Automatic Quantification of MS Lesions in 3D MRI Brain Data Sets: Validation of INSECT

Alex Zijdenbos[1], Reza Forghani[1,2], and Alan Evans[1]

[1]McConnell Brain Imaging Centre and [2]Department of Neurology and Neurosurgery,
Montreal Neurological Institute, Montréal, Canada
{alex,reza,alan}@bic.mni.mcgill.ca
http://www.bic.mni.mcgill.ca

Abstract. In recent years, the quantitative analysis of MRI data has become a standard surrogate marker in clinical trials in multiple sclerosis (MS). We have developed INSECT (Intensity Normalized Stereotaxic Environment for Classification of Tissues), a fully automatic system aimed at the quantitative morphometric analysis of 3D MRI brain data sets. This paper describes the design and validation of INSECT in the context of a multi-center clinical trial in MS. It is shown that no statistically significant differences exist between MS lesion load measurements obtained with INSECT and those obtained manually by trained human observers from seven different clinical centers.

1 Introduction

Although the use of magnetic resonance imaging (MRI) as a qualitative clinical diagnostic tool in the study of multiple sclerosis (MS) has been established for well over a decade, it is only in recent years that its quantitative analysis is attracting interest. This attention is driven by, among other things, the increased use of MRI as a surrogate marker in clinical trials aimed at establishing the efficacy of drug therapies [1, 11]. A landmark study in this respect was the interferon beta-1b trial [15], which showed a correlation between MRI measured lesion load (quantified using manual boundary tracing) and clinical findings. This study clearly shows that manual tracing can be used to measure MRI lesion load with sufficient accuracy to detect a clinical effect; however, the disadvantages of this method are that it is very labour-intensive and that it suffers from high intra- and interrater variabilities.

A number of researchers have shown that computer-aided techniques are able to not only reduce operator burden, but also the inter- and intrarater variability associated with the measurement [5, 13, 26]. However, considering that the amount of MRI data to analyze in present-day clinical trials is often on the order of hundreds or thousands of scans, even minor manual involvement for each scan is an arduous task. The development of fully automatic analysis techniques is desirable to further reduce both the operator time requirements and the measurement variability.

At the McConnell Brain Imaging Centre (BIC), we have developed INSECT (Intensity Normalized Stereotaxic Environment for Classification of Tissues), a system aimed at the fully automatic quantification of tissue types in medical image data. This system has been used to automatically quantify MS lesion load in over 2000 MRI scans, acquired in the context of a large-scale, multi-center clinical trial [22]. Clearly, the thorough validation of results obtained using such an automated technique is crucial for its acceptance into clinical practice. Crucial elements of such validation studies are the assessment of accuracy and reproducibility. In the case of INSECT, results obtained on the same data are perfectly reproducible, which is a considerable advantage over manual lesion delineation. The fact that the analysis is reproducible does however not imply that its results are also accurate. The main focus of this paper is the validation of the accuracy of INSECT for the quantification of MS lesion load in MRI.

2 Methods

This section gives a brief overview of the INSECT processing pipeline, followed by a description of the validation studies performed to assess its accuracy for the automatic quantification of MS lesion load.

2.1 Image Processing

Fig. 1 shows the general architecture of INSECT. The central module of this system is the registration of the data with, and resampling into, a standardized, stereotaxic brain space based on the Talairach atlas [2, 21]. The registration component is preceded by a number of preprocessing algorithms aimed at artifact reduction, and followed by postprocessing algorithms such as intensity normalization and tissue classification. Each of these image processing components is briefly described here.

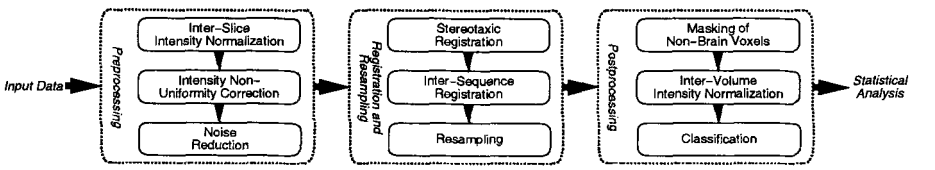

Fig. 1. INSECT flow diagram.

