NeuroImage: Clinical 13 (2017) 264–270

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation?

Christine Egger[a,*], Roland Opfer[a,b], Chenyu Wang[c,d], Timo Kepp[b], Maria Pia Sormani[e], Lothar Spies[b], Michael Barnett[c,d], Sven Schippling[a]

[a] Neuroimmunology and Multiple Sclerosis Research, Department of Neurology, University Hospital Zurich and University of Zurich, Frauenklinikstrasse 26, CH-8091 Zurich, Switzerland
[b] jung diagnostics GmbH, Hamburg, Germany
[c] Sydney Neuroimaging Analysis Centre, Sydney, Australia
[d] Brain and Mind Centre, University of Sydney, Sydney, Australia
[e] Biostatistics Unit, Department of Health Sciences, University of Genoa, Genoa, Italy

ARTICLE INFO

Article history:
Received 8 July 2016
Received in revised form 17 November 2016
Accepted 18 November 2016
Available online 20 November 2016

Keywords:
Magnetic resonance imaging
Fluid-attenuated inversion recovery
Multiple sclerosis
Automated lesion segmentation
Inter-rater variability
Dice coefficient

ABSTRACT

Introduction: Magnetic resonance imaging (MRI) has become key in the diagnosis and disease monitoring of patients with multiple sclerosis (MS). Both, T2 lesion load and Gadolinium (Gd) enhancing T1 lesions represent important endpoints in MS clinical trials by serving as a surrogate of clinical disease activity. T2- and fluid-attenuated inversion recovery (FLAIR) lesion quantification - largely due to methodological constraints – is still being performed manually or in a semi-automated fashion, although strong efforts have been made to allow automated quantitative lesion segmentation. In 2012, Schmidt and co-workers published an algorithm to be applied on FLAIR sequences. The aim of this study was to apply the Schmidt algorithm on an independent data set and compare automated segmentation to inter-rater variability of three independent, experienced raters.

Methods: MRI data of 50 patients with RRMS were randomly selected from a larger pool of MS patients attending the MS Clinic at the Brain and Mind Centre, University of Sydney, Australia. MRIs were acquired on a 3.0T GE scanner (Discovery MR750, GE Medical Systems, Milwaukee, WI) using an 8 channel head coil. We determined T2-lesion load (total lesion volume and total lesion number) using three versions of an automated segmentation algorithm (Lesion growth algorithm (LGA) based on SPM8 or SPM12 and lesion prediction algorithm (LPA) based on SPM12) as first described by Schmidt et al. (2012). Additionally, manual segmentation was performed by three independent raters. We calculated inter-rater correlation coefficients (ICC) and dice coefficients (DC) for all possible pairwise comparisons.

Results: We found a strong correlation between manual and automated lesion segmentation based on LGA SPM8, regarding lesion volume (ICC = 0.958 and DC = 0.60) that was not statistically different from the inter-rater correlation (ICC = 0.97 and DC = 0.66). Correlation between the two other algorithms (LGA SPM12 and LPA SPM12) and manual raters was weaker but still adequate (ICC = 0.927 and DC = 0.53 for LGA SPM12 and ICC = 0.949 and DC = 0.57 for LPA SPM12). Variability of both manual and automated segmentation was significantly higher regarding lesion numbers.

Conclusion: Automated lesion volume quantification can be applied reliably on FLAIR data sets using the SPM based algorithm of Schmidt et al. and shows good agreement with manual segmentation.

1. Introduction

Multiple sclerosis (MS) is an inflammatory and neurodegenerative disease of the central nervous system, in which focal and more widespread neuro-axonal loss culminates in neurological disability (Compston and Coles, 2008). Characteristic changes on magnetic resonance imaging (MRI) includes (symptomatic and/or asymptomatic) T2 hyperintense or T1 hypointense white matter (WM) lesions; and, using non-conventional MRI-techniques such as double inversion recovery (DIR) sequences (Filippi et al., 2012; Miller et al., 2014), focal hyperintense lesions within the grey matter (GM). Over the last two decades, MRI has become key not only for the diagnosis and monitoring of MS (Polman et al., 2011) but also as an endpoint for clinical trials, since T2 lesions next to Gadolinium enhancing T1 lesions are surrogates of clinical disease activity (Fahrbach et al., 2013; Sormani and Bruzzi, 2013). Against this background, the use of new and newly enlarging T2-lesions has recently been proposed to substitute relapses, under

* Corresponding author at: Neuroimmunology and Multiple Sclerosis Research, Department of Neurology, University Hospital Zurich and University of Zurich, Frauenklinikstrasse 26, CH-8091 Zurich, Switzerland.
E-mail address: christine.egger@usz.ch (C. Egger).

http://dx.doi.org/10.1016/j.nicl.2016.11.020

specific circumstances, as the primary endpoint in phase III clinical MS trials (Sormani and De Stefano, 2014).

Fluid attenuated inversion recovery (FLAIR) sequences suppress not only cerebrospinal fluid signal but also blood flow effects and thereby improve the detection of WM and also GM lesions (Gramsch et al., 2015), as compared to conventional T2 sequences. T2- and FLAIR lesion quantification, due to methodological constraints, is still largely being performed manually or in a semi-automated fashion; although both semi-automated and fully automated quantitative lesion segmentation approaches have been reported (Ashton et al., 2003; Filippi et al., 1995; Jain et al., 2015; Shiee et al., 2010; Udupa et al., 1997; Wicks et al., 1992). Semi-automated methods using seed-based region growing algorithms (Ashton et al., 2003), fuzzy connectedness (Udupa et al., 1997), or threshold-based methods (Filippi et al., 1995; Wicks et al., 1992) may reduce intra-rater variability. Still, they are time consuming and do not seem suited for large clinical trials. Another limitation of these methods is the dependency on the rater's input: for example, manual selection of lesions for seed-based analysis may lead to increased levels of inter-rater variability.

