NeuroImage: Clinical 23 (2019) 101918

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

FLAIR^2 improves LesionTOADS automatic segmentation of multiple sclerosis lesions in non-homogenized, multi-center, 2D clinical magnetic resonance images

M. Le[a], L.Y.W. Tang[a,b], E. Hernández-Torres[c,d,i], M. Jarrett[c,e], T. Brosch[a,f,g], L. Metz[h], D.K.B. Li[a,b,d], A. Traboulsee[i], R.C. Tam[a,b], A. Rauscher[c,j,k], V. Wiggermann[c,d,k,*]

[a] MS/MRI Research Group (Division of Neurology), University of British Columbia, Vancouver, BC, Canada
[b] Department of Radiology, University of British Columbia, Vancouver, BC, Canada
[c] Department of Pediatrics, University of British Columbia, Vancouver, BC, Canada
[d] UBC MRI Research Centre, University of British Columbia, Vancouver, BC, Canada
[e] Population Data BC, Vancouver, BC, Canada
[f] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada
[g] Philips Medical Innovative Technologies, Hamburg, Germany
[h] Department of Clinical Neurosciences, University of Calgary, Calgary, AB, Canada
[i] Department of Neurology (Division of Medicine), University of British Columbia, Vancouver, BC, Canada
[j] BC Children's Hospital Research Institute, Canada
[k] Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada

ARTICLE INFO

Keywords:
FLAIR
FLAIR^2
Segmentation
Lesion volume
Performance evaluation
Multi-center

ABSTRACT

Background: Accurate segmentation of MS lesions on MRI is difficult and, if performed manually, time consuming. Automatic segmentations rely strongly on the image contrast and signal-to-noise ratio. Literature examining segmentation tool performances in real-world multi-site data acquisition settings is scarce.
Objective: FLAIR^2, a combination of T_2-weighted and fluid attenuated inversion recovery (FLAIR) images, improves tissue contrast while suppressing CSF. We compared the use of FLAIR and FLAIR^2 in LesionTOADS, OASIS and the lesion segmentation toolbox (LST) when applied to non-homogenized, multi-center 2D-imaging data.
Methods: Lesions were segmented on 47 MS patient data sets obtained from 34 sites using LesionTOADS, OASIS and LST, and compared to a semi-automatically generated reference. The performance of FLAIR and FLAIR^2 was assessed using the relative lesion volume difference (LVD), Dice coefficient (DSC), sensitivity (SEN) and symmetric surface distance (SSD). Performance improvements related to lesion volumes (LVs) were evaluated for all tools. For comparison, LesionTOADS was also used to segment lesions from 3 T single-center MR data of 40 clinically isolated syndrome (CIS) patients.
Results: Compared to FLAIR, the use of FLAIR^2 in LesionTOADS led to improvements of 31.6% (LVD), 14.0% (DSC), 25.1% (SEN), and 47.0% (SSD) in the multi-center study. DSC and SSD significantly improved for larger LVs, while LVD and SEN were enhanced independent of LV. OASIS showed little difference between FLAIR and FLAIR^2, likely due to its inherent use of T_2w and FLAIR. LST replicated the benefits of FLAIR^2 only in part, indicating that further optimization, particularly at low LVs is needed. In the CIS study, LesionTOADS did not benefit from the use of FLAIR^2 as the segmentation performance for both FLAIR and FLAIR^2 was heterogeneous.
Conclusions: In this real-world, multi-center experiment, FLAIR^2 outperformed FLAIR in its ability to segment MS lesions with LesionTOADS. The computation of FLAIR^2 enhanced lesion detection, at minimally increased computational time or cost, even retrospectively. Further work is needed to determine how LesionTOADS and other tools, such as LST, can optimally benefit from the improved FLAIR^2 contrast.

* Corresponding author at: UBC MRI Research Centre, Room M10, Purdy Pavilion, 2221 Wesbrook Mall, University of British Columbia, V6T2B5 Vancouver, Canada.
E-mail address: vwiggerm@phas.ubc.ca (V. Wiggermann).

https://doi.org/10.1016/j.nicl.2019.101918
Received 4 October 2018; Received in revised form 18 June 2019; Accepted 30 June 2019
Available online 05 July 2019

1. Introduction

MRI plays an integral role for diagnosis and monitoring of MS due to its sensitivity for the depiction of focal lesions, which are characteristically present in the brain and spinal cord of MS patients (Rovira & León, 2008; Fazekas et al., 1999; Wattjes et al., 2015; Rovira et al., 2009). The ability to assess efficacy of new disease modifying treatments relies on the identification of new T_2-hyperintense lesions, and the detection of inflammatory lesion activity (Wattjes et al., 2015; Miller et al., 1996; Sormani & Bruzzi, 2013; Río et al., 2017). However, detection and accurate demarcation of MS lesions on MRI is challenging due to heterogeneity in lesion location, size and shape in addition to anatomical differences between subjects (García-Lorenzo et al., 2013) and therefore requires expert knowledge. Manual lesion segmentation is labor-intensive, time-consuming and subject to intra- and inter-expert variability (García-Lorenzo et al., 2013; Grimaud et al., 1996; Zijdenbos et al., 2002; Styner et al., 2008). Recent supervised automated lesion segmentation methods have shown potential to provide lesion masks that closely match the manual expert segmentations (Brosch et al., 2016; Valverde et al., 2017), e.g. by utilizing neuronal networks, but rely on large training data sets, which are often not readily available. The need for training in supervised segmentation approaches and empirical selection of tuning parameters have hampered widespread application and validation of these tools, and hindered their use particularly in small scale studies. In depth discussions of different supervised and unsupervised methods are presented in reviews by García-Lorenzo et al. (2013), Lladó et al. (2012) and Sweeney et al. (2014). For widespread clinical and research applicability, lesion segmentation tools should be publicly available and ideally function without manual fine-tuning of processing parameters to facilitate reproducibility.

Thus, publicly available, automated methods that require no or minimal training data are of interest, including LesionTOADS (Shiee et al., 2010), the Lesion Segmentation Toolbox (LST) (Schmidt et al., 2012), Salem Lesion Segmentation (SLS) (Roura et al., 2015), or Automated Statistical Interference for Segmentation (OASIS) (Sweeney et al., 2013), which have been widely applied in reference to other lesion segmentation approaches (Brosch et al., 2016; Valverde et al., 2017; Jain et al., 2015; Roy et al., 2014; Valcarcel et al., 2018; de Sitter et al., 2017; Shinohara et al., 2017). In particular, LesionTOADS and LST provide ample reference for their lesion segmentation performance. Both are available as cross-platform software packages; LesionTOADS as a plug-in to the Java-based MIPAV toolbox, and LST as a plug-in to SPM, run within MATLAB. Notably, LesionTOADS also provides a segmentation of brain tissues, a functionality that extends its use to cortical segmentations and atrophy assessments (Huo et al., 2016; Harrison et al., 2015). Therefore, LesionTOADS has also found widespread application in clinical research of MS (Sati et al., 2012; Ozturk et al., 2010) and beyond (Lampe et al., 2019).

Nonetheless, automated segmentation approaches are also challenged by the heterogeneous MR appearance of MS lesions and therefore generally unable to match manual or semi-automated lesion definitions (García-Lorenzo et al., 2013). Thus, semi-automatic segmentation methods, e.g. automated, user-controlled region growing approaches based on manually placed seed points (McAusland et al., 2010), continue to remain the standard in clinical studies and provide the reference for newer, automated techniques. Noteworthy, supervised as well as unsupervised segmentation methods are often tested on small, single-site, homogenized imaging data (García-Lorenzo et al., 2013), which do not correspond to the real-world application of these approaches to multi-center, often multi-vendor data sets.

Here, our objective was to test the performance of LesionTOADS in a multi-center clinical trial using non-homogenized 2D-imaging data. We investigated whether LesionTOADS segmentation benefited from the use of FLAIR^2, a new contrast recently suggested to aid automated lesion segmentation methods. FLAIR^2-images are obtained through multiplication of co-registered T_2-weighted (T_2w) and FLAIR-images, both standard in MS imaging protocols (Traboulsee et al., 2016). The combination of 3D-T_2w and 3D-FLAIR, referred to as FLAIR^2, has shown to improve tissue contrast-to-noise, while simultaneously suppressing CSF (Wiggermann et al., 2016). Thus, FLAIR^2 may aid automated lesion segmentation methods, potentially also in cases of lower field strengths and non-3D image acquisitions.

The performance of LesionTOADS, when applied to FLAIR^2 in comparison to FLAIR in the multi-center study, was compared in a secondary analysis with the segmentations obtained from OASIS, a segmentation package publicly available within R, and LST for the same data set. Lastly, we contrasted our findings with the application of LesionTOADS to data obtained from a single-center, homogenized imaging study, in the challenging setting of low lesion load volumes in patients with clinically isolated syndrome (CIS).

2. Methods & materials

2.1. Demographics

MRI scans from a cohort of 47 relapsing-remitting MS patients, randomly selected from a multi-centre clinical trial performed at 34 different scanning sites, were included in this study. The second cohort consisted of 40 CIS patients, scanned at baseline at a single-site, prior to treatment randomization in a clinical trial (NCT00666887).

All patients gave written informed consent. Due to the blinded nature of the data analysis in these trials, no further demographic information was available.

2.2. MR image acquisition and processing

2D-FLAIR, 2D-T_2w, 2D-Proton Density weighted (PDw), and 3D-T_1 weighted (T_1w) scans with a variety of acquisition parameters were selected in the multi-center study in order to reflect the range of values used in MS imaging studies. All scans were acquired at a voxel size of $0.94 \times 0.94\,mm^2$ and 3.00 mm slice thickness, at either 1.5 and 3 T, except for T_1w-images, which were acquired with voxel sizes varying between $0.94 \times 0.94\,mm^2$ to $1 \times 1\,mm^2$ and slices of 1 or 1.5 mm thickness. For all T_2w, FLAIR and PDw-scans, 60 slices were collected; T_1w-images had more slices due to their reduced slice thickness (between 116 and 208). Detailed scan parameters for the multi-center study are provided in Supplementary Table S1. All data concerning the single-center study were acquired at a 3 T Philips Achieva. The voxel size was the same as in the multi-center study, $0.94 \times 0.94 \times 3\,mm^3$, for all image contrasts. Dual-echo PD/T_2w-images were acquired at $TE_1 = 8.4$ ms, $TE_2 = 80$ ms, TR = 2800 ms; FLAIR-images used TE = 125 ms, TR = 11 s, TI = 2800 ms, refocusing flip angle = 125°; and T_1w-images were acquired at TE = 10 ms, TR = 657 ms and flip angle = 50°.

