Precision and reliability for measurement of change in MRI lesion volume in multiple sclerosis: a comparison of two computer assisted techniques

P D Molyneux, P S Tofts, A Fletcher, B Gunn, P Robinson, H Gallagher, I F Moseley, G J Barker, D H Miller

Abstract

***Objective*—The serial quantification of MRI lesion load in multiple sclerosis provides an effective tool for monitoring disease progression and this has led to its increasing use as an outcome measure in treatment trials. Segmentation techniques must display a high degree of precision and reliability if they are to be responsive to small changes over time. This study has evaluated the performance of two such techniques, the manual outlining and contour methods, in serial lesion load quantification.**

***Methods*—Sixteen patients with clinically definite multiple sclerosis were scanned at baseline and after two years. Scan analysis was performed twice, independently by three observers using each technique.**

***Results*—For the absolute lesion volumes the median intrarater coefficient of variation (CV) was 3.2% for the contour technique and 7.6% for the manual outlining method ($p<0.005$), the interrater CVs were 3.8% and 6.1% respectively ($p<0.01$) and the reliability of both techniques was very high. For the change in lesion volume the intrarater and interrater repeatability coefficients were respectively 2.6 cm^3 and 2.8 cm^3 for the contour technique, and 3.3 cm^3 and 3.7 cm^3 for the manual outlining method (lower values reflect higher precision). The values for intrarater and interrater reliability for measuring change in lesion volume were respectively, 0.945 and 0.944 for the contour technique, and 0.939 and 0.921 for the manual outline method (perfect reliability = 1.0).**

***Conclusions*—With such high values for reliability, the impact of measurement error in lesion segmentation on sample size requirements in multiple sclerosis treatment trials is minor. This study shows that a change in lesion volume can be measured with a higher level of precision and reliability with the contour technique and this supports its further application in serial studies.**

(*J Neurol Neurosurg Psychiatry* 1998;**65**:42–47)

Keywords: multiple sclerosis; magnetic resonance imaging; precision; reliability; lesion volume

NMR Research Unit, The Institute of Neurology, Queen Square, London, UK
P D Molyneux
P S Tofts
A Fletcher
B Gunn
P Robinson
H Gallagher
G J Barker
D H Miller

Lysholm Radiological Department, National Hospital for Neurology and Neurosurgery, Queen Square, London, UK
I F Moseley

Correspondence to:
Professor DH Miller, NMR Research Unit, The Institute of Neurology, Queen Square, London WC1N 3BG, UK. Telephone 0044 171 837 3611; fax 0044 171 919 5616.

Received 8 August 1997 and in revised form 5 November 1997
Accepted
11 November 1997

Magnetic resonance imaging (MRI) provides a powerful tool for measuring disease activity in patients with multiple sclerosis.[1 2] It is highly sensitive to subclinical pathology and provides an objective assessment of the extent of disease. These attributes have led to the use of serial MRI in monitoring treatment efficacy in clinical trials.[3–5] However, a strong correlation between change in MRI lesion load and clinical evaluation of disease progression has not yet been shown in serial studies and for phase III definitive treatment trials, clinical outcome remains the accepted primary outcome measure.[1 6 7] This weak relation in part reflects well known limitations of current clinical rating scales such as the expanded disability status scale (EDSS)[8 9] and the pathological heterogeneity of lesions on conventional brain MRI sequences. The contribution of measurement error in quantifying changes in T2 lesion load over time may also be significant. Random measurement error in the performance of lesion segmentation is only one of many potential sources of variability during image acquisition and analysis.[10–12] However, a high level of precision and reliability at this stage is essential if a technique is to be responsive to relatively small changes in lesion load over time. Precision, or reproducibility, is defined as the extent to which repeated measurements on the same object are in agreement. The reliability of a technique provides an assessment of measurement error as a proportion of variance between patients. Both precision and reliability are important factors to consider when investigating the utility of a measurement technique.