In 2013, Garcia-Lorenzo and colleagues presented an overview of automated supervised and unsupervised segmentation algorithms available at that time (Garcia-Lorenzo et al., 2011; Shiee et al., 2010; Van Leemput et al., 1999) based on which they concluded that a robust, accurate and fully-automated lesion segmentation tool was still not available (Garcia-Lorenzo et al., 2013). At the same time, Schmidt and co-workers (Schmidt et al., 2012) released the lesion segmentation toolbox (LST) (http://www.applied-statistics.de/lst.htm) which runs under the Statistical Parametric Mapping (SPM8) software package (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). They presented results from a validation cohort of 52 MS patients and 18 controls with satisfactory results. In 2015, the same group released an updated version of the lesion segmentation toolbox running under the 2014 release of SPM (SPM12, www.fil.ion.ucl.ac.uk/spm/software/spm12/).

In this study, we aimed to test different versions of LST on MRI scans from an independent MS patient cohort, acquired on a different scanner to verify or falsify Schmidt et al.'s results. In addition, we investigated whether the variability of the LST algorithm (including recent updates) would be lower than the inter-rater variability of manual segmentation. This would qualify such an algorithm applicable in clinical practice without performance loss compared to the gold standard of manual T2 lesion delineation.

2. Material and methods

2.1. Ethics

The study was approved by the human research and ethics committee at the University of Sydney, Sydney, Australia. All patients provided written, informed consent.

2.2. Subjects

MRI data of 50 patients with RRMS (7 males and 43 females with a mean age of 36.5 years (stdev. 9.0 years) and an average disease duration of 7.52 years (stdev. 7.01 years)) were randomly selected from a larger pool of MS patients attending the MS Clinic at the Brain and Mind Centre, University of Sydney, Australia.

2.3. MRI

All MRI scans were acquired on the same 3.0T GE scanner (Discovery MR750, GE Medical Systems, Milwaukee, WI) using an 8 channel head coil. A 0.9 mm isotropic 3DT1 (IR-FSPGR, TR/TI/TE = 7.2/2.8/450 ms, flip angle = 12°) and a 3D CUBE FLAIR (TR/TE/TI = 8000/165/2179 ms, flip angle = 90°, acquisition steps (Freq./Phase) = 256/224, FOV = 240 mm, slice thickness = 1.2 mm, slice spacing = 0.6 mm) sequence were acquired. To facilitate manual segmentation of T2 hyper-intense lesions, CUBE FLAIR and 3DT1 were co-registered and resampled to axial orientation with 2 mm slice thickness. T2 hyperintense contouring was performed on resampled FLAIR images slice by slice, with references from co-registered T1 images. Binary brain lesion masks were created automatically after all regions-of-interest were delineated in the brain. The increased slice thickness was justified based on results of a pilot study comparing manual and automated segmentation on five original and down sampled data sets. No significant difference could be detected between total lesion volumes of both reconstructions (The mean lesion volume on these 5 data sets was 5.7 ml; 5.68 ml by manual segmentation, 5.69 ml by automated segmentation on 2 mm slices and 5.83 ml on 0.9 mm slices. The mean volumetric difference between the original and down sampled images was 0.32 ml). However, automated segmentation was performed on original MRI images.

2.4. Manual and automated segmentation

Three experienced raters (in the following named Hamburg (TK), Sydney (CW), and Zurich (CE)) manually and independently segmented the full MRI data set. All raters performed axial slice by slice contouring using MeVisLab software (MeVis Solutions AG, Bremen, Germany) in Hamburg and Zurich and manual contouring as implemented in JIM6.0 (http://www.xinapse.com/) in Sydney. Raters were blinded regarding the results of the respective other manual raters, as well as the results of all automated segmentations.

Automated lesion detection was performed with the former and latest lesion segmentation toolbox (LST) (http://www.applied-statistics.de/lst.htm) published by Schmidt and coworkers (Schmidt et al., 2012), both running under the SPM software package (http://www.fil.ion.ucl.ac.uk/spm/). The original LST was developed for the SPM 8 software package (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). In October 2014, a major update of the SPM software (SPM 12) containing substantial algorithmic improvements was released. In July 2015, the Schmidt group released a new version of the LST running under SPM 12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12) platform. In this study, we deployed three different automated lesion segmentation algorithms:

1) Lesion growth algorithm based on SPM 8 (LGA SPM8): this is the original algorithm of the first release of the LST. The algorithm first segments the T1 image into the main compartments of grey and white matter and then combines the result with the FLAIR intensities in order to calculate lesion belief maps. By thresholding these grey matter lesion believe maps with a pre-chosen initial threshold (kappa), an initial binary lesion map is obtained which is subsequently grown along voxels that appear hyperintense in the FLAIR image.
2) Lesion growth algorithms based on SPM12 (LGA SPM12): the main steps of the LGA algorithms remained unchanged. However, the underlying segmentation of the T1 image was replaced by an SPM12 based algorithm.
3) Lesion prediction algorithms based on SPM12 (LPA SPM12): the updated lesion segmentation toolbox also contains a second completely and newly developed algorithm, referred to as lesion prediction algorithm. The LPA requires a FLAIR image only and does not require the initial thresholding parameter kappa.