For both studies, a bias correction was performed on all images using the revised N3 technique as described by Jones & Wong (2002), prior to further processing. The revised N3 techniques captures areas of steep inhomogeneity gradients, which are not fully corrected with N3 alone.

All images, T_1w, T_2w and FLAIR, were co-registered to the PDw-image space and brain extracted prior to LesionTOADS segmentation using FLIRT and BET, tools of the FSL software library (Jenkinson et al., 2002; Smith, 2002). For the purpose of the FLAIR^2-image computation, FLAIR-scans were also co-registered to the T_2w-images as described in Wiggermann et al. (2016). The aligned FLAIR and T_2w-scans were then multiplied, yielding the FLAIR^2-image, and subsequently mapped to PDw for lesion segmentation.

2.3. Semi-automated reference segmentation

Lesions identified by the semi-automatic method described in McAusland et al. (2010) were used in both studies for comparison with

LesionTOADS, OASIS or LST. A neuroradiologist identified lesions consistent with MS pathology on T_2w and PDw-scans and marked each lesion with a minimum of one lesion point. A technician then performed the semi-automatic growing process to create the reference lesion mask.

2.4. LesionTOADS

LesionTOADS is a topology-preserving segmentation tool (Shiee et al., 2010) designed to identify and segment white matter (WM) MS lesions while simultaneously classifying other brain tissues. Since LesionTOADS is optimized for FLAIR and T_1w, T_1w-images were included in addition to FLAIR or FLAIR^2, respectively. All LesionTOADS parameters remained at default, for both the FLAIR and FLAIR^2-image segmentation. The default parameters are listed in Inline Supplementary Table S2. For this work, the 2014 R4c version of TOADS CRUISE was downloaded from NITRC on September 28th 2018 and run within MIPAV version 7.0.1.

2.5. OASIS

OASIS uses logistic regression to estimate the voxel-level probability of lesion presence based on the voxel intensities on T_1w, T_2w, PDw and FLAIR-images (Sweeney et al., 2013). In contrast to LesionTOADS, OASIS uses the T_1w-image as reference and performs a non-linear registration to the MNI standard space using FSL tools. The user may provide original, unprocessed or pre-processed images to OASIS. In order to take full advantage of its pipeline, we used non co-registered, non-brain extracted images for OASIS. Although OASIS masks brain tissues from CSF prior to lesion detection, it does not yield further tissue class segmentations. The OASIS pipeline provides an already trained segmentation model, based on 15 MS patients and 5 healthy controls. However, study specific thresholding is recommended and study data training is possible. For best results, we trained OASIS on four data sets from our study, for FLAIR and FLAIR^2-data individually. OASIS version 3.0.4. was downloaded and installed within R 3.4.4 on February 13th 2019.

2.6. LST

Akin to LesionTOADS, the lesion segmentation toolbox (LST) (Schmidt et al., 2012) only requires T_1w and FLAIR or FLAIR^2-images as input. LST, like OASIS, pre-processes the FLAIR and T_1w-data. Tissue segmentation is performed on the T_1w-images and tissue probability labels are computed in combination with SPM's tissue probability map of WM. FLAIR-image intensities are subsequently used to create lesion belief maps for each tissue compartment, considering lesions to be intensity outliers within individual tissue classes. Finally, lesion growing is performed after thresholding of the gray matter lesion belief map with a pre-determined threshold. Based on the same four subjects used for OASIS training, we determined the thresholds of 0.19 and 0.24 for FLAIR and FLAIR^2, respectively. Note that although tissue segmentation on T_1w-images is performed, tissue masks are typically not provided as one of the LST outputs. LST version 2.0.15 was downloaded and installed within SPM12, downloaded on November 14th 2018.

2.7. Performance evaluation

We computed the commonly used relative lesion volume difference (LVD), the Dice coefficient (DSC), sensitivity (SEN) and the symmetric surface distance (SSD) to assess the performance of FLAIR versus FLAIR^2 for the different segmentation algorithms and studies. LVD represents the relative volume difference between the LesionTOADS, OASIS or LST and reference lesion segmentation. To assess the overlap between segmented lesion voxels, not captured by LVD, DSC and SSD were computed. SSD reflects the closeness of border voxels of the segmentation and the reference, while the DSC assesses the number of true positive lesion voxels compared to false positive and negative voxels. Additionally, SEN, another overlap measure which focuses only on the amount of true positive and false negative voxels, was estimated. All performance metrics are detailed in Table 1. In a secondary analysis, we stratified patients based on their absolute detected reference lesion volume (LV) and categorized them accordingly as patients of high ($> 15\,cm^3$), medium/low ($> 5\,cm^3$ & $< 15\,cm^3$) or minimal LV ($< 5\,cm^3$) to test for performance variation. For comparison with other studies, these volume thresholds were adapted from literature (Schmidt et al., 2012; Jain et al., 2015).

Table 1
Definition of the applied evaluation metrics, including the respective scoring systems and units. Note that the ratios for DSC and SEN are expressed in percentages. [Abbreviations: true positives (TP), false positives (FP), false negatives (FN), lesion volume of FLAIR or FLAIR^2 (VOL_{FLAIR}), lesion volume of reference segmentation (VOL_{REF}), Euclidean distance (d) on boundary voxels (∂).]

	Definition	Units	Best	Worst
LVD	$\frac{\lvert VOL_{FLAIR} - VOL_{REF}\rvert}{VOL_{REF}}$	a.u.	0	$+\infty$
DSC	$\frac{2 \cdot TP}{FP + FN + 2 \cdot TP}$	%	100	0
SEN	$\frac{TP}{TP + FN}$	%	100	0
SSD	$\frac{\sum_{v \in \partial_{SEG}} \min_{u \in \partial_{REF}} d(uv) + \sum_{v \in \partial_{REF}} \min_{u \in \partial_{SEG}} d(uv)}{card(SEG \cup REF)}$	mm	0	$+\infty$

2.8. Statistical analysis

All performance evaluation indices were calculated separately for each segmentation tool, comparing FLAIR and FLAIR^2 against the semi-automated reference. In addition, the mean difference between the FLAIR and FLAIR^2 scores as well as the relative improvement score ($mean_{diff}/mean_{FLAIR}$) for each metric were computed. Significance of the improvement was determined using a paired t-test. Furthermore, a mixed effects model was implemented in R (lme4 package (Bates et al., 2011)) to assess statistical differences in regard to the chosen input image (FLAIR/FLAIR^2) and the effect of LV for the multi-center study. The fixed model parameters were complemented with two random effects, addressing site and patient variability. The site parameter accounts for segmentation differences related to the site-specific imaging protocols in the multi-center study. Ultimately, pairwise multiple comparisons for the performance indiceswith respect to LV and the use of FLAIR or FLAIR^2 were carried out using the lsmeans package (Lenth, 2016), which performs a post-hoc Tukey's HSD test and p-value adjustment. Since the difference in segmentation performance between FLAIR and FLAIR^2 was mainly of interest, we did not statistically evaluate the performance differences between LesionTOADS, OASIS and LST.

3. Results

One patient's FLAIR-image was incorrectly reconstructed with non-zero signal outside of the brain, and thus no FLAIR^2-image could be obtained. Another patient's FLAIR-image exhibited strong motion-induced inter-slice misalignment (Tam et al., 2009), which prevented co-registration between FLAIR and T_2w. For the multi-center study, LesionTOADS failed to complete processing for two subjects on both the FLAIR and FLAIR^2-image. These data sets were excluded from further analysis, in addition to one data set for which the imaging parameters and field strength information were unavailable. Of the remaining 42 multi-center patients, 40 had been scanned at 1.5 T at 30 different sites and two data sets had been obtained from one 3 T site.

Fig. 1 shows representative lesion masks obtained from the semi-automated reference method (A) and the corresponding LesionTOADS segmentation based on FLAIR (B) and FLAIR^2 (C) on the respective