Inter-Slice Intensity Normalization: intensity variation between adjacent slices, an artifact generally attributed to eddy currents and crosstalk between slices, is a common problem in MRI [17, 24, 25]. INSECT employs a correction method in which the scaling factor between each pair of adjacent slices is estimated from

local (pixel-by-pixel) correction factors [25]. *Intensity Non-Uniformity Correction:* low-frequency spatial intensity variations, predominantly caused by electrodynamic interactions with the subject and inhomogeneity of the RF receiver coil sensitivity [8, 17, 18], are a major source of errors in the computer-aided, quantitative analysis of MRI data. A number of researchers have proposed algorithms to correct for this type of artifact [3, 10, 20, 23, 25]. INSECT employs a method developed in our institute [20], which was shown to be accurate and robust [19]. *Noise Reduction:* edge-preserving noise filters are often able to improve the accuracy and reliability of quantitative measurements obtained from MRI [12, 26]. INSECT relies on anisotropic diffusion, a filter commonly used for the reduction of noise in MRI [7, 16]. *Stereotaxic Registration*: this is accomplished by the registration of one of the image modalities to a stereotaxic target, using an automated 3D image registration technique [2]. This method minimizes, in a multi-stage, multi-scale approach, an image similarity measure between these two volumes as a function of a 9 parameter (3 translations, 3 rotations, 3 scales) linear geometric transformation. The stereotaxic target used at the BIC is an average T_1-weighted scan of 305 normal volunteers [2, 4, 6]. *Inter-Sequence Registration:* processing multi-modal (multi-feature) data typically requires the individual data volumes to be in exact spatial register, i.e., the feature values obtained from each modality at a specific voxel location should all reflect the same location in physical brain space. Since patient motion between different acquisitions is common, all scans of a patient or subject are explicitly registered with each other using the same technique as described for stereotaxic registration, with parameter values tailored to, in particular, the registration of a T_2-weighted scan to a T_1-weighted scan. *Resampling:* following stereotaxic and inter-scan registration, all data volumes are resampled onto the same voxel grid using trilinear interpolation. *Masking of Non-Brain Voxels*: for the application described herein, a standard brain mask, defined in stereotaxic space, is used. The fact that this 'average' brain mask may not accurately fit the individual brain does not affect the quantification of MS lesions, which are typically situated in the white matter well away from the cortical surface. *Inter-Volume Intensity Normalization:* given that all volumes at this stage are stereotaxically registered, they can be normalized using the same technique as described for the correction of inter-slice intensity variations (see [25]). In this case, a single, global intensity scale factor is estimated from the voxel-by-voxel comparison of each volume with a stereotaxic intensity model. *Tissue Classification:* for this application, INSECT employs a back-propagation artificial neural network (ANN) [26], which has been trained once to separate MS lesion from background (non-lesion). The classifier uses six input features, being the T_1-, T_2-, and PD weighted MRI volumes, as well as three (white matter, gray matter, CSF) SPAMs (Statistical Probability of Anatomy Maps), derived from normal human neuroanatomy (see [9]).

In order to process mass amounts of data, INSECT allows the user to specify this type of processing 'pipeline' using a high-level script language. During execution, each individual processing stage is submitted to a load-balancing queue, which distributes the job over a network of interconnected workstations. This

results in efficient mass-production using a high degree of (coarse-grain) parallelism.