We used version 2.0.11 of the updated LST. In the original paper, Schmidt and colleagues recommend a kappa value of 0.3 as an optimal default parameter for the LGA (Schmidt et al., 2012). For LGA SPM8 and LGA SPM12, we also tested kappa values of 0.2 and 0.4.

2.5. Qualitative analysis

All three algorithms generate lesion probability maps as an output with voxel values between 0 and 1 that were down sampled (using tri-linear interpolation) to a 2 mm slice thickness in order to allow a voxel by voxel comparison with the manual lesion masks. We generated binary lesion masks for various probability thresholds *t* by setting all voxels above that threshold to 1 and to 0 otherwise. Using a connected component analysis (Thurfjell et al., 1992) we decomposed the resulting binary lesion maps into connected components representing separate lesions. Further, we removed lesion clusters containing <8 voxels (=2.56 mm^3) from the binary lesion map, because lesions below that size are not well defined and are very likely to be false positive findings of the algorithm. The threshold was chosen based on the smallest lesion size consistently detected by all three manual raters (3.2 mm^3). We defined the threshold of 2.56 mm^3 based on a conservative approach (80% of the smallest consistently detected lesion in our cohort), which is still less than the minimal lesion size suggested to define a new or enlarging lesion in a recent publication by Rovira et al. (Rovira et al., 2010). In the next step, we calculated the total lesion volume as well as the number of lesions for all 50 data sets. We compared three different measures with respect to the lesions masks generated by the three raters to the results of the automated algorithms: the absolute volume difference (in ml), the absolute difference in lesion number, and the dice coefficient (DC), which measures the degree of overlap between two binary lesion maps (Dice, 1945). More precisely, if *A* and *B* were two binary lesion masks, then the DC would be defined as $2 \cdot \frac{|A \cap B|}{|A|+|B|}$. We calculated these matrices for all the three rater pairs and used the mean of the three values (in the following called "raters") to simplify the comparison between raters and the results of the automated algorithms. Additionally, we also calculated sensitivity ($\frac{|A \cap B|}{|A|}$, if A is considered as the ground truth) and false positive rate ($\frac{B-|A \cap B|}{|B|}$) for all manual rater pairings (six values since each rater was taken as ground truth) and manual rater – algorithm pairings (only manual raters were taken as ground truth). To characterize the dependency of the deployed metrics on absolute lesion volumes, we performed the analysis for the whole cohort as well as for three subgroups with different levels of total lesion load. As suggested in a recent lesion segmentation validation study by Jain et al., 2015, we considered a subgroup of patients with a total lesion volume < 5 ml as low, between 5 and 15 ml as medium, and a lesion volume > 15 ml as high lesion load (Jain et al., 2015).

2.6. Statistical analysis

The total lesion volume and lesion number between different raters and algorithms was compared by means of intra-class correlation coefficients (ICC) (Koch, 1982), calculated using the *psy* package in R, and Bland-Altman analyses. Since ICCs are generally meaningful in case of normally distributed data and T2 lesion volumes are usually skewed, logarithmic values of all lesion volumes were calculated before entering the ICC and Bland-Altman analyses (Adams et al., 1999; Gasperini et al., 2001). To determine whether the variability between each rater differed from the variability between the raters and the algorithms, we deployed an unpaired Student's *t*-test to compare the DC between each rater with the DC between the raters and the algorithm. Since absolute volume differences and absolute differences in lesion numbers are not normally distributed, the Wilcoxon rank-sum test was used for comparisons. The Bland-Altman and ICC analyses were performed with the statistical software package R (version 3.2.0). All other tests were performed using the MATLAB 2014a Statistics and Machine Learning Toolbox.

3. Results

MRI scans showed a broad range of total FLAIR lesion volume with a median lesion volume (average of three independent raters) of 4.82 ml, with a minimum of 0.20 ml and maximum of 48.97 ml. For some of the following analyses, patients have been split put into three groups: Low lesion load (<5 ml, n = 26), medium lesion load (5-15 ml, n = 14), and high lesion load (>15 ml, n = 10).

3.1. Dice coefficients

To determine the accuracy of manual raters as well as automated lesion segmentation tools, we calculated DC for several lesion map comparisons. Fig. 1 shows the mean DC (across all 50 data sets) between raters and segmentation algorithms depending on a probability threshold *t* (*t* = 0–0.8) and kappa (0.2, 0.3, and 0.4) for LGA SPM8/SPM12 and LPA SPM12. The recommended default kappa value (k = 0.3) (Schmidt et al., 2012) delivered not only higher but also more stable DCs compared to k = 0.2 and k = 0.4 (Fig. 1) and was therefore used for all further comparisons. While LGA SPM12 and SPM8 were less susceptible to a threshold change, LPA showed an increased DC with higher thresholds. We found a probability threshold of 0.5 to deliver maximal DCs and therefore to be optimal for LPA, and $t = 0.4$ to be optimal for both LGA algorithms. These probability thresholds were used for all further analyses.

