



A fully automated pipeline for brain structure segmentation in multiple sclerosis



Sandra González-Villà^{a,b,*}, Arnau Oliver^a, Yuankai Huo^b, Xavier Lladó^a, Bennett A. Landman^b

^a Institute of Computer Vision and Robotics, University of Girona, Ed. P-IV, Campus Montilivi, University of Girona, 17003 Girona, Spain

^b Electrical Engineering, Vanderbilt University, Nashville, TN 37235, USA

ARTICLE INFO

Keywords:

Brain structures
Parcellation
Multiple sclerosis lesions
Segmentation
MRI
Multi-atlas
Label fusion

ABSTRACT

Accurate volume measurements of the brain structures are important for treatment evaluation and disease follow-up in multiple sclerosis (MS) patients. With the aim of obtaining reproducible measurements and avoiding the intra-/inter-rater variability that manual delineations introduce, several automated brain structure segmentation strategies have been proposed in recent years. However, most of these strategies tend to be affected by the abnormal MS lesion intensities, which corrupt the structure segmentation result. To address this problem, we recently reformulated two label fusion strategies of the state of the art, improving their segmentation performance on the lesion areas. Here, we integrate these reformulated strategies in a completely automated pipeline that includes pre-processing (inhomogeneity correction and intensity normalization), atlas selection, masked registration and label fusion, and combine them with an automated lesion segmentation method of the state of the art. We study the effect of automating the lesion mask acquisition on the structure segmentation result, analyzing the output of the proposed pipeline when used in combination with manually and automatically segmented lesion masks. We further analyze the effect of those masks on the segmentation result of the original label fusion strategies when combined with the well-established pre-processing step of lesion filling. The experiments performed show that, when the original methods are used to segment the lesion-filled images, significant structure volume differences are observed in a comparison between manually and automatically segmented lesion masks. The results indicate a mean volume decrease of $1.13\% \pm 1.93$ in the cerebrospinal fluid, and a mean volume increase of $0.13\% \pm 0.14$ and $0.05\% \pm 0.08$ in the cerebral white matter and cerebellar gray matter, respectively. On the other hand, no significant volume differences were found when the proposed automated pipeline was used for segmentation, which demonstrates its robustness against variations in the lesion mask used.

1. Introduction

Multiple sclerosis (MS) is a chronic immune-mediated demyelinating disease of the central nervous system. It is characterized by the formation of lesions (also called plaques), inflammation, and the destruction of myelin sheaths of neurons. As in other neurodegenerative diseases, the neuronal and axonal loss that MS patients experience as the disease progresses can be quantitatively evaluated from magnetic resonance (MR) images. This quantification is very useful for practical treatment evaluation, since it has been demonstrated that there is a correlation between brain tissue atrophy measurements and MS disability status (Filippi et al., 2013; Fisher et al., 2008). Furthermore, a number of clinical observations as well as neuropathologic and neuroimaging studies have clearly demonstrated extensive involvement of

the thalamus, basal ganglia, and neocortex in patients with MS (Minagar et al., 2013). In the concrete case of the thalamus, its atrophy has been proved to be a clinically relevant biomarker of the neurodegenerative disease process (Houtchens et al., 2007).

The most extended procedure to obtain this quantification for the different structures or regions that compose the brain is structure segmentation. This technique consists of delineating the brain structures/regions in MR images acquired at different time points and then, computing their volume differences. This segmentation is usually performed on the T1-weighted sequence, due to its good contrast between tissues, where MS plaques appear as focal low signal intensity areas (hypo-intense with respect to white matter (WM)). In order to automate this process and to avoid the intra-/inter-rater variability that manual segmentations introduce, a large number of brain structure

* Corresponding author.

E-mail address: sgonzalez@eia.udg.edu (S. González-Villà).

<https://doi.org/10.1016/j.nicl.2020.102306>

Received 14 August 2019; Received in revised form 31 May 2020; Accepted 1 June 2020

Available online 04 June 2020

2213-1582/ © 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Table 1
MICCAI MSSEG 2016 Challenge dataset acquisition details.

Scanner	Modality	Matrix	Slices	Voxel resolution (mm)
Philips Ingenia 3T	Sagittal 3D FLAIR	336 × 336	261	0.74 × 0.74 × 0.7
	Sagittal 3D T1	336 × 336	200	0.74 × 0.74 × 0.85
	Axial 2D PD-T2	512 × 512	46	0.45 × 0.45 × 3
Siemens Aera 1.5T	Sagittal 3D FLAIR	256 × 224	128	1.03 × 1.03 × 1.25
	Sagittal 3D T1	256 × 256	176	1.08 × 1.08 × 0.9
	Axial 2D PD-T2	320 × 320	25	0.72 × 0.72 × 4 Gap: 1.2
Siemens Verio 3T	Sagittal 3D FLAIR	512 × 512	144	0.5 × 0.5 × 1.1
	Sagittal 3D T1	256 × 256	176	1 × 1 × 1
	Axial 2D PD-T2	240 × 320	44	0.69 × 0.69 × 3

segmentation algorithms have been proposed during the last two decades (Patenaude et al., 2011; Kushibar et al., 2018; Fischl et al., 2002; González-Villà et al., 2016). However, most of them are not designed to deal explicitly with MS lesions, which makes oscillate their accuracy when applied to MS patient images (González-Villà et al., 2017).

A commonly used and successful technique to overcome this issue in automated tissue segmentation consists of replacing the lesion intensities on the T1-w sequence with signal intensities of the normal-appearing WM before segmentation (Valverde et al., 2014). This pre-processing step is commonly referred to in the literature as *lesion filling*, and has achieved a significant reduction in the associated errors of WM lesions in tissue volume measurements (Popescu et al., 2014).

In our previous work (González-Villà et al., 2019), we reformulated two structure segmentation approaches from the literature (Huo et al., 2017; Wang et al., 2013) to segment the brain structures on MR images of MS patients containing lesions. We compared the segmentation results obtained with those proposals to the ones obtained with the original methods when segmenting the lesion-filled images, concluding that with the reformulated strategies the pre-processing step of lesion filling can be disregarded, obtaining similar or even more accurate segmentation results. Both of the analyzed strategies require a previous delineation of the lesions, either to perform lesion filling (in the case of the original methods) or as input of the segmentation algorithm (in the reformulated version). However, in all of the experiments performed, the lesion masks used were annotated manually, which in practice are rarely available, since obtaining them is a highly time-consuming task. Furthermore, the use of manually annotated masks requires expert interaction before applying the structure segmentation algorithms, which is impractical if our objective is to automate the whole brain-parcellation process. Fortunately, an increasing number of automated MS lesion segmentation algorithms have been proposed in recent years with very promising results (Guizard et al., 2015; Deshpande et al., 2015; Roura et al., 2015; Tomas-Fernandez and Warfield, 2015; Harmouche et al., 2015; Valverde et al., 2017). The use of these methods allows us to automate the lesion mask acquisition, with the final objective of segmenting the brain structures in a completely automated manner.

