

# A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus

Žiga Lesjak<sup>1</sup> · Alfiia Galimzianova<sup>1</sup> · Aleš Koren<sup>2</sup> · Matej Lukin<sup>2</sup> · Franjo Pernuš<sup>1</sup> · Boštjan Likar<sup>3</sup> · Žiga Špiclin<sup>1</sup>

Published online: 4 November 2017  
© Springer Science+Business Media, LLC 2017

**Abstract** Quantified volume and count of white-matter lesions based on magnetic resonance (MR) images are important biomarkers in several neurodegenerative diseases. For a routine extraction of these biomarkers an accurate and reliable automated lesion segmentation is required. To objectively and reliably determine a standard automated method, however, creation of standard validation datasets is of extremely high importance. Ideally, these datasets should be publicly available in conjunction with standardized evaluation methodology to enable objective validation of novel and existing methods. For validation purposes, we present a novel MR dataset of 30 multiple sclerosis patients and a novel protocol for creating reference white-matter lesion segmentations based on multi-rater consensus. On these datasets three expert raters individually segmented white-matter lesions, using in-house developed semi-automated

lesion contouring tools. Later, the raters revised the segmentations in several joint sessions to reach a consensus on segmentation of lesions. To evaluate the variability, and as quality assurance, the protocol was executed twice on the same MR images, with a six months break. The obtained intra-consensus variability was substantially lower compared to the intra- and inter-rater variabilities, showing improved reliability of lesion segmentation by the proposed protocol. Hence, the obtained reference segmentations may represent a more precise target to evaluate, compare against and also train, the automatic segmentations. To encourage further use and research we will publicly disseminate on our website <http://lit.fe.uni-lj.si/tools> the tools used to create lesion segmentations, the original and preprocessed MR image datasets and the consensus lesion segmentations.

**Keywords** Clinical image dataset · White matter lesion · Image segmentation · Intra- and inter-rater variability · Gold standard

This research was supported by the Slovenian Research Agency under grants J2-5473, L2-5472, J7-6781 and J2-8173.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s12021-017-9348-7>) contains supplementary material, which is available to authorized users.

✉ Žiga Lesjak  
ziga.lesjak@fe.uni-lj.si

- <sup>1</sup> Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia
- <sup>2</sup> Institute of Radiology, University Medical Center Ljubljana, Zaloška cesta 2, 1000 Ljubljana, Slovenia
- <sup>3</sup> Sensum, Computer Vision Systems, Tehnološki park 21, 1000 Ljubljana, Slovenia

## Introduction

Quantification of white-matter lesions based on magnetic resonance (MR) images in terms of volume and count represents important neuroimaging biomarkers, which may be used as predictive factors or surrogates of clinical signs in a number of neurological and cerebrovascular diseases, and mental disorders (DeBette and Markus 2010). In multiple sclerosis (MS) patients, for instance, inflammatory activity in the white-matter is visible as hyperintense lesions in T2-weighted (T2w), proton density weighted (PDw) and fluid attenuated inversion recovery (FLAIR) MR sequences, some of which are chronic lesions and appear hypointense

on T1-weighted (T1w) sequence. Uher et al. (Uher et al. 2016) established that at the baseline MR exam, T2w lesion number, T1w and T2w/FLAIR lesion volumes were among the best predictors of sustained disability progression in MS patients over the 12-year observation period, while they also play an important role in monitoring disease progression and response to treatment (Popescu et al. 2013; Stangel et al. 2015). To quantify neuroimaging biomarkers like lesion volume and count, accurate and reliable detection and segmentation of the lesions in MR images is required.

Lesions can be segmented manually, however, this task is tedious and time-consuming. Even more critical is the subjective nature inherent to both the process of lesion detection and lesion contouring that leads to large intra- and inter-rater variabilities, a notorious characteristic of manual lesion segmentations (Grimaud et al. 1996; Zijdenbos et al. 2002; Styner et al. 2008). Hence, it is long known that routine manual segmentations are not accurate and reliable enough for the extraction of biomarkers (Filippi et al. 1995). To provide more objective and consistent lesion segmentations, automated methods have been intensively developed over the last two decades (Garcia-Lorenzo et al. 2013; Llado et al. 2012). Main challenges for automated methods lie in their robustness to MR acquisition imperfections (MR bias field, partial volume effect and image noise), and high biological variability of brain anatomy and lesion pathology. Despite substantial methodological advancements that aim to address the aforementioned challenges, it is not yet clear, which method or even a class of automated methods (e.g. unsupervised and supervised) can be considered a standard for the biomarkers' extraction (Vrenken et al. 2013).

A standard lesion segmentation method can only be established based on objective and rigorous validation on *gold standard* datasets with highly accurate lesion segmentations. In spite of known deficiencies most researchers still use manual lesion contouring (Llado et al. 2012). To mitigate the influence of errors in single-rater segmentations, some researchers validate the automated methods against multiple rater segmentations (Styner et al. 2008). Another approach is to merge manual segmentations through a consensus of multiple raters (Anbeek et al. 2004). One can also fuse the segmentations automatically, for instance, using the STAPLE algorithm (Warfield et al. 2004; Commowick and Warfield 2009). Potential practical drawbacks are that a sufficiently large number of rater segmentations are required, especially since rater performance may vary substantially according to anatomical location of lesions and across time due to rater fatigue or other subjective reasons.

Besides engaging multiple raters, a clinical dataset should also involve a substantial number of patient images so as to capture as much biological and pathological variability.

Clearly, production of such a gold standard dataset is a difficult and laborious enterprise, often beyond the capability of each individual research team. For this reasons, public dissemination of gold standard datasets along with protocols and tools for their creation is of high importance, since it enables other research teams to produce new datasets in a consistent manner. In conjunction with standardized evaluation methodology such datasets are indispensable tools that enable objective and rigorous comparison of multiple lesion segmentation algorithms.

### Public Datasets

A public MR simulator called BrainWeb (Cocosco et al. 1997) enables the creation of synthetic images of T1w, T2w and PDw MR sequences, but there exists only one brain template that contains simulated MS lesions. The BrainWeb dataset can therefore only be used to provide proof-of-concept for new segmentation methods.

The first publicly available dataset of clinical MR images was created for the purpose of a challenge on MS lesion segmentation (Styner et al. 2008). The dataset consists of 52 cases of MS patients imaged with conventional brain MR sequences T1w, T2w and FLAIR on two 3T Siemens scanners at different sites. Each case was independently segmented by expert raters at the two sites and the resulting segmentations were used as gold standard. Unfortunately, the rater segmentations exhibit rather large inter-rater variabilities with typical values of relative volume difference of 68%, mean surface distance of 4.85 mm and overlap error of 75% (Garcia-Lorenzo et al. 2013). It is not clear whether the raters agreed upon a common segmentation protocol, which generally helps to improve rater agreement (Rovaris et al. 1999). Considering the large variabilities, it is arguable if such a gold standard segmentations provide for a reliable validation of automated methods.

Another dataset of clinical MR images of MS patients was disseminated within a challenge on longitudinal lesion segmentation (Pham 2015). The dataset consists of 20 cases of MS patients, each imaged in 3–5 time points by a 3T MR scanner using conventional T1w, T2w, PDw, and FLAIR sequences. Altogether there were 80 datasets and in each the lesions were manually segmented by two trained raters. As of yet, a public website for data dissemination is under development.

The most recent challenge on MS lesion segmentation (Barillot et al. 2016) provides 53 datasets from 4 different sites and 4 different 3T/1.5T MRI scanners. Each set contains FLAIR, pre- and post-contrast T1w and DP/T2w sequences. In each case seven independent experts manually segmented the lesions and a consensus segmentation was created by fusing the segmentations using an automatic

LOP STAPLE algorithm (Akhondi-Asl et al. 2014). The obtained consensus segmentation is intended for evaluation of automated methods.

