

Accuracy and reproducibility of manual and semi-automated quantification of MS lesions in MRI

Running Title: Reproducibility of MS lesion measurement

¹Edward A. Ashton, PhD, ¹Chihiro Takahashi, MD, ²Michel J. Berg, MD,
²Andrew Goodman, MD, ¹Saara Totterman, MD, PhD, ¹Sven Ekholm, MD, PhD

¹Dept. of Radiology, University of Rochester Medical Center, Rochester, NY

²Dept. of Neurology, University of Rochester Medical Center, Rochester, NY

Corresponding Author: Edward A. Ashton
487 Granger Circle
Webster, NY 14580
Phone: (585) 671-4413
Fax: (585) 383-1294
Email: edashton@ece.rochester.edu

ABSTRACT

Purpose: To evaluate the accuracy, reproducibility and speed of two semi-automated methods for quantifying total white matter lesion burden in multiple sclerosis patients with respect both to manual tracing and to other methods presented in recent literature.

Materials and Methods: Two methods for semi-automated quantification of total lesion burden in multiple sclerosis patients using MRI were examined. The first method, geometrically constrained region growth, requires user specification of lesion location. The second, directed multi-spectral segmentation, requires only the location of a single exemplar lesion. Test data sets included both clinical MS data and MS brain phantoms.

Results: Mean processing time was 60 minutes for manual tracing, 10 minutes for region growth, and 3 minutes for directed segmentation. Intra- and inter-operator coefficients of variation were 5.1% and 16.5% for manual tracing, 1.4% and 2.3% for region growth, and 1.5% and 5.2% for directed segmentation. The average deviations from manual were 9% for region growth and 5.7% for directed segmentation.

Conclusions: Both semi-automated methods were shown to provide significant advantage over manual tracing in terms of speed and precision. Accuracy for both methods was acceptable, given the high variability of manual results.

Key Words: multiple sclerosis; volume measurement; lesion quantification; multi-spectral MRI; reproducibility

INTRODUCTION

Accurate and reproducible quantification of brain lesion count and volume in multiple sclerosis (MS) patients using magnetic resonance imaging is a vital tool for evaluation of disease progression and patient response to therapy (1). Current standard methods for obtaining these data points are largely manual and subjective, and are therefore both error-prone and subject to inter- and intra-operator variability. In addition, manual tracing of white matter lesions requires as much as 60 minutes of expert intervention time per case. Clearly, there is a need for a rapid automated lesion quantification method, but as yet there is no generally accepted solution to this problem. A useful review of techniques developed in the past to address this problem has been presented by Zijdenbos and Dawant (2).

Several groups have recently evaluated the speed, precision and accuracy of both automated and semi-automated novel MS lesion identification and quantification techniques. Brunetti *et al.* have described an automated method for white matter lesion quantification, reporting a sensitivity of 87.3% and coefficient of variation over repeated measurements of 9.4%. (3) Rovaris *et al.* have reported an intra-observer coefficient of variation for repeated measurements using a semi-automated local thresholding technique of 2.6%. (4) Dastidar *et al.* have described a semi-automated method for lesion burden measurement with coefficients of variation for inter-observer variability of 7% and for intra-observer variability of 4%, and a total measurement time of 10 minutes per scan. (5) Ballester *et al.* have estimated the measurement confidence interval associated with measurement of MS lesions using four automated methods, reporting results ranging from 28.3% to 44.9% of total lesion volume. (6) In addition, Van Leemput *et al.* have reported an inter-operator coefficient of variation for manual measurement of MS lesions using two expert raters of 18.3% (7), and Tan *et al.* have reported intra-operator coefficients of variation for manual measurement in scan-rescan studies on the order of 50% (8). These numbers will serve as useful benchmarks for the results presented in this paper.

The goals of this paper are to present two novel solutions to the problem of white matter lesion in MS patients, and to compare these solutions to both the currently prevalent manual identification method as well as the techniques noted above in terms of three variables: precision, accuracy, and speed. Both automated and manual processes were evaluated in terms of required operator time and intra- and inter-operator variability using clinical MRI scans of MS patients. In addition, both automated processes were compared to the mean manual values as a preliminary estimate of

global accuracy. Finally, both automated and manual analyses were applied to simulated brain data obtained from the McConnell Brain Imaging Centre. (9 – 12) Because this data set included ground truth, it enabled us to obtain a realistic estimate of the overall accuracy of both manual and automated lesion measurement.