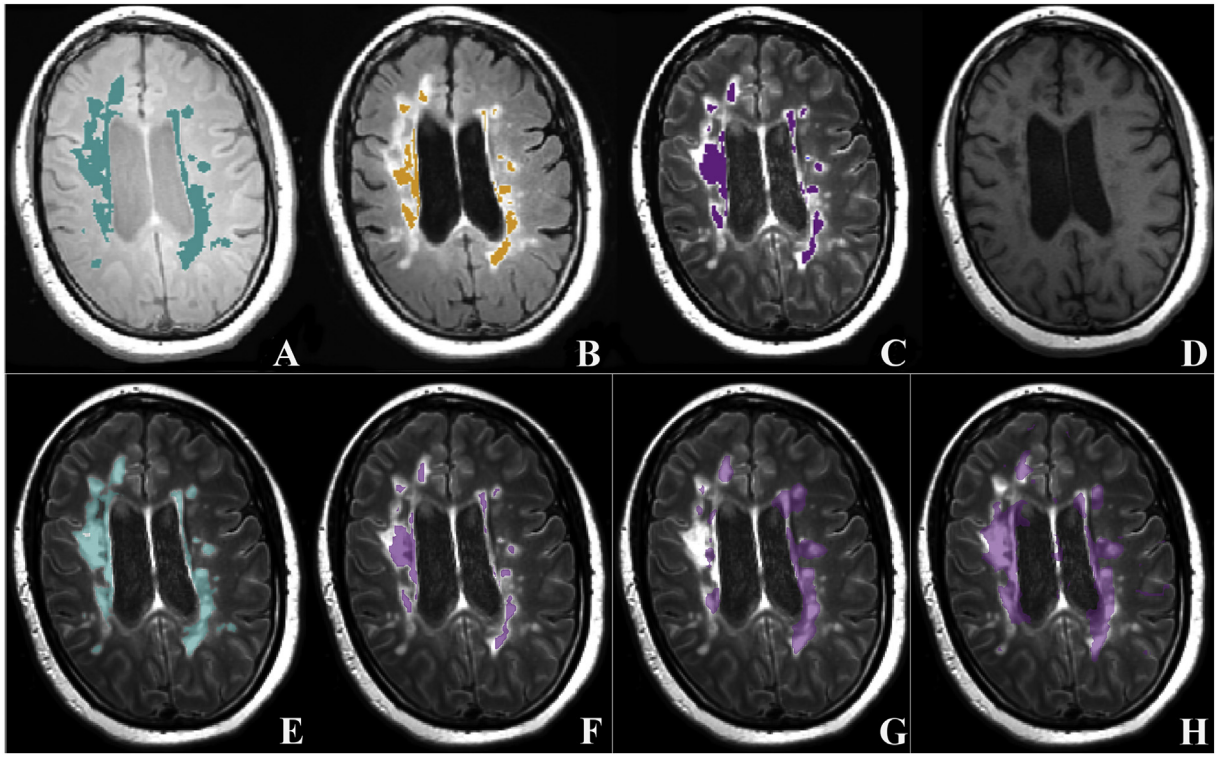

Fig. 1. Example LesionTOADS segmentations shown overlaid on the images that were used to obtain them (top row); Comparison of FLAIR^2-based segmentations obtained by different tools (bottom row). **Top row**: A) 2D-PDw and the lesion mask (green) obtained by semi-automated seed point selection and lesion growing; B) 2D-FLAIR and the default LesionTOADS segmentation (yellow); C) 2D-FLAIR^2 and the respective LesionTOADS segmentation (purple); D) matching 3D-T_1w used as additional input to LesionTOADS, OASIS and LST. **Bottom row**: Segmentations shown overlaid on FLAIR^2: E) PDw-reference; F) FLAIR^2-LesionTOADS segmentation; G) FLAIR^2-OASIS segmentation; H) FLAIR^2-LST segmentation. LesionTOADS segmented a larger lesion volume when using FLAIR^2 compared to FLAIR alone, achieving a better agreement with the PDw-based reference segmentation, albeit still underestimating the LV. OASIS and LST captured lesions, but also diffuse abnormalities, to a greater extent.

image contrasts for one patient in the multi-center study. The FLAIR^2-based lesion segmentation allowed for better capture of confluent lesions compared to the FLAIR-only segmentation, facilitated by the improved tissue contrast of FLAIR^2 between MS lesions and surrounding WM as well as between gray and subcortical WM. The bottom row displays the FLAIR^2-based segmentations of LesionTOADS (F), OASIS (G) and LST (H). Compared to LesionTOADS, both OASIS and LST segmented larger volumes, thereby capturing lesions more fully, but also fluently connected lesion areas. All tools failed to segment smaller lesions, particularly when they were not periventricular and presented with lesser hyperintensity.

The estimated LVs for all MS patients in the multi-center study were: semi-automated reference segmentation: mean/median/range = 11.1/7.4/0.14–48.7 cm^3; LesionTOADS FLAIR segmentation: mean/median/range = 4.2/3.2/0.01–15.9 cm^3; LesionTOADS FLAIR^2 segmentation: mean/median/range = 5.8/4.3/0.30–24.3 cm^3. In line with the example segmentation shown in Fig. 1, both FLAIR and FLAIR^2 LesionTOADS underestimated LVs compared to the semi-automated reference segmentation (Fig. 2A). However, FLAIR^2 significantly improved LV estimates compared to using FLAIR alone (p = .018, relative improvement 31.6%). In particular, we noticed an improved segmentation for LVs > 10 cm^3 when using FLAIR^2 compared to FLAIR, although a correct estimation of large LVs appears to be more challenging as noted by the increasing discrepancy between the LesionTOADS and reference segmentation estimated LVs.

The LesionTOADS results of all evaluation indices are presented in the top part of Table 2. For comparison, OASIS results are shown in the middle and LST results in the bottom third. For LesionTOADS, segmentation based on FLAIR^2 scored significantly higher than FLAIR in three of the four indices: LVD: $\text{mean}_{\text{diff}} = -0.2$ a.u. (p = .018); DSC: $\text{mean}_{\text{diff}} = 5.2\%$ (p = .19) with a relative improvement of 14%; SEN: $\text{mean}_{\text{diff}} = 6.98\%$ (p = .048) with a relative improvement of 25.1%; SSD: $\text{mean}_{\text{diff}} = -3.5$ mm (p = .0097), with the largest relative improvement of 47%.

To assess whether these improvements are specific to the use of FLAIR^2 within LesionTOADS, we applied the OASIS segmentation pipeline as well as LST to the same data set. A side-by-side comparison of the LesionTOADS, OASIS and LST segmentation performance, relative to the semi-automated reference, is shown in Fig. 3. Mean scores and relative improvements are summarized in Table 2.

DSC showed little discrepancy between the different segmentation approaches as well as between FLAIR and FLAIR^2, hence yielding the lowest relative improvements of all performance scores. OASIS showed more variability in LVD ad SSD, achieving on average lower performance than LesionTOADS in these two scores. LST similarly varied more in LVD and SSD, but this variability and the lower performance were limited to FLAIR^2. One exception is SEN, which was notably higher for both OASIS and LST. Comparison of FLAIR and FLAIR^2 with LST replicated the significant improvement in SEN seen with LesionTOADS (p = .031), however, SSD was significantly higher when using LST with FLAIR^2 compared to FLAIR (p = .0002).

Pairwise comparison of the OASIS scores obtained using FLAIR and FLAIR^2 yielded no significant differences (LVD p = .38, DSC p = .91, SEN p = .43, SSD p = .32), although LVD and SSD showed relative improvements for FLAIR^2 over FLAIR similar to LesionTOADS (LVD 55.8%, SSD 21.8% improvement).

A visual comparison of the resulting FLAIR^2-based segmentations from the three segmentation tools is provided in Fig. 4 for two subjects. LesionTOADS (left) generated more conservative segmentation results than OASIS (middle) and LST (right), consistently sparing lesion edges

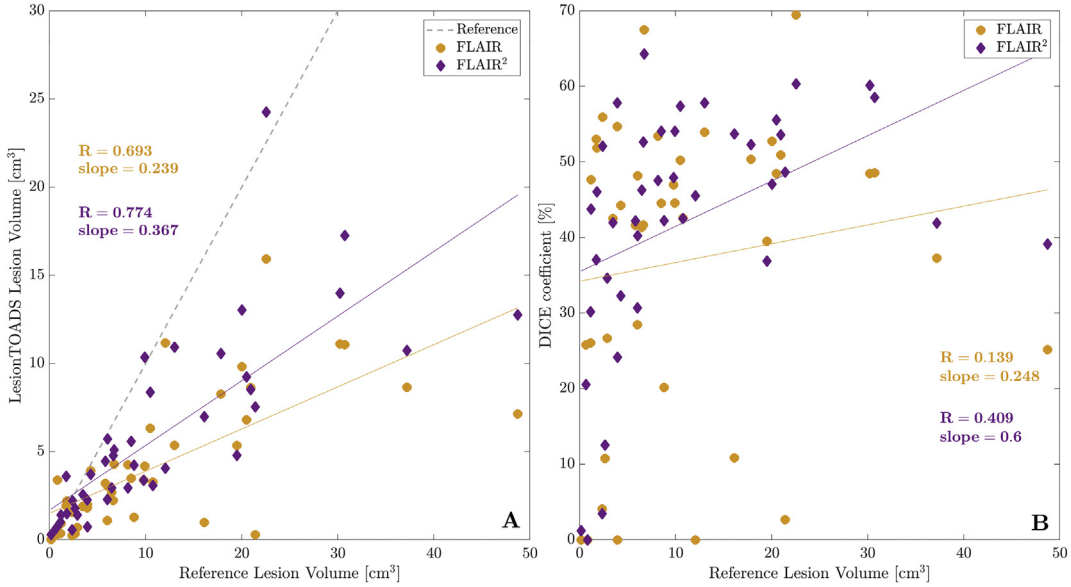

Fig. 2. (A) Comparison of segmented absolute LVs with respect to the reference segmentation volume. The gray dashed line indicates the level of the reference, expert segmentation. FLAIR^2 improved LesionTOADS segmentation results and gained in particular at LV $> 10\,cm^3$. (B) DSC improved significantly when segmenting FLAIR^2 images with larger LVs, less so for FLAIR. The data points suggest that the relationship between LV and DSC may be non-linear, therefore, the linear fit is only a general indicator of improvement and not meant to model the exact relationship between these parameters.

in line with the reference segmentation. This leads to reduced false positive (FP), but also lower true positive (TP) rates. In contrast, OASIS over-segmented LV by including diffuse FLAIR-hyperintensities in between focal lesions into the lesion mask. LST reduced the false detection of diffuse FLAIR-hyperintense regions, but still identified hyperintensities adjacent to the ventricles as lesion tissue. Notably, neither technique captured small lesions that were identified in the reference segmentation. As the only tool, LST identified hyperintensities in the corpus callosum just above the ventricles. The over-segmentation in both OASIS and LST allowed a nearly complete capture of the reference segmentation, yielding high SEN scores via high TP voxel rates, despite increased FP rates compared to LesionTOADS.

Table 2
Summary of the mean and standard errors of the lesion segmentation performance scores for all three segmentation tools applied to the multi-centre study data set: LVD – relative lesion volume difference, DSC – DICE score, SEN – sensitivity, SSD – symmetric surface distance, CI – confidence interval. The relative improvement is assessed for FLAIR^2 with respect to FLAIR (paired *t*-test, $n = 42$ for all scores). The right-hand column quotes the *p*-values extracted from the mixed effects model for the performance comparison of FLAIR versus FLAIR^2, when separately modeling the effect of LV.