Several techniques are available for performing lesion volume quantification[4 7 12–18] and they vary in the amount of human interaction required. Automated techniques offer the potential for high precision and speed, but care must be taken to ensure that such methods are responsive to genuine changes in lesion volume before their application to treatment trials. More operator dependent techniques have successfully been used to detect treatment effect.[4] However, their high level of human interaction may cause measurement error to be a significant problem.[16 18] Defining the precision and reliability of such methods as part of their validation is therefore important. Several studies have assessed the precision of measurement of lesion load in a cross sectional manner.[7 14 16–19] In treatment trials, however, it is not the absolute lesion volume but change over serial studies that is the outcome measure, and the precision and reliability of lesion load quantification in identifying such change has not previously been defined. Furthermore, the

J Neurol Neurosurg Psychiatry: first published as 10.1136/jnnp.65.1.42 on 1 July 1998. Downloaded from http://jnnp.bmj.com/ on June 28, 2022 by guest. Protected by copyright.

measurement of reliability also provides a means for assessing the impact of measurement error on sample size requirements using T2 weighted lesion load as an outcome measure. The present study considers these issues by evaluating the performance of two quantitative techniques—the manual outlining and contours methods—in measuring change in lesion volume over time.

Patients and methods

The scans of 16 patients with clinically definite multiple sclerosis according to Poser criteria[20] at baseline and after an interval of two years were used for this study. These patients represented part of a larger cohort of patients that participated in the North American interferon β-1b study on patients with relapsing-remitting multiple sclerosis;[4] eight patients were randomly selected from the placebo arm and eight from the group treated with high dose (8 million IU) interferon β-1b to provide a set of images actually used in a previous treatment trial. Scores on the EDSS were between 1.0 and 5.5. All patients gave informed written consent.

MRI SEQUENCES

All scans were performed on a 1.5 T Signa system (General Electric, Milwaukee, WI). Twenty two contiguous axial images were obtained from foramen magnum to vertex using a slice thickness of 5 mm. Each patient had dual echo proton density and T2 weighted conventional spin echo sequences at baseline and after two years (repetition time (TR) of 2000 ms, echo times (TE) of 30 and 70 ms). The field of view was 200 mm and the matrix was 128×256.

LESION IDENTIFICATION

Three experienced observers (DHM, IFM, PDM) identified and marked the multiple sclerosis lesions on the hard copies by consensus. The baseline and year 2 studies of each patient were assessed together to allow consistent decisions to be made on inclusion or exclusion of equivocal lesions on serial scans. Lesion identification and subsequent delineation (see below) was performed on the proton density images.

QUANTIFICATION OF LESION VOLUME

Three experienced raters (AF, BG, PR) performed the lesion volume quantification on Sun workstations (Sun Microsystems, USA). Only those lesions marked on the hard copy were segmented. Analysis of all 32 scans was performed twice, independently by three raters, using both techniques. This provided a means of assessing both intrarater and interrater precision and reliability. The potential for any memory of the images to introduce systematic bias was minimised by randomising the scan order and ensuring a delay of at least one week between repeated measurements on the same scan.

(1) The manual outline technique was performed on the computer display (Sun Microsystems, Mountain View, CA, USA) by tracing the lesion outline with a mouse controlled cursor.[4]

(2) The contour method incorporates a local thresholding algorithm to trace the lesion boundary and runs as part of the Dispimage package.[18 21] A point on the lesion edge is identified by the rater. The algorithm finds the lesion edge by searching for the strongest local intensity gradient. The lesion is delineated by following the contour of isointensity and this is displayed to allow expert review. Manual editing of part of the lesion boundary to delete regions of increased signal not corresponding to lesion is sometimes necessary, particularly where lesion/background contrast is poor.

Lesion volumes were calculated automatically for both techniques as the lesion area on each slice multiplied by the slice thickness. The time consumption of the two techniques is similar in experienced hands.

STATISTICS

Several statistical methods can be used to define the precision of a measurement technique and care must be taken to ensure that an appropriate descriptive statistic is employed. The coefficient of variation (CV)has been used as a measure of precision in several previous cross sectional studies.[4 18 22–24] This was therefore calculated for the repeated measurements on the baseline scans to allow comparison with other studies. The coefficient of variation was calculated as the standard deviation (SD) of the replicated measurements divided by their mean.[25] The intrarater CV averaged across the three raters was calculated for all 16 baseline scans with each technique. The interrater CV for each baseline scan was averaged across the two repeats performed by each rater.