All three publicly available clinical datasets are disseminated without an exact protocol specification of how raters performed manual segmentations and which tools they used, while even a more critical deficiency is that the resulting gold standard segmentations were not subject to any objective quality assurance. Any gold standard method, even the consensus, should itself be carefully validated and the quality assurance step should answer the question: "What is the variability of the consensus segmentation?" Given that the idea of creating the consensus is to integrate, and possibly even harmonize, expert knowledge of different raters, it seems reasonable to assume that the variability between two consensus segmentations, obtained by the exact same protocol, would exhibit smaller variability compared to inter- and intra-rater variability. If this is confirmed, the so obtained reference lesion segmentations would represent a more precise and reliable target to evaluate and compare against the segmentations obtained from automated methods.

## Contributions

In this paper, we describe a protocol for creating gold standard white-matter lesion (WML) segmentations applied on MR datasets of 30 MS patients, which were acquired on a 3T MR scanner with conventional sequences. On these datasets three expert raters performed segmentations of WMLs, which they then revised in several joint sessions to create a consensus-based gold standard segmentation. The idea was to let the raters critically re-evaluate their segmentations and reach an agreement on expert opinion of what is and what is not a lesion.

To delineate WMLs raters used in-house developed semi-automated medical image visualization and segmentation tools. The protocol was executed twice, with a six months break, whereas the second time consensus segmentations were recreated partially on a subset of axial slices of the MR images in order to evaluate the variability of rater and consensus segmentations. The obtained intra-consensus variability was substantially lower compared to the intra- and inter-rater variabilities, showing improved lesion segmentation consistency by the proposed lesion segmentation protocol.

To encourage other researchers to reproduce and expand the results of this study, we will publicly disseminate on our website <http://lit.fe.uni-lj.si/tools> the tools used to create lesion segmentations, the original and preprocessed MR image datasets and the consensus lesion segmentations. This will also allow interested researchers to train, test and objectively and reliably evaluate existing and novel state-of-the-art (automated) lesion segmentation methods.

## Materials and Methods

### Image Acquisition

A cohort of 30 MS patients were imaged by a 3T Siemens Magnetom Trio MR system at the University Medical Center Ljubljana (UMCL). Each patient's MR scans consisted of a 2D T1-weighted (turbo inversion recovery magnitude, repetition time (TR)=2000 ms, echo time (TE)=20 ms, inversion time (TI)=800 ms, flip angle (FA)=120°, sampling=0.42×0.42×3.30 mm), a 2D T2-weighted (turbo spin echo, TR=6000 ms, TE=120 ms, FA=120°, sampling=0.57×0.57×3.00–3.30 mm) and a 3D FLAIR image (TR=5000 ms, TE=392 ms, TI=1800 ms, FA=120°, sampling=0.47×0.47×0.80 mm). These MR sequences, which are part of a clinical protocol for imaging the MS patients at the UMCL, adhere to current MAGNIMS<sup>1</sup> consensus guidelines (Rovira et al. 2015) for baseline and follow-up evaluation of the MS patients.

All 30 subjects have given written informed consent at the time of enrollment for imaging and the UMCL approved the use of MRI data for this study. The authors confirm that the data were anonymized prior to analysis. Table 1 gives a summary of patient demographic and treatment information at the time of imaging.

### Image Preprocessing

Prior to performing lesion segmentation, each subject's T1w, T2w, and FLAIR images were preprocessed. First, the brain region was masked in the T1w image (Iglesias et al. 2011), followed by mutual-information based registration of the T1w and T2w images onto the FLAIR image using affine transformations (Klein et al. 2010). Based on the computed transformations, the T1w and T2w images, and the T1w brain mask were resampled into FLAIR image space using cubic and nearest neighbor interpolation, respectively. Finally, intensity inhomogeneity correction (Tustison et al. 2010) was performed on each of the masked T1w, T2w and FLAIR images. The voxels lying within the brain mask were considered in lesion segmentation. Using this preprocessing steps the lesion segmentations were obtained in native FLAIR image space.

### Semi-Automated Lesion Segmentation

In order to facilitate the segmentation of lesions we have developed a specialized software named BrainSeg3D that enables their accurate and efficient delineation in 3D MR images. For this purpose the BrainSeg3D, which is based on an open-source medical image processing and visualization

<sup>1</sup>Magnetic resonance imaging in MS [www.magnims.eu](http://www.magnims.eu)

**Table 1** Patient demographic and treatment information

Gender	Age	MS phenotype	Therapy
23 female	25 to 64 years	24 RR	1 Avonex
7 male	Median: 39 years (female 39, male 33)	2 SP 1 PR 2 CIS 1 unspecified	3 Betaferon 1 Copaxone 1 Extavia 5 Gilenya 1 Myfenax 2 Rebif 6 Tysabri 1 unspecified 9 no therapy

RR–relapsing remitting; SP–secondary progressive; PR–progressive relapsing; CIS–clinically isolated syndrome

platform Seg3D (CIBC 2016), provides an interactive local semi-automated segmentation tool. The BrainSeg3D is freely available for download at <http://lit.fe.uni-lj.si/tools>.

For improved detection of lesions the rater could visualize the co-registered T1w, T2w and FLAIR images in side-by-side view. The semi-automated lesion segmentation required a rater to inspect the MR images slice-by-slice, position the mouse cursor over a lesion and define the radius of a local circular region such that the lesion was completely or partially captured together with some of its surrounding structures (Fig. 1b). In the rater-highlighted circular region an automated segmentation was executed in real-time and the MR images were interactively overlaid with the segmentation result, which the rater could then either accept or reject by clicking the mouse (Fig. 1c). For more details please refer to BrainSeg3D manual in the e-Supplements.

The local automated segmentation extracted hyperintense lesions from the FLAIR image by K-means clustering of the intensity values within the local region. Three clusters were found by default, whereas the initial cluster centers were obtained by the K-means++ algorithm (Arthur and Vassilvitskii 2007). Alternatively, a rater could also manually select seed points in the MR images, by which he or she determined the number of clusters and the initial cluster centers for the K-means. The clusters obtained by K-means produced a multi-label segmentation, which was post-processed by the  $3 \times 3$  median filter and connected component analysis, followed by extraction of the component that included the center of the local region. This component represented a tentative binary segmentation, which was interactively overlaid onto the MR images for the rater to accept or reject. Figure 1 shows three examples of the semi-automated lesion segmentation.

The described segmentation approach was found very efficient for creating accurate lesion segmentations and, compared to manual segmentation of WMLs, its use was shown to reduce both intra- and inter-rater segmentation variability (Lesjak et al. 2015).

