Multiple Sclerosis 2004; **10:** 402–406
www.multiplesclerosisjournal.com

Has your patient's multiple sclerosis lesion burden or brain atrophy actually changed?

Xingchang Wei[1], Charles RG Guttmann[2], Simon K Warfield[3], Michael Eliasziw[4] and J Ross Mitchell[*,1]*

[1]Seaman Family MR Research Center, Departments of Radiology and Clinical Neurosciences, University of Calgary, Calgary, AB, Canada; [2]Center for Neurological Imaging, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [3]Computational Radiology Laboratory, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [4]Departments of Community Health Sciences and Clinical Neurosciences, University of Calgary, Calgary, AB, Canada

Changes in mean magnetic resonance imaging (MRI)-derived measurements between patient groups are often used to determine outcomes in therapeutic trials and other longitudinal studies of multiple sclerosis (MS). However, in day-to-day clinical practice the changes within individual *patients may also be of interest. In this paper, we estimated the measurement error of an automated brain tissue quantification algorithm and determined the thresholds for statistically significant change of MRI-derived T2 lesion volume and brain atrophy in individual patients. Twenty patients with MS were scanned twice within 30 min. Brain tissue volumes were measured using the computer algorithm. Brain atrophy was estimated by calculation of brain parenchymal fraction. The threshold of change between repeated scans that represented statistically significant change beyond measurement error with 95% certainty was 0.65 mL for T2 lesion burden and 0.0056 for brain parenchymal fraction. Changes in lesion burden and brain atrophy below these thresholds can be safely (with 95% certainty) explained by measurement variability alone. These values provide clinical neurologists with a useful reference to interpret MRI-derived measures in individual patients.*
Multiple Sclerosis *(2004)* **10,** 402–406

Key words: *brain atrophy; computer-assisted; image processing, magnetic resonance imaging; multiple sclerosis*

Introduction

Interactive as well as automated computerized techniques have been developed to quantify pathologies that appear on magnetic resonance (MR) images in patients with multiple sclerosis (MS). Of the many measurements these techniques can provide, T2 lesion volume and brain atrophy are among the most important and most commonly used.[1,2] MRI-derived measurements serve as the primary outcome in phase I and phase II trials and as a secondary outcome in phase III trials.[1] It has been recommended that T2 lesion volume measures should be obtained in all clinical subgroups while brain atrophy measures should be obtained in progressive MS trials.[1]

An automated brain MR image segmentation program that classifies image elements (voxels) into predefined brain tissue classes based on their signal intensity characteristics and anatomical location has been reported.[3,4] Highly accurate T2 lesion volume quantification was demonstrated with this method, named template-driven segmentation plus (TDS+), by comparing the automated segmentation results with the manual segmentation results by experts.[4] Excellent reliability (expressed in terms of repeatability coefficients and intraclass correlation coefficients) was demonstrated by comparing measurements from paired scans.[4] These results support the application of the automated segmentation program in MS therapeutic trials and longitudinal studies where the overall mean lesion burden is compared between groups of patients. However, in day-to-day clinical practice, the changes of MR-derived measures in individual patients, instead of the mean of a group of patients, may be of concern. In particular, it is important to know if any difference in lesion burden or brain atrophy between two MR exams can be explained by measurement variability alone. This paper presents a simple methodology to determine thresholds beyond which changes in lesion burden and brain atrophy can no longer be explained by variability in the MR examination and volume measurement processes. These thresholds are reported for TDS+, as an example of image segmentation methodology.