	FLAIR	FLAIR^2	Difference (FL^2-FL) 95% CI	Relative improvement [%]	p-value (holm adj.)	p-value mixed effects model
LesionTOADS						
LVD [a.u.]	0.636 ± 0.071	0.436 ± 0.044	−0.201 ± 0.543 (−0.365) – (−0.037)	31.56	0.018	0.021
DSC [%]	36.98 ± 3.00	42.16 ± 2.47	5.19 ± 13.70 1.04–9.33	14.03	0.190	0.011
SEN [%]	27.85 ± 2.60	34.84 ± 2.32	6.98 ± 10.26 3.88–10.09	25.07	0.048	3.5e-5
SSD [mm]	7.40 ± 1.07	3.92 ± 0.76	−3.48 ± 5.15 (−5.04) – (−1.92)	46.99	0.010	5.5e-5
OASIS						
LVD [a.u.]	6.07 ± 3.78	2.69 ± 0.72	−3.39 ± 21.04 (−9.75) – (2.97)	55.78	0.380	0.322
DSC [%]	38.79 ± 2.92	38.32 ± 2.80	−0.47 ± 5.71 (−2.19) – 1.26	−1.2	0.910	0.518
SEN [%]	63.0 ± 2.64	59.9 ± 2.85	−3.11 ± 11.53 (−6.60) – 0.38	−4.93	0.430	0.065
SSD [mm]	11.95 ± 2.11	9.34 ± 1.49	−2.61 ± 7.07 (−4.75) – (−0.47)	21.83	0.320	0.025
LST						
LVD [a.u.]	1.84 ± 1.45	1.43 ± 0.30	−0.408 ± 9.643 (−3.32) – 2.51	22.19	0.780	0.785
DSC [%]	40.91 ± 2.75	38.22 ± 2.97	−2.69 ± 14.55 (−7.09) – 1.71	−6.58	0.510	0.242
SEN [%]	43.62 ± 3.64	54.26 ± 3.21	10.63 ± 14.48 6.25–15.02	24.38	0.031	2.6e-5
SSD [mm]	4.73 ± 1.45	15.15 ± 2.19	10.42 ± 11.99 6.80–14.05	−220.3	0.0002	2.0e-7

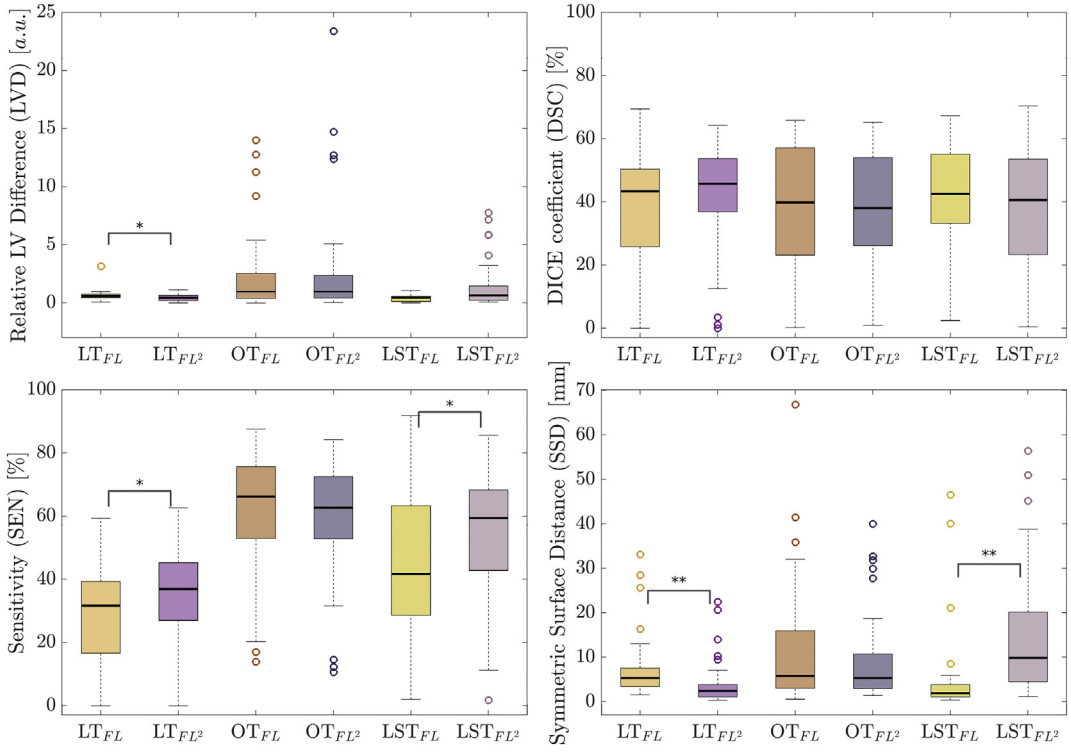

Fig. 3. Comparison of the performance of LesionTOADS (LT), OASIS$_{trained}$ (OT) and LST when using FLAIR or FLAIR^2 as input. Displayed significances correspond to the paired t-test results with * indicating $p < .05$ and ** $p < .01$. OASIS did not replicate the FLAIR vs. FLAIR^2 difference in performance seen with LesionTOADS. Despite similar DSC and higher SEN, OASIS demonstrated more variability, including higher values for LVD and SSD. LST performance was comparable to LesionTOADS, but also showed greater LVD and SSD variability with FLAIR^2. LST FLAIR^2 vs. FLAIR SEN was significantly improved, in line with LesionTOADS.

Since the average performance of LesionTOADS, as well as OASIS and LST, was lower than reported in previous studies (e.g. mean DSC < 42.2%, see Table 2, compared to (Shiee et al., 2010; Sweeney et al., 2013)), we further investigated the effect of LV under varying acquisition parameters in the multi-center imaging study. The performance scores for LesionTOADS are summarized in Fig. 5. LVD (Fig. 5A) and SEN (C) improved in FLAIR^2-based segmentations largely independent of LV. This was confirmed by the linear mixed effects model, which showed no significant effect of LV stratification nor an interaction between the two grouping factors (input MR sequence and LV, $p > .09$). Pairwise post-hoc comparisons demonstrated a significantly improved LVD when using FLAIR^2 over FLAIR at minimal LV ($p = .043$) as well as significant improvements in SEN at low/medium and high LV ($p = .0007$ and $p = .002$, respectively).

In contrast, DSC (B) and SSD (D), albeit insignificantly so for SSD, changed with LV for FLAIR^2. Similar, but insignificant trends were observed for FLAIR. Both, DSC and SSD, indicated that FLAIR^2-segmented data sets with larger LVs showed greater comparability to the

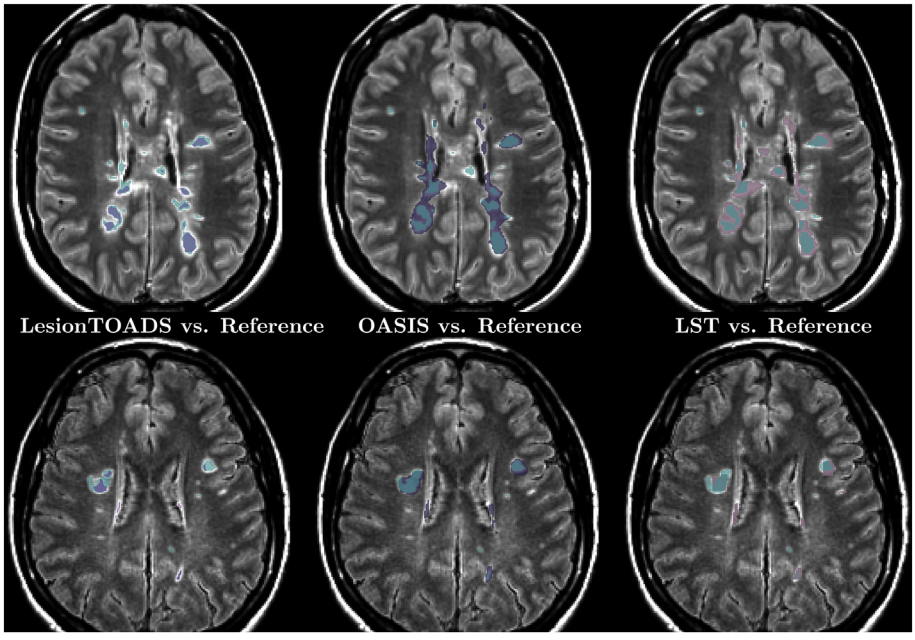

Fig. 4. Visual comparison of the FLAIR^2-based LesionTOADS (left, purple), OASIS (middle, purple) and LST (right, purple) segmentations in two subjects (top and bottom), with respect to the semi-automated reference (green). Similar to the examples in Fig. 1 F-H, LesionTOADS underestimated the LV in both cases shown here, while OASIS and LST more fully captured, and partly over-determined, lesion voxels. Thereby, LesionTOADS generated fewer FP voxels, while OASIS and LST achieved high rates of TP lesion voxels. All tools also segmented unrelated FLAIR-hyperintensities, such as present adjacent to the ventricles, but in turn failed to detect small hyperintense lesions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