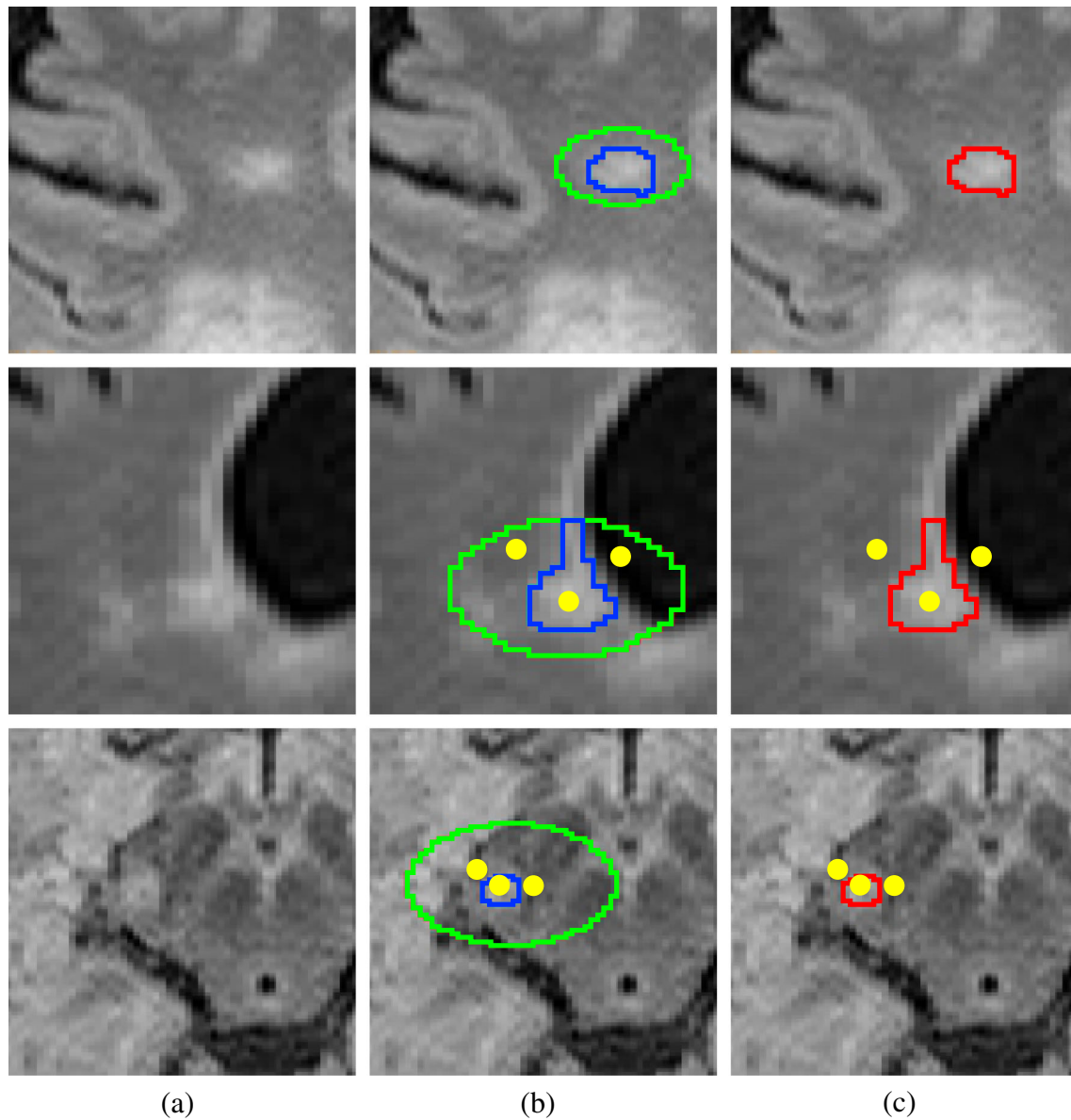
### Lesion Segmentation Protocol

The segmentation protocol is illustrated in Fig. 2. Lesion segmentations were created by three raters. One rater was a second-year radiology intern, while the other two raters were senior neuroradiologists with more than 10 years of experience in assessing MR scans of MS patients.

Prior to segmenting lesions the raters agreed to a common segmentation protocol. Lesion segmentation was to be created in the space of FLAIR image and mainly performed on the axial cross-sections, whereas the co-registered T1- and T2-weighted images were displayed side-by-side to the FLAIR image. The axial plane was selected since T1- and T2-weighted images were acquired in axial planes with 3 mm slice thickness and thus had best resolution in the axial cross-sections. Raters focused on the detection and segmentation of T2w/FLAIR hyperintense lesions within the white-matter. Since the definition of “hyperintense” is somewhat subjective, the raters agreed that a FLAIR hyperintense location is characterized by the FLAIR intensity greater than the FLAIR intensity of closest gray-matter region. Additional criteria for detecting a lesion were the pattern of abnormality, location, and enhancement features, which should be characteristic for MS.

To deal with MR artifacts such as MR signal overshoots around the lateral ventricles, which lead to lesion-like periventricular FLAIR hyperintensities, the presence of true lesions with somewhat ovoid shape stemming from the ventricle wall was to be confirmed on sagittal and coronal cross-sections. Furthermore, due to pulsation of the cerebrospinal fluid (CSF) lesion-like hyperintense FLAIR artifacts may also appear within the ventricles, but which were rejected based on cross-checking the lesion location on the T1w image. To reduce false lesion detection due to partial voluming, for instance at the white- and gray-matter border in FLAIR, the potential lesion locations were to be inspected in coronal and sagittal views before confirming the lesion presence. Similarly, the presence of juxtacortical lesions was to be confirmed on sagittal and coronal cross-sections, as regions, which are hyperintense compared to the nearest cortical FLAIR intensity and extend across the cortical tissue border observed on the T1w image. The hyperintense regions within the cortex on FLAIR were not to be segmented, as well as hypointense lesions in the cortex and white matter on FLAIR images that might be observed on MRI scans due to other possible brain pathology.



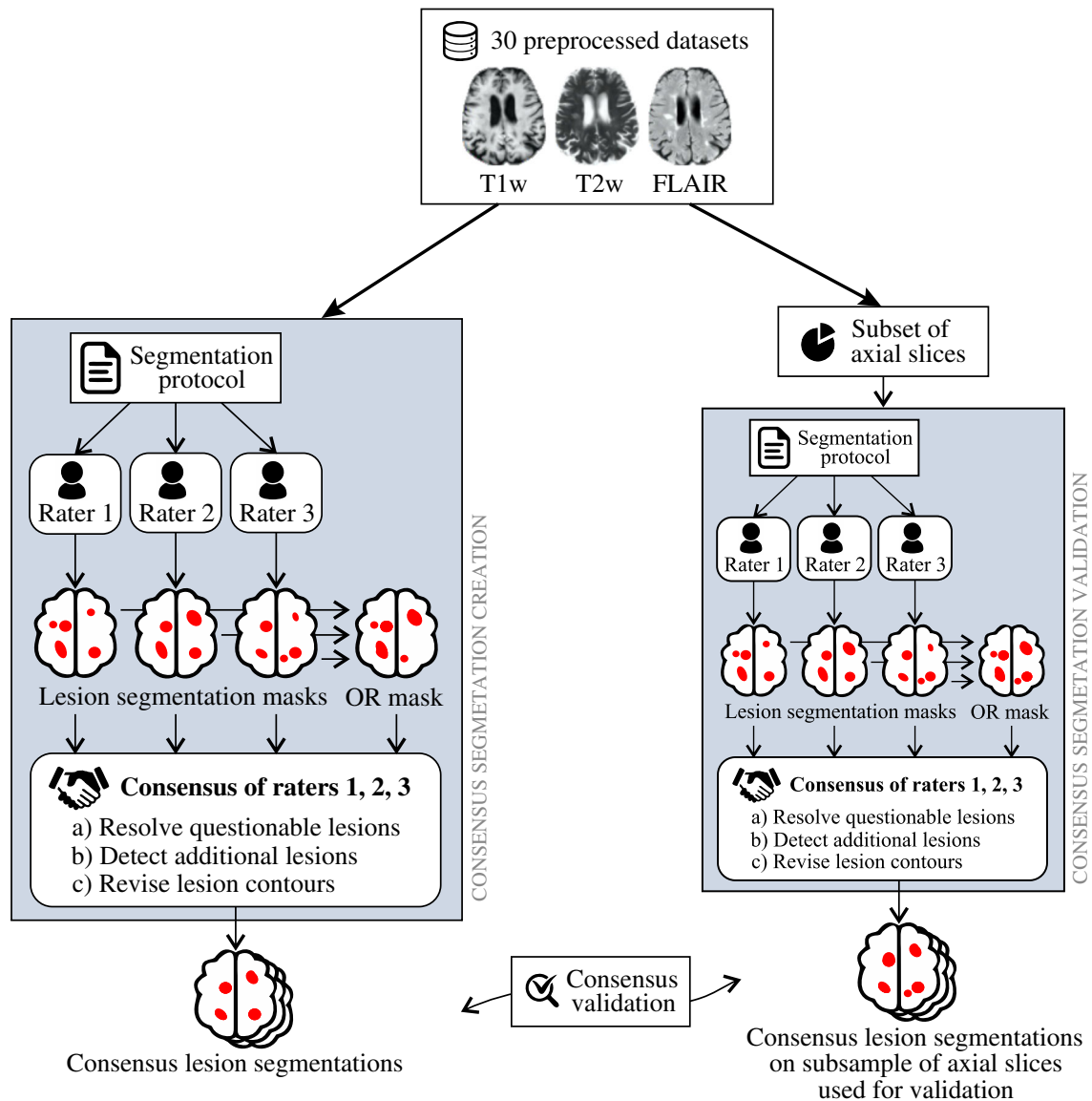


**Fig. 1** Three examples (*in rows*) of using the interactive semi-automated segmentation tool: **a** axial cross-section of FLAIR image with superimposed **b** circular region (*green*) and overlaid result of

a local automated segmentation (*blue*). Rater-selected seed points are shown as *yellow* dots. **c** Upon rater acceptance the local segmentation is propagated to lesion segmentation mask (*red*)

Hyperintense white-matter lesions may also appear due to reasons unrelated to MS, e.g. white-matter ischemia resulting from aging. Distinguishing these lesions from the lesions related to MS based solely on the MR images is difficult and patient's clinical symptoms and background (e.g. cerebrovascular risk factors) may help resolve the lesion origin. Since the majority of patients were young (median age was 39 years, cf. Table 1), with clinically-definite diagnosis of MS and without clinical evidence of other diseases associated to white-matter lesions, the raters did not explicitly consider the lesion origin. Judgement whether a hyperintense white-matter lesion is related to MS or not was based on their personal knowledge and experience.