**Correspondence: J Ross Mitchell, Seaman Family MR Research Center, Foothills Medical Center, University of Calgary, 1403-29th Street NW, Calgary, Alberta T2N 2T9, Canada.*
E-mail: rmitch@ucalgary.ca
Received 13 September 2003; revised 18 February 2004; accepted 18 March 2004

© Arnold 2004
10.1191/1352458504ms1061oa

Patients and Methods

Eleven patients with secondary-progressive MS (SPMS) and nine patients with relapsing–remitting MS (RRMS) were randomly selected from two therapeutic MS trials.[5] The first trial consisted of 30 patients with SPMS, while the second trial consisted of 34 patients with RRMS. Interleaved dual-echo spin-echo images (TR/TE1/TE2/NEX = 3000 ms/30 ms/80 ms/0.5, FOV = 24 cm × 24 cm, acquisition matrix = 256 × 192) covering the whole brain with contiguous 3-mm thick slices were acquired on a 1.5 T GE Signa MR scanner (GE Medical Systems, Milwaukee, WI). Variable receiver bandwidth was used for the first (10.67 kHz) and second (4.27 kHz) echoes. Each patient was imaged twice within 30 min. Between the two imaging sessions, the patients exited and re-entered the scanner room. Two radiology technologists alternated in positioning the patients to better simulate realistic conditions of a follow-up study. Informed consent was obtained from all patients in agreement with institutional policies.

Image voxels were classified into predefined tissue classes, i.e., lesion (white matter (WM) signal abnormalities), normal appearing WM, grey matter and cerebrospinal fluid (CSF), using the segmentation pipeline named TDS+ that combines expectation–maximization tissue segmentation, template-driven segmentation (TDS), and partial volume effect correction algorithms.[3,4,6] Tissue volumes were the numbers of voxels multiplied by the nominal volume of a single voxel. Brain parenchymal fraction (BPF) was defined as the ratio of brain parenchymal tissue volume (sum of lesion, normal appearing WM and grey matter) to the intracranial cavity volume. This definition is the same as the definition of BICCR (brain to intracranial capacity ratio) given by Collins *et al*.[7] and the definition of percentage brain parenchyma volume (PBV) given by Ge *et al*.,[8] but different from the definition given by Fisher *et al*.[2,9] in that the BPF is the ratio of brain parenchymal volume to the total volume within the brain surface contour. BPF can be used as a normalized measurement of brain atrophy. A more recent study demonstrated that the two brain volume metrics are highly correlated.[10]

The data were analysed by calculating the standard error of measurement (SEM) for the repeated measurements of T2 lesion volumes and BPF. The SEM is also referred to in the literature as measurement error.[11] It is defined as the inherent variability among measurements of the same quantity within the same individual,[12,13] and is calculated as the square root of the pooled common variance. In the special case of only two repeated scans per subject, the calculation is simplified to

$$\mathrm{SEM} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n}(x_{i1} - x_{i2})^2} \qquad (1)$$

where x_{i1} and x_{i2} are the measurements for the first and second scan for subject i, and n is the number of subjects. Threshold values (δ) that represent 95% certainty of change beyond chance (i.e., statistically significant change) were computed as $\sqrt{2} \times 1.96 \times \mathrm{SEM}$.[12,13] If the observed change in tissue volumes between two scans for an individual patient is greater than δ or less than $-\delta$, then one is 95% certain that the change is too large to be explained by variability in the measurement process alone. Conversely, observed changes in the range of $-\delta$ to δ can be 'safely' (with 95% certainty) explained by measurement variability alone. These threshold values are also referred to in the literature as reliable change indices (RC).[14]

Equation (1) relies upon two assumptions: a) that we can pool individual variances to determine the overall variance; and b) that there is no relationship between the standard deviation of the measurements (or the absolute value of the difference between repeated measurements in case of only two measurements per subject) and the magnitude or mean value of the measurements. Assumption a) was tested using Bartlett's test for equality of variances. Assumption b) was tested by plotting the magnitude of the differences between the repeated measures versus the mean of the repeated measures then calculating Kendall's rank correlation coefficient.[15]

Results

The T2 lesion volumes and BPF of each scan for each MS patient obtained using the TDS+ pipeline are listed in Table 1. Bartlett's test showed no statistically significant differences among the individual variances for T2 lesion volumes ($P = 0.35$) or for BPF ($P = 0.06$). Thus, pooling variances was deemed appropriate. Figure 1 shows the absolute difference between repeated measures versus the mean of the repeated measures for both T2 lesion volume and BPF. The variability in the repeated measures appears to be independent of the magnitude of the measurements.