Table 1 summarizes means and standard deviations of all DCs between different lesion masks. The DCs for LGA SPM12 (0.53) and LPA SPM12 (0.57) comparisons with the average of all raters were significantly (p = 0.05) lower than the DCs between manual raters (0.66), whereas LGA SPM8 (DC = 0.60) showed no significant difference to manual raters. Notably, the significantly weaker performance of LGA SPM12 and LPA SPM12 appeared only in those groups of patients with low or medium but not high lesion load. Overall, LGA SPM8 clearly outperformed LGA SPM12 and LPA SPM12 when assessed by DCs. Independently of total lesion volumes, LGA SPM12 delivered the lowest DCs compared to both manual raters and the remaining automated algorithms.

3.2. Absolute volume differences and ICC

To investigate how precisely manual raters and automated tools are able to segment lesion volumes in MS patients, we calculated absolute volume differences and ICC for several pairs and compared Bland-Altman plots. Table 2 summarizes absolute volume differences [ml] between different lesion masks. The median absolute volume difference between the three raters ranged from 0.4 ml to 0.75 ml. The median absolute volume differences between the manual raters and the LGA SPM8 (0.68 ml), LGA SPM12 (0.93 ml), and LPA SPM12 (0.85 ml) was not significantly (p = 0.05) higher or lower than differences between manual raters (0.66 ml). LGA SPM12 and LPA SPM12 showed numerically higher volume differences (0.93 and 0.86 ml) than LGA SPM8 (0.68) when compared to the average manual rater.

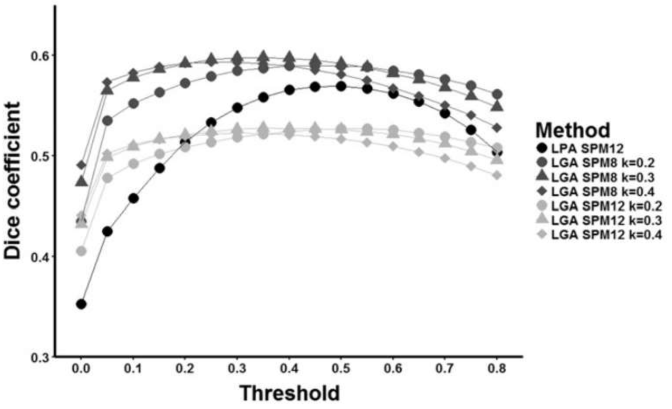

Fig. 1. Mean dice coefficients (over n = 50 data sets) depending on threshold (0–0.8) and kappa (0.2, 0.3, and 0.4) for LGA SPM8/SPM12 and LPA SPM12. For each data set and each algorithm the mean of the dice to the three raters was computed.

Table 1
Mean and standard deviation of dice coefficients (DC).

	All (n = 50)	<5 ml (n = 26)	5–15 ml (n = 14)	>15 ml (n = 10)
Ham-Syd	0.67 (0.12)	0.63 (0.15)	0.69 (0.06)	0.74 (0.08)
Ham-Zur	0.66 (0.13)	0.61 (0.16)	0.68 (0.06)	0.74 (0.07)
Zur-Syd	0.67 (0.12)	0.63 (0.14)	0.69 (0.07)	0.72 (0.09)
Raters	0.66 (0.12)	0.62 (0.14)	0.69 (0.06)	0.73 (0.08)
Raters-LGA SPM 8	0.60 (0.15)	0.53 (0.16)	0.65 (0.08)	0.70 (0.09)
Raters-LGA SPM 12	0.53* (0.16)	0.45* (0.18)	0.59* (0.09)	0.63 (0.10)
Raters-LPA	0.57* (0.16)	0.49* (0.17)	0.63 (0.10)	0.68 (0.11)

Means and standard deviations (in brackets) of DC between lesion masks generated by manual or automated segmentation. The numbers marked with an asterisk indicate DC which are significantly (p = 0.05) different from the DC of the manual raters ("Raters"). In columns 3–5 the same analysis was performed but restricted to groups with different total lesion volumes.

Fig. 2 shows ICC and Bland-Altman plots of all possible manual rater pairs with an ICC of 0.970 (Fig. 2A). When comparing the Bland-Altman plots, Sydney and Zurich showed the highest level of agreement (Fig. 2B).

We visualized the comparisons between the three automated segmentation tools and the average manual rater in Fig. 3. LGA SPM8 delivered the highest ICC (0.959) and the narrowest Bland-Altman plot with values that are comparable to those of manual rating. LPA SPM12 and LGA SPM12 showed lower ICCs of 0.949 and 0.927, respectively.

The averaged values of sensitivity and false positive rate are presented in Table 3.

3.3. Lesion number differences

Table 4 shows median and 95th percentiles of absolute lesion number differences. Manual raters showed a median difference in lesion numbers of 5.5–8.0 (mean = 8.0) with 95th percentiles of 20.0–52.0 (mean = 40). The performance of LGA SPM8 was comparable to those of the three manual raters (8.67 (54.00)) while LGA SPM12 and LPA SPM12 delivered significantly higher values. The calculated ICCs (figures not shown), with 0.901 for manual raters, 0.743 for LGA SPM8, 0.596 for LGA SPM12, and 0.701 for LPA SPM12, were clearly lower than the ICCs for lesion volume comparisons. The Bland-Altman plots indicate that all automated segmentation methods lost precision with increasing lesion numbers (figures not shown) and the inter-rater variability of manual segmentation increased.