In this work, we present a fully automated pipeline to segment the brain structures on MR images of MS patients, that uses both T1-w and FLAIR modalities. The pipeline follows a multi-atlas strategy (Iglesias and Sabuncu, 2015) in which a set of MR images of healthy subjects with available manual segmentation, i.e. atlases, are non-rigidly registered to the target MS patient T1-w image, masking out the lesion areas

to reduce the effect of the lesion intensities on the registration result. After that, the deformation fields obtained from these registrations are applied to the corresponding segmentations in such a way that new pairs of images (structural image and segmentation) are obtained, which are similar to target. Then, these candidate segmentations of the target are fused (i.e. label conflicts between the candidate segmentations are resolved voxel-wise) with one of the reformulated strategies presented in our previous work (González-Villà et al., 2019), to obtain the final segmentation. The lesion delineations used in this pipeline are obtained automatically from the FLAIR sequence by means of a representative unsupervised method of the state of the art (Roura et al., 2015). In opposition to our previous study, where the lesion masks were annotated manually, here, we combine the reformulated label fusion strategies (González-Villà et al., 2019) with automatically obtained lesion delineations (Roura et al., 2015; Valverde et al., 2017), performing an analysis of the influence of the lesion masks used on the structure segmentation result of the proposed pipeline. Moreover, a detailed analysis of the effect of false positive and false negative lesions on the pipeline result is also performed.

2. Materials and methods

2.1. Data

The images used for evaluation are from the MICCAI MS segmentation (MSSEG 2016) Challenge database (Commowick et al., 2018). This dataset consists of 15 MS patients with lesion loads ranging from 0.91 to 68.94 cm^3 . The images of this database are from three different MRI scanners and different manufacturers including those using 3T and 1.5T magnets. For each MS patient four different MR sequences (3D FLAIR, 3D T1 weighted sequence pre- (T1-w) and post-Gadolinium injection (T1-w GADO) and axial dual PD-T2 weighted sequence) are provided. More details about the acquisition parameters of those images can be found in Table 1. Furthermore, manual lesion delineations from seven different trained experts are also available. From these segmentations, a consensus ground truth segmentation was built for evaluation with the LOP STAPLE algorithm (Akhondi-Asl et al., 2014). Demographics are shown in Table 2.

The atlases used in our experiments consist of 45 T1-w MR images obtained from the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database (Landman and Warfield, 2012). The images were obtained from Open Access Series on Imaging Studies

Table 2
MICCAI MSSEG 2016 Challenge dataset demographics.

Scanner	Patient age		Patient gender			Lesion load (ml)	
	Mean	Std	Male	Female	Male:female	Mean	Std
Siemens Verio 3T	35.00	10.10	1	4	0.25:1	17.40	24.86
Siemens Aera 1.5T	43.80	8.32	2	3	0.67:1	6.68	17.90
Philips Ingenia 3T	46.00	9.14	4	1	4.00:1	9.41	14.92
Overall	41.60	9.85	7	8	0.88:1	10.30	20.15