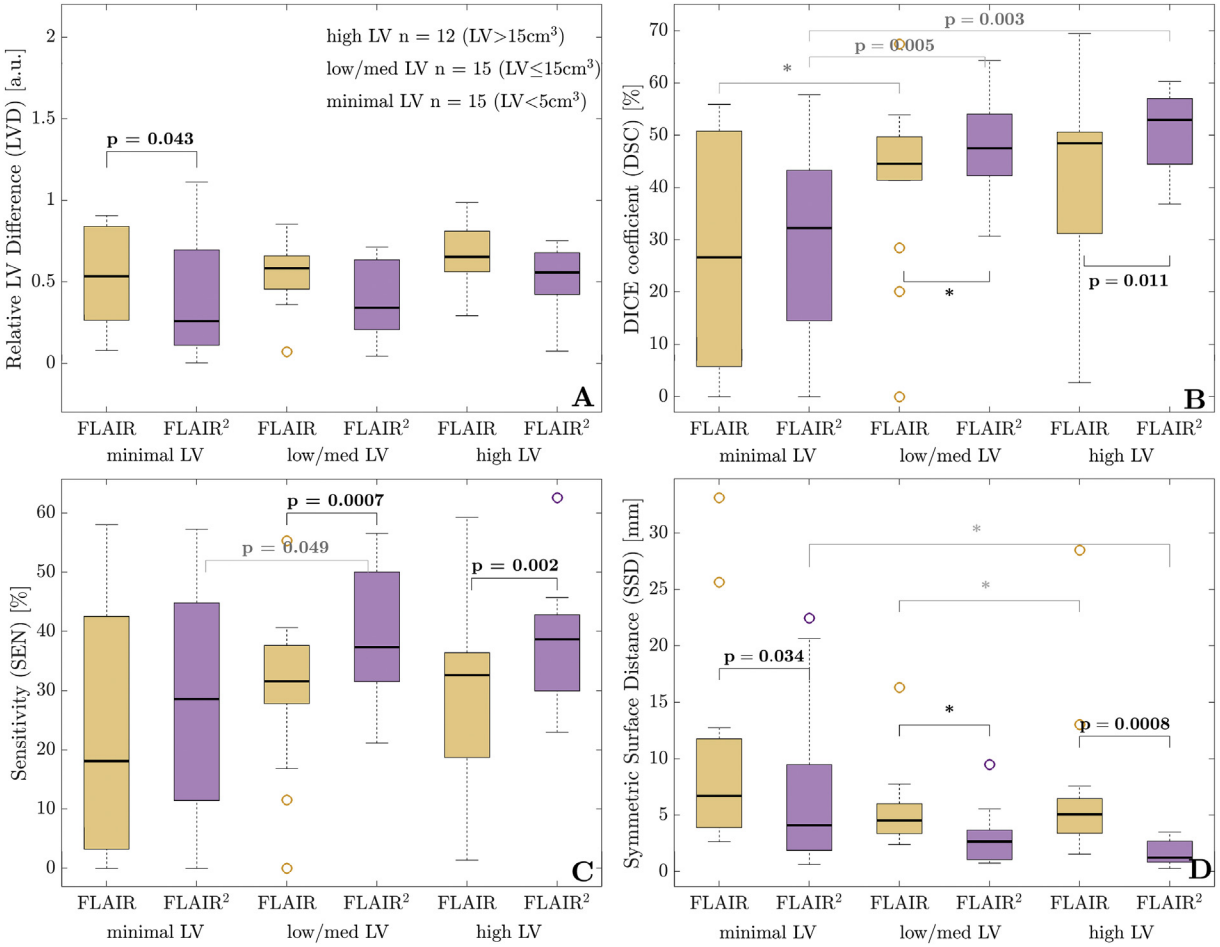

Fig. 5. FLAIR^2-based lesion segmentations showed a clear improvement over FLAIR-based estimates, in particular at high lesion volumes (LVs). The performance improvement was independent of LV with respect to the estimated relative volume difference (LVD) and sensitivity (SEN). There was a significant improvement in DSC for segmenting images with medium or higher LV, compared to LVs $< 5\,cm^3$. FLAIR^2-based DSC improved from a mean of 29.2% to 50.7%, while SSD dropped from 6.84 mm to 1.65 mm. The corresponding FLAIR SSD-score remained unchanged. For minimal LVs, FLAIR^2-based segmentations significantly improved LVD (p = .043) and SSD (p = .044). p-values indicating differences between FLAIR and FLAIR^2 performance are shown in black, LV related p-values in gray. * indicates $p < .1$.

reference segmentation. This is reflected in the smaller standard deviations for FLAIR^2 at greater LVs noted for DSC and SSD in Fig. 5, but also represented in Fig. 2B, where FLAIR^2 segmentation for LVs $> 5\,cm^3$ consistently achieved DSC scores > 30%, while some FLAIR-based LesionTOADS lesion masks continued to exhibit little similarity to the reference.

Mean and standard errors for all LesionTOADS performance metrics, when stratified by LV, are summarized in Table 3. In addition, examples of the segmentation performance of LesionTOADS at different LVs are displayed in Fig. 6.

Similar dependencies on LV were detected with both OASIS and LST, which are displayed in Fig. 7. OASIS (A-D) demonstrated LV-dependent improvements for DSC and SSD, while LVD and SEN were less affected by LV. In contrast to LesionTOADS, OASIS showed little difference between FLAIR and FLAIR^2-based performance scores, even after accounting for LV. Notably, the significant improvement in SSD for minimal LV was maintained, although data sets with minimal LVs were overall poorly segmented. LST (E-H) showed similar LV-dependent improvements in DSC, and a stronger effect of LV on SEN than observed with LesionTOADS or OASIS. LVD and SSD were LV-independent for FLAIR-based segmentations, but showed large improvements for FLAIR^2. Although FLAIR^2 performed similar or better (SEN) than FLAIR when LVs passed $5\,cm^3$, small LVs segmentations with LST performed significantly poorer with FLAIR^2 compared to FLAIR.

Within the limited scope of 3 T data available within the multi-center study (n = 2 versus n = 40 for 1.5 T), the mixed effects model suggested no significant effect of field strength on the outcomes, except for SSD (p = 0.02), with worse performance at 3 T.

To test the performance of LesionTOADS at higher field strength and in the challenging setting of lower LVs, we used LesionTOADS to segment data in a CIS single-site data set, again using FLAIR and FLAIR^2. The performance is summarized in Fig. 8. The mean/median/range LV, obtained from the reference segmentation, of this cohort was = 4.42/2.67/0.13–20.98 cm^3. 29 of the 40 patients had minimal LVs ($LV < 5\,cm^3$), eight patients had LVs $5\,cm^3 < LV < 15\,cm^3$ and only three patients counted toward the high LV group in this cohort. There were no significant differences between FLAIR and FLAIR^2 ($p > 0.48$) with the exception of SSD (p = 0.012). The average scores were LVD = 9.73 a.u. and 8.98 a.u.; DSC = 28.38% and 25.92%; SEN = 49.72% and 46.30%; SSD = 12.66 mm and 25.75 mm, for FLAIR and FLAIR^2, respectively. Note that although the DSC values were approximately on the order of the DSC scores of the multi-center study for $LVs < 5\,cm^3$, LVD and SSD were on average higher, indicating worse performance, despite approximately doubled SEN. Visual assessment of the segmentation performance, however, seemed to suggest that LesionTOADS with FLAIR^2 should perform better than FLAIR (see lesion [1]), if not for misclassification of some FLAIR^2-hyperintense regions (regions [2] and [3]).

Table 3
Mean and standard errors for all LesionTOADS segmentation performance indices after stratification by LV (< 5 cm^3 (n = 15), 5–15 cm^3 (n = 15), > 15 cm^3 (n = 12)): LVD – relative lesion volume difference, DSC – DICE coefficient, SEN – sensitivity, SSD – symmetric surface distance. The mixed effects model p-values are reported in the second part of the table. Significant values are highlighted in bold font.

	LV < 5 cm^3	5 cm^3 < LV < 15 cm^3	LV > 15 cm^3
LVD [a.u.]: FLAIR	0.688 ± 0.190	0.551 ± 0.049	0.678 ± 0.196
$FLAIR^2$	0.402 ± 0.097	0.393 ± 0.061	0.531 ± 0.153
DSC [%]: FLAIR	29.55 ± 7.63	41.68 ± 10.8	40.37 ± 11.7
$FLAIR^2$	29.18 ± 7.53	48.36 ± 12.5	50.65 ± 14.6
SEN [%]: FLAIR	25.28 ± 6.53	30.07 ± 7.76	28.30 ± 8.17
$FLAIR^2$	27.61 ± 7.13	39.46 ± 10.2	38.08 ± 11.0
SSD [mm]: FLAIR	9.75 ± 2.52	5.37 ± 1.39	7.01 ± 2.02
$FLAIR^2$	6.84 ± 1.77	2.83 ± 0.73	1.65 ± 0.48
LVD p-values	**FL - FL^2: 0.043**	FL - FL^2: 0.258	FL - FL^2: 0.343
	Mini - low (FL): 0.582	Low - high (FL): 0.625	Mini - high (FL): 1.000
	Mini - low (FL^2): 0.996	Low - high (FL^2): 0.583	Mini - high (FL^2): 0.626
DSC p-values	FL - FL^2: 0.93	FL - FL^2: 0.059	**FL - FL^2: 0.011**
	Mini - low (FL): 0.100	Low - high (FL): 0.977	Mini - high (FL): 0.192
	Mini - low (FL^2):0.005	Low - high (FL^2): 0.933	**Mini - high (FL^2): 0.003**
SEN p-values	FL - FL^2: 0.365	**FL - FL^2: 0.0007**	**FL - FL^2: 0.002**
	Mini - low (FL): 0.469	Low - high (FL): 0.954	Mini - high (FL): 0.703
	Mini - low (FL^2): 0.05	Low - high (FL^2): 0.971	Mini - high (FL^2): 0.115
SSD p-values	**FL - FL^2: 0.034**	FL - FL^2: 0.062	**FL - FL^2: 0.0008**
	Mini - low (FL): 0.084	Low - high (FL): 0.554	Mini - high (FL): 0.583
	Mini - low (FL^2): 0.122	Low - high (FL^2): 0.961	Mini - high (FL^2): 0.09

4. Discussion

We demonstrated that by combining FLAIR and T_2w-images prior to selecting them as input for LesionTOADS, MS lesion segmentation in lower-field strength, multi-center studies may be enhanced by 14–47%, depending on the performance evaluation score. Since FLAIR and T_2w-scans are commonly acquired as part of clinical and research MR protocols for MS (Rovira et al., 2009; Traboulsee et al., 2016), the computation of $FLAIR^2$ and the subsequently improved LesionTOADS segmentation are realized without the need for additional scanning and at minimal organizational and computational time or financial investment. The improvement of using $FLAIR^2$ over FLAIR, however, was limited to the use of LesionTOADS in the multi-center study, and could not be replicated at lower LVs. Other segmentation tools, e.g. OASIS and LST, only partially captured the benefits of $FLAIR^2$, indicating that these tools need further optimization in order to gain from the $FLAIR^2$ contrast.

4.1. Multi-center study

Previously, the value of $FLAIR^2$ was demonstrated using 3D-FLAIR and 3D-T_2w acquired at 3 T (Wiggermann et al., 2016). Based on 5 healthy controls and 7 MS patients, a 133% increase in contrast-to-noise between gray matter vs. WM and 158% for lesions vs. WM was observed when using $FLAIR^2$ over FLAIR. It was furthermore described that the gain in signal-to-noise ratio (SNR) in 3D-acquisitions compared to 2D-scans allowed for the acquisition of isotropic voxels, providing robust image registration and permitting image reformatting. For the present study, $FLAIR^2$ was computed from 2D-scans as 2D-data acquisitions have been, until recently, the standard in clinical trials and clinical practice. We demonstrated that 2D-$FLAIR^2$ provides significantly improved lesion segmentation with LesionTOADS, even when most data were collected at 1.5 T using non-homogenized image acquisition protocols. While the 2D-results do not necessarily predict the success of LesionTOADS for 3D-imaging data, the SNR gain suggests further improvements.