MATERIALS AND METHODS

Measurement Techniques

One manual and two automated measurement techniques were evaluated in this study. Manual measurements were carried out by expert observers, who were required to trace the boundaries of each MS lesion in each experimental data set using a computer mouse. Tracing was carried out using a software tool which allowed the user to change his or her view at will among any of the three images (T1, T2, PD) or a pseudo-colored composite image in which the T1 values were mapped to the red channel, the T2 values were mapped to the green channel, and the PD values were mapped to the blue channel. This was done with the intention of providing the manual tracers with the maximum possible amount of diagnostic information. Once the tracing was completed, volume measurements for each lesion were calculated automatically.

The first automated method examined was geometrically constrained region growth (GEORG) (13,14). This technique requires a user to place a “seed” within each MS lesion in the volume using a single mouse click. The seed region then expands into neighboring voxels provided that two constraints are satisfied: the spectral signature of the neighboring voxel must have a high probability of falling within the statistical distribution defined by all current included voxels, and inclusion of the neighboring voxel must not cause the shape of the included region to deviate excessively from the *a priori* regional shape model. It is the first constraint that distinguishes this approach from deformable template techniques such as that described by Carlbom *et al.* (15), and the second constraint that distinguishes it from competitive region growth algorithms such as that described by Taylor and Barrett (16). In the case of MS lesions, the *a priori* model is a generalized ovoid with flexible boundaries. The expansion process continues until a stable boundary has been established. The operation of this algorithm is illustrated in Figure 1.

The second automated method, directed multi-spectral segmentation (DMSS), requires a user to identify one lesion within the volume. In this study this exemplar lesion was identified using GEORG, reducing user interaction to a single mouse click. In principle, however, the exemplar could be identified using manual tracing or some other semi-automated or automated method. Given an exemplar, DMSS then identifies the statistical characteristics of the apparent normal tissue using an adaptive multivariate Bayesian classifier (17,18). Finally, an iterated conditional modes (ICM) (19) – based classification pass is made, using the background statistics supplied by the Bayesian classifier and the target statistics supplied by the exemplar. The operation of this algorithm is illustrated in Figure 2.

It should be noted that both of the algorithms described in this paper are designed to operate on multi-spectral data. Each of the data sets used in this work consists of three MRI studies: one T1 weighted, one T2 weighted, and one proton density weighted. These individual scans were co-registered, so that each voxel in the volume can be described by a three-point vector consisting of the T1, T2, and PD weighted values rather than by a single grayscale value. This formulation results in a class discriminant function given by:

$$g_i(x) = -\ln|R_i| - (x - m_i)^t R_i^{-1} (x - m_i) \quad (1)$$

where R is the class covariance matrix, m is the class mean, and x is the signature of the voxel under consideration. Note that this assumes a multivariate normal class model. In the event that the data is scalar (CT or single pulse sequence MRI) R is simply replaced by class variance and m becomes the scalar class mean. This discriminant function is used by GEORG in the identification of tissue boundaries, and by DMSS in the selection of voxel classification.

Experimental Procedure

The experiments carried out in this study were intended to assess the performance of the two algorithms under consideration with respect to manual tracing and to the studies cited in Section 1 in terms of three parameters: speed, precision, and accuracy. Two data sets, one clinical and one synthetic, were used to assess required processing time for each algorithm, inter- and intra-operator measurement variability, and global accuracy.

The clinical data used in this work consisted of T1, T2, and proton density weighted MRI scans for three multiple sclerosis patients, with 2 to 4 repeats for each patient taken over the course of 4 to 8 days. A total of 9 data sets were analyzed. These data sets were originally collected in 1997 for use in an unrelated study, and were provided to our group without patient identification information. The data were acquired on General Electric Signa 1.5 Tesla Horizon MR Scanners. The images were obtained using the standard GE head coil. The use of 2d spin echo axial images and 2d variable echo multi-planar imaging pre- and post-contrast was used. All data sets analyzed in this study were collected pre-contrast. The thickness was 5mm with a slice center spacing of 5mm and a 256x192 matrix. In-plane resolution in these scans was 0.86mm. Sample images from this data set are shown in Figure 3.