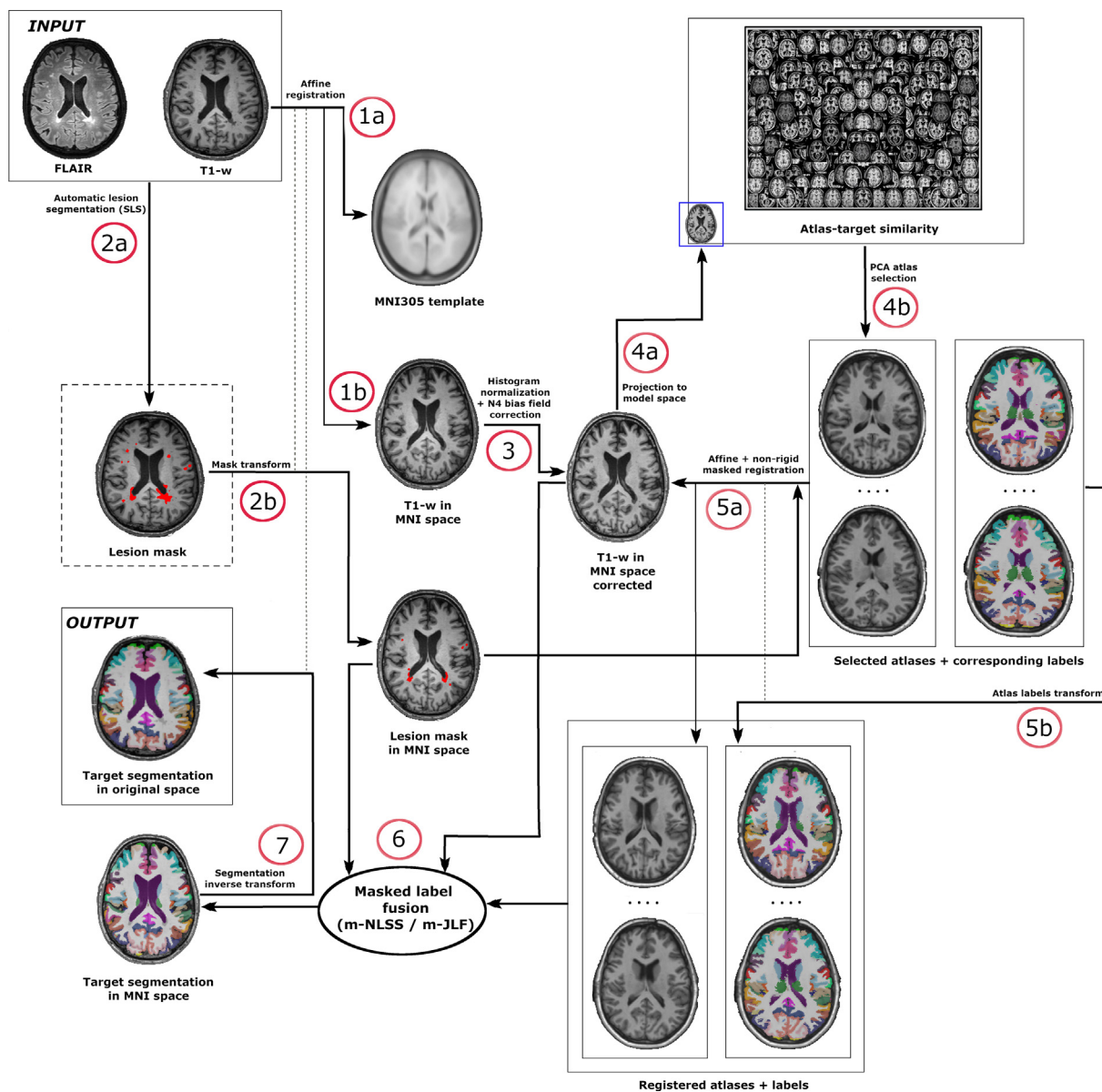


Fig. 1. Fully automated pipeline for structure segmentation of MS patients. As input of this pipeline, the T1-w and the FLAIR sequences of the patient are required. The automatic segmentation of the lesions (step 2a) is performed by means of the SLS algorithm (Roura et al., 2015). Then, the T1-w image and the lesion mask are moved to MNI space (steps 1b and 2b), where the segmentation is performed. Once in MNI space, the patient T1-w image is bias field corrected (Tustison et al., 2010) and intensity normalized to the atlases model space (Asman et al., 2015) (step 3). After that, the target is projected to the model space (step 4a) and the 15 most similar atlases are selected to participate in the segmentation (step 4b). Those atlases are affine (Ourselin et al., 2001) and non-rigid (Avants et al., 2008) registered to the target, masking out the lesion voxels labeled by SLS (step 5a). The deformation fields obtained from these registrations are applied to the corresponding atlas labels (step 5b). Finally, the atlas labels are fused by means of one of the reformulated strategies (González-Villà et al., 2019) (m-NLSS or m-JLF), for which information from the target, the lesion mask and the atlas intensities is required (step 6). The obtained segmentation result is then back-propagated to its original space (step 7).

(OASIS) dataset (Marcus et al., 2007) and labeled according to Brain-COLOR protocol (Klein et al., 2010), including 133 labels that cover the whole brain: sub-cortical structures, ventricles, cerebral WM, cerebellum, brainstem and 98 regions in the cortex.

2.2. The pipeline

In this section, we explain in detail the different steps involved in the proposed automated pipeline. A graphical representation of these steps is presented in Fig. 1. Note from this figure that, besides the brain structure segmentation result, the automatically obtained lesion mask can also be considered as an output of this pipeline in case it was necessary for medical purposes, such as lesion quantification or follow-up.

The proposed pipeline requires as input images the T1-w and FLAIR sequences. Both image modalities have to be co-registered, being the target image space the one in which the segmentation results would be given. The election of the registration method is up to the user, however, for all the experiments presented in this work, we used the niftyreg software (Ourselin et al., 2001) to affinely register the FLAIR image to the T1-w sequence.

In a first step, the co-registered T1-w and FLAIR sequences of the patient are given as input to the automated lesion segmentation algorithm, which generates the lesion mask (step 2a). Here, we have used the Salem Lesion Segmentation algorithm (SLS) (Roura et al., 2015), since it is an unsupervised strategy, and hence it does not require any specific training on the used dataset. Moreover, it has been shown to

provide good segmentation results (i.e. obtained the second best position in both lesion segmentation and detection tasks on the MICCAI MSSEG Challenge (Commowick et al., 2018)). However, any lesion segmentation method of the state of the art can be used instead, as we will see later in our experiments. SLS is an outlier-segmentation-based approach that uses brain tissue labeling and post-processing rules to provide a lesion mask. The authors consider the lesions as intensity outliers, that appear as hyper-intense regions in the FLAIR sequence. First, they perform tissue segmentation on the T1-w sequence, which they use to compute the intensity distribution of the gray matter (GM) in the FLAIR image, since it is the brightest healthy tissue in this modality. Then, since the lesions are even brighter than the GM, their intensities are considered to be outliers of this distribution. Thereafter, some post-processing steps are applied to remove false positive lesions that remain after thresholding the FLAIR volume.

After the lesion segmentation, the T1-w image is affinely registered (Ourselin et al., 2001) to the MNI305 template (step 1a) and the resulting transformation is applied to the obtained lesion mask (step 1b). This is done in such a way that both the T1-w image and the lesion mask are moved to a standardized space (MNI) where the brain structure segmentation takes place. Once in MNI space, the patient T1-w image is N4-bias-field corrected (Tustison et al., 2010), and intensity normalized to a previously built “atlas model” space (Asman et al., 2015) (step 3). Normalizing the target intensities with the atlases is very important since our label fusion strategies depend on correspondence search models based on target-atlas patch-intensity similarity.

In this pipeline, only the 15 atlases more similar to the target, from a cohort of 45 (Landman and Warfield, 2012), are used for segmentation. Those 15 atlases are selected by performing a PCA based atlas-selection strategy (Asman et al., 2015). To obtain the PCA manifold from all the 45 atlases, which is done offline, the 3D intensities within the same MNI brain mask of each atlas are converted to a 1D vector. Then, a naïve PCA projection is performed on 1D vectors from all atlases to learn the PCA manifold. The first fifteen modes of variation in the PCA are used for both projection and measuring the similarity between the atlases and the target image.

After intensity normalization, the patient T1-w image is projected into the same PCA manifold (step 4a) and the 15 atlases with the smallest Euclidean distance to the patient scan are selected to perform the segmentation (step 4b). Then, the selected atlases are registered to the normalized patient image, using an initial affine registration (Ourselin et al., 2001) followed by a non-rigid procedure (Avants et al., 2008) (step 5a). In all the registrations performed, the automatically segmented lesion mask, that is already in MNI space (step 2b), is used to mask out the lesion areas in such a way that we avoid their intensities interfering in the similarity metric calculation. The deformation fields obtained from the registration are then applied to the corresponding atlas labels, which are propagated to the patient space (step 5b), becoming potential brain structure segmentations of the target.