2.2 Validation

The accuracy of INSECT has been assessed by means of two validation studies, in which automatic results were compared against those obtained from trained observers. *I. Independent Validation*: a total of 10 axial slice triplets (T_1-, T_2- and PD-weighted), each acquired from a different MS patient and at a different scanner, were selected from the data acquired for a multi-center clinical trial. Selection was such that the data reflect a reasonable range of lesion load and spatial distribution. The slices were extracted from data which was registered with and resampled into the BIC standard Talairach $1mm^3$ isotropic brain space. These data were distributed to seven different institutes for evaluation (see the acknowledgements), with the specific request to manually label all MS lesion pixels in each image triplet and return the binary label maps to the BIC. *II. BIC validation*: MS lesions were identified, using manual tracing, by four raters from the BIC community on a total of 29 MRI volume triplets (T_1-, T_2-, and PD-weighted). The raters, who all had substantial previous familiarity with neuroanatomy, were supervised by R.F. and trained for at least a month prior to data analysis. The criteria for lesion delineation were established by R.F. based on existing literature (e.g. [14]) and in collaboration with local neurologists and neuroradiologists.

The primary objective of this validation study is to test the hypothesis that there is no statistically significant difference between automatic and manual measurements, i.e., that the automatic lesion quantification can be seen as yet another expert manual measurement. In the following, the total lesion load (TLL) obtained automatically (INSECT) is compared with those obtained manually (human expert) using z-scores, correlation coefficients, and analysis-of-variance (ANOVA). For the ANOVA, the treatments, or groups, of the analysis are the various manual and the automatic measurements, and the MRI data sets are the subjects. A one-way ANOVA shows whether the mean measurements (over subjects) of the treatments are equal.

3 Results and Discussion

3.1 Independent Validation

An example of the labeled lesion scans obtained in the independent validation study is shown in Fig. 2. This figure clearly illustrates the variability present amongst different expert observers.

Fig. 3 shows the total lesion load, for each of the 10 slices included in the independent validation study, both obtained manually (mean $\pm$ sd over 7 raters) and automatically. Since the measurement variance increases with lesion size, the vertical axis shows the cubic root of the TLL, which converts the volumetric TLL

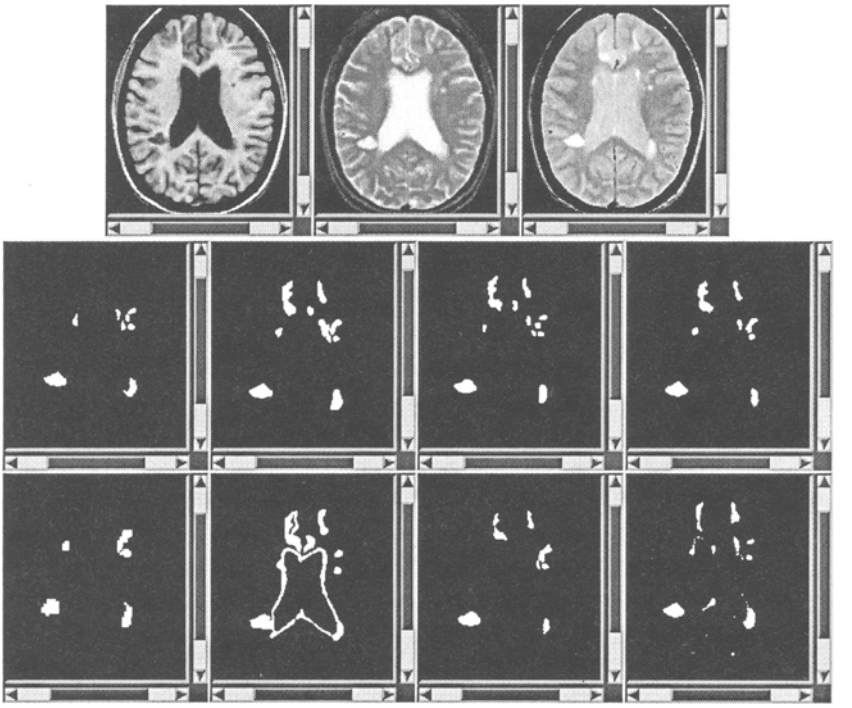

Fig. 2. Source MRI data (*top row, from left to right*: T_1-, T_2-, and PD-weighted image), (*second and third row*) labeled lesions obtained from 7 sites, and INSECT-obtained lesion map (*bottom right*), for slice data set # 10 (cf Fig. 3 and Table 1).