However, the CV has major limitations as a measure of precision, the most important of which is its dependence on the magnitude of the measured value; an inverse relation exists between the lesion volume and the CV of replicated measurements. This implies that a single mean or median value for CV cannot fully describe precision across a wide range of lesion volumes. Furthermore, care must be taken when comparing different studies on precision that use the CV as the descriptive statistic, as widely differing lesion volumes have been used in such studies. The CV was not used for assessing precision in measuring the change in lesion volume, as its value is too heavily dependent on the size of the measured change. In view of the limitations of the CV, repeatability coefficients were used to describe precision for measurements of the change in lesion volume.[26 27] The difference between two measurements for the same subject is expected to be less than the repeatability coefficient in 95% of observations. Precision is therefore expressed in terms of the unit of measurement. The assumptions inherent in the repeatability coefficient are that there should be no systematic bias between replicated measurements and no relation between the SD of the replicated measurements and the mean. For the baseline measurements, the second of these criteria was not met (the SD was positively correlated with the

J Neurol Neurosurg Psychiatry: first published as 10.1136/jnnp.65.1.42 on 1 July 1998. Downloaded from http://jnnp.bmj.com/ on June 28, 2022 by guest. Protected by copyright.

Table 1 Lesion volume measurements on baseline scans and for changes in lesion volume with the two techniques

	Technique	Mean	Median	Min	Max
Baseline lesion volume (cm^3)	Manual outlining	18.8	12.2	2.0	77.1
	Contour method	18.2	12.1	2.1	74.7
Change in lesion volume (cm^3)	Manual outlining	+2.5	+0.8	−3.4	+14.9
	Contour method	+1.8	+0.5	−4.4	+14.8

The values were obtained by combining the results of the three raters for the 16 patients. However, the minimum and maximum values reflect the results of all observations regardless of rater. The intraclass correlation coefficient for agreement between the lesion volumes obtained with the two techniques was 0.996 for measurements on the baseline scans (bias 3%), and 0.910 for measurements of the change in lesion volume.

Table 2 Intrarater and interrater precision (CV) and reliability (ICC) for absolute lesion volumes (16 baseline scans)

	Manual outlining					Contouring				
	CV (%)					CV (%)				
	Mean	Median	Min	Max	ICC	Mean	Median	Min	Max	ICC
Intrarater	8.6	7.6	0.1	26.1	0.995	4.2	3.2	0.1	13.4	0.998
Interrater	7.2	6.1	1.1	21.1	0.997	4.7	3.8	1.0	9.7	0.998

ICC=Intraclass correlation coefficient, a statistical measure of reliability that assesses the ability of a technique to discriminate between different patients. Its range is from zero (representing no reliability) to 1.0 (perfect reliability). CV=coefficient of variation, a measure of precision over replicated measurements. The intrarater precision and reliability were calculated by pooling data acquired from the repeated measurements of all three raters on the 16 patients. The interrater precision and reliability were averaged across the two sets of measurements with each technique.

magnitude of the lesion volumes) and repeatability coefficients were therefore not calculated. However, for the replicated measurements of the change in lesion load, both criteria were fulfilled by the data in this study and this statistic was therefore used to describe precision in measurement of change in lesion volume.

An intraclass correlation coefficient (ICC) was calculated as a measure of intrarater and interrater reliability for both absolute lesion volumes and the change in lesion volume.[28 29] Analysis of variance (ANOVA) was used to calculate the ICC using a model treating raters as a fixed factor. The ICC gives the proportion of total variance including measurement errors, in measurements from a number of subjects, arising from the true variance between the subjects. It varies from zero (no reliability) to one (perfect reliability). An ICC was also used as a measure of agreement between the results obtained with the two techniques.[31]

Differences between lesion volumes and CVs obtained with the two techniques were evaluated by means of the Wilcoxon signed ranks test. All calculations were performed using the SPSS package.

Table 3 Intrarater and interrater precision and reliabilty for measurements of change in lesion volume for all 16 patients

	Manual outlining		Contour	
	Repeatability coefficient (cm^3)	ICC	Repeatability coefficient (cm^3)	ICC
Intrarater	3.3	0.939	2.6	0.945
Interrater	3.7	0.921	2.8	0.944

ICC=Intraclass correlation coefficient, a measure of reliability. The intrarater precision (repeatability coefficient) and reliability for the change in lesion volume were calculated by pooling data acquired from the replicated measurements of all three raters on the 16 patients. The interrater precision and reliability for the change in lesion volume were averaged across the two sets of measurements with each technique. The coefficient of variation was not used to define precision in measuring change in lesion volume as its value is too heavily dependent on the magnitude of the measured change. Repeatability coefficients were therefore used as a measure of precision. The difference between two measurements for the same subject is expected to be less than the repeatability coefficient in 95% of observations.