Then, the propagated atlas labels are fused with one of the reformulated strategies proposed in our previous work (González-Villà et al., 2019), i.e. masked Non-local Spatial STAPLE (m-NLSS) and masked Joint Label Fusion (m-JLF), to obtain the final brain structure segmentation. These strategies exploit the target-atlas similarity under the assumption that images with similar appearance are more likely to have similar segmentations. Assuming that the one-to-one mapping obtained from the atlas-target registration is not perfect, they re-compute the correspondences for every voxel of the target image and the atlases before segmentation, based on patch-intensity similarity. Since the target image contains MS lesions, these methods assume that their abnormal intensities may affect the correspondence finding on the healthy atlases, obtaining more inaccurate matches than the ones obtained after a masked registration to the atlas. For this reason, they force the correspondence imposed by the registration result on the lesion areas, whereas they redefine the patch shape on the surroundings of the lesion to prevent these abnormal intensities from interfering in

the correspondence search.

Both m-NLSS and m-JLF require information from the patient T1-w image, the lesion mask and the atlas structural images to obtain the atlas-target correspondences. Thus, in the last step of this pipeline, both the atlas labels and intensity images, combined with the patient structural image and its corresponding lesion mask are fed to the fusion algorithm, that computes the final segmentation (step 6). Note that the obtained segmentation is in MNI space, therefore, the inverse of the transformation resulting from affine registration of the original patient image to the MNI305 template is applied to the fusion result to move it back to the original patient space (step 7).

2.3. Evaluation

In our experiments, we evaluate how the automatically segmented lesion masks affect the output of the proposed pipeline. To do this, we compare the segmentation result of the label fusion methods when using the consensus masks described in Section 2.1, and the automatically-obtained SLS masks. Besides, in order to better analyze the effect that the lesion mask has on the proposed pipeline, we also evaluate a second automated lesion segmentation method (Valverde et al., 2017). This second method follows a supervised strategy based on Convolutional Neural Networks (CNN), in contrast to SLS, which is an unsupervised method. This approach was also presented to the MSSEG 2016 Challenge, obtaining the best position in lesion detection, and the third best position in lesion segmentation. To perform such experiments, the 15 images are segmented with both methods (m-NLSS and m-JLF) tree times (consensus, SLS and CNN), applying the same lesion masks also for the atlas masked registration.

Furthermore, to be consistent with our previous work (González-Villà et al., 2019), we also compare how the original strategies, i.e. Non-local Spatial STAPLE (NLSS) (Huo et al., 2017) and Joint Label Fusion (JLF) (Wang et al., 2013), behave when we perform lesion filling (Valverde et al., 2014) with the consensus, the SLS, and the CNN masks. The filling strategy used (Valverde et al., 2014) is the same as in González-Villà et al. (2019). This method replaces the lesion voxel intensities by random values of a normal distribution generated from the mean WM signal intensity of each two-dimensional slice. As stated by their authors, this technique is a compromise between local methods – which use local intensities from the surrounding neighboring voxels of lesions – and global methods – which use global WM intensities from the whole brain–, reducing the bias caused by refilled voxels on GM and WM tissue distributions by means of global information from the whole slice, whereas aims to reproduce more precisely the signal variability between slices by means of re-computing the mean signal intensity of the normal-appearing WM at each slice.

In this second case, the pipeline presented in Fig. 1 is slightly modified. In particular, before registering the target T1-w image to the MNI305 template, lesion filling (Valverde et al., 2014) is applied on this image, as indicated by the lesion mask. Then, the lesion mask is no longer used in the rest of the pipeline, neither in the atlas registration (non-masked) nor in the label fusion (original strategies).

Since brain structure ground truth is not available for the MSSEG 2016 database, we quantitatively evaluate the effect of the automatic lesion masks on the segmentation using the structure volume percentage increase with respect to the manual lesion mask execution, as follows:

$$Vol_{\%increase} = \frac{100 \cdot (Vol_{automatic} - Vol_{manual})}{Vol_{manual}} \quad (1)$$

where $Vol_{automatic}$ and Vol_{manual} are the structure volumes when the automatic and consensus lesion masks are fed to the pipeline, respectively.

In order to provide global results, the 133 structures obtained from the segmentation are combined into the following regions: cortical GM, cerebral WM, sub-cortical GM, CSF, cerebellum GM, cerebellum WM and brainstem.