Initially, each rater independently segmented the lesions on all 30 datasets. For this purpose raters used the BrainSeg3D's semi-automated tool described in the previous section. During the segmentation process, rater had to capture each lesion within a circular region around the mouse cursor and then accept the overlaid tentative segmentation with a mouse click. Segmentation in the circular region was based on clustering contained FLAIR intensity values into three classes. Three classes were sometimes not sufficient to segment lesions in certain brain regions, e.g. typically around gyri, sulci and the brainstem, therefore, in such situations, the rater manually selected seed point locations to determine the class number and corresponding class intensities.



**Fig. 2** Protocol used to create and validate consensus WML segmentations. To create consensus WML segmentations (*left*) three raters independently segment each of the FLAIR images. Their segmentations are then merged using a “boolean OR” operation creating the initial consensus segmentation. Afterwards, the raters jointly refined the initial consensus segmentation by 1) removing falsely detected lesions (eg. segmented MR artifacts), 2) adding previously missed

lesions and 3) by revising and refining lesion contours (see text in “[Lesion Segmentation Protocol](#)” for details). To validate the consensus segmentation, the same segmentation protocol is executed on a subset of axial slices (*right*) to obtain a second set of consensus segmentations. These are then compared to the first consensus segmentations to assess the overall variability of the consensus segmentation protocol (see text in “[Validation of Consensus Segmentation](#)” for more details)

In the following, all three raters held several joint sessions, in which they mutually resolved discrepancies between the individual lesion segmentations on each of the 30 datasets. The goal of these joint sessions was to create the so-called *consensus segmentation* of the lesions. For each subject’s dataset, the independent lesion segmentations of all three raters were first merged using boolean OR operation (ie. each voxel that was identified as belonging to a lesion by any rater, was marked as belonging to a lesion in the merged segmentation) to obtain the most sensitive WML

segmentation, which represented an initial consensus segmentation. This segmentation was then carefully revised by raters, who used the same tools as for creating individual segmentations. The first joint session focused on resolving questionable lesion locations, in which there was no overlap between a connected component in the consensus segmentation and at least one of the three independent segmentations. The second session focused on detecting lesions that were not included in the consensus segmentation. In the last session the raters revised the consensus segmentations,

making sure that the WML contour borders were as accurate as possible and that none of the image slices were skipped during previous segmentation steps. Figure 3 shows two examples of the final consensus lesion segmentations, while an animated preview of consensus lesion segmentation visualized in 3D and in corresponding axial slices of the FLAIR image with overlaid masks is available as part of the [Supplementary Material](#).

All 30 MR datasets were segmented according to the lesion segmentation protocol.

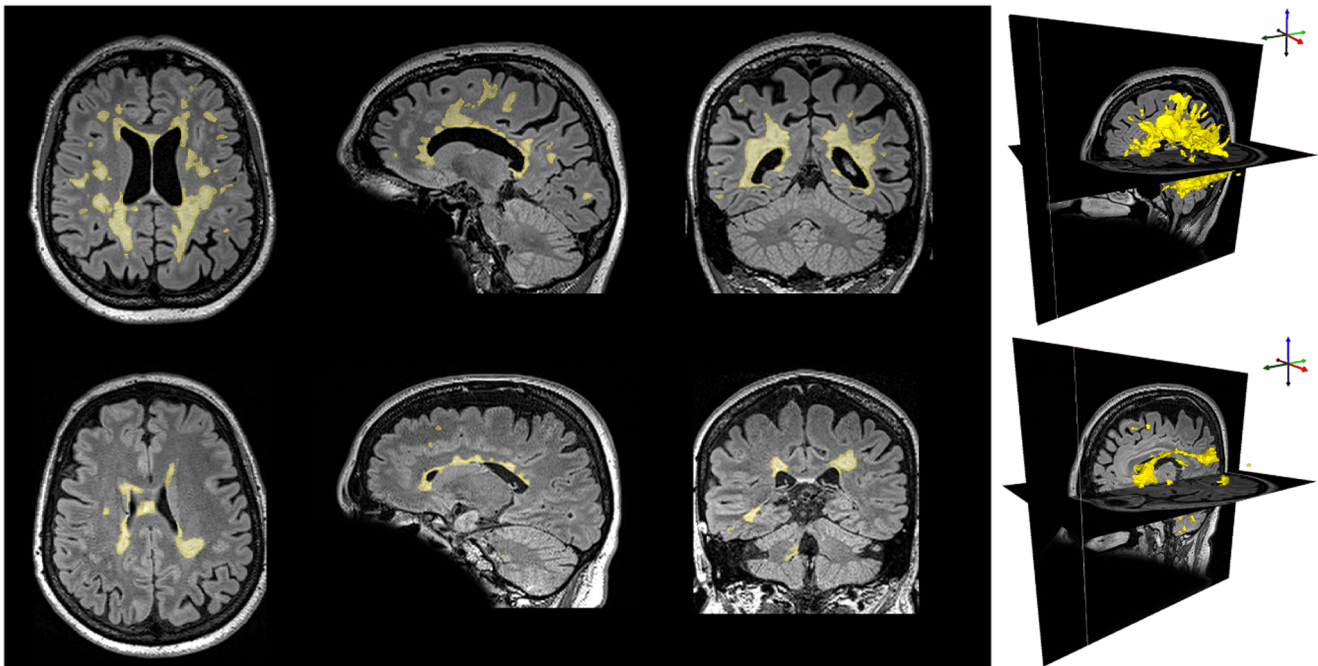
### Validation of Consensus Segmentation

In order to evaluate the quality of consensus segmentations, the same three raters were asked to execute the proposed segmentation protocol (Fig. 2) in two independent rounds on each of the 30 MR datasets. In the first round the raters were required to segment lesions in the whole FLAIR image, while in the second round each case was segmented only in a subset of five axial FLAIR cross-sections so as to speed-up the experiment. Furthermore, to remove possible rater bias due to the effect of learning the second round was performed six months after the first round was completed.

The subset of axial cross-sections considered in the second round was selected automatically in all datasets such that the axial coordinate of each of the five cross-sections per dataset lied in the interval between 5% and 95% of the axial range of subject's brain mask. In this way, the cross-sections at

the extreme supra- and infra-tentorial part of the brain that rarely contain lesions were not considered. On 15 datasets, the cross-sections were selected completely at random from the mentioned interval. On the other 15 datasets, stacks of five consecutive cross-sections were considered such that the central axial coordinate of the stack was selected randomly from the interval. The use of both scattered and stacked axial cross-sections ensured that the segmentations were not biased in favor of either lesion detection or lesion delineation, thereby harmonizing their impact on the assessment of accuracy and reproducibility of consensus segmentation.