Results

LESION VOLUMES OBTAINED BY THE TWO TECHNIQUES

The baseline lesion volumes (table 1) showed excellent agreement between the two techniques (ICC=0.996), but the mean volume obtained with the manual outlining method was slightly higher (p=0.01) with a bias of 3%. Agreement between the techniques for the change in lesion volume was also high (ICC=0.910).

REPEATED MEASUREMENTS ON THE BASELINE SCANS BY THE THREE OBSERVERS

Table 2 shows the intrarater and interrater performances. The median intrarater CVs averaged across the three raters for the contour and manual outlining methods were 3.2% and 7.6% respectively (p<0.005). The median interrater CVs for the contour and manual outlining methods were 3.8% and 6.1% (p<0.01). There was no significant difference between intrarater and interrater CV for the manual outlining (p=0.1) or contour methods (p=0.2) The intrarater and interrater reliability values for both techniques were >0.99 (table 2).

REPEATED MEASUREMENTS OF CHANGE IN LESION VOLUME

Table 3 shows the values for precision (repeatability coefficients) and reliability (ICC) for the change in lesion volume. Intrarater and interrater precision and reliability were better for the contour method than the manual outlining technique.

Discussion

Lesion load quantification on serial MRI provides a sensitive and objective technique for assessing disease activity in multiple sclerosis. It has provided important insights into the natural history of the disease and is increasingly being used as a surrogate marker in treatment trials,[1 2] offering several benefits over clinical indices such as the EDSS. One major advantage is a high level of precision. Several cross sectional studies have confirmed this with newer quantitative techniques.[14 16 18] However, it is not the absolute lesion volume but the difference between serial estimates of lesion load that is measured to provide an end point in definitive treatment trials. To our knowledge, no previous work has defined the precision and reliability of such techniques in measuring this change. Clearly, measurement of any change requires a technique with a high level of precision, as random errors in measuring lesion load at each time point may have a cumulative effect on differences over serial MRI investigations. This is particularly important given that changes in T2 lesion load measured on annual MRI are often small.

In this study we have examined the efficacy of two quantitative techniques for measuring lesion load. Lesion segmentation is only one of many potential sources of measurement error

J Neurol Neurosurg Psychiatry: first published as 10.1136/jnnp.65.1.42 on 1 July 1998. Downloaded from http://jnnp.bmj.com/ on June 28, 2022 by guest. Protected by copyright.

and the overall accuracy and precision in measurement is affected by errors at each stage. Our results therefore ignore the impact of variable scanner performance arising from inconsistent coil loading, receiver attenuation setting, and scanner preamplifier gain. Furthermore, the effects of suboptimal repositioning[23] and inconsistency in lesion identification have not been considered, because the aim was to define and compare the precision and reliability of the quantitative techniques themselves.

Many statistical methods are available for describing the precision of a measurement technique and no single approach has been universally accepted. We have used the CV to describe precision in measuring the absolute lesion volumes because this is the most commonly used statistic in recent studies.[4 18 22–24] It has the advantage of expressing the measurement error as a proportion of the actual lesion volume and is therefore easy to comprehend. The values for intrarater and interrater CV obtained in this study are similar to previous reports and we have confirmed that the contour technique offers significantly greater precision than is possible with manual outlining. A high level of agreement was found between lesion volumes obtained with the two segmentation techniques used in this study. The manual outline technique has shown a treatment effect in a large multicentre trial[4] and it can be regarded as a gold standard measure. The contour method produces very similar lesion volumes with the significantly higher precision afforded by computer assisted lesion delineation and this strongly supports its use in lesion load quantification.

Furthermore, our results confirm that the contour method also has higher precision than the manual outlining technique in identifying differences in lesion volume between serial studies. This implies that, being less subject to random error, it represents a more powerful technique for identifying any effect of treatment on change in lesion load.