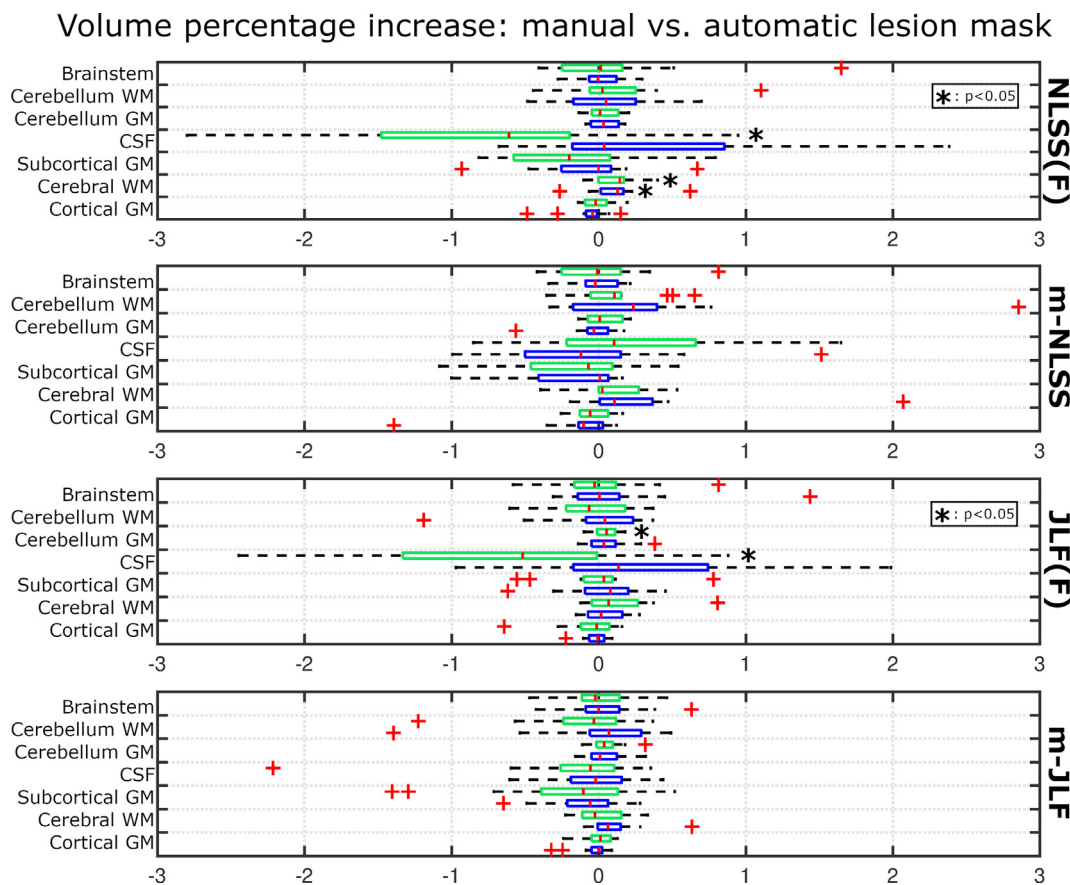


Fig. 2. Structure volume percentage increase and/or decrease. Comparison of the structure volume change when utilizing manual vs. automatic (SLS/CNN) lesion masks on the analyzed brain structure segmentation strategies: (1) NLSS on the lesion-filled target image (NLSS(F)), (2) masked NLSS on the original target image (m-NLSS), (3) JLF on the lesion-filled target image (JLF(F)) and (4) masked JLF on the original target image (m-JLF). Green boxplots represent the structure volume percentage change when comparing SLS to manual lesions masks, whereas blue boxplots represent the change between CNN and manual lesion masks. Significance assessed for each method and structure independently, with box-plots significance independent to each other.

Statistical analysis is performed using the Matlab software package. We test for significant differences in the structure volumes obtained with the analyzed methods, considering the Bonferroni correction to counteract the multiple comparisons problem. These differences are computed using one-sample t-tests of the structure volume percentage increase. Moreover, the Pearson's linear correlation coefficient is used to compute the correlation between the lesion volume percentage increase and the structure volume percentage increase achieved when comparing manual vs. automatic lesion masks.

3. Results

3.1. Quantitative results

In the following, we present a comparative analysis of the structure volume changes observed on the analyzed methods, when using the two different lesion masks (manual/automatic). Fig. 2 shows the structure volume percentage increase and/or decrease obtained for the four parcellation algorithms: (1) Non-local Spatial STAPLE segmentation of the lesion-filled target image (NLSS(F)), (2) masked Non-local Spatial STAPLE of the original target image (m-NLSS), (3) Joint Label Fusion of the lesion filled target image (JLF(F)), and (4) masked Joint Label Fusion of the original target image (m-JLF).

Each of the four box-plots shows the differences of executing the same pipeline, i.e. one among (1)–(4), when using manual vs. automatically segmented lesion mask as input. Hence, for example, the NLSS(F) sub-figure shows the comparison of the result of NLSS when segmenting the images filled according to the manual lesion masks, to

the segmentation obtained with NLSS when segmenting the images filled according to the automatically obtained lesion segmentation (for both CNN and SLS separately). The same applies for the rest of sub-figures.

The two boxplot colors represent the structure volume changes when comparing the results using manual lesion masks and those obtained using the SLS (green) and CNN (blue) lesion masks.

Notice that significance is assessed for each box-plot independently, since they are independent to each other. Each of them represents the volume percentage increase and/or decrease of the analyzed structure when the presented pipeline is fed with automatic lesion masks (SLS/CNN) with respect to the same pipeline fed with manual (consensus) lesion masks. Negative values indicate a volume decrease.

From the figure we can observe that with the reformulated methods (m-NLSS and m-JLF), the volume change of the analyzed structures was not significant, when using either manual or automatic lesion masks, for both SLS and CNN lesion segmentation methods. On the other hand, when comparing the result of segmenting the lesion-filled images with the original methods (NLSS(F) and JLF(F)), we observed that some brain structures presented a significant volume change. In the case of NLSS(F), the cerebral WM presented a $0.13\% \pm 0.14$ ($p < 0.05$) volume increase when using SLS lesion masks, and a $0.11\% \pm 0.19$ ($p < 0.05$) when using the CNN masks. In the case of the CSF, it overcame a $1.13\% \pm 1.93$ ($p < 0.05$) volume decrease when comparing the NLSS(F) results of using the SLS masks to perform the filling vs. the manually annotated lesion masks. Furthermore, in the case of JLF(F), the cerebellum GM presented a $0.05\% \pm 0.08$ ($p < 0.05$) volume increase, whereas the CSF showed a $0.74\% \pm 1.03$ ($p < 0.05$) volume decrease

when comparing the use of SLS and consensus lesion masks.

In a second analysis, we compared the structure segmentation results of the four label fusion strategies when using the automatically segmented lesion masks as input of the pipeline (SLS vs. CNN). In the case of the reformulated methods (m-NLSS and m-JLF), no significant structure volume changes were observed when using either SLS or CNN lesion masks. Contrarily, when the original strategies were used to segment the lesion-filled images (NLSS(F) and JLF(F)), a significant volume change of the CSF was observed in both label fusion strategies. In the case of NLSS(F), the CSF volume decreased a $1.59\% \pm 2.12$ ($p < 0.05$) when using the SLS lesion masks with respect to the CNN masks, whereas in the case of JLF(F), the decrease was a $1.00\% \pm 1.21$ ($p < 0.05$).

Lastly, we analyzed the extent to which volume changes in the lesion load (manual vs. automatic) affected the observed structure volume changes of the evaluated methods. In the case of the CNN lesion volume, significant correlations were found on the volume changes of the CSF when segmented with NLSS(F) ($r = -0.73$, $p < 0.05$) and with JLF(F) ($r = -0.73$, $p < 0.05$); on the cortical GM ($r = 0.53$, $p < 0.05$), the cerebral WM ($r = -0.55$, $p < 0.05$), and the cerebellum WM ($r = -0.54$, $p < 0.05$) when segmented with m-NLSS; and on the sub-cortical GM ($r = -0.52$, $p < 0.05$) when segmented with JLF(F). On the other hand, in the case of SLS, no significant correlations were found between the lesion volume percentage increase and the structure volume percentage increase, except for the cerebral WM, which showed a weak correlation ($r = 0.53$, $p < 0.05$) with the lesion load when the method analyzed was NLSS(F).