It is worthwhile at this point to mention the reasons for using the pulse sequences given above, and for excluding other frequently used sequences such as FLAIR, diffusion tensor imaging, and magnetization transfer imaging. Clinical multi-center studies to evaluate the effect of a specific treatment in MS have been going on for many years. One problem with multi-center studies like these is related to the equipment. Different sites frequently have different makes of scanner and also different versions of hardware and software. To allow for comparative studies under those circumstances it is important to use imaging sequences that are as comparable as possible, independent of make or version. Because of this it was for a long time more or less mandatory to use conventional spin-echo (CSE) sequences to obtain PD- and T2-weighted images as well as T1-weighted images before and after contrast agent enhancement. In later years the so called fast spin-echo (turbo spin-echo) sequences have often been accepted since they seem to be rather comparable, independent of equipment used and also more sensitive than CSE for posterior fossa. Other imaging sequences differ far more from one make to another. This limits the sites that can be involved in a given study since one would have to use the same kind of equipment and about the same version of hardware and software in order to obtain comparable results. It is rather common to add one or more protocols to the basic protocol using spin-echo imaging to find better means for treatment evaluation in the future. However, these added sequences are usually limited to a few of the sites involved with the intent to answer specific questions, e.g. to evaluate the sensitivity and specificity of these newer techniques when compared with the old-fashioned but sturdy spin-echo sequences.

The synthetic data set, which was used to obtain an objective assessment of global accuracy for both manual and automated measurement techniques, was obtained from the McConnell Brain Imaging Centre (MBIC) (9 - 12). This data is available with various levels of simulated noise and distortion. We examined data sets with noise levels

ranging from 0 – 5% and with distortion levels ranging from 0 – 20%. In-plane resolution for the experimental data sets was 1.0mm. Slice thickness was 3.0mm. Slice thickness for the ground truth map was 1.0mm. Sample images from this data set are shown in Figure 4.

The clinical data sets shown in Figure 3 were used to evaluate required operator time and reproducibility for both manual and automated processing, as well as to establish correlation between manual and automated results. First, each of the three baseline studies was evaluated using manual tracing, GEORG and DMSS. Ten separate measurements of lesion burden in each data set were made by a single operator using both GEORG and DMSS. In addition, ten manual measurements of each data set were made by a trained neuro-radiologist. Each of the six revisit scans was then evaluated using manual tracing as well as both automated algorithms. These experiments served to establish intra-operator variability for all three processes.

In order to establish inter-operator variability, each baseline scan was then evaluated by four trained operators using GEORG and DMSS, and by four expert observers using manual tracing. It should be noted that none of the participants in this study had access to the results obtained by any of the others prior to evaluating the data. Per-scan evaluation time was noted and reported by all participants in this study, allowing the calculation of potential time savings resulting from the adoption of automated processing techniques. Finally, mean results for all three techniques were compared, in order to establish the level of correlation between manual and automated measurements.

The MBIC synthetic data was used primarily to determine some estimate of the global accuracy of both manual tracing and the automated measurement methods. Because these data sets contained various levels of noise and distortion they also allowed an examination of the sensitivity of both manual and automated measurement to these parameters. Four data sets were evaluated: noise=0, distortion=0; noise=3%, distortion=0; noise =3%, distortion=20%; and noise=5%, distortion=20%. Each of these studies was evaluated once by a single operator using GEORG and once using DMSS. Each study was also evaluated by a neuro-radiologist using manual tracing. The results of each of these measurements were compared to the provided ground truth map in order to estimate global accuracy. Also, results were examined relative to one another in order to determine sensitivity to noise and distortion in the data.

RESULTS

Results of the first intra-operator variability trials, measuring precision over repeated measurements of the same data sets, are given in Table 1. These results indicate that both automated techniques provide substantial improvement over manual tracing in terms of precision. The automated methods correlate well in terms of mean lesion burden for Patients 1 and 2, although GEORG appears to deviate somewhat for Patient 3. In addition, correlation between automated and manual volumes is reasonable, with a consistent bias of between one and two standard deviations for all cases other than the measurement of Patient 3 using GEORG. This bias is most likely explained by a tendency for manual tracers to over-estimate lesion margins. This tendency will be discussed in greater detail in the following section.