The purpose of this experiment was to determine whether the obtained consensus lesion segmentations exhibit better agreement (or lower variability) compared to the individual rater segmentations. Hence, we assessed the variability between the two consensus segmentations, which were obtained on same datasets on two separate occasions, by comparing the lesion masks on the selected axial cross-sections. The inter- and intra-rater variabilities were assessed based on the individual segmentations of the three raters. For comparison, the inter-rater variability was also assessed based on the 2008 MS lesion segmentation challenge training datasets (Styner et al. 2008), in which 10 cases have manual reference segmentations made by two independent raters. To assess the variability improvements of consensus segmentations over that of intra- and inter-rater, intra-consensus variability (ie. variability between the



**Fig. 3** Axial, sagittal and coronal cross-sections of the FLAIR image, and 3D visualization (from left to right) of consensus segmentation of white-matter lesions for two subjects, one with severe (*top*) and the other with moderate (*bottom*) total lesion load

original consensus segmentations and the segmentations of the second consensus created on the subset of axial slices six months later) was also computed.

The variability between any two lesion segmentations was assessed by Dice similarity coefficient (DSC) (Dice 1945) and average symmetric surface distance (SSD). Intra-consensus variability was obtained by averaging DSC and SSD across all 30 datasets, while to measure inter- and intra-rater variabilities, all pair-wise comparisons were made between the three raters and then the obtained DSCs and SSDs were averaged across all 30 datasets. Another measure of variability was Pearson's coefficient ( $r$ ) (Pearson 1895), which was computed between total lesion loads (TLLs) from all 30 datasets from any pair of segmentations methods. The corresponding inter- and intra-rater variabilities were measured by averaging all pair-wise Pearson's coefficients for the three raters.

Furthermore, we analyzed the anatomical distribution of the differences between the TLL and lesion count as measured on the individual rater segmentations and the consensus segmentation on all 30 datasets. On each subject's FLAIR image the individual lesions were obtained by connected component analysis of the segmentations and then assigned into one of four anatomical regions: periventricular, juxtacortical, infratentorial and deep white-matter. These anatomical regions, with the exception of the spinal cord that was not imaged, were chosen as they play an important role in determining the dissemination of lesions in the brain, which is a key factor in the 2010 McDonald diagnostic criteria for MS (Polman et al. 2011).

To obtain consistent assignments of lesions into the four anatomical regions, the corresponding labels were determined automatically based on a patient-specific atlas computed for the T1w image. The patient-specific atlas was obtained using a publicly available method based on a multi-atlas co-registration and label fusion (Cardoso et al. 2015). The patient-specific atlas contained 145 different brain labels, a subset of which were grouped into cortex, lateral ventricles and the cerebellum and brainstem. Individual lesions were classified as periventricular or juxtacortical if there were lesion voxels within 2 mm of the lateral ventricles or cortex, respectively. Infratentorial lesions were those located in the cerebellum or brainstem, while other lesions were marked as deep white-matter lesions.

### Validation of Segmentation Tools

To evaluate the variability of semi-automated vs. manual segmentation tools, we performed an additional experiment, in which we randomly selected 413 of the total of 3316 lesions (approximately  $1/8$ ) from all 30 datasets. Then two independent raters segmented one axial slice per lesion using FLAIR images. The axial slices were chosen in

such way that they intersected the center of lesions (or as close as possible). Each rater segmented the lesions' axial slices twice using manual and twice using semi-automated segmentation tools in the BrainSeg3D. Finally, intra- and inter-rater variability of segmentation obtained by manual or semi-automated tools were assessed by computing the DSC.

## Results and Discussion

### Consensus-based vs. Individual Lesion Segmentations

Table 2 shows that according to DSC, Pearson's  $r$  and SSD the intra-consensus variability was lower compared to both the inter- and intra-rater variability on the same datasets. This indicates that the employed lesion segmentation protocol produced very consistent and reproducible lesion segmentations.

The inter- and intra-rater variabilities on our datasets were (DSC in the range 0.67 – 0.73) comparable to a previous report (0.71–0.81) by (Zijdenbos et al. 1994). Therein the authors also used semi-automated segmentation tools, but the MR datasets were acquired using 2D sequences on 1.5T scanner as compared to our 3D sequences on a 3T scanner. The slightly higher variability observed on our datasets might be because there are generally more lesions visible on 3T than on 1.5T MR images (Di Perri et al. 2009) and because the use of 3D versus 2D FLAIR imaging also increases the sensitivity to lesions (Patzig et al. 2014).

The 2008 MS lesion segmentation challenge (Styner et al. 2008) employed comparable MR datasets to ours, namely 3D acquisition mode on a 3T Siemens MR scanner, however, there the inter-rater variability was 0.237 in terms of DSC. This rather low DSC value indicates a very poor agreement between the raters. Nevertheless, this was one of the first publicly available MR datasets with reference lesion segmentations and even today it is still extensively used to

**Table 2** Variability in terms of average Dice similarity coefficient (DSC) and average symmetric surface distance (SSD) assessment within each rater (intra-rater), between the raters (inter-rater) and between consensus segmentations (intra-consensus)

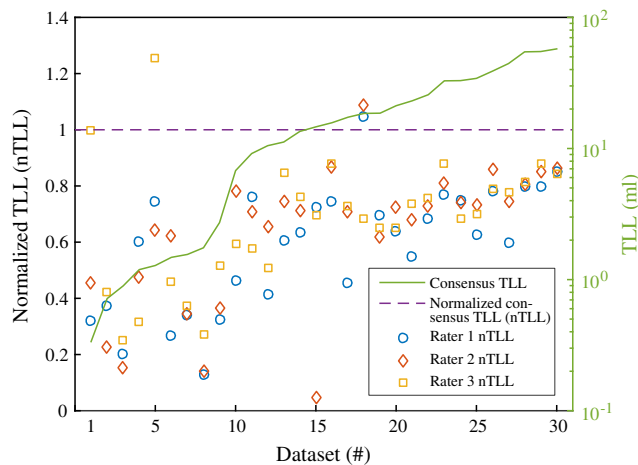
Segmentation variability	DSC	Pearson's $r$	SSD
Intra-consensus	<b>0.776</b>	<b>0.779</b>	<b>1.06</b>
Intra-rater	0.676	0.710	1.73
Inter-rater	0.724	0.731	1.32
Inter-rater (MS challenge 2008)	0.237	0.275	5.86

The reported Pearson's  $r$  is given between total lesion loads of the segmentations. Values are averaged across all 30 datasets and, in case of intra- and inter-rater variabilities also across all three raters. Inter-rater variability for MS Challenge datasets was computed for 10 training cases

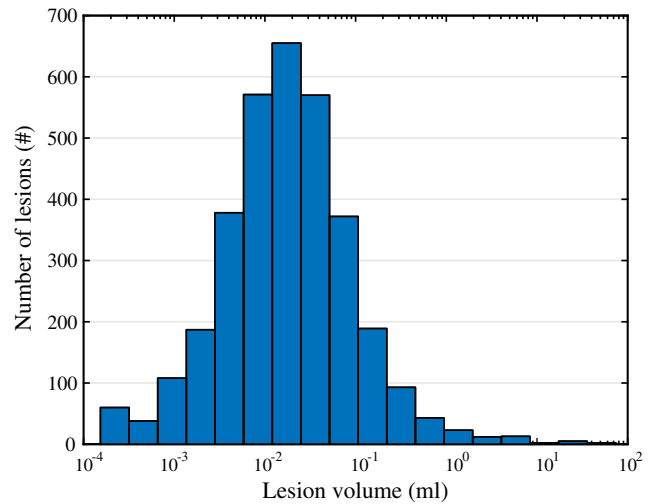


train and tune, quantitatively evaluate and compare state-of-the-art automated lesion segmentation methods. Clearly, public datasets with more accurate and reliable reference segmentations are needed to advance the development and objectively evaluate, compare and rank the automated lesion segmentation methods.