measurement to a 'lesion diameter' (this was only done for illustration purposes; all calculations are done on the original, non-transformed data). Table 1 shows these data in tabular form including the z-scores and associated p-values, for each slice, of the automatic TLL with respect to the mean and sd of the manual TLL. Clearly, the INSECT TLL is not significantly different from the average manual TLL, and is within one standard deviation from the mean on 9 out of 10 data sets. This is confirmed by the ANOVA on these data: F=1.03, p=0.42, indicating that none of the treatment groups is significantly different from any of the others. The interrater coefficient of variation (sd/mean over 10 slices) for the manual measurements is 44±20% (mean±sd).

As expected from these data, the correlation coefficient calculated between INSECT TLL and the average manual TLL is also very high: $r = 0.93$, $p < 0.0001$. It is also interesting to look at the correlations, over these 10 slices, between each pair of measurements. This is done in Table 2, which shows the significance levels of these correlations. From this table, it is clear that INSECT TLL measurements correlate significantly which the measurements made by any and all of the sites, whereas this is true for only 3 out of 7 sites. In other words,

INSECT measurements correlate on average better with manual measurements than most manual measurements correlate with each other.

The high interrater coefficient of variation obtained from this study is in part due to the fact that each of the sites used their own criteria for lesion selection. This illustrates that there is considerable disagreement among experts as to the identification of MS lesions on MRI, which in general confounds the accuracy assessment of computer-aided techniques.

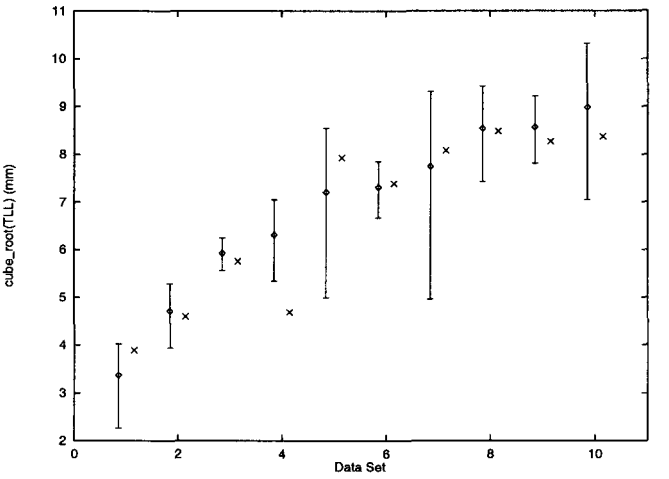

Fig. 3. Manual (mean ± sd) and automatic TLL measurements for each slice. The slices have been ordered with respect to mean manual TLL, which corresponds to the order used in Table 1. In order to obtain uniform variance accross data sets, the vertical axis shows 'lesion diameter' ($\sqrt[3]{\text{TLL}}$).

3.2 BIC Validation

Fig. 4 shows the data obtained from the 29-volume BIC validation (cf Fig. 3). effiche z-scores calculated from the data shown in Fig. 4 show that the INSECT TLL is within two standard deviations from the mean manual for 21 out of 29 (72%) volumes, and within one sd for 14 volumes (48%). Although there are a number of volumes for which the INSECT TLL deviates from the average manual TLL, overall this is not significant, as is shown by ANOVA: F=0.09 (p=0.97). Similar to the results obtained in the independent validation, the correlation coefficient calculated between INSECT TLL and average manual TLL is high: $r = 0.95$, $p < 0.0001$. In this case, the pairwise correlations between measurements are all highly significant ($p < 0.0001$). The inter-rater coefficient of variation (sd/mean) of the manual TLLs over the 29 volumes is 27±16% (mean±sd), showing the reduction in variability (down from 44±20%) when all raters adhere to the same lesion selection criteria.