Our second intra-operator variability study involved measurement of lesion burden in repeat scans taken of each of the three patients over a period of 4 – 8 days. Examination by an expert radiologist showed no noticeable change in lesion number or size in any of the three patients over this period. Results of this experiment are given in Table 2.

Clearly, this experiment demonstrates that manual measurements are less repeatable than Table 1 seems to indicate. The primary difference between this experiment and the previous one is that in the first experiment the repeated measurements were performed on the same data sets, allowing the tracer to use memory of the previous tracings to guide each new one. In the second experiment, although the underlying lesion burden was presumably consistent over such a short time period, the appearance of a particular lesion in a given slice of one scan might differ considerably from that in another scan due to imprecision in patient positioning. This is illustrated in Figure 5. Consequently, the second experiment shows a considerable increase in the variability of the manual measurements. The primary source of variation in the automated measurements in the second experiment was changes in the apparent margins of small lesions due to partial volume effects. This effect is most pronounced for Patient 3, whose lesion burden was considerably lower and whose average lesion size was significantly smaller.

Inter-operator variability was evaluated by measuring lesion burden in each of the initial scans using four different operators for each algorithm. Manual tracing was carried out by qualified experts in either neurology or neuro-radiology. Automated processing was carried out by trained MR technicians. Results of this experiment are given in Table 3.

As expected, inter-operator variability is considerably higher than intra-operator variability for all three measurement techniques. The most significant increase is seen in manual tracing, while GEORG shows the least sensitivity to different operators. This may seem counter-intuitive, as GEORG actually requires more operator interaction than DMSS. However, recall that DMSS is dependent on an operator-selected exemplar lesion. The variability seen in Table 3 is the result of the various operators selecting different exemplars, each of which might have somewhat different statistics.

Our final experiment involved data taken from the MBIC brain database. This experiment was designed to assess global accuracy, as the MBIC data was thoroughly ground-truthed. Experimental results were generated for data sets with noise levels of 3% and 5%, and distortion levels of 0% and 20%. Results of this experiment are given in Table 4.

DISCUSSION

The results presented in the previous section raise a number of interesting questions. In comparing automated results to one another and to manual measurements, one of the most notable facts is that both automated techniques appear to have a negative bias with respect to manual tracing. In every case presented in Tables 1 – 4 the mean manual measurement exceeds the mean of either semi-automated measurement. At first glance this would appear to indicate a systematic flaw in both GEORG and DMSS. However, visual examination of the lesion identifications made by the three algorithms indicates that in many cases the manual tracers have over-estimated the lesion boundaries in an effort to mark the entire lesion. This is especially apparent in the simulated MRI data whose results are presented in Table 4. This phenomenon is illustrated in Figure 6.

This effect is even more pronounced in the MBIC simulation data. Lesions in this data set tend to be quite small, magnifying the effect of even single-pixel overestimation of lesion boundary locations. In addition, the ground truth data provides us the opportunity to objectively assess which technique provides a more accurate measurement. Lesion over-estimation in MBIC data is illustrated in Figure 7.

Another point of interest in the data presented in the previous section is the increased variability seen in measurements of studies taken of the same patient over a period of a few days, shown in Table 2, with respect to the variability seen in repeated measurements of the same data sets, shown in Table 1. This result is consistent with that

reported by Guttman *et al.* (20) and Tan *et al.* (8), who have shown that variability introduced by MRI acquisition and image registration procedures is typically large with respect to variability introduced by measurement techniques. More significantly, the increase in variability is much greater for manual tracing than for the automated techniques. This provides some indication that the relatively low coefficients of variation seen for manual tracing in Table 1 are the result of the use of local landmarks. It is therefore arguable that the variability seen in Table 2 is likely to be more representative of what might be expected in a realistic situation.

The final question that this study was intended to answer, of course, is that of the relative utility of the three measurement techniques examined here, as well as the others published methods cited previously, in terms of our three parameters of interest: speed, precision, and accuracy. In terms of speed, as expected, DMSS provided the best results, with a mean processing time of 3 minutes per study. This compares to a mean of 10 minutes per study for GEORG, 10 minutes per study for the method presented by Dastidar *et al.*, and 60 minutes per study for manual tracing. In addition, DMSS requires significantly less operator expertise than either of the other measurement methods. DMSS requires only that the operator should be able to identify a single large lesion somewhere within the brain – an easily trainable skill. Both GEORG and manual tracing require that the operator should be able to identify every lesion throughout the brain. For obvious reasons, this requires a much higher level of operator expertise.