3.2. Qualitative results

Some qualitative results of the segmentation outputs obtained in our experiments are shown in this section. Fig. 3 presents an example of a false positive lesion on the left hippocampus (right side of the image), and an over-segmented lesion on the right hippocampus (left side of the image). The figure shows the (a) FLAIR and (b) T1-w sequences of the target image and the superimposed lesion masks, in red, of (c) the experts' consensus delineation and (d) the automatic segmentation obtained with SLS. In addition to this, the resulting images of applying lesion filling to the target image with (e) the manual and (f) the SLS masks are depicted in the image. As we can observe in this image, the hippocampus area, which is highlighted with red circles on the T1-w sequence, i.e. Fig. 3 (f), looks brighter than the normal-appearing GM on the FLAIR sequence (Fig. 3 (a)). For this reason, in this particular case, the automated lesion segmentation method has mis-classified some voxels of this area as outliers (lesions). In both lesions shown in the figure, the over-segmented voxels belong to the hippocampus, which is a GM structure. Therefore, the resulting "extra" filled lesion voxels (Fig. 3 (f)) obtained with the automatic mask show abnormal WM-like intensities on that area. On the other hand, when the manual mask is used to fill the lesions, both hippocampi seem to conserve their original intensities and shapes.

Regarding the effect of this mis-classification on the segmentation output of the analyzed methods, we can observe from Fig. 3 that independently of the lesion mask used, the proposed strategies (Fig. 3 (i), (j), (m), and (n)) present similar structure classification results. However, in the case of the filled images we can observe a similar trend with both segmentation methods (NLSS and JLF). In the case of the automatic lesion mask, the abnormal WM-like intensities are producing an under-segmentation of both hippocampi (Fig. 3 (h) and (l)) when compared to the manual lesion mask (Fig. 3 (g) and (k)).

Another example is illustrated in Fig. 4. This figure shows a case of false negative lesions, two of them surrounded by WM and one periventricular lesion, i.e. attached to the lateral ventricle. From Fig. 4 (d) we can observe that some lesions have not been detected by the automated segmentation method (highlighted with red circles on Fig. 4 (f)). Hence, after applying the filling on the target image with the SLS mask

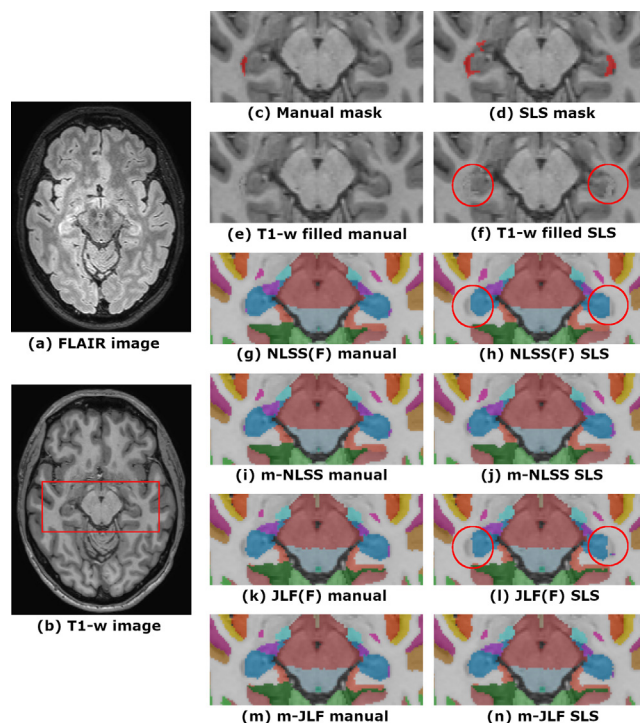


Fig. 3. False positive and over-segmentation example. Axial slice of the original (a) FLAIR and (b) T1-w sequences, (c) the super-imposed consensus lesion mask, and (d) the automatic lesion mask. T1-w image with lesion filling on the (e) the manual and (f) automatic lesion masks. NLSS segmentation on the target image filled (NLSS(F)) with (g) the manual and (h) automatic lesion masks. Segmentation result of m-NLSS for the target image with (i) the manual and (j) the automatic lesion mask as input. JLF segmentation on the target image filled (JLF(F)) with (k) the manual and (l) automatic lesion masks. Segmentation result of m-JLF for the target image with (m) the manual and (n) the automatic lesion mask as input. Note that (g-h) and (k-l) are the result of segmenting the lesion-filled images, however the original target is shown under the segmentation in order to allow an easier comparison with (i-j) and (m-n), respectively.

(Fig. 4 (f)), their abnormal intensities have not been replaced by normal-appearing WM intensities, contrary to Fig. 4 (e).

When analyzing the behavior of the multi-atlas segmentation strategies (Fig. 4 (g)–(n)) we observed that both families of methods (NLSS and JLF) achieved similar results. In this case, the three highlighted lesions should be classified as WM. However, if we pay attention to the peri-ventricular one (right-inferior circle), we realize that when using the automatic lesion mask, in which that lesion has not been detected, the four analyzed strategies (Fig. 4 (h), (j), (l), and (n)) classify it as part of the lateral ventricles. On the other hand, when the consensus manual mask is used, both original methods previous lesion filling (Fig. 4 (g) and (k)) and the proposed strategies (Fig. 4 (i) and (m)), properly classify the lesion as WM.

Regarding the other two lesions (two circles on the top), they are totally surrounded by WM. In this case, when the manual lesion mask is used, either to fill the lesion intensities before performing the structure segmentation with the original methods (Fig. 4 (g) and (k)), or as input to the proposed strategies (Fig. 4 (i) and (m)), the lesions are correctly classified as WM. On the other hand, when the automatic lesion mask is used, where the lesions have not been detected, all the methods also segment the lesions as part of the WM. Even though this behavior may seem disconcerting, it is totally normal, since the original methods (NLSS and JLF) do not always mis-classify the lesion areas. In particular, they tend to succeed when lesions appear surrounded by WM. Given that NLSS and JLF correctly classify these two lesions in the original (not filled) T1-w image, the analyzed methods, which behave exactly as their originals when no lesion mask applies, also succeed in