4.2. LesionTOADS

Overall, LesionTOADS underestimated the LV in our data set, even with $FLAIR^2$. We performed the segmentation with default parameters; optimizing the parameters and/or modifying the software may increase performance for both scans, however, this was not the purpose of this work. If modified, LesionTOADS may also arguably not be classified as an unsupervised segmentation method, although much fewer data sets would be required to optimize performance than needed for deep learning machines. $FLAIR^2$-LesionTOADS with default parameters seemed to detect the central, most hyperintense parts of MS lesions, but

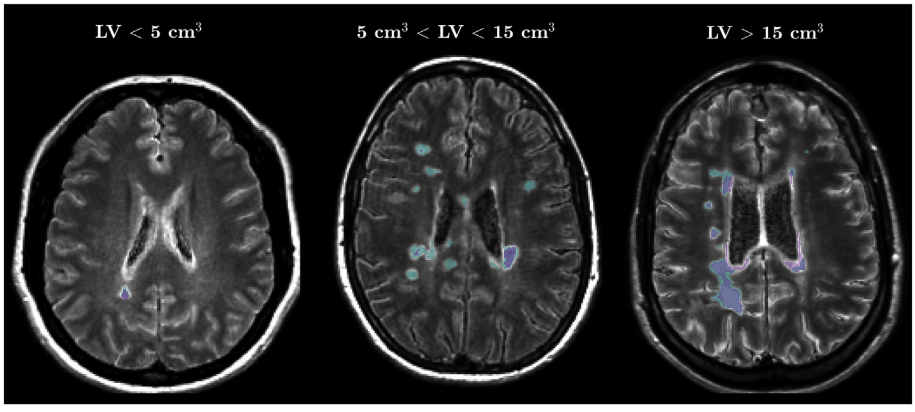

Fig. 6. Example $FLAIR^2$-based lesion segmentations obtained from LesionTOADS at different LVs. Note that besides the LV, the image contrast of $FLAIR^2$ varied strongly between scans from different sites, affecting the segmentation quality. Lesions of sufficient hyperintensity in the periventricular area are well captured, while lesser hyperintensities and lesions in the subcortical WM may be missed.

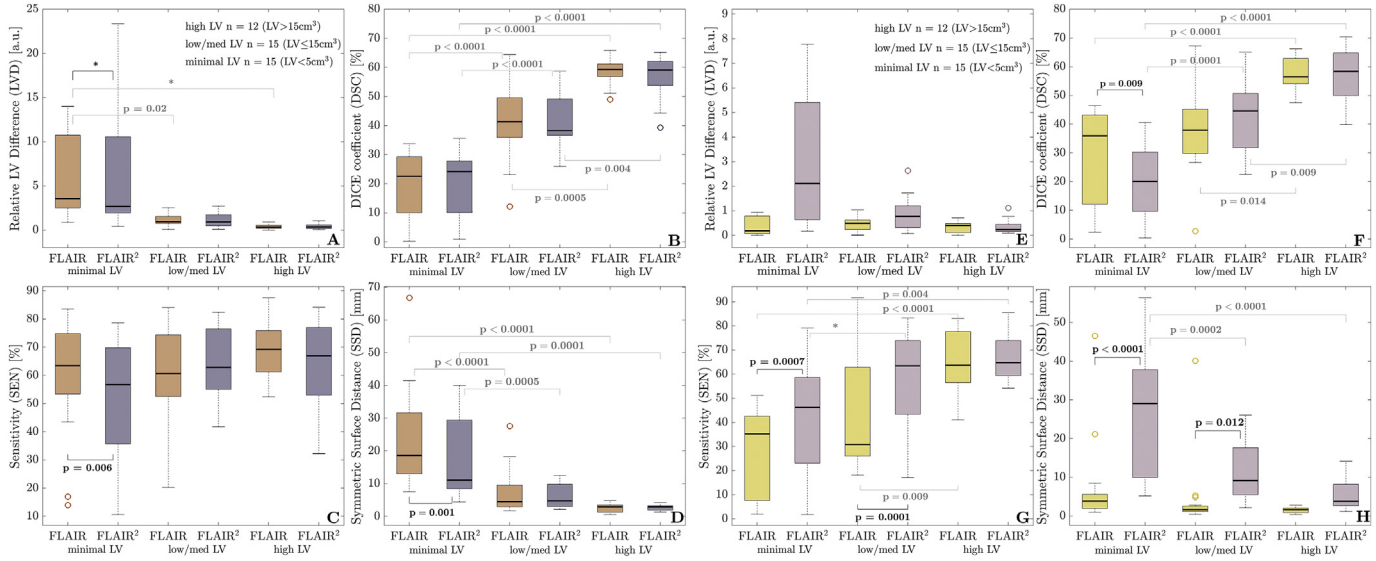

Fig. 7. LV-dependency of performance scores for OASIS (A-D) and LST (E-H). Similar to LesionTOADS, OASIS and LST performance improved significantly at higher LVs, increasing DSC and decreasing SSD. For LST, SEN also strongly depended on LV (G). LST-LVD remained statistically unchanged (E), although a clear improvement in the LVD estimates for FLAIR^2 is apparent between LV $< 5\,cm^3$ and larger $5\,cm^3$. After accounting for LV, OASIS showed little difference between FLAIR and FLAIR^2-based segmentations. With LST, FLAIR^2 typically performed equal and in some cases better than FLAIR, however, FLAIR^2 segmentations failed in terms of SSD estimation, particularly at minimal LVs.

omitted areas that appeared more diffusely damaged as well as lesion boundaries and lesions that were not periventricular and appeared less hyperintense (Fig. 1). In contrast, the T_2w and PDw-based semi-automated segmentation, commonly used for clinical trials, provided a reference closer to the ground truth (McAusland et al., 2010) by more fully capturing the extent of MS lesions. Our data showed that although large LVs will be incompletely captured, even by FLAIR^2, the areas that are detected by LesionTOADS are in good correspondence to the reference. The extent of smaller LVs was better captured (lower LVD), but lacked in overlap to the reference as noted by low DSC and large SSD scores. By using FLAIR^2, LVD improved to under 0.5 ($LVD_{FL^2} = 0.4$), which is better than LesionTOADS achieved in the lesion segmentation challenge (Shiee et al., 2008). The low DSC and SEN scores agree with other multi-center studies (Roy et al., 2014), while a single-site, homogenized imaging study previously achieved better mean DSC of up to 61% (Jain et al., 2015). However, improvements with respect to LV were comparable. The aforementioned study reported a 15% change in DSC between small and larger LVs, in line with the 20% improvement

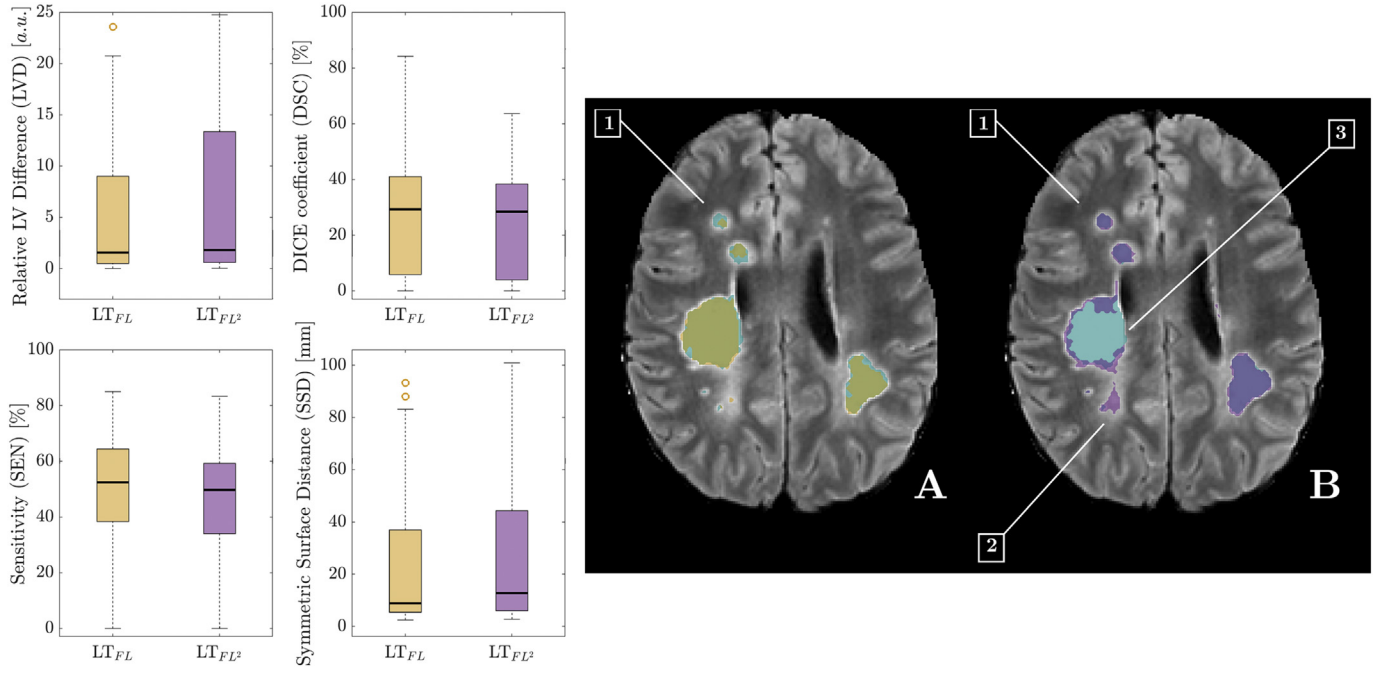

Fig. 8. Performance of LesionTOADS in low LV, 3 T MR data. **Left:** The DSC scores were approximately in agreement with the minimal LV scores shown in Fig. 5. While SEN was higher, LVD and SSD indicated worse performance in this data set. Due to the large variance in achieved scores, there was no difference between the use of FLAIR and FLAIR^2 with the exception for SSD. **Right:** Although the FLAIR^2-based segmentation (B, purple) better captured the extent of lesions than FLAIR (A, yellow) alone (1), FP voxels were detected in FLAIR^2-hyperintense regions (2). The reference segmentation is shown in green. In addition, LesionTOADS may fail to capture areas of high FLAIR^2 signal intensity, such as the center of large lesions (3).

in our study when using FLAIR^2. Note also that deep learning tools currently achieve sub-optimal DSC scores of 63% (Birenbaum & Greenspan, 2016). Thus, automated segmentation approaches, like LesionTOADS, may be presently favored given their minimal need for training data.

