume 1} + \text{volume 2})/2$$

we obtained the following median values for our data. The median intraobserver agreement index was 99.1% [40 studies measured twice by each of two operators ($N = 80$)] compared with their 96.3% (with independent choice of threshold among raters). Our median interobserver agreement index was 98.8% [two measurements of two studies per patient by one operator compared with the corresponding measurements by the other operator ($N = 160$)] ,whereas their index was 93.7% (with independent choice of threshold among raters). A median interstudy agreement index created by comparing the volumes of the two MR exams from each patient and randomly selecting only one measurement for each exam was 94.9%. Unfortunately, Filippi et al (31) did not present data on the reproducibility of the SAT on repeated MRI of the same patients. Another recent study of measurement variability of MS lesion volumes assessed the fully manual and computer-assisted outlining of five MS lesions in one patient but did not report any results for total lesion volumes (30). These variability estimates therefore cannot be readily compared with ours.

Previous reports of reproducibility with a MS lesion quantitation method, in particular the very few repeating MRI exams on the same subjects, include few subjects. Thus they are likely to be less representative for the range of lesion patterns and heterogeneity present in the general MS population. Our study design reflects the desire to validate our methods on a sample somewhat representative of the range of lesion shapes and distributions (size, anatomical) encountered in the MS population. In contrast to most previous work, we also applied stringent criteria in acquiring repeated studies of the same patient within a very short time and controlling for the randomness of patient head positioning.

We calculated variability of lesion volume measurements by simulating a trial in which only one observer would evaluate each study. The finding of a sub-milliliter range for the 95% confidence interval (−0.17–0.28 cm^3) suggests that this method is well suited for evaluation of clinical trials, even those with the small patient cohorts typical of Phase II studies. The limits of agreement (−1.35 and 1.46 cm^3) indicate the precision that could be expected in the evaluation of lesion burden changes in serial exams of an individual MS patient. We estimate that this is within the range of the yearly rate of plaque growth in progressive MS patients. Therefore we believe that yearly follow-up with this technique has the potential to provide valuable information on disease progression and response to treatment in individual patients.

All measurements in this assessment of reproducibility were performed using the same MR system. It was not the goal of this work to assess the consistency of measurements across systems or when varying the response characteristics of a given imager. However, this is a relevant factor in long-term as well as in multicenter clinical trials. Our case study of an individual MS patient followed over 3 years underlines the need for further assessment of potential inconsistencies. We also found that assessing the ratio of abnormal to total white matter may counteract this artifact. It is encouraging that the observed changes in lesion burden appeared to be consistent with the clinical history in this patient.

The method presented has adequate overall reproducibility for the evaluation of MS progression in patient populations such as those typically needed for clinical drug trials. The next challenge, follow-up of individual patients in a clinical context, appears feasible but still elusive, in light of the precision of this method and the anecdotal data on one patient. Inconsistencies, presumably due to changes in MRI equipment characteristics, appeared to be successfully corrected, but further attention to this problem is needed.

ACKNOWLEDGMENTS

The authors thank Dr. Andre Robatino, Mr. Jan Horsky, and Mrs. Diane Doolin for technical assistance. Ms. Sandra Cook, Ms. Maureen Ainslie, Dr. Michael J. Olek, and Dr. Marika J. Hohol, as well as the rest of the clinical staff in the departments of Radiology and Neurology, were indispensable in recruiting the patients and performing the MR examinations and clinical assessments. We also thank Dr. Kenneth Jones for critical review of an early version of this manuscript and Ms. Sally Edwards for extensive editorial comments in the final redaction phase. We are deeply grateful to the 20 patients who participated in this study. C.R.G.G. was supported in part by a fellowship from the Swiss National Science Foundation; R.K. and F.A.J. acknowledge partial funding through the National Institutes of Health (grant P01 CA67165).