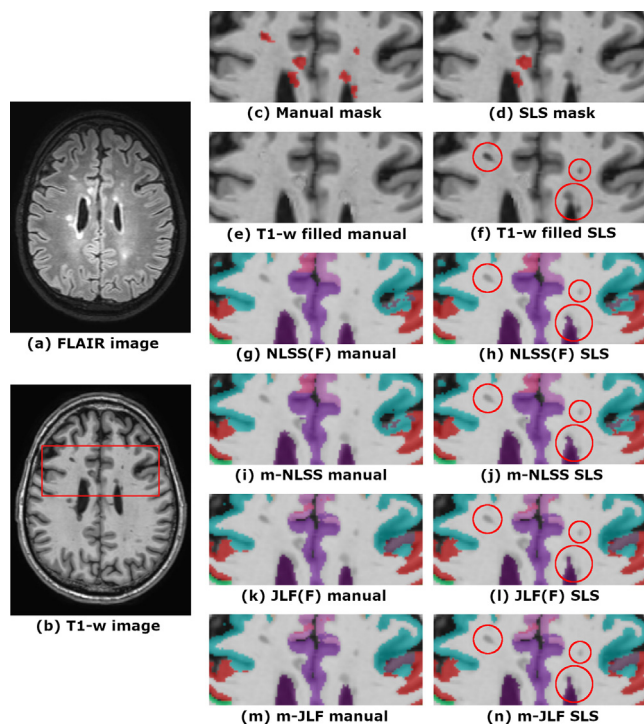


Fig. 4. False negative example. Axial slice of the original (a) FLAIR and (b) T1-w sequences, (c) the super-imposed consensus lesion mask, and (d) the automatic lesion mask. T1-w image with lesion filling on the (e) the manual and (f) automatic lesion masks. NLSS segmentation on the target image filled (NLSS(F)) with (g) the manual and (h) automatic lesion masks. Segmentation result of m-NLSS for the target image with (i) the manual and (j) the automatic lesion mask as input. JLF segmentation on the target image filled (JLF(F)) with (k) the manual and (l) automatic lesion masks. Segmentation result of m-JLF for the target image with (m) the manual and (n) the automatic lesion mask as input. Note that (g-h) and (k-l) are the result of segmenting the lesion-filled images, however the original target is shown under the segmentation in order to allow an easier comparison with (i-j) and (m-n), respectively.

their classification.

4. Discussion

We have evaluated the effect of using automatic lesion masks, from both SLS and CNN, on the brain structure segmentation results of two proposed masked strategies (m-NLSS and m-JLF) and on their corresponding originals combined with lesion filling (NLSS(F) and JLF(F)). The results obtained showed that with the proposed strategies, no significant structure volume differences were found with respect to the use of manual lesion masks. On the other hand, when the original methods were used, significant volume differences on the CSF, the cerebellum GM and the cerebral WM appeared, when comparing the segmentation result obtained for both the lesion-filled images (manual vs. automatic lesion masks).

Therefore, based on the analysis performed, we can conclude that the most robust procedure to segment the brain structures in MRI images of patients with MS visible lesions would be the combination of one of the proposed masked label fusion strategies, i.e. m-NLSS or m-JLF, with any of the analyzed lesion segmentation strategies. Particularly, we would suggest the use of m-JLF combined with SLS. Combining m-JLF with any of the automated lesion segmentation methods we obtain structure volumes significantly comparable to the ones obtained when using manually segmented lesion masks, which allows us to automate the lesion segmentation task, without the need of human expert intervention. Even though the results obtained with either SLS or CNN are comparable, we suggest the use of SLS, since it is an

unsupervised strategy and does not need to be trained for a specific dataset, obtaining at the same time slightly better segmentation performance (69.03 ± 9.55 vs. 67.37 ± 13.18 Dice Similarity Coefficient (DSC) on the analyzed database). Regarding to the label fusion strategy, m-JLF showed less variation in volumetric measurements in comparison to m-NLSS, which makes this method more robust when combined with automatic lesion masks. Besides, based on the quantitative analysis performed in our previous work (González-Villà et al., 2019), where both methods were tested on a database of 45 simulated MS patients, m-JLF showed better global segmentation results (86.03 ± 1.16 DSC) in comparison to m-NLSS (79.35 ± 1.34 DSC).

Regarding the nature of the lesions, when false positives are found by the automated algorithm, either in the form of a new lesion or as an over-segmentation of an existing lesion, the use of lesion filling might be risky. Note that SLS segments lesions as hyper-intense outliers in the FLAIR sequence, and therefore, since the GM is the brightest healthy tissue in FLAIR, it is probable that the false positives belong to this tissue. If this is the case, filling the healthy GM tissue with normal-appearing WM intensities might be self-defeating, causing the automated brain structure segmentation methods to mis-classify those areas, as seen in Section 3.2. On the other hand, with the reformulated methods, the structure classification in the “fake lesion areas” relies only on the label information of the healthy atlases. Since we do not take the intensities into account, it becomes a “traditional” label fusion problem, i.e. fusion without structural image information, which has been demonstrated to achieve very competitive results on healthy subjects (Artachevarria et al., 2009).

Alternatively, the effect of false-negative lesions will depend mostly on the behavior of the original algorithms (NLSS and JLF) against that particular lesion, but also on the (masked) registration result. In particular, when the lesion is totally surrounded by the structure to which it belongs, for example WM, the original approaches tend to correctly classify it as part of this structure, as we saw in Section 3.2. Note that the effect will be the same when a lesion of this kind is detected as a false positive. However, in the case of juxta-cortical (attached to the cerebral cortex) or peri-ventricular lesions (abutting the lateral ventricles), the result is more uncertain, since the lesion intensity and morphology play an important role in the segmentation result. In any case, false negative lesions will have the same effect on both the original previous lesion filling and the proposed strategies, given that the proposed methods behave as their originals in the absence of lesions.

Note that automated lesion segmentation methods are not free of producing segmentation errors (i.e. misclassification of dirty white matter, periventricular lesions, imaging artifacts, etc). If desired, those segmentation errors could be corrected or refined by manual intervention of physicians and radiologists, which could help to improve the final brain segmentation results. Nevertheless, without any manual intervention, the proposed pipeline, i.e. automated lesion segmentation (Roura et al., 2015; Valverde et al., 2017) and masked label fusion (González-Villà et al., 2019), has shown similar results to those obtained with manually annotated masks.

The present study is not free of limitations. The most important one is the limited number of analyzed images. However, in spite of being obtained from only 15 patients, we believe that the database was very heterogeneous in terms of scanner manufacturers, magnetic fields and lesion loads, providing, in addition, reliable lesion masks, which were also necessary for the study. Another limitation of this study is the lack of gold standards for the volume of the analyzed structures. Even though this would make the demonstrated robustness of our pipeline more compelling, as far as we know, there is not any publicly available database with both MS lesion and structure segmentation ground truth where we can evaluate our methods.

In conclusion, the proposed pipeline with any of the reformulated segmentation strategies, m-NLSS and m-JLF, has shown to be robust against variations in the lesion mask used. The robustness of these strategies when used in combination with automatically segmented

lesion masks makes the presented fully automated pipeline totally suitable for medical practice, obtaining similar results to the ones achieved with the use of expert segmented lesion masks. However, the same conclusion cannot be applied to the original strategies when segmenting the lesion-filled images. Even though the volume change observed for most of the structures is low, the results show that these changes are significant when comparing the segmentation of the target image filled with the manual and automatic lesion masks.