4.3. OASIS

In contrast to LesionTOADS, performance of OASIS did not improve when using FLAIR^2 instead of FLAIR for the segmentation. OASIS did perform less consistent than LesionTOADS for the same data, particularly at lower LVs, introducing larger heterogeneity in the achieved LVD and SSD segmentation scores. Regardless of the larger variance in segmentation performance that might limit statistically significant findings, OASIS likely benefited less from FLAIR^2, since it already utilized the input of T_2w and FLAIR in addition to T_1w and PDw. Thus, various T_2-weightings were already included in OASIS, even when using FLAIR. Using multiple imaging modalities for MS lesions segmentation agrees with the heterogeneous presentation of MS lesions. Automated lesion segmentation approaches, however, rely often on few modalities, e.g. T_1w and FLAIR only (Shiee et al., 2010; Schmidt et al., 2012; Roura et al., 2015; Jain et al., 2015). Incorporating other, possibly quantitative modalities, as attempted in the MS Lesion Segmentation Challenge, which provided diffusion tensor imaging data (Styner et al., 2008), may further enhance our ability to segment MS lesions. On the other hand, as the imaging protocol for MS evolves (Traboulsee et al., 2016), not all clinical images will continue to be acquired, possibly posing a challenge for existing segmentation tools that require now optional scans, such as PDw. The need for data specific thresholding or training adds to the analysis complexity. Initial trials of using the default threshold of 0.16 (data not shown) yielded largely over-segmented lesion masks for OASIS. Threshold adjustment alleviated some of these concerns, however, manual selection of the threshold enforces ultimately a trade-off between capturing lesion extent and loosing small lesions. In most cases, sensitivity toward small lesions will be sacrificed. Nevertheless, even after training and threshold adjustment, OASIS continued to incorporate diffuse FLAIR-hyperintense regions into the segmented lesion masks. The over-segmentation with OASIS resulted in lower LVD and SSD. SEN scores, however, were high, since SEN only considers TP and false negative (FN) voxels, not FP. Thus, relying on SEN alone as an indicator of segmentation performance, may lead to erroneous assessments. Note also that OASIS was the only tool which performed consistently worse at lower LVs, independent of the use of FLAIR or FLAIR^2.

4.4. LST

LST also over-segmented lesions, albeit less so than OASIS, largely sparing diffuse hyperintense regions. Thus, SEN was higher than for LesionTOADS, but all other scores indicated similar performance. Notably, LST replicated the significant improvement in SEN of FLAIR^2 over FLAIR. In contrast, SSD for FLAIR^2-LST was significantly higher than for FLAIR and higher than observed with LesionTOADS. This pattern was more apparent when investigating the performance of LST relative to LV. The deficiency for accurate segmentation of small LVs with FLAIR^2 was expressed by large LVD and SSD, despite high SEN. Both scores normalized compared to FLAIR at higher LVs. This shortfall was only noted for FLAIR^2, not LST-FLAIR, which achieved comparable results to FLAIR-based LesionTOADS, showing no LV-dependency in LVD and SSD. Since the performance difference between FLAIR and FLAIR^2 was primarily noted on SSD, it is likely that LST segments hyperintensities at the gray matter – WM boundary, possibly due to CSF leakage from image co-registration errors. Further FP may be detected adjacent to the ventricles, increasing LVD and decreasing DSC, while maintaining SEN. If the tissue probability and lesion belief maps can be adjusted to account for possible artefacts at tissue boundaries, LST-performance could improve, and FLAIR^2-based segmentations could possibly achieve similar or better results than FLAIR.

The multi-site data acquisition, while a drawback, is also a strength of our study. For the 2013 review (García-Lorenzo et al., 2013) of 47 MS lesion segmentation approaches, 11 used the two-site data provided for the MS lesion segmentation challenge (Styner et al., 2008) and only two validated their methods using multi-centre data. Moreover, the 42 and 40 data sets included here exceed the cohort size of most of these publications; 29 studies had 20 data sets or less in their analyses and the largest cohort comprised 41 patients. We demonstrated in this larger, multi-center cohort that by changing the input image contrast used by LesionTOADS, lesion segmentations were significantly improved in this real-world setting. While other tools need to be optimized to handle amplified hyperintensities appearing on FLAIR^2, FLAIR^2 already performed similar or better in some cases than FLAIR.

4.5. CIS single-center study

To test whether LesionTOADS in combination with FLAIR^2 also enhanced segmentation performance at higher field strength, we applied LesionTOADS to 40 data sets of CIS patients acquired at 3 T. Note that this data did not overlap with the 3 T data from the multi-center study. In this data set, LesionTOADS-FLAIR^2 did not demonstrate benefits over FLAIR. Notably, the data set included primarily patients with minimal LVs, on average less than half of the LV detected in the multi-center MS cohort. As shown, LesionTOADS performance improved with increasing LV. Thus, DSC was low and LVD and SSD were high in this data set, in agreement with the scores presented for the minimal LV group in Fig. 5. Notably, using LesionTOADS un-optimized for small LVs, may lead to an overestimation of LV, if the true LV is small (Roy et al., 2014). Despite similar performance scores, visual image inspection demonstrated that FLAIR^2-based segmentations captured better the lesion extent (Fig. 8). However, amplified hyperintensities unrelated to lesion tissue may lead to mis-segmentation in FLAIR^2. Mis-segmentation may in part occur as over-segmentation due to enhanced brightness of voxels in periventricular areas, where diffuse FLAIR-hyperintensities are present, or in cortical areas, where the WM - gray matter contrast difference is enhanced, similar to the contrast of double inversion recovery images. Moreover, mis-registration can lead to CSF leaking into the FLAIR^2-image. Parameter optimization in LesionTOADS may be able to address these issues partly by specifying the distance between lesion voxels and cortex as well as ventricles. However, improving the WM segmentation that is performed within LesionTOADS rather than relying on parameter tuning will provide a more generalized and standardized approach to segmentation improvement. Note that, although considered the reference ground truth, manual or semi-manual expert segmentation will also be image contrast dependent and can be prone to errors. Finally, enhanced FLAIR^2-hyperintensities may be missed by LesionTOADS, as shown in Fig. 8, where the center of the large lesion remained undetected. This may be considered an inherent property of LesionTOADS and its fuzzy C-means clustering. In case that a voxel presents very high intensity on FLAIR^2, it may be far from any centroid in the clustered space and may thus produce equal, unpredictable memberships. This may in part explain the continued under-segmentation of LV by means of LesionTOADS segmentation, despite the improved contrast-to-noise of FLAIR^2. A pre-processing step could be applied to threshold the intensities of FLAIR^2, prior to LesionTOADS segmentation.

4.6. Limitations

We focused our study on LesionTOADS, one of the publicly available automated segmentation tools that had been previously evaluated in the 2008 MS Grand Segmentation Challenge (Styner et al., 2008). Other segmentation tools may benefit even more, or possibly less, from the combination of FLAIR and T_2w as shown here for example with OASIS

and LST. Publicly available, state-of-the-art lesion segmentation tools have been widely compared, however, depending on the evaluation score, assessments vary broadly.

Among the publicly available tools, Souplet's approach (Souplet et al., 2008) ranked highest in the initial challenge just above LesionTOADS, indicating high similarity in their performance (Lladó et al., 2012). Although Cabazas et al. (Cabezas et al., 2014) showed that LST outperformed Souplet's method, as assessed by DSC and SSD, which would therefore suggest a better performance than LesionTOADS, Jain et al. (Jain et al., 2015) in fact, showed that on average LesionTOADS had greater precision and achieved higher DSC than LST, with equal SEN. LST yielded considerably lower DSC scores at LV $< 5\,cm^3$ than LesionTOADS. Noting that approximately one third of our patients in the multi-center study and 72.5% of patients in our single-center study had LVs $< 5\,cm^3$, using LST may be suboptimal, although it was shown to provide the most comparable LV estimates. Notably, our data showed comparable LVD and DSC scores at small LVs between the two segmentation tools. However, LesionTOADS-FLAIR^2 improved LVD in our multi-center study compared to FLAIR significantly at LVs $< 5\,cm^3$ ($p = 0.043$), while retaining performance on all other scores. SLS performed better than LesionTOADS in the detection of FP lesion voxels and in terms of volume differences, while LST achieved superior segmentation accuracy (Brosch et al., 2016). For this study, LesionTOADS was chosen because it was readily available, its comparable and favorable performance, and because it provides ample reference for comparisons (Brosch et al., 2016; Sweeney et al., 2013; Jain et al., 2015). LesionTOADS moreover facilitates simultaneous tissue segmentation, not available using most other segmentation methods, and does not require costly licenses, a possibly limiting factor for accessing some tools, e.g. such as implemented within SPM/Matlab.

4.7. Evaluation metrics

A wide range of parameters exists that can be used to assess the quality of segmentations with respect to a reference. DSC and SEN are the most frequently used (Lladó et al., 2012), but other scores such as the lesion-wide TP or FP rates may need to be considered when determining the segmentation precision of small lesions in the cortex or the deep WM, where partial volume effects play a larger role (García-Lorenzo et al., 2013). Although the same T_1w-images were used in combination with FLAIR and FLAIR^2, the contrast combination of the two input images will ultimately determine the success of LesionTOADS. The default parameters, which were employed in our study, that work optimally for the T_1w-FLAIR input, may not be ideal for T_1w-FLAIR^2. Double inversion recovery (DIR) is currently suggested to be the most suitable sequence for the detection of cortical lesions, which are generally not captured by lesion segmentation algorithms, but DIR suffers from limited SNR and relatively long data acquisition times (Geurts et al., 2011). FLAIR^2 provides image contrast similar to DIR at a higher spatial resolution and with improved SNR and contrast-to-noise ratio (Wiggermann et al., 2016). Fig. 1C demonstrates the enhanced contrast of the cortical gray matter, which could provide a starting point for cortical lesion segmentation, particularly at higher field strength.