REFERENCES

1. Kikinis R, Guttmann CRG, Metcalf D, et al. Quantitative follow-up of patients with multiple sclerosis using MRI: technical aspects. J Magn Reson Imaging 1998;9:519–530.
2. Hohol M, Guttmann C, Olek M, et al. Roquinimex (Linomide®) treatment in secondary progressive and relapsing remitting multiple sclerosis: results from 48 week randomized, double-blind, placebo-controlled pilot studies. Neurology 1997;48:A174.
3. Guttmann CRG, Ahn SS, Hsu L, Kikinis R, Jolesz FA. The evolution of multiple sclerosis lesions on serial MR. AJNR 1995;16:1481–1491.
4. Kikinis R, Shenton ME, Jolesz FA, et al. Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging. J Magn Reson Imaging 1992; 2:619–629.
5. Gerig G, Kübler O, Kikinis R, Jolesz FA. Nonlinear anisotropic filtering of MRI data. IEEE Trans Med Imaging 1992;11:221–232.
6. Ettinger GJ, Grimson WEL, Lozano-Pérez T, Wells WM III, White SJ, Kikinis R. Automatic registration for multiple sclerosis change detection. In: Proceedings of the IEEE Workshop on Biomedical Image Analysis, Seattle, WA, June 1994. p 297–306. IEEE Computer Society Press, Los Alamitos, CA.
7. Wells WM III, Grimson WEL, Kikinis R, Jolesz FA. Adaptive segmentation of MRI data. IEEE Trans Med Imaging 1996;15:429–442.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–310.
9. Jackson EF, Narayana PA, Wolinsky JS, Doyle TJ. Accuracy and reproducibility in volumetric analysis of multiple sclerosis lesions. J Comput Assist Tomogr 1993;17:200–205.
10. Microsoft Excel version 5 user's guide. Redmond, WA: Microsoft Corporation; 1993. p 589–606.
11. Guttmann CRG, Jolesz FA, Kikinis R, et al. White matter changes with normal aging. Neurology 1998;50:972–978.
12. Guttmann CRG, Weiner HL, Kikinis R, et al. A prospective study of multiple sclerosis: correlation between MRI, clinical and immunological findings. Proc Soc Magn Reson 1995;2:1292.
13. Khoury SJ, Guttmann CRG, Orav EJ, et al. Longitudinal MRI in multiple sclerosis: correlation between disability and lesion burden. Neurology 1994; 44:2120–2124.
14. Matsumae M, Kikinis R, Mórocz IA, et al. Age-related changes in intracranial compartment volumes in normal adults assessed by magnetic resonance imaging. J Neurosurg 1996;84:982–991.
15. Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. IEEE Pattern Anal Machine Intell 1990;12:629–639.
16. Mitchell JR, Karlik SJ, Lee DH, Eliasziw M, Rice GP, Fenster A. Quantification of multiple sclerosis lesion volumes in 1.5 and 0.5 T anisotropically filtered and unfiltered exams. Med Phys 1996;23: 115–126.
17. Wang L, Thorpe JW, Tofts PS. Evaluation of cluster analysis in multiple sclerosis lesion segmentation. Proc Soc Magn Reson 1995; 2:709.
18. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 1994;13:716–724.
19. Kapouleas I. Automatic detection of white matter lesions in magnetic resonance brain images. Comput Methods Programs Biomed 1990;32:17–35.
20. Kamber M, Shinghal R, Collins DL, Francis GS, Evans AC. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. IEEE Trans Med Imaging 1995;14:442–453.
21. Warfield S, Dengler J, Zaers J, et al. Automatic identification of grey matter structures from MRI to improve the segmentation of white matter lesions. J Image Guided Surg 1996;1:326–338.
22. Gawne-Cain ML, Webb S, Tofts P, Miller DH. Lesion volume measurement in multiple sclerosis: how important is accurate repositioning? J Magn Reson Imaging 1996;6:705–713.
23. The IFNB Study Group, University of British Columbia MS/MRI Analysis Group. Interferon beta-1b in the treatment of multiple sclerosis: final outcome of the randomized, controlled trial. Neurology 1995;45:1277–1285.
24. Oshio K, Panych LP, Guttmann CRG. A simple noninvasive stereotactic device for routine MR head examinations. J Comput Assist Tomogr 1996; 20:588–591.
25. Zijdenbos A, Evans A, Riahi F, Sled J, Chui H-C, Kollokian V. Automatic quantification of multiple sclerosis lesion volume using stereotactic space. In: Proceedings of the Fourth International Conference on Visualization in Biomedical Computing (VBC), Hamburg, Germany, 1996.
26. Zijdenbos AP, Dawant BM. Brain segmentation and white matter lesion detection in MR images. Crit Rev Biomed Eng 1994;22:401–465.
27. Filippi M, Horsfield MA, Tofts PS, Barkhof F, Thompson AJ, Miller DH. Quantitative assessment of MRI lesion load in monitoring the evolution of multiple sclerosis. [Review] Brain 1995;118:1601–1612.
28. Herman GT. Quantitation using 3D images. In: Udupa JK, Herman GT, editors. 3D imaging in medicine. Boca Raton: CRC Press; 1991. p 145–161.
29. Pannizzo F, Stallmeyer MJ, Friedman J, et al. Quantitative MRI studies for assessment of multiple sclerosis. Magn Reson Med 1992;24:90–99.
30. Mitchell JR, Karlik SJ, Lee DH, Eliasziw M, Rice GP, Fenster A. The variability of manual and computer assisted quantification of multiple sclerosis lesion volumes. Med Phys 1996;23:85–97.
31. Filippi M, Horsfield MA, Bressi S, et al. Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis. A comparison of techniques. Brain 1995;118:1593–1600.
32. Stone LA, Albert PS, Smith ME, et al. Changes in the amount of diseased white matter over time in patients with relapsing-remitting multiple sclerosis. Neurology 1995;45:1808–1814.