The estimation of reliability is an alternative approach to assessing the impact of random measurement error, and it is in some ways a more useful statistic than assessment of precision. Reliability provides a measure of the ability of a measurement technique to discriminate between the different members of a sample population.[28 29] It defines the proportion of variance in the repeated measurements that is attributable to differences between patients. If a technique has perfect reliability, all the variance in repeated measurements arises from systematic differences between subjects. Even a technique that is highly precise may not be able to distinguish between patients if the population range of the measured value is narrow. As the aim of serial lesion volume quantification is to discriminate between subjects and identify trends, its reliability is an important consideration. Reliability in part depends on the heterogeneity of the sample chosen. The very high values of reliability for measurements of baseline lesion volumes are perhaps not surprising, given the wide range of lesion volumes on these scans. More significantly, however, reliability for measuring relatively small changes in lesion volume was excellent with both techniques. This suggests that variance due to random measurement error is small compared with that due to wide biological variability in changes in lesion load across the patient population. To exclude the possibility that sample variability had been increased by including eight patients treated with interferon β-1b, the variance between patients for the change in lesion load in the placebo group and for the group as a whole was subsequently analysed. Variance between patients was actually reduced by inclusion of the treated group and the values we have obtained for reliability were not therefore increased by the choice of sample. The sample size was too small to allow any meaningful assessment of treatment effect.

The impact of less than perfect reliability on sample size estimations for treatment trials is illustrated by the following equation[28]:

$$n = n^{\star}/R$$

where $n^{\star}$ is the sample size per group based on a perfect measurement technique, R is the reliability defined as the ICC, and n is the sample size per group incorporating the effects of measurement error. With values we have found for reliability >0.9, the effect of measurement error on sample size requirements is clearly small with both segmentation techniques (measurement error would necessitate an increased sample size in each arm of <11%). This reflects the wide distribution within the sample for the change in lesion load and might suggest that optimal precision may not be an imperative. However, in a more homogenous population or with a shorter interval between serial studies, the significantly higher precision of the contour method might be reflected in a more substantial difference in reliability between the two techniques, and using the more reliable segmentation method is clearly appropriate.

It is also important to stress that additional sources of variance such as image acquisition methodology and lesion identification have not been considered in the above equation and their impact on the overall reliability of measurements of the change in lesion volume is likely to be appreciable. An accurate estimate of sample size requirements must reflect the influence of all potential sources of variation in measurements. More work is needed to define the contribution of each factor on the reliability of the whole process of image acquisition and analysis.

A major disadvantage of both quantitative techniques used in this study is the high level of human interaction that they necessitate. Definitive phase III treatment trials may require analysis of many images and both lesion identification and segmentation can take months to perform. Several automated quantitative techniques have recently been developed using multiparametric approaches to perform lesion segmentation.[12 13 15] These offer the potential for considerably greater efficiency, but the significant presence of motion artefacts, field

J Neurol Neurosurg Psychiatry: first published as 10.1136/jnnp.65.1.42 on 1 July 1998. Downloaded from http://jnnp.bmj.com/ on June 28, 2022 by guest. Protected by copyright.

inhomogeneity within images and partial volume effects can cause errors in classification of lesions with such automated techniques. Any inconsistency in classification of lesions on serial images will result in inaccurate assessment of the change in lesion volume. Such techniques must therefore be validated by showing that they can tolerate the presence of artefact and remain responsive to real changes in lesion load over time. Despite the considerable time requirements that lesion identification on serial images demands with the contour technique, human intervention at this stage minimises the risk of misclassification. Furthermore, if serial images are assessed together, consistent decisions can be made on whether or not to classify equivocal areas of high intensity as lesions. The contour method therefore utilises both the ability of an experienced observer to discriminate between lesion, artefact and normal anatomy, and a higher degree of precision in lesion delineation than is possible with the fully manual technique.