CRedit authorship contribution statement

Sandra González-Villà: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Arnau Oliver:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition. **Yuankai Huo:** Conceptualization, Software, Writing - review & editing. **Xavier Lladó:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Bennett A. Landman:** Conceptualization, Supervision, Resources, Writing - review & editing.

Acknowledgements

This work has been supported by “La Fundació la Marató de TV3”, and by Retos de Investigación TIN2015-73563-JIN and DPI2017-86696-R.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.nicl.2020.102306>.

References

- Akhondi-Asl, A., Hoyte, L., Lockhart, M., Warfield, S., 2014. A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights. *IEEE Trans. Med. Image.* 33 (10), 1997–2009.
- Artaechevarría, X., Muñoz-Barrutia, A., Ortiz-de Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. *IEEE Trans. Med. Image.* 28 (8), 1266–1277.
- Asman, A.J., Huo, Y., Plassard, A.J., Landman, B.A., 2015. Multi-atlas learner fusion: an efficient segmentation approach for large-scale data. *Med. Image Anal.* 26 (1), 82–91.
- Avants, B., Epstein, C., Grossman, M., Gee, J., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12 (1), 26–41.
- Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Amélie, R., Ferré, J.-C., Kerbrat, A., Tournias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., Mahbod, A., Wang, C., McKinley, R., Wagner, F., Muschelli, J., Sweeney, E., Roura, E., Lladó, X., Santos, M.M., Santos, W.P., Silva-Filho, A.G., Tomas-Fernandez, X., Urien, H., Bloch, I., Valverde, S., Cabezas, M., Vera-Olmos, F., Malpica, N., Guttmann, C., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S.K., Cotton, F., Barillot, C., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Reports* 8 (1), 1–17.
- Deshpande, H., Maurel, P., Barillot, C., 2015. Classification of multiple sclerosis lesions using adaptive dictionary learning. *Comput. Med. Imaging Graph.* 46, 2–10.
- Filippi, M., Preziosa, P., Copetti, M., Riccitelli, G., Horsfield, M., Martinelli, V., Comi, G., Rocca, M., 2013. Gray matter damage predicts the accumulation of disability 13 years later in MS. *Neurology* 81 (20), 1759–1767.
- Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33 (3), 341–355.
- Fisher, E., Lee, J., Nakamura, K., Rudick, R., 2008. Gray matter atrophy in multiple sclerosis: a longitudinal study. *Ann. Neurol.* 64 (3), 255–265.
- González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., Lladó, X., 2016. A review on brain structures segmentation in magnetic resonance imaging. *Artif. Intell. Med.* 73, 45–69.
- González-Villà, S., Valverde, S., Cabezas, M., Pareto, D., Vilanova, J., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X., 2017. Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation. *NeuroImage: Clinical* 15, 228–238.
- González-Villà, S., Oliver, A., Huo, Y., Lladó, X., Landman, B.A., 2019. Brain structure segmentation in the presence of multiple sclerosis lesions. *NeuroImage: Clinical* 22, 101709.
- Guizard, N., Coupé, P., Fonov, V., Manjón, J., Arnold, D., Collins, D., 2015. Rotation-invariant multi-contrast non-local means for MS lesion segmentation. *NeuroImage: Clinical* 8, 376–389.
- Harmouche, R., Subbanna, N., Collins, D., Arnold, D., Arbel, T., 2015. Probabilistic multiple sclerosis lesion classification based on modeling regional intensity variability and local neighborhood information. *IEEE Trans. Biomed. Eng.* 62 (5), 1281–1292.
- Houtchens, M., Benedict, R., Killiany, R., Sharma, J., Jaisani, Z., Singh, B., Weinstock-Guttman, B., Guttmann, C., Bakshi, R., 2007. Thalamic atrophy and cognition in multiple sclerosis. *Neurology* 69 (12), 1213–1223.
- Huo, Y., Asman, A., Plassard, A., Landman, B.A., 2017. Simultaneous total intracranial volume and posterior fossa volume estimation using multi-atlas label fusion. *Hum. Brain Mapp.* 38 (2), 599–616.
- Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: a survey. *Med. Image Anal.* 24 (1), 205–219.
- Klein, A., Canton, T.D., Ghosh, S., Landman, B., Lee, J., Worth, A., 2010. Open labels: Online feedback for a public resource of manually labeled brain images. *Organization of Hum. Brain Mapp.*
- Kushibar, K., Valverde, S., González-Villà, S., Bernal, J., Cabezas, M., Oliver, A., Lladó, X., 2018. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Med. Image Anal.* 48, 177–186.
- Landman, B., Warfield, S., 2012. MICCAI 2012 workshop on multi-atlas labeling. In: *MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling*.
- Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* 19 (9), 1498–1507.
- Minagar, A., Barnett, M.H., Benedict, R.H., Pelletier, D., Pirko, I., Sahraian, M.A., Frohman, E., Zivadinov, R., 2013. The thalamus and multiple sclerosis: modern views on pathologic, imaging, and clinical aspects. *Neurology* 80 (4), 210–219.
- Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N., 2001. Reconstructing a 3D structure from serial histological sections. *Image Vis. Comput.* 19 (1), 25–31.
- Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage* 56 (3), 907–922.
- Popescu, V., Ran, N., Barkhof, F., Chard, D., Wheeler-Kingshott, C., Vrenken, H., 2014. Accurate GM atrophy quantification in MS using lesion-filling with co-registered 2D lesion masks. *NeuroImage: Clinical* 4, 366–373.
- Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Lladó, X., 2015. A toolbox for multiple sclerosis lesion segmentation. *Neuroradiology* 57 (10), 1031–1043.
- Tomas-Fernandez, X., Warfield, S.K., 2015. A Model Of Population and Subject (MOPS) intensities with application to multiple sclerosis lesion segmentation. *IEEE Trans. Med. Image.* 34 (6), 1349–1361.
- Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Image.* 29 (6), 1310–1320.
- Valverde, S., Oliver, A., Lladó, X., 2014. A white matter lesion-filling approach to improve brain tissue volume measurements. *NeuroImage: Clinical* 6, 86–92.
- Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage* 155, 159–168.
- Wang, H., Suh, J.W., Das, S., Pluta, J., Craige, C., Yushkevich, P., 2013. Multi-atlas segmentation with joint label fusion. *IEEE Trans. Pattern Anal. Machine Intell.* 35 (3), 611–623.