Table 1 T2 lesion volume and BPF measurements with TDS+ segmentation pipeline

Patient	*T2 lesion volume (mL)*		*BPF*	
	Scan 1	*Scan 2*	*Scan 1*	*Scan 2*
1	1.63	1.50	0.8337	0.8346
2	2.46	2.20	0.9124	0.9116
3	4.70	4.43	0.7504	0.7438
4	1.67	1.31	0.7850	0.7916
5	11.18	11.36	0.8095	0.8112
6	7.08	6.76	0.8739	0.8703
7	1.85	1.68	0.8979	0.8979
8	5.36	5.23	0.8831	0.8845
9	2.68	2.60	0.8819	0.8842
10	3.08	2.79	0.8361	0.8367
11	2.35	2.24	0.8451	0.8454
12	2.50	2.60	0.8310	0.8293
13	3.08	3.00	0.8127	0.8130
14	4.40	4.73	0.8951	0.8931
15	7.93	7.95	0.7282	0.7306
16	5.06	5.62	0.7549	0.7558
17	6.58	6.59	0.8241	0.8229
18	10.87	11.88	0.8307	0.8366
19	3.31	2.95	0.8567	0.8578
20	1.55	1.72	0.8318	0.8316

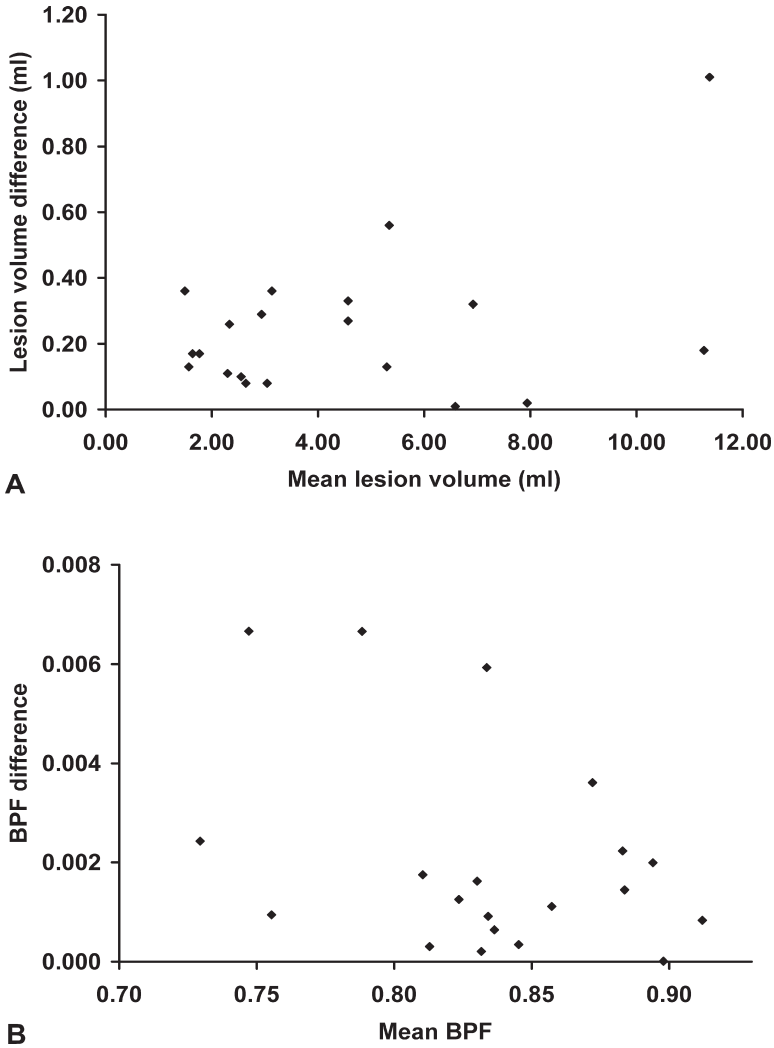

Figure 1 *Absolute difference in T2 lesion volume (A) and BPF (B) between repeated scans against their mean for data in Table 1. The variability in the repeated measures appears to be independent of the magnitude of the measurements, which is confirmed analytically by Kendall's rank correlation coefficient test.*

Kendall's test confirmed this observation for both T2 lesion volume (Kendall's $\tau = 0.08$, $P = 0.62$) and BPF (Kendall's $\tau = -0.23$, $P = 0.15$). The SEM and threshold value for 95% certainty of statistically significant change for both T2 lesion volume and BPF are shown in Table 2. The SEMs were small compared with the mean measurements of the 40 scans.

Table 2 Standard error of measurements and threshold values representing statistically significant (95% certainty) change for T2 lesion volume and BPF measurement

	T2 lesion volume (mL)	*BPF*
Mean of 40 scans	4.46	0.8339
Standard error of measurement (SEM)	0.23	0.0020
Threshold value of real change (δ)	0.65	0.0056

Discussion

It is well known that in longitudinal studies, the variation of tissue volume measurements can be caused by inconsistencies in any step of image acquisition and postacquisition analysis. When automated segmentation methods are used, the inherent operator-related reliability is close to perfection and reliability over time is dominated by image acquisition variables. Our scan–rescan study design mirrors the practical situations of longitudinal studies and attempts to incorporate most of the possible sources of variation.