4. Discussion

Defining and validating a gold standard for automated lesion segmentation in MS would be of utmost relevance, not only regarding the expected decrease in time needed for manual evaluation but also regarding the notable reduction of bias when data is analyzed by different manual raters. Previous studies have reported improvement in manual intra- and inter-rater variability with semi-automated segmentation tools (Ashton et al., 2003; Filippi et al., 1995). However, the use of semi-automated segmentation tools is still time consuming and largely relies on manual input. The validation of a fully automated, fast, and reliable tool, applicable in MS clinical trial settings, therefore appears paramount. Our aim was to validate a previously published algorithm of Schmidt and colleagues and to compare variability of fully automated versus manual segmentation. Within this comparison, we did not restrict our analysis to lesion volumes, but also analyzed the performance of automated lesion segmentation with respect to lesion numbers.

Table 2
Median and 95th percentiles of absolute volume differences in ml.

	All (n = 50)	<5 ml (n = 26)	5–15 ml (n = 14)	>15 ml (n = 10)
Ham-Syd	0.64 (5.45)	0.27 (2.57)	0.86 (3.82)	3.75 (16.94)
Ham-Zur	0.75 (3.13)	0.47 (2.12)	1.02 (3.70)	1.65 (3.29)
Zur-Syd	0.40 (4.49)	0.28 (1.25)	0.39 (3.13)	1.97 (15.61)
Raters	0.66 (3.63)	0.39 (1.85)	0.76 (2.71)	2.63 (11.29)
Raters-LGA SPM 8	0.68 (7.13)	0.38 (3.14)	0.93 (3.75)	2.94 (9.53)
Raters-LGA SPM 12	0.93 (7.61)	0.47 (3.67)	1.55 (5.16)	5.56 (16.15)
Raters-LPA	0.85 (8.13)	0.49 (2.27)	1.40 (5.05)	3.77 (16.58)

Medians and 95th percentiles (in brackets) of absolute volume difference in ml. The numbers marked with an asterisk indicate values which are significantly (p = 0.05) different from the between differences of the manual raters ("Raters"). In columns 3–5 the same analysis was performed but restricted to groups with different levels of total lesion volume.

The default kappa value of 0.3 proved to be optimal in our study. However, as shown in Fig. 1, the results for different kappa values were comparable. This supports the assumption that the results presented here are generalizable and an optimization of such a parameter might not be essential in a multicenter setting in which different scanners are being used. Non-surprisingly, LGA SPM12 and SPM8 were relatively unsusceptible to changes of the probability threshold parameter, since approximately 90% of all voxels on probability maps feature a value of 1, as already shown in the original work by Schmidt and coworkers. However, The LPA SPM12 showed higher DCs with higher thresholds (Fig. 1). The DC reached a maximum value for $t = 0.5$ and then decreased again for t values > 0.5. It remains unclear whether a t-value of 0.5 would also result in optimal DC if MRI images were acquired on different MRI scanners.

4.1. Dice coefficient – accuracy of manual raters and automated tools

When assessed by DC the former SPM8 algorithm clearly outperformed the more recent SPM12 based algorithms. This might appear surprising since users of the LST would expect better performance in the updated version of the toolbox. Numerically, manual raters reached a slightly higher level of agreement between each other as compared to the LGA SPM8 algorithm. However, these differences were not statistically significant. The differences were much smaller in the group of patients with medium or high lesion load. As reported previously (Schmidt et al., 2012), we found increasing DCs with increasing lesion loads. This was the case both for the between-rater comparisons and for the comparison between manual and automated segmentation. A possible explanation might be that a disagreement on a single voxel has a higher impact on overall agreement if the total segmented voxel number is low. Therefore, one might question whether DCs are an appropriate parameter to describe the accuracy of lesion segmentation in patients with a very low lesion load.

Previous studies (Jain et al., 2015) reported DCs in a similar range compared to our study (0.55 for LGA SPM 8, 0.67 for MSmetrix, and 0.61 for Lesion TOADS). As opposed to the study by Jain et al., >50% of patients in our study had a lesion load below 5 ml whereas in their study only 15% of patients (3/20) had similarly low lesion volumes. Against this background, the DC presented in our study might even be underestimated in comparison to the study by Jain et al. due to the high number of patients with very low lesion loads.

4.2. Absolute volume differences and ICC - precision of manual raters and automated tools

LGA SPM8 clearly outperformed the new SPM12 based LGA and LPA algorithms for determination of lesion volumes. Both, LGA SPM12 and LPA SPM12 showed significantly higher volume differences to manual raters than LGA SPM8.

The between-rater variability in assessing total lesion volumes (range 0.40–0.75, median 0.66 ml) was as high as between raters and LGA SPM8 (0.68). Therefore, the assessment of total lesion volume by manual rating and automated detection with LGA SPM8 appears interchangeable. The fact that LGA SPM8 showed ICCs, false positive rates, and absolute volume