Table 1. Manual (min, max, mean, sd, cv) and INSECT-based lesion volume measurements. Also reported are the z-scores and associated p-values of INSECT versus manual.

Slice	Manual (mm^3)					INSECT		
Data Set	min	max	mean	sd	cv	mm^3	z	p
1	0	82	38	27	70	59	0.77	0.44
2	43	168	105	43	41	97	-0.17	0.86
3	161	251	208	36	17	191	-0.48	0.63
4	140	442	252	99	39	103	-1.50	0.13
5	153	919	374	250	67	499	0.50	0.62
6	295	555	391	94	24	404	0.14	0.89
7	256	1235	465	343	74	528	0.18	0.86
8	403	893	624	214	34	611	-0.06	0.95
9	485	929	630	153	24	567	-0.41	0.68
10	394	1495	725	375	52	588	-0.37	0.71

Table 2. The significance of the correlation, over 10 slice data sets, between all combinations of manual and automatic measurements.

Site	INSECT	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7
1	***	-	*	***	****	****	*	***
2	**	*	-	*	*	n.s.	n.s.	**
3	**	***	*	-	****	**	*	**
4	****	****	*	****	-	***	*	**
5	**	****	n.s.	**	***	-	n.s.	**
6	*	*	n.s.	*	*	n.s.	-	n.s.
7	**	***	**	**	**	**	n.s.	-

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$

Based on the visual inspection of the 'outlier' volumes, where the INSECT TLL measurement was significantly different from the distribution of the manual measurements, a number of points can be made: 1) Many of the outliers occur in volumes with small amounts of lesion (see Fig. 4). If the image volume contains many small lesions and/or the image quality is poor, the human raters tended to be somewhat conservative in their measurements. 2) In the same circumstances, different human raters often identified different lesions on the same data set. 3) The criteria for lesion identification that INSECT implicitly uses, are necessarily not identical to those used by the human raters. By studying individual cases and longitudinal scan sequences, we were able to confirm that, although its measurements may deviate from the average manual measurements, INSECT was very consistent (see Fig. 5). In clinical trials, this type of reproducibility is crucial for determining treatment efficacy.

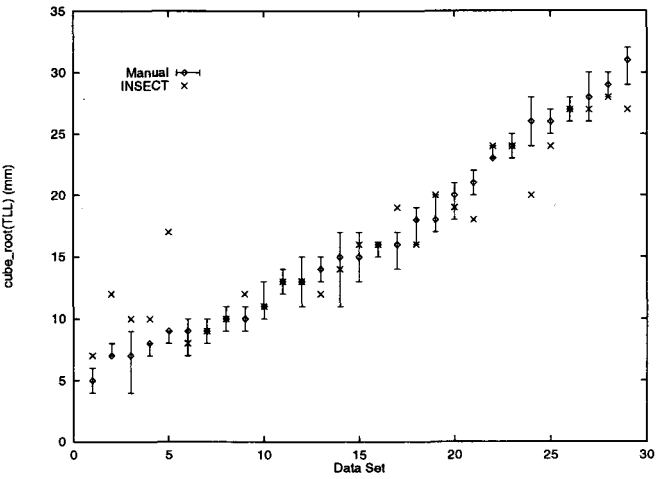

Fig. 4. Manual (mean ± sd) and automatic TLL measurements for each volume. The volumes have been ordered with respect to mean manual TLL. In order to obtain uniform variance accross data sets, the vertical axis shows 'lesion diameter' ($\sqrt[3]{\text{TLL}}$)

4 Conclusion

Validation studies of INSECT, a fully automatic system for the mass quantitative analysis of MRI data, focused on the detection of multiple sclerosis lesions, have been presented. These studies consistently show that there is high degree of agreement between automatically and manually detected lesion load in MRI. Using ANOVA, no statistically significant differences between automatic and manual lesion volume measurements were found. Since INSECT is accurate as compared with expert opinion and eliminates intra- and interobserver variability, it is a valuable tool for the evaluation of drug therapies in large-scale clinical studies.