Much of the existing literature about MRI-based measurements in MS has focused on the reliability of the measurements and their application to groups of patients. This paper concerns the variability of lesion burden and brain atrophy measurements in individual patients with MS.

Variability in MRI-based measures is often expressed as a coefficient of variation (COV), that is the standard

deviation in the measurements expressed as a percentage of their mean. However, the COV implies that there is a direct linear relationship between the variability of a measurement and the grand mean of the measured values. For example, suppose a researcher develops a new technique to measure lesion volumes and quantifies its performance after applying it to a group of patients. If the variability in the measurements is 1 mL, and the mean lesion volume in the patients is 20 mL, then the researcher can report a COV of 5% and conclude that the new technique has good performance. Suppose the researcher then uses the tool in a clinical trial where the mean patient lesion burden is 40 mL; should he or she now conclude that since the COV is 5% the variability must be 2 mL? The only time the variability was actually measured it was 1 mL. Must the variability change in direct proportion with the mean from a group of patients with heterogeneous disease and large interpatient differences?

Evidence from our current work (Figure 1) and previous work indicates that the variability of lesion volume and BPF measurement is independent of their means.[16] MRI-based brain tissue volume measurement *is* highly dependent upon the shape and complexity of the lesions or brain tissue being measured.[16,17] For example, imagine two patients with identical lesion burden volumes. Patient 1 has a few large lesions each of which is highly spherical. Patient 2 has many small lesions each with a highly convoluted and irregular border. Partial volume effects and small errors identifying the location of lesion edges will have a more dramatic effect on measured volume in patient 2 than in patient 1, leading to increased variability in patient 2 through repeated scans and measurements.