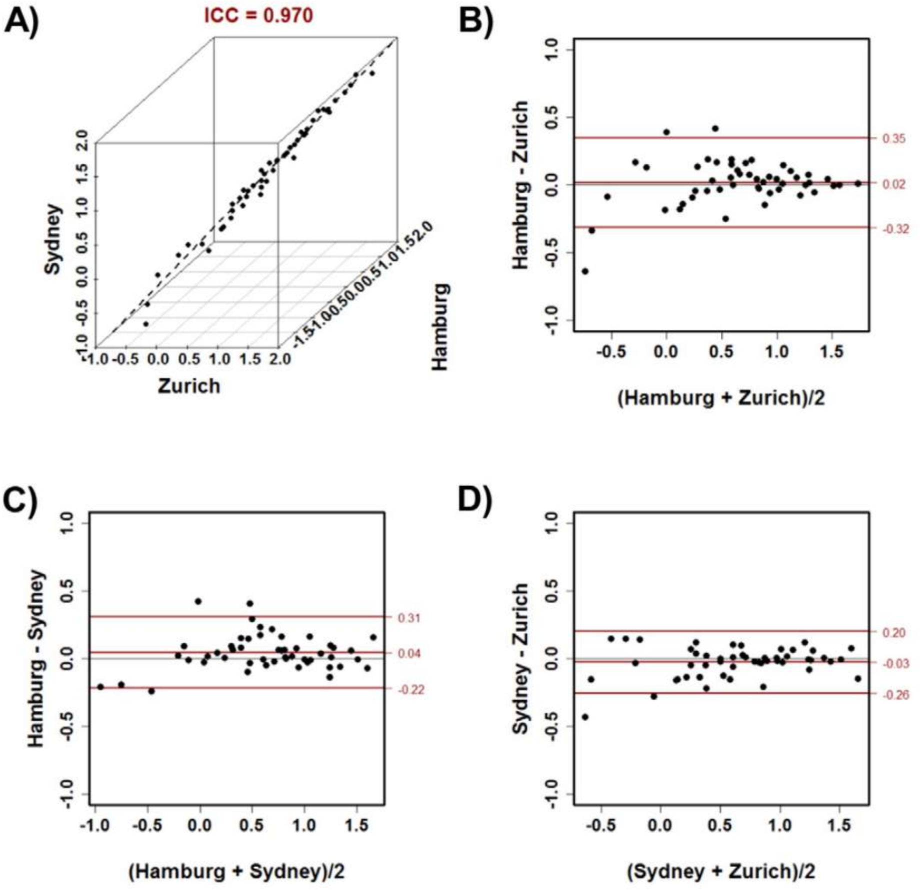

Fig. 2. Precision of three independent manual raters regarding total lesion volumes [log(ml)] visualized by ICC (absolute agreement) (A) and Bland-Altman plots for each pair (B–D). Total lesion volumes [ml] are shown as logarithmic values.

differences close to the inter-rater comparison supports our observation, whereas the sensitivity readouts suggest that LGA SPM8 might perform a more conservative segmentation than manual raters. Interestingly, the absolute volume differences we received in our study were higher than those reported in the study by Jain et al. in 2015 (mean 4.75 ml + − 3.63) (Jain et al., 2015), whereas ICCs in our study (0.927–0.958) were higher than in the Jain paper (0.63–0.80) (Jain et al., 2015). The Pearson's squared correlation coefficient between manual raters and LGA SPM8 ($R^2 = 0.94$) was comparable to the value reported by Schmidt et al. ($R^2 = 0.93$) (Schmidt et al., 2012).

In multicenter MS clinical trials, as much as in routine clinical practice, MRI T2/FLAIR lesion load is assessed longitudinally. It is against this background that measures of inter-rater variability in T2 lesion segmentation are crucial to understand the intrinsic variability of the methodology used for segmentation. The 95th percentile of the absolute volume differences between manual raters in our study showed a range from 3.13 to 5.45 ml (mean 3.63 ml); a minimum volume difference in total lesion load detectable in longitudinal or cross-sectional studies with an error probability of 5% can therefore be estimated to be in the range of 3.60 ml.

In longitudinal studies the level of variability can be reduced significantly when two scans are assessed simultaneously using co-registration/subtraction approaches as has been suggested by others (Battaglini et al., 2014; Moraal et al., 2010).

4.3. Lesion number differences – lack of precision by manual and automated segmentation

Quantification of lesion numbers by both manual and automated segmentation performed worse than lesion volume quantification. Even if two manual raters showed a high agreement, comparison to the third rater underlines the variability of manual segmentation, which is comparable to the disagreement between algorithms and average manual performance. Again, LGA SPM8 showed values close to manual segmentation, while SPM12 based LGA and LPA showed significantly higher lesion number differences.

Of interest, all algorithms underestimated total lesion numbers in patients with a high lesion load (figures not shown). This might be explained by the difficulties in assessing confluent lesions, e.g. in the periventricular area.

5. Conclusion

In this study, we investigated the inter-rater variability of manual versus automated MRI FLAIR lesion segmentation in a set of MRI scans from 50 patients with MS. We further analyzed whether the variability of the LST algorithm published by Schmidt et al. (including recent updates) is smaller than the inter-rater variability of manual segmentation. We estimated the inter-rater variability of measuring total lesion load in FLAIR images to be 3.63 ml accepting an error probability of 5%.

LGA SPM8 shows variabilities for measuring volumetric lesion load and lesion number, which are comparable to the inter-rater variabilities. This qualifies LGA SPM8 to be used for volumetric lesion segmentation in clinical applications without performance loss compared to manual lesion volume segmentation when applied on FLAIR images.

Disclosures

Christine Egger and Chenyu Wang have nothing to disclose.