The employed process of merging the WML segmentations of individual raters into final consensus segmentation notably increased the overall sensitivity of lesion detection (Fig. 4). This somehow indicates the subjectiveness of the individual lesion detection and segmentation process. Figure 4 shows another compelling evidence of rater subjectiveness (and possibly rater fatigue); for MR scans with small lesion load (up to 2 ml), some raters were too sensitive, and some less, compared to the consensus, while for larger lesion loads (above 2 ml) the raters consistently segmented less lesions, thus leading to lower TLL. Overall, the TLL of consensus segmentation was significantly higher in comparison to TLL of any single rater’s segmentation ( $p < 0.05$ , Wilcoxon signed-rank test). According to consensus segmentations, the dataset contains a total of 3316 segmented WMLs with an overall TLL of 567 ml. The median TLL per patient was 15.2 ml (min: 0.337 ml, max: 57.5 ml, inter-quartile range: 31.1 ml). The distribution of WML volumes across the 30 datasets is shown in Fig. 5, while their spatial distribution, obtained by mapping all the final consensus segmentations onto the T1w images of the MNI152 atlas (Fonov et al. 2009) through a two-stage affine



**Fig. 4** Total lesion load (TLL) based on individual rater and consensus lesion segmentations on each of the 30 datasets. The TLL obtained from each rater’s segmentation was normalized by consensus-based TLL. For easier interpretation the patients are ordered by increasing consensus TLL. The consensus segmentation exhibits higher overall sensitivity and higher TLL of lesion detection compared to any single rater. For MR scans with TLL above 2 ml the raters consistently segmented less lesions, possibly indicating rater fatigue while contouring cases with large TLL, while with TLL below 2 ml, subjective errors dominated the TLL in both directions

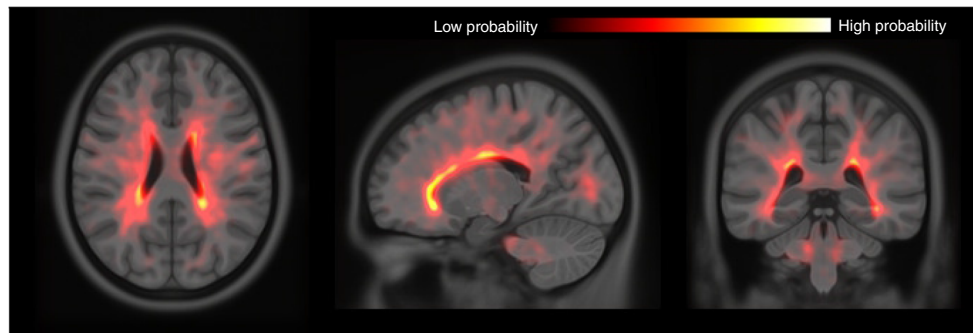


**Fig. 5** Distribution of white-matter lesion number with respect to the volume of individual lesion across 30 datasets. According to consensus segmentations the dataset contains a total of 3316 WMLs with sizes ranging from  $2\mu\text{l}$  to  $250\mu\text{l}$  ( $5^{\text{th}}$  and  $95^{\text{th}}$  percentile respectively), median size of  $17\mu\text{l}$  and an overall TLL of 567 ml

and deformable B-spline T1w-to-T1w registration (Klein et al. 2010), is shown in Fig. 6.

A rater’s judgement about the presence of a lesion is clearly subjective and is probably a process too complex to define rigorously by a set of simple rules. Namely, it is heavily dependent on prior knowledge and experience of the rater and involves a number of other more or less subjective judgements. For instance, the rater has to simultaneously consider the intensity, the spatial location and context of the lesion and its shape, and the spatial distribution and shapes of other lesions, and also the demographic and clinical data in order to successfully differentiate a demyelinating MS lesions from an MRI artifact or lesions due to small-vessel ischemia, Susac’s or CADASIL syndrome, progressive multifocal leukoencephalopathy (PML), etc. The reason we have decided to include a hyperintensity threshold is to impose a somehow more objective criterion onto the raters so as to better differentiate between the focal white-matter lesions, which are typically hyperintense with respect to the gray matter, and the dirty white-matter. These are areas of the white-matter, where demyelination is already in progress, but which show only a slight increase of the intensity. If the decision to include or exclude the dirty white-matter was left to the raters, without any differentiation criterion, then there could potentially be even more discrepancy between the raters’ segmentations. We feel that the use of a hyperintensity threshold is valid since the role of dirty white-matter and whether it should be considered as demyelinated tissue is not yet clear in the MS community.

The main benefit of performing consensus segmentation, although still based on subjective judgement, is the



**Fig. 6** Spatial probability map of white-matter lesion (WML) appearance as a heat-map overlaid on the T1w template image. The probability map was created from co-registered consensus segmentations of the

30 datasets, obtained through non-linear registration of corresponding T1w image of to each dataset to T1w template image of the MNI2009c atlas (Fonov et al. 2009)

increased sensitivity and repeatability of WML detection and their quantification. A revision of segmentations performed during the creation of consensus mostly identified new lesions previously missed by some raters and, in far fewer cases (especially those with small lesion load), differences in lesion contours or false positives due to erroneously segmented MRI artifacts. The consensus segmentation protocol is otherwise a complex process that was made efficient and relatively simple by delegating one task at a time to the raters. At the time consensus, when the majority of lesions were already delineated, the raters focused on three specific tasks: 1) resolve inter-rater segmentation discrepancies, 2) detect new lesions and 3) perform fine lesion contouring. Using this directed focus allowed for high efficiency, but also contributed to higher quality of the gold standard, since questionable and additional lesions, and lesion contours were resolved one-by-one through expert agreement.

Table 3 shows the differences between the individual rater segmentations and the final consensus segmentation in terms of number of missed WMLs and their total volumes grouped by specific anatomical location. Clearly, the raters' sensitivity to detect lesions varies greatly between different anatomical locations. The lesions, which are most likely to be missed are those in the infratentorial and juxtacortical

regions. In the respective regions an individual rater missed 46–68% and 9–49% of lesions with respect to the consensus segmentations. This means that on average each rater missed about 15.6 juxtacortical and 3.8 infratentorial lesions per patient MR dataset, which is clearly alarming if one wants to confirm the diagnosis or assess disease progression according to the current MS diagnostic criteria (Polman et al. 2011).

This study did not attempt to directly address certain interesting questions, such as *What is the optimal number of raters to create the gold standard?* and *How the number of raters affect gold standard quality?* However, from Fig. 4 and Table 3 we can infer that, while segmentations of a single rater are insufficient, since he or she tends to miss a substantial number of lesions, each additional rater adds valuable information. This further implies that more raters would likely produce a more accurate and reliable gold standard. However, with higher number of raters one would sooner or later reach a plateau, where due to inter-rater variabilities of individual lesion contours the variability of multi-rater consensus would no longer decrease steadily.

We have engaged three expert raters in our study, while the majority of previously published researches engaged at most two raters for validation purposes (Garcia-Lorenzo

**Table 3** Total lesion load (TLL) and count of white-matter lesions (WMLs) with respect to their anatomical location. The table contains both the consensus values as well as the absolute values and relative differences for each of the raters