Overall, our study suggests that the computation of FLAIR^2 is particularly beneficial if the image quality of FLAIR itself is lower, possibly due to lower SNR at lower field strength as observed in our multi-center study. Whether FLAIR^2 will be beneficial at higher field strength remains to be shown. Improved WM delineation and adjustments of the FLAIR^2-image intensity may be needed to take full advantage of the benefits of FLAIR^2 in LesionTOADS as well as with other segmentation tools.

5. Conclusion

The computation of FLAIR^2 from FLAIR and T_2w-images increases the performance of automated lesion segmentation with LesionTOADS at minimal additional scan time or computational cost, in the setting of multi-center, lower field strength non-homogenized 2D-data acquisitions. As long as both FLAIR and T_2w-scans are available, whether acquired 2D or 3D, FLAIR^2 can be obtained, and used for automated segmentation of MS lesions. Further work is needed to determine how segmentation tools can ideally benefit from the improved FLAIR^2 contrast.

Disclosures

ML, LYWT, EHT, MJ and VW have nothing to disclose. TB is an employee of Philips Medical. LM has received grant support from Hoffman La Roche. AT has received research funding from Chugai, Roche, Novartis, Genzyme and Biogen and acted as a consultant for Genzyme, Roche, Teva, Biogen and Serono. DKBL reports research grants from Genzyme, Merck-Serono, Novartis and Roche, consulting fees from Vertex Pharmaceuticals and speaker fees from Novartis, Biogen-Idec, Sanofi-Genzyme and Teva. DKBL has served as a member of the data and safety advisory board to Opexa Therapeutics as well as on the scientific advisory boards of the Adelphi group, Celgene, Novartis and Roche and is the Emeritus director of the UBC MS/MRI research group, which has been contracted to perform central analysis of MRI scans for therapeutic trials with Novartis, Perceptives, Roche and Sanofi-Aventis. AR has received speaking fees from Philips Medical. RCT has received research support as part of sponsored clinical studies from Novartis, Roche and Sanofi Genzyme.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2019.101918.

Acknowledgments and grant support

Data acquisition for this study was in part supported by the MS Society of Canada (21569). ML was a holder of a Student Research Scholarship from the Foundation of the Consortium of Multiple Sclerosis Centers (FCMSC). MJ was supported by the National MS Society (RG-1507-05301) and NSERC (402039-2001, 2016-05371). VW and EHT are supported by the National MS Society (RG-1507-05301). VW was holder of a graduate award of the MS Society of Canada (EGID 2002). AR is supported by Canada Research Chairs. RCT received support from NSERC and the Milan and Maureen Ilich Foundation. We wish to thank Carolyn Taylor for helpful discussion in regard to the statistical analysis.

References

Bates, D., Mächler, M., Dai, B., 2011. Lme4: Linear Mixed-Effects Models Using S4 Classes (R Package Version 1.1-12).

Birenbaum, A., Greenspan, H., 2016. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., JMRS, Tavarres, Belagiannis, V. ... Cornebise, J. (Eds.), Deep Learning and Data Labeling for Medical Applications, Volume 10008 of the Series Lecture Notes in Computer Science. Springer, New York, pp. 58–67.

Brosch, T., Tang, L.Y.W., Yoo, Y., et al., 2016. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imaging 35 (5), 1229–1239.

Cabezas, M., Oliver, A., Roura, E., et al., 2014. Automatic multiple sclerosis lesion detection in brain MRI by FLAIR thresholding. Comput. Methods Programs Biomed. 115, 147–161.

Fazekas, F., Barkhof, F., Filippi, M., et al., 1999. The contribution of magnetic resonance imaging to the diagnosis of multiple sclerosis. Neurology 53, 448–456.

García-Lorenzo, D., Francis, S., Narayanan, S., et al., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. 17 (1), 1–18.

Geurts, J.J.G., Roosendaal, S.D., Calabrese, M., et al., 2011. Consensus recommendations for MS cortical lesion scoring using double inversion recovery MRI. Neurology 76 (5), 418–424.

Grimaud, J., Lai, M., Thorpe, J., et al., 1996. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn. Reson. Imaging 14 (5), 495–505.

Harrison, D.M., Roy, S., Oh, J., et al., 2015. Association of cortical lesion burden on 7-T

magnetic resonance imaging with cognition and disability in multiple sclerosis. JAMA. Neurol. 72 (9), 1004–1012.
Huo, Y., Plassard, A.J., Carass, A., et al., 2016. Consistent cortical reconstruction and multi-atlas brain segmentation. NeuroImage 138, 197–210.
Jain, S., Sima, D.M., Ribbens, A., et al., 2015. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. NeuroImage Clin. 8, 367–375.
Jenkinson, M., Bannister, P., Brady, M., et al., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.
Jones, C., Wong, E., 2002. Multi-scale application of the N3 method for intensity correction of MR images. Proc SPIE Medical Imaging 4684, 1123–1130. https://doi.org/10.1117/12.467069. Image Processing.
Lampe, L., Kharabian-Masouleh, S., Kynast, J., et al., 2019. Lesion location matters: the relationships between white matter hyperintensities on cognition in the healthy elderly. J. Cereb. Blood Flow Metab. 39 (1), 36–43.
Lenth, R.V., 2016. Least-squares means: the R package lsmeans. J. Stat. Softw. 69 (1), 1–33. https://doi.org/10.18637/jss.v069.i01.
Lladó, X., Oliver, A., Cabezas, M., et al., 2012. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. Inf. Sci. 186, 164–185.
McAusland, J., Tam, R.C., Wong, E., et al., 2010. Optimizing the use of radiologist seed points for improved multiple sclerosis segmentation. IEEE Trans. Biomed. Eng. 57 (11), 2689–2698.
Miller, D.H., Albert, P.S., Barkhof, F., et al., 1996. Guidelines for the use of magnetic resonance techniques in monitoring the treatment of multiple sclerosis. Ann. Neurol. 39 (1), 6–16.
Ozturk, A., Smith, S.A., Gordon-Lipkin, E.M., et al., 2010. MRI of the corpus callosum in multiple sclerosis: association with disability. Mult. Scler. 16 (2), 166–177.
Río, J., Auger, C., Rovira, À., 2017. MR imaging in monitoring and predicting treatment response in multiple sclerosis. Neuroimaging Clin. N. Am. 27 (2), 277–287.
Roura, E., Oliver, A., Cabezas, M., et al., 2015. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology 57 (10), 1031–1043.
Rovira, À., León, A., 2008. MR in the diagnosis and monitoring of multiple sclerosis: an overview. Eur. J. Radiol. 67 (3), 409–414.
Rovira, À., Swanton, J., Tintor, M., et al., 2009. A single, early magnetic resonance imaging study in the diagnosis of multiple sclerosis. Arch. Neurol. 66 (5), 587–592.
Roy, S., He, Q., Carass, A., et al., 2014. Example based lesion segmentation. In: *Proc SPIE Medical Imaging*: Image Processing. 9034. pp. 90341Y.
Sati, P., George, I.C., Shea, C.D., et al., 2012. FLAIR*: A combined MR contrast technique for visualizing white matter lesions and parenchymal veins. Radiology 265 (3), 926–932.
Schmidt, P., Gaser, C., Arsic, M., et al., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. NeuroImage 59 (4), 3774–3783.
Shiee, N., Bazin, P.-L., Pham, D.L., 2008. Multiple sclerosis lesion segmentation using statistical and topological atlases. *Grand Challenge Work.:*. Mult. Scler. Lesion Segm. Challenge. MIDAS Journal 1–10.
Shiee, N., Bazin, P.L., Ozturk, A., et al., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49 (2), 1524–1535.
Shinohara, R.T., Oh, J., Nair, G., et al., 2017. Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. AJNR Am. J. Neuroradiol. 38, 1501–1509.
de Sitter, A., Steenwijk, M.D., Ruet, A., et al., 2017. Performance of five research-domain automated WM lesion segmentation methods in a multi-center MS study. NeuroImage 163, 106–114.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.
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.
Souplet, J.C., Lebrun, C., Ayache, N., et al., 2008. An automatic segmentation of T2-flair multiple sclerosis lesions. In: Grand Challenge Work.: Mult. Scler. Lesion Segm, pp. 1–11 New York, NY, USA, United States.
Styner, M., Lee, J., Chin, B., et al., 2008. 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. MIDAS Journal 1–6.
Sweeney, E.M., Shinohara, R.T., Shiee, N., et al., 2013. OASIS is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in MRI. NeuroImage Clin 2, 402–413.
Sweeney, E.M., Vogelstein, J.T., Cuzzocreo, J.L., et al., 2014. A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structure MRI. PLoS One 9 (4), e95753.
Tam, R.C., Riddehough, A., Li, D.K.B., 2009. Detection and measurement of coverage loss in interleaved multi-acquisition brain MRIs due to motion-induced inter-slice misalignment. Med. Image Anal. 13 (3), 381–391.
Traboulsee, A., Simon, J.H., Fischer, E., et al., 2016. Revised recommendations of the consortium of MS centers task force for a standardized MRI protocol and clinical guidelines for the diagnosis and follow-up of multiple sclerosis. AJNR Am. J. Neuroradiol. 37 (3), 394–401.
Valcarcel, A.M., Linn, K.A., Vandekar, S.N., et al., 2018. MIMoSA: an automated method for intermodal segmentation analysis of multiple sclerosis brain lesions. J. Neuroimaging 28 (4), 389–398.
Valverde, S., Cabezas, M., Roura, E., et al., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage 155, 159–168.
Wattjes, M.P., Rovira, À., Miller, D., et al., 2015. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis – establishing disease prognosis and monitoring patients. Nat. Rev. Neurol. 11 (10), 597–606.
Wiggermann, V., Hernandez-Torres, E., Traboulsee, A., et al., 2016. FLAIR^2: a combination of FLAIR and T2 for improved MS lesion detection. AJNR Am. J. Neuroradiol. 37, 259–265.
Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic “pipeline” analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans. Med. Imaging 21 (10), 1280–1291.