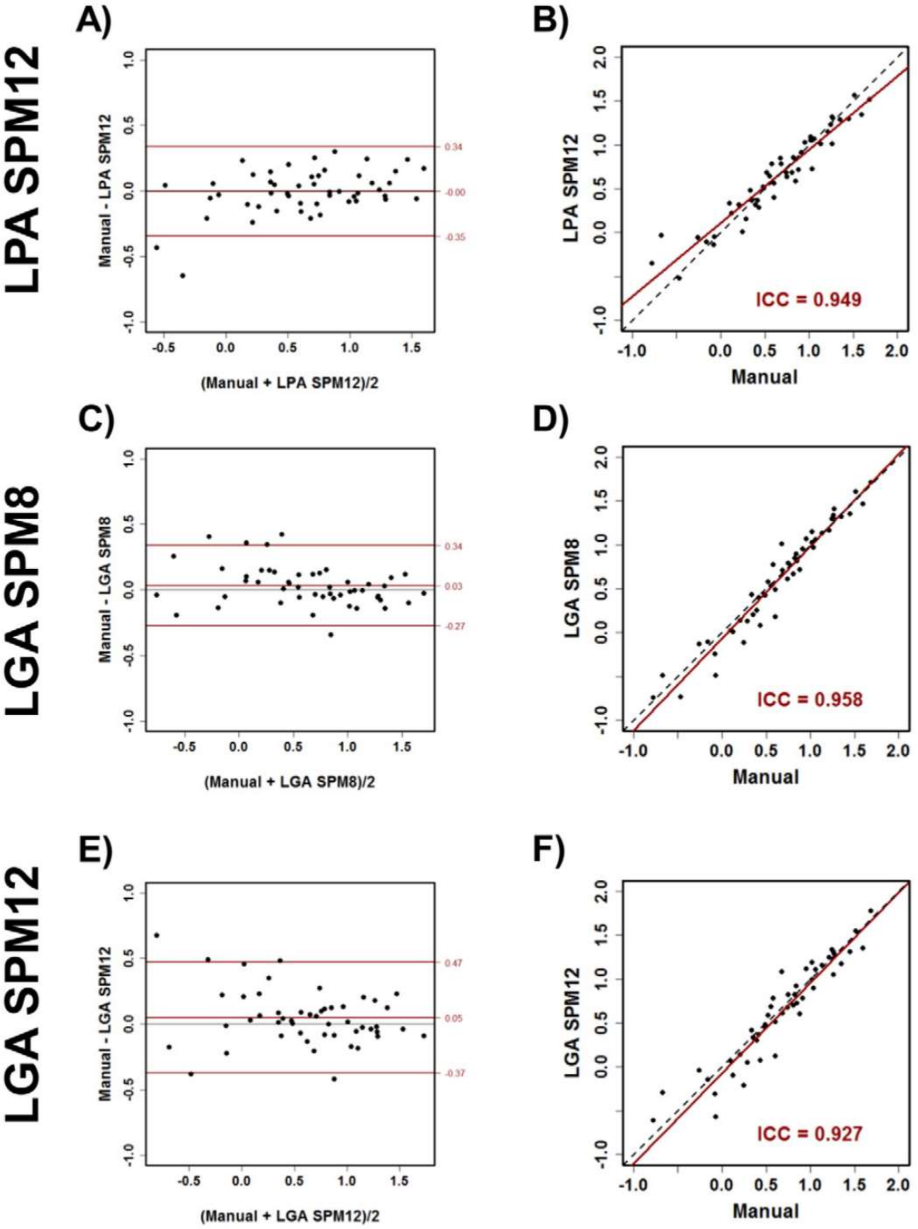

Fig. 3. Precision of three automated segmentation tools regarding total lesion volumes [log(ml)]; visualized by Bland-Altman plots for each pair (A,C,E) and ICC (absolute agreement) (B,D,F). Total lesion volumes are shown as logarithmic values and are compared to the averaged values of three manual raters.

Roland Opfer, Timo Kepp, and Lothar Spies are employees of jung diagnostics GmbH.

Maria Pia Sormani has received personal compensation for consulting services and for speaking activities from Merck Serono, Teva, Novartis, Roche, Genzyme and Biogen.

Table 3
Sensitivity and false positive rate.

	Sensitivity	False positive rate
Manual rater	0.676	0.324
LPA SPM12	0.628	0.457
LGA SPM8	0.566	0.335
LGA SPM12	0.505	0.393

Averaged sensitivity and false positive rate of all six possible manual rater pairings ("Manual rater"; each rater was taken as ground truth for each comparison) and algorithm-to-rater comparison (only manual rater were taken as ground truth).

Michael Barnett has received research support from Biogen, Novartis, and Sanofi Genzyme; and institutional support for participation in advisory boards from Biogen, Novartis, and Sanofi Genzyme.

Table 4
Median and 95th percentiles of absolute differences in lesion numbers.

	All (n = 50)
Ham-Syd	8.00 (49.00)
Ham-Zur	5.50 (20.00)
Zur-Syd	6.00 (52.00)
Raters	8.00 (40.00)
Raters-LGA SPM 8	8.67 (54.00)
Raters-LGA SPM 12	10.50* (63.00)
Raters-LPA	16.83* (45.67)

Median and 95th percentiles (in brackets) of absolute differences in lesion numbers. The numbers marked with an asterisk indicate the values which are significantly ($p = 0.05$) different from the differences between the manual raters ("Raters").

Sven Schippling has received research grants from Biogen, Bayer Healthcare, Novartis and Sanofi Genzyme, and consulting/speaker fees as well as travel support from Bayer Healthcare, Biogen, Merck Serono, Novartis, Teva, and Sanofi-Genzyme.

Acknowledgements

Sven Schippling is supported by the Clinical Research Priority Program of the University of Zurich, the Swiss Multiple Sclerosis Society, and the Betty and David Koetser Foundation for Brain Research.