Variability, expressed as an SEM, allows one to estimate the smallest statistically significant change between two measurements. In our experiments, the threshold value for significant T2 lesion burden change (0.65 mL) was approximately half of the average yearly change (1.5 mL) seen in RRMS, and approximately equal to the average yearly change (0.7 mL) seen in SPMS.[18] Thus, it is likely that we would have to follow a single SPMS patient for more than one year to observe a change in lesion burden large enough that it could not be explained by measurement variability alone. In terms of the ratio of SEM to the mean, our method has less measurement variability for BPF than for T2 lesion volume. This may be explained by the smaller surface to volume ratio of brain parenchyma than that of lesions.[17]

Variability in MRI-based measures will also depend upon the scanner, pulse sequence and analysis method used. For this work, we used a common 1.5 T MR system and standard pulse sequences.[19] The analysis method is completely automatic with the exception of minimal operator interaction for the identification of the intracranial cavity,[4,6,20] and as such will help reduce variability and the minimum statistically significant change. Therefore, these results should provide useful benchmarks for researchers at institutions using similar equipment and techniques. The threshold value results in the current study are restricted to repeated measurements of the same individual imaged on the same scanner without major hardware and software upgrades. In our image segmentation pipeline, great efforts were made to overcome the inconsistence in signal intensity between serial image sets,[3,21,22] but the possible variability arising from different scanners or major hardware and software upgrades of the same scanner has not been quantitatively evaluated. The threshold values may be larger if follow-up images are acquired in either situation. In practice, the source of variation from different scanner characteristics can be limited by following the guidelines for using quantitative measures of brain MRI abnormalities in monitoring the treatment of MS, i.e., the same scanner should be used for all follow-up scans, and follow-up scans should be obtained prior to major machine upgrades if possible.[19]

In the current study, MR images were obtained with 3-mm slice thickness. A previous study using 'lesion' phantoms shown that volume estimations based on 3-mm-thick sections were more accurate and less variable than those based on 5-mm thick sections.[16] Another study using MS patient data demonstrated that acquisition of 3-mm-thick sections increased both the sensitivity and precision of MS lesion detection and quantification compared to acquisition of 5-mm-thick sections.[23] In fact, recent guidelines for using quantitative measures of brain MRI abnormalities in monitoring the treatment of MS recommend acquisition of 3-mm thick slices even though more time is needed for data acquisition and analysis.[19] Although using 5-mm slice thickness may also be justified for other clinical reasons, we believe using slice thickness of 3 mm is appropriate if monitoring disease evolution and treatment efficacy *quantitatively* is attempted.

In conclusion, we estimated the measurement error and the threshold of statistically significant change of an automated brain tissue quantification method. For repeated MR scans, changes in T2 lesion burden less than 0.65 mL in magnitude can be safely (with 95% certainty) explained by measurement variability alone. Changes in brain parenchyma fraction less than 0.0056 can be safely explained by measurement variability alone. These values provide clinical neurologists with a reference to interpret MR-derived measurements.

Acknowledgements

This study was partially supported by grants from National Institutes of Health (Contract grant numbers: P41RR013218, R01NS035142, R44MH057200, P01AG004953, R01CA086879, R21 MH67054), The Foundation for Neurological Diseases, Whitaker Foundation, Multiple Sclerosis Society of Canada, The Alberta Heritage Foundation for Medical Research and The Canadian Institutes of Health Research Interdisciplinary Health Research Team on MMPs in MS.