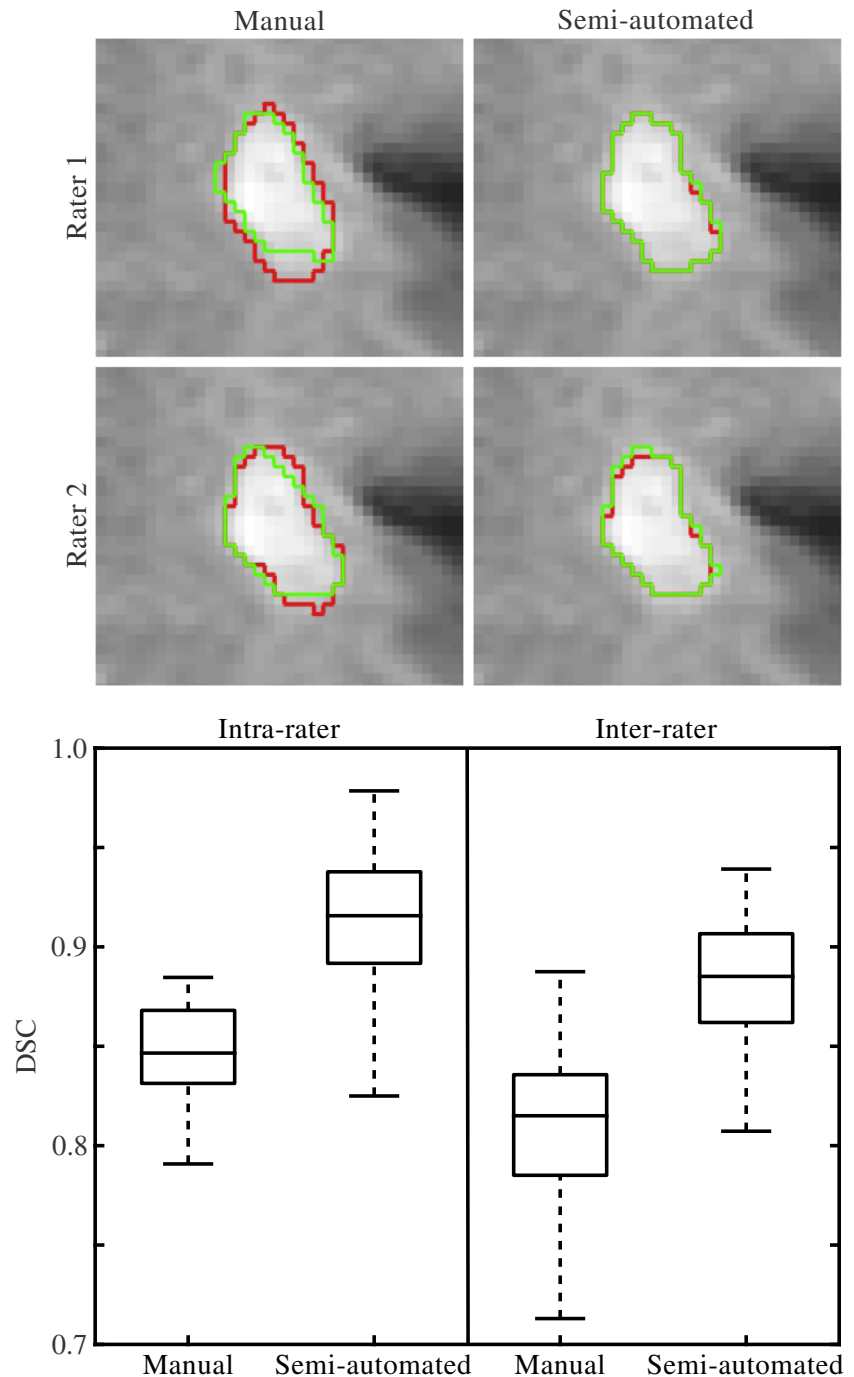
Segmentation	Periventricular		Juxtacortical	
	TLL (ml)	Count	TLL (ml)	Count
Consensus	397	421	110	959
Difference to rater 1	114 (29 %)	101 (24 %)	33.7 (30 %)	469 (49 %)
Difference to rater 2	100 (25 %)	62 (15 %)	22.7 (21 %)	367 (38 %)
Difference to rater 3	99.8 (25 %)	21 (5 %)	21.2 (19 %)	82 (9 %)
Segmentation	Infratentorial		Deep white-matter	
	TLL (ml)	Count	TLL (ml)	Count
Consensus	13.8	167	46.3	1769
Difference to rater 1	5.06 (37%)	109 (65 %)	11.0 (24 %)	538 (30 %)
Difference to rater 2	4.51 (33 %)	114 (68 %)	9.16 (20 %)	536 (30 %)
Difference to rater 3	8.46 (61 %)	77 (46 %)	4.54 (10 %)	-4 (0.2 %)

et al. 2013). It should be noted, that the number of raters to engage is inevitably a compromise between the amount of available resources (experts, time, costs) and the quality of gold standard. In our case, for instance, the use of semi-automated compared to manual tools reduced the time required to perform the segmentation by 25%, however, a single rater still needed around 300 hours (37 days if working 8 hours per day) to segment the lesions of all 30 patient datasets. This shows that production of (consensus-based) gold standard segmentations is very labor intense. From our

experience and the reported results engaging three raters seems to be a good compromise between the obtained gold standard quality and the resources required.

One possible shortcoming of the present study is that the reproducibility of consensus lesion segmentation was evaluated with the same team of expert raters. It would be interesting to compare the consensus segmentation made by another team of experts, using the same semi-automated tools and following the proposed lesion segmentation protocol. Hence, to encourage other researchers to reproduce and

**Fig. 7** Variability of semi-automated vs. manual segmentation tools. Lesion contours were created twice (*top* figure in *red* and *green*) by two raters either using manual or semi-automated segmentation tools. These contours were used to compute intra- and inter-rater DSC values. Box plots (*bottom*) show the median values, 1<sup>st</sup> and 3<sup>rd</sup> quantiles and minimum and maximum values of DSC computed over all 413 segmented lesions as described in “Validation of Segmentation tools”



expand the results of this study, we will publicly disseminate our semi-automated tools and MR image datasets on our website <http://lit.fe.uni-lj.si/tools>. Furthermore, to train, test and objectively and reliably evaluate novel or state-of-the-art (automated) lesion segmentation methods we will also disseminate the consensus lesion segmentations along with the original and preprocessed MR image datasets. In the future, we aim to expand this public repository with more MR datasets acquired on different scanners and sites along with corresponding reference lesion segmentations and establish an on-line system for an objective evaluation of (automated) lesion segmentation methods.

### Manual vs. Semi-automated Lesion Segmentation

The obtained median DSC values were 0.85 and 0.82 for intra- and inter-rater variability using manual tools, while the respective values obtained with the semi-automated tools were 0.92 and 0.89. Note that the DSC values obtained in this experiment are somewhat higher than those reported over all lesion segmentations (Table 2). This is because in this experiment the raters knew the lesion locations in advance, hence, there were no missed lesions, which otherwise can have a strong impact on DSC. Corresponding distributions of the DSC values across all 413 lesion segmentations and an example of the actual difference in lesion contours in Fig. 7 clearly show the advantage of using semi-automated instead of manual lesion contouring tools.

The conclusion of this experiment is that, instead of manual tools, the semi-automated tools should be used to reduce intra- and inter-rater variability. Hence, to reduce error propagation within a multi-rater consensus-based segmentation protocol the raters were required to use semi-automated tools for lesion contouring.

### Conclusions

In this paper, we presented a novel dataset for validation of lesion segmentation methods in MR images and a novel protocol for creating reference white-matter lesion segmentations based on multi-rater consensus. The dataset consists of MR images of 30 patients with MS acquired on a 3T Siemens MRI machine using conventional MR imaging sequences. The reference lesion segmentations were created for each case by three independent raters, who used in-house developed MR image visualization and segmentation tools. The segmentation protocol required from each rater to segment each of the 30 cases. The obtained segmentations were later jointly merged into a consensus segmentation by the same three raters, thereby integrating and harmonizing their expertise. To evaluate the variability of rater and consensus segmentations, and as a quality assurance

step, the segmentation protocol was executed twice on the same MR images, with a six months break. The obtained intra-consensus variability was substantially lower compared to the intra- and inter-rater variabilities, showing improved reliability of lesion segmentation by the proposed protocol. Hence, we conclude that the obtained reference segmentations represent a more precise and reliable target to evaluate and compare against the segmentations obtained from automated white-matter lesion segmentation methods.

### Information Sharing Statement

The MR images of 30 MS patients for this study were acquired on a 3T Siemens Magnetom Trio MR system at the University Medical Centre Ljubljana (UMCL). All 30 subjects have given written informed consent at the time of enrollment for imaging and the UMCL approved the use of MRI data for this study. The authors, who have obtained approval from the UMCL to use the data, confirm that the data was anonymized (patient information removed, defacing of MR images). The semi-automated tools and the anonymized MR datasets in original and preprocessed form as used in this study, and the consensus-based reference segmentations of the lesions, will be made publicly available on our website <http://lit.fe.uni-lj.si/tools>.

**Acknowledgements** The authors would like to acknowledge Nuška Pečarič, MD, from the University Medical Centre Ljubljana for performing the lesion segmentations.