Acknowledgements

The authors would like to express their gratitude to the people who have manually labeled MS lesions on the validation data sets. For the independent validation, these are: *Dr. Frederik Barkhof*, Dept. of Radiology, Free University Hospital, Amsterdam, The Netherlands; *Dr. Massimo Filippi*, Clinica Neurologica, Ospedale San Raffaele, Milan, Italy; *Dr. Joseph Frank*, Laboratory of Diagnostic Radiology Research, NIH, Bethesda, MD, U.S.A.; *Dr. Robert Grossmann*, Dept. of Radiology, University of Pennsylvania, Philadelphia, PA, U.S.A.; *Dr. Charles Guttmann*, Dept. of Radiology, Brigham and Women's Hospital, Boston, MA, U.S.A.; *Dr. Stephen Karlik*, Dept. of Diagnostic Radiology & Nuclear Medicine, University of Western Ontario, London, Ontario, Canada; and *Dr. David Miller*, Institute of Neurology, Queens Square, University of London, London, U.K. For the BIC validation: *Hooman Farhadi*, *Reza Forghani*, *Dana Small*, and *Dr. Selva Tekkök*. The authors also thank *Dr. Gordon Francis* and *Dr. Denis Melançon* for their assistance in establishing the criteria for lesion delineation.

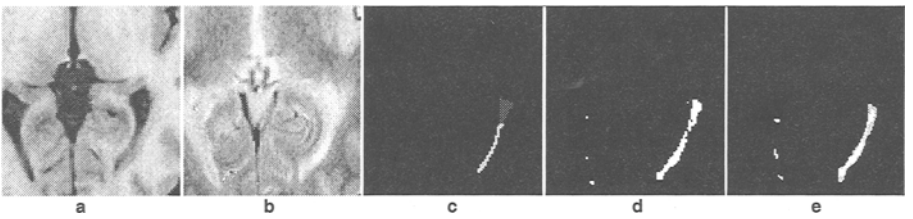

Fig. 5. *a and b*: T_1- and PD-weighted MRI; *c*: manual labeling at baseline scan (average of 4 raters); *d*: INSECT labeling at baseline; *e*: average INSECT labeling (over a series of 9 scans, taken 3 months apart). The brightness in the 'average' lesion labelings is proportional to the frequency of voxel selection. Panel *c* shows that, on the baseline scan, at most 2 of the 4 raters identified what INSECT labeled as a large, elongated lesion (*d*). However, panel *e* shows that INSECT consistently identified that same lesion on 9 consecutive quarterly scans. This volume corresponds to data set # 5 in Fig. 4.

References

1. F. Barkhof, M. Filippi, D. H. Miller, P. Tofts, L. Kappos, and A. J. Thompson. Strategies for optimizing MRI techniques aimed at monitoring disease activity in multiple sclerosis treatment trials. *Journal of Neurology*, 244(2):76–84, Feb 1997.
2. D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, 18(2):192–205, Mar./Apr. 1994.
3. B. M. Dawant, A. P. Zijdenbos, and R. A. Margolin. Correction of intensity variations in MR images for computer-aided tissue classification. *IEEE Transactions on Medical Imaging*, 12(4):770–781, Dec. 1993.
4. A. Evans, M. Kamber, D. Collins, and D. MacDonald. An MRI-based probabilistic atlas of neuroanatomy. In S. D. Shorvon et al., editors, *Magnetic Resonance Scanning and Epilepsy*, chapter 48, pages 263–274. Plenum Press, 1994.
5. A. C. Evans, J. A. Frank, J. Antel, and D. H. Miller. The role of MRI in clinical trials of multiple sclerosis: Comparison of image processing techniques. *Annals of Neurology*, 41(1):125–132, Jan. 1997.
6. A. C. Evans, S. Marrett, P. Neelin, et al. Anatomical mapping of functional activation in stereotactic coordinate space. *NeuroImage*, 1:43–53, 1992.
7. G. Gerig, O. Kübler, R. Kikinis, and F. A. Jolesz. Nonlinear anisotropic filtering of MRI data. *IEEE Transactions on Medical Imaging*, 11(2):221–232, June 1992.
8. R. M. Henkelman and M. J. Bronskill. Artifacts in magnetic resonance imaging. *Reviews of Magnetic Resonance in Medicine*, 2(1):1–126, 1987.
9. M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis, and A. C. Evans. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. *IEEE Transactions in Medical Imaging*, 14(3):442–453, Sept. 1995.
10. C. R. Meyer, P. H. Bland, and J. Pipe. Retrospective correction of intensity inhomogeneities in MRI. *IEEE Transactions on Medical Imaging*, 14(1):36–41, Mar. 1995.
11. D. H. Miller, P. S. Albert, F. Barkhof, G. Francis, J. A. Frank, S. Hodgkinson, F. D. Lublin, D. W. Paty, S. C. Reingold, and J. Simon. Guidelines for the use of magnetic resonance techniques in monitoring the treatment of multiple sclerosis. *Annals of Neurology*, 39:6–16, 1996.