References

1 Filippi M, Dousset V, McFarland HF, Miller DH, Grossman RI. Role of magnetic resonance imaging in the diagnosis and

monitoring of multiple sclerosis: consensus report of the White Matter Study Group. *J Magn Reson Imaging* 2002; **15**: 499–504.
2 Fisher E, Rudick RA, Simon JH, Cutter G, Baier M, Lee JC *et al*. Eight-year follow-up study of brain atrophy in patients with MS. *Neurology* 2002; **59**: 1412–20.
3 Warfield SK, Kaus M, Jolesz FA, Kikinis R. Adaptive, template moderated, spatially varying statistical classification. *Med Image Anal* 2000; **4**: 43–55.
4 Wei X, Warfield SK, Zou KH, Wu Y, Li X, Guimond A *et al*. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. *J Magn Reson Imaging* 2002; **15**: 203–209.
5 Hohol M, Guttmann C, Olek M, Cook S, Gjorstrup P, Linde A *et al*. Roquinimex (LinomideR) treatment in secondary progressive and relapsing remitting multiple sclerosis: results from 48 week randomized, double-blind, placebo-controlled pilot studies. *Neurology* 1997; **48**: A174.
6 Guttmann CR, Kikinis R, Anderson MC, Jakab M, Warfield SK, Killiany RJ *et al*. Quantitative follow-up of patients with multiple sclerosis using MRI: reproducibility. *J Magn Reson Imaging* 1999; **9**: 509–18.
7 Collins DL, Montagnat J, Zijdenbos AP, Evans AC, Arnold DL. Automated estimation of brain volume in multiple sclerosis with BICCR. In *Proceedings of the 17th International Conference on Information Processing in Medical Imaging*, London. Springer-Verlag, 2001: 141–47.
8 Ge Y, Grossman RI, Udupa JK, Wei L, Mannon LJ, Polansky M *et al*. Brain atrophy in relapsing–remitting multiple sclerosis and secondary progressive multiple sclerosis: longitudinal quantitative analysis. *Radiology* 2000; **214**: 665–70.
9 Rudick RA, Fisher E, Lee JC, Simon J, Jacobs L. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing–remitting MS. Multiple Sclerosis Collaborative Research Group. *Neurology* 1999; **53**: 1698–704.
10 Caramanos Z, Arnold DL, Collins DL. Measuring brain volume in patients with multiple sclerosis: a comparison of three common approaches. *Proc Int Soc Magn Reson Med* 2003; **11**: 922.
11 Bland JM, Altman DG. Measurement error. *BMJ* 1996; **313**: 744.
12 Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys Ther* 1994; **74**: 777–88.
13 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.
14 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991; **59**: 12–19.
15 Bland JM, Altman DG. Measurement error proportional to the mean. *BMJ* 1996; **313**: 106.
16 Mitchell JR, Karlik SJ, Lee DH, Fenster A. Computer-assisted identification and quantification of multiple sclerosis lesions in MR imaging volumes in the brain. *J Magn Reson Imaging* 1994; **4**: 197–208.
17 Kikinis R, Shenton ME, Gerig G, Martin J, Anderson M, Metcalf D *et al*. Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging. *J Magn Reson Imaging* 1992; **2**: 619–29.
18 Weiner HL, Guttmann CR, Khoury SJ, Orav EJ, Hohol MJ, Kikinis R *et al*. Serial magnetic resonance imaging in multiple sclerosis: correlation with attacks, disability, and disease stage. *J Neuroimmunol* 2000; **104**: 164–73.
19 Filippi M, Horsfield MA, Ader HJ, Barkhof F, Bruzzi P, Evans A *et al*. Guidelines for using quantitative measures of brain magnetic resonance imaging abnormalities in monitoring the treatment of multiple sclerosis. *Ann Neurol* 1998; **43**: 499–506.
20 Kikinis R, Guttmann CR, Metcalf D, Wells WM, Ettinger GJ, Weiner HL *et al*. Quantitative follow-up of patients with multiple sclerosis using MRI: technical aspects. *J Magn Reson Imaging* 1999; **9**: 519–30.
21 Warfield S, Dengler J, Zaers J, Guttmann CR, Wells WM, 3rd, Ettinger GJ *et al*. Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. *J Image Guid Surg* 1995; **1**: 326–38.
22 Wells WM, Grimson WEL, Kininis R, Jolesz FA. Adaptive segmentation of MRI data. *IEEE Trans Med Imaging* 1996; **15**: 419–42.
23 Molyneux PD, Tubridy N, Parker GJ, Barker GJ, MacManus DG, Tofts PS *et al*. The effect of section thickness on MR lesion detection and quantification in multiple sclerosis. *AJNR Am J Neuroradiol* 1998; **19**: 1715–20.