Although the contour technique has been shown to be more precise than manual tracing of the lesion boundary, the algorithm still requires an observer to place the cursor at a point on the lesion edge. Lesions may have poorly defined edges due to the effects of partial volume. Several possible boundaries can be produced by the contour algorithm for less discrete lesions, depending on the exact position of cursor placement, and this significantly contributes to inconsistency in derived lesion volumes. Two approaches may further improve precision in serial studies. The first is to optimise lesion/background contrast and therefore reduce the amount of manual editing that is required. The fast FLAIR sequence utilises an inversion pulse to suppress high CSF signal intensity and is reported in some[22 31] but not all[32] cross sectional studies to improve precision with the contour method. Further studies are needed to consider the impact of this approach in serial studies. The second approach is to use a smaller slice thickness to minimise partial volume effects. One effect of finite slice thickness is to cause tissue mixing at the perimeter of lesions and produce loss of edge definition. As slice thickness is reduced, partial volume effects are less apparent and this may improve precision in quantification of lesion volume.[19 24 33] The increased acquisition time that imaging with smaller slice thickness requires can perhaps be offset by using faster pulse sequences such as fast spin echo.

In summary, we have shown that the contour technique represents a major improvement over manual outlining for lesion load quantification in terms of precision. Furthermore, the reliability was found to be better with the contour method, and in a more homogenous population this difference is likely to be even more apparent. These results support its further use in quantification on serial MRI, in which precision and reliability are essential requirements. Errors in measurement of the change in lesion load due to inconsistent scanner performance, suboptimal repositioning, and variability in lesion identification are likely to be more important than that due to the quantitative technique itself and the impact of these factors requires further evaluation.

We gratefully acknowledge the Multiple Sclerosis Society of Great Britain and Northern Ireland for financial support. We thank Dr L. Masuoka from Berlex Laboratories and the University of British Columbia multiple sclerosis/MRI Study Group for providing the images used in this study. PDM, AF, BG, and PR are supported by a grant from Schering AG.

1 Miller DH, Albert PS, Barkhof F, *et al.* Guidelines for the use of magnetic resonance techniques in monitoring the treatment of multiple sclerosis. *Ann Neurol* 1996;**39**:6–16.
2 Barkhof F, Filippi M, Miller D, *et al.* Strategies for optimising MRI techniques aimed at monitoring disease activity in multiple sclerosis treatment trials. *J Neurol* 1997;**244**:76–84.
3 Zhao GJ, Li DKB, Wolinsky JS, *et al.* Clinical and magnetic resonance imaging changes correlate in a clinical trial monitoring cyclosporine therapy for multiple sclerosis. *J Neuroimaging* 1997;**7**:1–7.
4 Paty DW, Li DKB, UBC MS/MRI Study Group. Interferon β-1b is effective in relapsing-remitting multiple sclerosis. MRI analysis results of a multcenter, randomised, double-blind, placebo-controlled trial. *Neurology* 1993;**43**:662–7.
5 Edan G, Miller D, Clanet M, *et al.* Therapeutic effect of mitoxantrone combined with methylprednisolone in multiple sclerosis: a randomised multicentre study of active disease using MRI and clinical criteria. *J Neurol Neurosurg Psychiatry* 1997;**62**:112–8.
6 Miller DH, Barkhof F, Berry I, *et al.* MRI in monitoring the treatment of MS: concerted action guidelines. *J Neurol Neurosurg Psychiatry* 1991;**54**:683–8.
7 Filippi M, Horsfield MA, Tofts PS, *et al.* Quantitative assessment of MRI lesion load in monitoring the evolution of multiple sclerosis. *Brain* 1995;**118**:1601–12.
8 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983;**33**:1444–52.
9 Noseworthy JH, Vandervoort MK, Wong CJ, *et al.* Interrater variability with the expanded disability status scale (EDSS) and functional systems (FS) in a multiple sclerosis clinical trial. *Neurology* 1990;**40**:971–5.
10 Plante E, Turkstra L. Sources of error in the quantitative analysis of MRI scans. *Magn Reson Imaging* 1991;**9**:589–95.
11 Goodkin DE, Ross JS, Medendorp SM, *et al.* MRI lesion enlargement in MS. Disease-related activity, chance occurrence, or measurement artifact? *Arch Neurol* 1992;**49**:261–3.
12 Clarke LP, Velthuizen RP, Camacho MA, *et al.* MRI segmentation: methods and applications. *Magn Reson Imaging* 1995;**13**:343–68.
13 Cline HE, Lorensen WE, Kikinis R, *et al.* Three-dimensional segmentation of MR images of the head using probability and connectivity. *J Comput Assist Tomogr* 1990; **14**:1037–45.
14 Wicks DAG, Tofts PS, Miller DH, *et al.* Volume measurement of multiple sclerosis lesions with magnetic resonance images. A preliminary study. *Neuroradiology* 1992:**34**;475–9.
15 Mitchell JR, Karlik SJ, Lee DH, *et al.* Computer-assisted identification and quantification of multiple sclerosis lesions in MR imaging volumes in the brain. *J Magn Reson Imaging* 1994;**4**:197–208.
16 Filippi M, Horsfield MA, Bressi S, *et al.* Intra- and inter-observer variability of brain MRI lesion volume measurements in multiple sclerosis. A comparison of techniques. *Brain* 1995;**118**:1593–600.
17 van Walderveen MAA, Barkhof F, Hommes OR, *et al.* Correlating MRI and clinical disease: relevance of hypointense lesions on short-TR/short-TE (T1-weighted) spin-echo images. *Neurology* 1995;**45**:1684–90.
18 Grimaud J, Lai M, Thorpe J, *et al.* Quantification of MRI lesion load in MS: a comparison of three computer-assisted techniques. *Magn Reson Imaging* 1996;**14**:495–505.
19 Filippi M, Rovaris M, Baratti C, *et al.* Intra- and interobserver variability in measuring brain MRI lesion volumes in multiple sclerosis: the contribution of 3 mm thick slices and fast FLAIR [abstract]. *Proceedings of the International Society for Magnetic Resonance in Medicine* 1996;**1**:540.
20 Poser CM, Paty DW, Scheinberg L, *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol* 1983;**13**:227–31.
21 Plummer DL. Dispimage: a display and analysis tool for medical images. *Rivista di Neuroradiologia* 1992;**5**:489–95.
22 Filippi M, Yousry T, Baratti C, *et al.* Quantitative assessment of MRI lesion load in multiple sclerosis: a comparison of conventional spin echo with fast fluid attenuated inversion recovery. *Brain* 1996;**119**:1349–55.
23 Gawne-Cain ML, Webb S, Tofts P, *et al.* Lesion volume measurement in MS: how important is accurate repositioning? *J Magn Reson Imaging* 1996;**6**:705–13.
24 van Waesberghe JHTM, Filippi M, Gawne-Cain M, *et al.* Reproducibility of measuring MRI lesion load postmortem in multiple sclerosis: influence of the scanners and of slice thickness [abstract]. *Proceedings of the International Society for Magnetic Resonance in Medicine* 1996;**1**:532.