References

Adams, H.P., Wagner, S., Sobel, D.F., Slivka, L.S., Sipe, J.C., Romine, J.S., Beutler, E., Koziol, J.A., 1999. Hypointense and hyperintense lesions on magnetic resonance imaging in secondary-progressive MS patients. Eur. Neurol. 42, 52–63.

Ashton, E.A., Takahashi, C., Berg, M.J., Goodman, A., Totterman, S., Ekholm, S., 2003. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. J. Magn. Reson. Imaging 17, 300–308.

Battaglini, M., Rossi, F., Grove, R.A., Stromillo, M.L., Whitcher, B., Matthews, P.M., De Stefano, N., 2014. Automated identification of brain new lesions in multiple sclerosis using subtraction images. J. Magn. Reson. Imaging 39, 1543–1549.

Compston, A., Coles, A., 2008. Multiple sclerosis. Lancet 372, 1502–1517.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Fahrbach, K., Huelin, R., Martin, A.L., Kim, E., Dastani, H.B., Rao, S., Malhotra, M., 2013. Relating relapse and T2 lesion changes to disability progression in multiple sclerosis: a systematic literature review and regression analysis. BMC Neurol. 13, 180.

Filippi, M., Horsfield, M.A., Bressi, S., Martinelli, V., Baratti, C., Reganati, P., Campi, A., Miller, D.H., Comi, G., 1995. Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis. A comparison of techniques. Brain 118 (Pt 6), 1593–1600.

Filippi, M., Rocca, M.A., Barkhof, F., Bruck, W., Chen, J.T., Comi, G., DeLuca, G., De Stefano, N., Erickson, B.J., Evangelou, N., Fazekas, F., Geurts, J.J., Lucchinetti, C., Miller, D.H., Pelletier, D., Popescu, B.F., Lassmann, H., Attendees of the Correlation between Pathological, M.R.I.f.i.M.S.w, 2012. Association between pathological and MRI findings in multiple sclerosis. Lancet Neurol. 11, 349–360.

Garcia-Lorenzo, D., Prima, S., Arnold, D.L., Collins, D.L., Barillot, C., 2011. Trimmed-likelihood estimation for focal lesions and tissue segmentation in multisequence MRI for multiple sclerosis. IEEE Trans. Med. Imaging 30, 1455–1467.

Garcia-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. 17, 1–18.

Gasperini, C., Rovaris, M., Sormani, M.P., Bastianello, S., Pozzilli, C., Comi, G., Filippi, M., 2001. Intra-observer, inter-observer and inter-scanner variations in brain MRI volume measurements in multiple sclerosis. Mult. Scler. 7, 27–31.

Gramsch, C., Nensa, F., Kastrup, O., Maderwald, S., Deuschl, C., Ringelstein, A., Schelhorn, J., Forsting, M., Schlamann, M., 2015. Diagnostic value of 3D fluid attenuated inversion recovery sequence in multiple sclerosis. Acta Radiol. 56, 622–627.

Jain, S., Sima, D.M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., De Mey, J., Barkhof, F., Steenwijk, M.D., Daams, M., Maes, F., Van Huffel, S., Vrenken, H., Smeets, D., 2015. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. Neuroimage Clin. 8, 367–375.

Koch, G.G., 1982. Intraclass correlation coefficient. Encyclopedia of Statistical Sciences.

Miller, T.R., Mohan, S., Choudhri, A.F., Gandhi, D., Jindal, G., 2014. Advances in multiple sclerosis and its variants: conventional and newer imaging techniques. Radiol. Clin. N. Am. 52, 321–336.

Moraal, B., Wattjes, M.P., Geurts, J.J., Knol, D.L., van Schijndel, R.A., Pouwels, P.J., Vrenken, H., Barkhof, F., 2010. Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. Radiology 255, 154–163.

Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F.D., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A.J., Waubant, E., Weinshenker, B., Wolinsky, J.S., 2011. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann. Neurol. 69, 292–302.

Rovira, A., Tintore, M., Alvarez-Cermeno, J.C., Izquierdo, G., Prieto, J.M., 2010. Recommendations for using and interpreting magnetic resonance imaging in multiple sclerosis. Neurologia 25, 248–265.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Forschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Muhlau, M., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. NeuroImage 59, 3774–3783.

Shiee, N., Bazin, P.L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49, 1524–1535.

Sormani, M.P., Bruzzi, P., 2013. MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. Lancet Neurol. 12, 669–676.

Sormani, M.P., De Stefano, N., 2014. MRI measures should be a primary outcome endpoint in phase III randomized, controlled trials in multiple sclerosis: yes. Mult. Scler. 20, 280–281.

Thurfjell, L., Bengtsson, E., Nordin, B., 1992. A new three-dimensional connected components labeling algorithm with simultaneous object feature extraction capability. CVGIP: Graphical Models and Image Processing. 54, pp. 357–364.

Udupa, J.K., Wei, L., Samarasekera, S., Miki, Y., van Buchem, M.A., Grossman, R.I., 1997. Multiple sclerosis lesion quantification using fuzzy-connectedness principles. IEEE Trans. Med. Imaging 16, 598–609.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18, 885–896.

Wicks, D.A., Tofts, P.S., Miller, D.H., du Boulay, G.H., Feinstein, A., Sacares, R.P., Harvey, I., Brenner, R., McDonald, W.I., 1992. Volume measurement of multiple sclerosis lesions with magnetic resonance images. A preliminary study. Neuroradiology 34, 475–479.