### References

- Akhondi-Asl, A., Hoyte, L., Lockhart, M.E., Warfield, S.K. (2014). A logarithmic opinion pool based staple algorithm for the fusion of segmentations with associated reliability weights. *IEEE Transactions on Medical Imaging*, 33(10), 1997–2009.
- Anbeek, P., Vincken, K.L., van Osch, M.J.P., Bisschops, R.H.C., van der Grond, J. (2004). Probabilistic segmentation of white matter lesions in MR, imaging. *NeuroImage*, 21(3), 1037–1044.
- Arthur, D., & Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete algorithms, SODA '07* (pp. 1027–1035). Philadelphia: Society for Industrial and Applied Mathematics.
- Barillot, C., Commowick, O., Guttmann, C., Styner, M., Warfield, S. (2016). MS Segmentation challenge. Last accessed: 20 oct, 2016. <https://portal.fli-iam.irisa.fr/msseg-challenge/overview>.
- Cardoso, M.J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., Ourselin, S. (2015). Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion. *IEEE Transactions on Medical Imaging*, 34(9), 1976–1988. <https://doi.org/10.1109/TMI.2015.2418298>.
- CIBC (2016). Seg3d: Volumetric image segmentation and visualization. Scientific computing and imaging institute (SCI), download from: <http://www.seg3d.org>.
- Cocosco, C.A., Kollokian, V., Kwan, R.K.S., Pike, G.B., Evans, A.C. (1997). BrainWeb: online interface to a 3d MRI simulated brain database. *NeuroImage*, 5, 425.



- Commowick, O., & Warfield, S. (2009). A continuous STAPLE for scalar, vector, and tensor images: an application to DTI analysis. *IEEE Transactions on Medical Imaging*, 28(6), 838–846. <https://doi.org/10.1109/TMI.2008.2010438>.
- Debette, S., & Markus, H.S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ*, 341, c3666. <https://doi.org/10.1136/bmj.c3666>.
- Di Perri, C., Dwyer, M.G., Wack, D.S., Cox, J.L., Hashmi, K., Saluste, E., Hussein, S., Schirda, C., Stosic, M., Durfee, J., Poloni, G.U., Nayyar, N., Bergamaschi, R., Zivadinov, R. (2009). Signal abnormalities on 1.5 and 3 Tesla brain mri in multiple sclerosis patients and healthy controls. A morphological and spatial quantitative comparison study. *NeuroImage*, 47(4), 1352–1362.
- Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302. <https://doi.org/10.2307/1932409>. ArticleType: research-article / Full publication date: Jul., 1945 / Copyright ©1945 Ecological Society of America.
- 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: A Journal of Neurology*, 118 (Pt 6), 1593–1600.
- Fonov, V., Evans, A., McKinstry, R., Almlí, C., Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage*, 47(Supplement 1), S102.
- 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. *Medical Image Analysis*, 17(1), 1–18. <https://doi.org/10.1016/j.media.2012.09.004>.
- Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G.J., Plummer, D.L., Tofts, P.S., McDonald, W.I., Miller, D.H. (1996). Quantification of MRI lesion load in multiple sclerosis: A comparison of three computer-assisted techniques. *Magnetic Resonance Imaging*, 14(5), 495–505. [https://doi.org/10.1016/0730-725X\(96\)00018-5](https://doi.org/10.1016/0730-725X(96)00018-5).
- Iglesias, J., Liu, C.Y., Thompson, P., Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9), 1617–1634. <https://doi.org/10.1109/TMI.2011.2138152>.
- Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.P.W. (2010). Elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1), 196–205.
- Lesjak, Z., Galimzianova, A., Likar, B., Pernuš, F., Špiclin, Z. (2015). Increased accuracy and reproducibility of MS lesion volume quantification by using publicly available BrainSeg3d image analysis software. In *ECTRIMS Online Library*, no. 116236. <http://onlinelibrary.ectrims-congress.eu/ectrims/2015/31st/116236/iga.lesjak.increased.accuracy.and.reproducibility.of.ms.lesion.volume.html?f=p6m3e891o11460>.
- Llado, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramio-Torrent, A., Rovira, A. (2012). Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. *Information Sciences*, 186(1), 164–185. <https://doi.org/10.1016/j.ins.2011.10.011>.
- Patzig, M., Burke, M., Brückmann, H., Fesl, G. (2014). Comparison of 3d cube FLAIR with 2d FLAIR for multiple sclerosis imaging at 3 Tesla. *RoFo: Fortschritte Auf Dem Gebiete Der Rontgenstrahlen Und Der Nuklearmedizin*, 186(5), 484–488. <https://doi.org/10.1055/s-0033-1355896>.
- Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.
- Pham, D. (2015). Longitudinal MS lesion segmentation challenge. Last accessed: 20 oct, 2016. <http://iacl.ece.jhu.edu/index.php/MSChallenge>.
- 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., Weinstenker, B., Wolinsky, J.S. (2011). Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Annals of Neurology*, 69(2), 292–302. <https://doi.org/10.1002/ana.22366>.
- Popescu, V., Agosta, F., Hulst, H.E., Sluimer, I.C., Knol, D.L., Sormani, M.P., Enzinger, C., Ropele, S., Alonso, J., Sastre-Garriga, J., Rovira, A., Montalban, X., Bodini, B., Ciccirelli, O., Khaleeli, Z., Chard, D.T., Matthews, L., Palace, J., Giorgio, A., De Stefano, N., Eisele, P., Gass, A., Polman, C.H., Uitdehaag, B.M.J., Messina, M.J., Comi, G., Filippi, M., Barkhof, F., Venken, H., MAGNIMS Study Group (2013). Brain atrophy and lesion load predict long term disability in multiple sclerosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, 84(10), 1082–1091.
- Rovaris, M., Rocca, M.A., Sormani, M.P., Comi, G., Filippi, M. (1999). Reproducibility of brain MRI lesion volume measurements in multiple sclerosis using a local thresholding technique: effects of formal operator training. *European Neurology*, 41(4), 226–230.
- Rovira, A., Wattjes, M.P., Tintoré, M., Tur, C., Youstry, T.A., Sormani, M.P., De Stefano, N., Filippi, M., Auger, C., Rocca, M.A., Barkhof, F., Fazekas, F., Kappos, L., Polman, C., Miller, D., Montalban, X., MAGNIMS Study group (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-clinical implementation in the diagnostic process. *Nature Reviews Neurology*, 11(8), 471–482. <https://doi.org/10.1038/nrneuro.2015.106>.
- Stangel, M., Penner, I.K., Kallmann, B.A., Lukas, C., Kieseier, B.C. (2015). Towards the implementation of 'no evidence of disease activity' in multiple sclerosis treatment: the multiple sclerosis decision model. *Therapeutic Advances in Neurological Disorders*, 8(1), 3–13.
- Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S. (2008). 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. *MIDAS Journal*, 2008, 1–6.
- Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6), 1310–1320.
- Uher, T., Vaneckova, M., Sobisek, L., Tyblova, M., Seidl, Z., Krasensky, J., Ramasamy, D., Zivadinov, R., Havrdova, E., Kalincik, T., Horakova, D. (2016). *Combining clinical and magnetic resonance imaging markers enhances prediction of 12-year disability in multiple sclerosis*. Houndmills: Multiple Sclerosis.
- Vrenken, H., Jenkinson, M., Horsfield, M.A., Battaglini, M., Schindler, R.A.V., Rostrup, E., Geurts, J.J.G., Fisher, E., Zijdenbos, A., Ashburner, J., Miller, D.H., Filippi, M., Fazekas, F., Rovaris, M., Rovira, A., Barkhof, F., Stefano, N.D., Group, M.S. (2013). Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis. *Journal of Neurology*, 260(10), 2458–2471. <https://doi.org/10.1007/s00415-012-6762-5>.
- Warfield, S.K., Zou, K.H., Wells, W.M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7), 903–921.
- Zijdenbos, A., Dawant, B., Margolin, R., Palmer, A. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging*, 13(4), 716–724. <https://doi.org/10.1109/42.363096>.
- 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 Transactions on Medical Imaging*, 21(10), 1280–1291.