In terms of precision, GEORG and DMSS showed similar performance in all cases other than the inter-operator variability experiment. In that case, GEORG demonstrated a notably lower coefficient of variation. This is easily explained, since the four DMSS operators were not constrained to select the same exemplar, and choosing different exemplar lesions might be expected to produce somewhat different segmentation results. GEORG, however, gives results that are quite consistent over different interior seed points, as demonstrated in Table 3 and in (14). GEORG also demonstrated less sensitivity to changes in image parameters, as shown in Tables 2 and 4. Both automated algorithms, as expected, demonstrated significantly greater precision than manual tracing.

Both algorithms described here provide a level of precision comparable to or better than that of the techniques described in the literature. Intra-operator coefficients of variation for GEORG ranged from 0.9% to 2.4%, while those for DMSS ranged from 0.9% to 2.1%. Comparable measures from the literature are 9.4% for the technique presented by Brunetti *et al.* (3), 4% for the technique presented by Dastidar *et al.* (5), and 2.6% for the technique presented by

Rovaris *et al.* (4). Additionally, Wei *et al.* have reported scan-rescan coefficients of variation for three automated methods ranging from 2.57% to 7.5% (21).

Inter-operator coefficients of variation for GEORG ranged from 0.9% to 4.9%, while those for DMSS ranged from 3% to 6.8%. Comparable values from the literature include 7%, reported by Dastidar *et al.* (5), and 0.9%, reported by Udupa *et al.* for a semi-automated method making use of fuzzy-connectedness principles and manual false positive and false negative mitigation (22).

Accuracy is, of course, the most difficult parameter to evaluate. In general, DMSS showed somewhat better correlation with manual tracing than GEORG across all experiments. However, the results given in Table 4 indicate that correlation with manual tracing may be less than ideal as a gold standard. Across the three phantom data sets, GEORG showed better correlation with ground truth than DMSS, and both automated algorithms performed significantly better than manual tracing.

In the final analysis, the optimal system for quantification of MS lesions in multi-spectral MRI will most likely prove to be a combination of the two algorithms presented here, with a first pass using DMSS followed by correction of false positives and false negatives using GEORG. A system of this sort would provide much of the speed advantage offered by DMSS, while also incorporating the accuracy and precision provided by GEORG.

REFERENCES

1. Filippi M, Dousset V, McFarland H, Miller D, Grossman R. 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. Zijdenbos A, Dawant B. Brain segmentation and white matter lesion detection in MR images. *Critical Reviews in Biomedical Engineering* 1994; 22: 401 – 465.
3. Brunetti A, Larobina A, Quarantelli M, Tedeschi *et al.* Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. *J Magn Reson Imaging* 2000; 12: 799 – 807.
4. Rovaris M, Inglese M, van Schijndel R *et al.* Sensitivity and reproducibility of volume change measurements of different brain portions on magnetic resonance imaging in patients with multiple sclerosis. *J Neurol* 2000; 247: 960 – 65.
5. Dastidar P, Heinonen T, Lehtimäki T *et al.* Volumes of brain atrophy and plaques correlated with neurological disability in secondary progressive multiple sclerosis. *J Neurol Sci* 1999; 165: 36 – 42.
6. Ballester M, Zisserman A, Brady M. Segmentation and measurement of brain structures in MRI including confidence bounds. *Medical Image Analysis* 2000; 4: 189 – 200.
7. Van Leemput K, Maes F, Vandermeulen D, Colchester A, Suetens P. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans Med Imaging* 2001; 20: 677 – 88.
8. Tan I, van Schijndel R, van Walderveen M *et al.* Magnetic resonance image registration in multiple sclerosis: Comparison with repositioning error and observer-based variability. *J of MRI* 2002; 15: 505 – 510.
9. Cocosco C, Kollokian V, Kwan R, Evans A. BrainWeb: Online interface to a 3D MRI simulated brain database. *Proceedings of 3-rd International Conference on Functional Mapping of the Human Brain* 1997; 5: S425.
10. Kwan R, Evans A, Pike G. MRI simulation-based evaluation of image-processing and classification methods. *IEEE Trans Med Imaging* 1999; 18: 1085 – 97.
11. Kwan R, Evans A, Pike G. An extensible MRI simulator for post-processing evaluation. *Visualization in Biomedical Computing* 1996; 1131: 135 – 140.
12. Collins D, Zijdenbos A, Kollokian V *et al.* Design and construction of a realistic digital brain phantom. *IEEE Trans Med Imaging* 1998; 17: 463 – 468.