J Neurol Neurosurg Psychiatry: first published as 10.1136/jnnp.65.1.42 on 1 July 1998. Downloaded from http://jnnp.bmj.com/ on June 28, 2022 by guest. Protected by copyright.

25 Zar J. Measures of dispersion and variability. In: Zar J, ed. *Biostatistical analysis. 2nd ed*. Englewood Cliffs, NJ: Prentice-Hall, 1984:27–39.
26 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10.
27 Bland JM, Altman DG. Measurement error. *BMJ* 1996;**312**: 1654.
28 Fleiss JL, ed. *The design and analysis of clinical experiments*. New York: John Wiley, 1985.
29 Streiner DL, Norman GR. In: Streiner DL, Norman GR, ed. *Reliability in health and measurement scales: a practical guide to their development and use*. New York: Oxford University Press 1995:104–27.
30 Armitage P, Berry G. *Statistical methods in medical research*. Oxford: Blackwell,1994.
31 Bastianello S, Bozzao A, Paolillo A, *et al.* Fast spin-echo and fast fluid-attenuated inversion-recovery versus conventional spin-echo sequences for MR quantification of multiple sclerosis lesions. *AJNR Am J Neuroradiol* 1997;**18**:699–704.
32 Gawne-Cain M, O'Riordan JI, Coles A, *et al*. MRI lesion volume measurement in MS and its correlation with disability: a comparison of fast flair and spin echo sequences. *J Neurol Neurosurg Psychiatry* 1998;**64**:197–203.
33 Filippi M, Horsefield MA, Campi A, *et al*. Resolution-dependent estimates of lesion volumes in MRI studies of the brain in multiple sclerosis. *Ann Neurol* 1995;**38**: 749–54.

J Neurol Neurosurg Psychiatry: first published as 10.1136/jnnp.65.1.42 on 1 July 1998. Downloaded from http://jnnp.bmj.com/ on June 28, 2022 by guest. Protected by copyright.