12. J. R. Mitchell, S. J. Karlik, D. H. Lee, M. Eliasziw, G. P. Rice, and A. Fenster. Quantification of multiple sclerosis lesion volumes in 1.5 and 0.5T anisotropically filtered and unfiltered MR exams. *Medical Physics*, 23(1):115–126, Jan. 1996.
13. J. R. Mitchell, S. J. Karlik, D. H. Lee, and A. Fenster. Computer-assisted identification and quantification of multiple sclerosis lesions in MR imaging volumes in the brain. *Journal of Magnetic Resonance Imaging*, pages 197–208, Mar./Apr. 1994.
14. I. E. C. Ormerod, D. H. Miller, W. I. McDonald, et al. The role of NMR imaging in the assessment of multiple sclerosis and isolated neurological lesions. *Brain*, 110:1579–1616, 1987.
15. D. W. Paty, D. K. B. Li, UBC MS/MRI Study Group, and IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. *Neurology*, 43:662–667, 1993.
16. P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, July 1990.
17. A. Simmons, P. S. Tofts, G. J. Barker, and S. R. Arridge. Sources of intensity nonuniformity in spin echo images. *Magnetic Resonance in Medicine*, 32:121–128, 1994.
18. J. G. Sled and G. B. Pike. Standing-wave and RF penetration artifacts caused by elliptic geometry: an electrodynamic analysis of MRI. *IEEE Transactions on Medical Imaging*, 1997. (submitted).
19. J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A comparison of retrospective intensity non-uniformity correction methods for MRI. In *Proceedings of the 15th International Conference on Information Processing in Medical Imaging (IPMI)*, pages 459–464, Poultney, VT, USA, June 1997.
20. J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1), Feb. 1998.
21. J. Talairach and P. Tournoux. *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging.* Thieme Medical Publishers, New York, NY, 1988.
22. H. L. Weiner. Oral tolerance for the treatment of autoimmune diseases. *Annual Review of Medicine*, 48(48):341–51, 1997. 75 refs, Review.
23. W. M. Wells III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15(4):429–442, Aug. 1996.
24. D. A. G. Wicks, G. J. Barker, and P. S. Tofts. Correction of intensity nonuniformity in MR images of any orientation. *Magnetic Resonance Imaging*, 11(2):183–196, 1993.
25. A. P. Zijdenbos, B. M. Dawant, and R. A. Margolin. Intensity correction and its effect on measurement variability in the computer-aided analysis of MRI. In *Proceedings of the 9th International Symposium and Exhibition on Computer Assisted Radiology (CAR)*, pages 216–221, Berlin, Germany, June 1995.
26. A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer. Morphometric analysis of white matter lesions in MR images: Method and validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724, Dec. 1994.