13. Ashton E, Totterman S, Takahashi C, Tamez-Pena J, Parker K. Automated measurement of structures in CT and MR imagery: A validation study. Proc 14th IEEE Symposium on Computer-Based Medical Systems 2001; 300 – 306.
14. Ashton E, Parker K, Berg M, Chen C. A novel volumetric feature extraction technique, with applications to MR images. IEEE Trans Med Imaging 1997; 16: 365 – 71.
15. Carlbom I, Terzopoulos D, Harris K. Computer assisted registration, segmentation and 3D reconstruction from images of neuronal tissue sections. IEEE Trans Med Imaging 1994; 13: 351 – 362.
16. Taylor D, Barrett W. Image segmentation using globally optimum growth in three dimensions with an adaptive feature set. Visualization in Biomedical Computing 1994; 2359: 98 – 107.
17. Ashton E. Detection of subpixel anomalies in multispectral infrared imagery using an adaptive Bayesian classifier. IEEE Transactions on Geoscience and Remote Sensing 1998; 36: 506 – 17.
18. Ashton E. Multialgorithm solution for automated multispectral target detection. Optical Engineering 1999; 38: 717 – 24.
19. Besag J. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society 1986; 48: 259 – 302.
20. Guttman C, Kikinis R, Anderson M *et al.* Quantitative follow-up of patients with multiple sclerosis using MRI: reproducibility. J Magn Reson Imaging 1999; 9: 509 – 18.
21. Wei X., Warfield S, Zou K *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.
22. Udupa J, Wei L, Samarasekera S *et al.* Multiple sclerosis lesion quantification using fuzzy-connectedness principles. IEEE Trans Med Imaging 1997; 16: 598 – 609.

Tables

Table 1

Results of Intra-operator Variability Study 1. This table shows the results for 10 measurements of each patient's total lesion burden, along with mean, standard deviation, and coefficient of variation. Values given are volumes in cubic centimeters.

	<i>T1</i>	<i>T2</i>	<i>T3</i>	<i>T4</i>	<i>T5</i>	<i>T6</i>	<i>T7</i>	<i>T8</i>	<i>T9</i>	<i>T10</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>C.V.</i>
Patient1													
Manual	20.0	21.2	20.8	20.6	20.2	19.1	21.0	20.4	19.2	19.2	20.3	0.7	3.4%
GEORG	19.5	19.5	19.6	19.7	19.3	19.1	19.1	19.3	19.2	19.5	19.4	0.2	1.0%
DMSS	19.4	19.5	19.4	19.4	19.4	20.4	19.4	19.4	19.5	19.4	19.5	0.3	1.5%
Patient2													
Manual	26.8	26.5	22.5	23.1	24.3	24.1	26.0	26.8	24.9	27.7	25.3	1.7	6.7%
GEORG	22.1	21.9	22.0	22.1	21.9	21.8	21.7	21.7	21.7	21.8	21.9	0.2	0.9%
DMSS	21.4	22.2	22.0	22.0	22.2	22.3	22.1	22.1	22.2	22.2	22.1	0.2	0.9%
Patient3													
Manual	9.6	10.5	10.6	9.2	10.4	10.4	10.1	8.0	10.1	8.9	9.8	0.8	8.2%
GEORG	8.5	8.5	8.3	8.3	8.3	8.0	8.0	8.0	8.0	8.1	8.2	0.2	2.4%
DMSS	9.7	9.8	9.8	9.7	9.4	9.8	9.7	9.8	9.7	9.2	9.7	0.2	2.1%

Table 2

Results of repeated scan study. This table shows the results of total lesion burden measurement of studies temporally separated by several days. All values are given in cubic centimeters. Note that the variability of the manual measurements has increased greatly with respect to that of the automated measurements.

	<i>Repeat 1</i>	<i>Repeat2</i>	<i>Repeat3</i>	<i>Repeat4</i>
Patient 1				
Manual	20.0	29.3	34.9	N/A
GEORG	19.4	19.4	20.2	N/A
DMSS	19.4	19.6	19.1	N/A
Patient 2				
Manual	26.8	34.8	N/A	N/A
GEORG	22.1	20.7	N/A	N/A
DMSS	21.4	20.0	N/A	N/A
Patient 3				
Manual	9.6	9.8	8.4	9.4
GEORG	8.5	8.0	8.2	7.6
DMSS	9.7	10.5	11.3	8.1

Table 3:

Results of inter-observer variability trials. This table shows the results of total lesion burden measurement at a single time point for each patient by four different observers. As expected, the automated methods show good correlation with each other and with manual results. The automated methods also show significantly lower coefficients of variation.

	<i>Obs. 1</i>	<i>Obs. 2</i>	<i>Obs. 3</i>	<i>Obs. 4</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>C.V.</i>
Patient 1							
Manual	27.0	18.8	20.6	18.4	21.2	4.0	18.9%
GEORG	19.5	19.6	19.7	19.3	19.5	0.2	1.0%
DMSS	20.0	22.6	19.4	20.0	20.5	1.4	6.8%
Patient 2							
Manual	22.8	25.3	27.3	21.0	24.1	2.8	11.5%
GEORG	22.1	22.2	21.8	21.8	22.0	0.2	0.9%
DMSS	20.9	22.5	21.4	19.5	21.1	1.2	5.9%
Patient 3							
Manual	9.9	9.7	10.8	6.7	9.3	1.8	19.2%
GEORG	8.5	8.4	7.7	8.1	8.2	0.4	4.9%
DMSS	9.7	10.1	9.7	9.4	9.7	0.3	3%

Table 4

Results of phantom experiments. This table shows the results of total lesion burden measurement for the MBIC phantom under three different noise conditions. Values are given in cubic centimeters. Note that in addition to showing significantly better precision, automated results are much closer to ground truth than those obtained via manual tracing. Poor manual results are primarily due to the large number of very small and difficult to trace lesions present in this data set.

	<i>Noise 3 Dist. 0</i>	<i>Noise 3 Dist. 20</i>	<i>Noise 5 Dist. 20</i>
Manual	10.2	8.5	8.7
GEORG	3.2	3.2	3.1
DMSS	4.5	3.4	4.1
Ground Truth	3.5	3.5	3.5

CAPTIONS

Figure 1: This image series shows the operation of GEORG as it identifies lesion boundaries, from initial seed placement (a) to stable boundary identification (e).

Figure 2: Operation of DMSS algorithm. (a) Initial operator-placed seed on slice 17/30. (b) Exemplar lesion as identified by GEORG algorithm. (c) Original slice 22/30. (d) Lesions on slice 22/30 as identified by DMSS.

Figure 3: Sample images from the baseline T2-weighted scans for (a) Patient 1 – (c) Patient 3.

Figure 4: Sample images from the MBIC synthetic data set. (a) Noise = 0, distortion = 0. (b) Noise = 3%, distortion = 20%. (c) Noise = 5%, distortion = 20%. (d) Ground truth map.

Figure 5: (a) A section from a T2 weighted scan for Patient 3 containing several white matter lesions. (b) The subsequent slice from the same scan. (c) A section corresponding to that shown in (a) from a scan taken 24 hours later. (d) The subsequent slice from the same scan. Note that partial volume effects have caused the apparent size of several of the lesions shown here to change considerably, despite the fact that this patient's lesion burden was determined by an expert radiologist to be stable over this time period.

Figure 6: (a) A white matter lesion found in Patient 1. (b) Manual tracing of the lesion shown in (a) as identified by Observer 1. (c) Identification of the lesion shown in (a) using GEORG. Clearly, manual tracing has over-estimated the volume of this lesion, while GEORG provides a very close approximation of the lesion boundaries.

Figure 7: (a) A white matter lesion in the MBIC simulation data set with noise = 3%, distortion = 20%. (b) Manual tracing of the lesion in (a) performed by Observer 2. (c) Identification of the lesion in (a) using DMSS. (d) The lesion in (a) as identified in the ground truth file. Although both methods appear to have over-estimated lesion volume in this case, the automated identification in (c) is clearly significantly more accurate. This impression is reflected in the data in Table 4.