NeuroImage: Clinical 15 (2017) 228–238

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation

Sandra González-Villà[a,*], Sergi Valverde[a], Mariano Cabezas[a], Deborah Pareto[b], Joan C. Vilanova[c], Lluís Ramió-Torrentà[d], Àlex Rovira[b], Arnau Oliver[a], Xavier Lladó[a]

[a] Institute of Computer Vision and Robotics, University of Girona, Ed. P-IV, Campus Montilivi, 17073 Girona, Spain
[b] Magnetic Resonance Unit, Dept of Radiology, Vall d'Hebron University Hospital, Spain
[c] Girona Magnetic Resonance Center, Spain
[d] Multiple Sclerosis and Neuroimmunology Unit, Dr. Josep Trueta University Hospital, Spain

ARTICLE INFO

Keywords:
Brain structures
Multiple sclerosis lesions
Segmentation
Magnetic resonance imaging

ABSTRACT

In recent years, many automatic brain structure segmentation methods have been proposed. However, these methods are commonly tested with non-lesioned brains and the effect of lesions on their performance has not been evaluated. Here, we analyze the effect of multiple sclerosis (MS) lesions on three well-known automatic brain structure segmentation methods, namely, FreeSurfer, FIRST and multi-atlas fused by majority voting, which use learning-based, deformable and atlas-based strategies, respectively. To perform a quantitative analysis, 100 synthetic images of MS patients with a total of 2174 lesions are simulated on two public databases with available brain structure ground truth information (IBSR18 and MICCAI'12). The Dice similarity coefficient (DSC) differences and the volume differences between the healthy and the simulated images are calculated for the subcortical structures and the brainstem. We observe that the three strategies are affected when lesions are present. However, the effects of the lesions do not follow the same pattern; the lesions either make the segmentation method underperform or surprisingly augment the segmentation accuracy. The obtained results show that FreeSurfer is the method most affected by the presence of lesions, with DSC differences (generated − healthy) ranging from -0.11 ± 0.54 to 9.65 ± 9.87, whereas FIRST tends to be the most robust method when lesions are present (-2.40 ± 5.54 to 0.44 ± 0.94). Lesion location is not important for global strategies such as FreeSurfer or majority voting, where structure segmentation is affected wherever the lesions exist. On the other hand, FIRST is more affected when the lesions are overlaid or close to the structure of analysis. The most affected structure by the presence of lesions is the nucleus accumbens (from -1.12 ± 2.53 to 1.32 ± 4.00 for the left hemisphere and from -2.40 ± 5.54 to 9.65 ± 9.87 for the right hemisphere), whereas the structures that show less variation include the thalamus (from 0.03 ± 0.35 to 0.74 ± 0.89 and from -0.48 ± 1.08 to -0.04 ± 0.22) and the brainstem (from -0.20 ± 0.38 to 1.03 ± 1.31). The three segmentation approaches are affected by the presence of MS lesions, which demonstrates that there exists a problem in the automatic segmentation methods of the deep gray matter (DGM) structures that has to be taken into account when using them as a tool to measure the disease progression.

1. Introduction

Neurodegenerative disorders are frequently associated with structural changes in the brain, such as variations in the volume or shape of the deep gray matter (DGM) structures (Debernard et al., 2015; Mak et al., 2014; Lee et al., 2015). In multiple sclerosis (MS), it has been demonstrated that gray matter (GM) atrophy is relevant to disease progression (Jacobsen et al., 2014); however global GM volume measurement approaches are insufficiently sensitive during the early stages of disease (Bergsland et al., 2012). Thus, an increasing number of studies have investigated patients with clinically isolated syndrome and early relapsing-remitting MS to study the atrophy effect on GM substructures in order to identify the specific structures that are more susceptible to this disease (Audoin et al., 2010; Calabrese et al., 2011; Schoonheim et al., 2012; Štecková et al., 2014).

Some studies have analyzed the effect of MS on the subcortical structures and have concluded that volume loss is predominant in this region compared with that of the periphery (Bishop et al., 2017) and

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

http://dx.doi.org/10.1016/j.nicl.2017.05.003
Received 20 February 2017; Received in revised form 5 May 2017; Accepted 7 May 2017
Available online 08 May 2017

that its atrophy is closely related to the magnitude of inflammation (Minagar et al., 2013). Furthermore, the effect of isolated structures has also been studied, such as the thalamus, in which atrophy is a clinically relevant biomarker of the neurodegenerative disease process (Houtchens et al., 2007), or the corpus callosum, which undergoes atrophy and becomes thinned out as the disease progresses (Kazi et al., 2013).

The most common procedure to measure the volume of a structure is to segment or delineate it in a T1-weighted magnetic resonance image (MRI), which is a non-intrusive and painless technique that achieves accurate results due to the significant contrast between tissues. Although manual segmentation is still used, its results are poorly reproducible (subject to inter- and intra- expert variability), and it is a very time-consuming task. For this reason, automatic brain structure segmentation methods have been widely studied in recent years (González-Villà et al., 2016; Fischl et al., 2002; Heckemann et al., 2006; Patenaude et al., 2011; Weisenfeld and Warfield, 2011; Iglesias et al., 2012; Wang and Yushkevich, 2013). However, these automatic methods are designed to segment non-lesioned brains, either from healthy subjects or from patients with schizophrenia, Alzheimer's, epilepsy and other diseases, and these patients typically do not have white matter (WM) lesions such as those found in MS patients. These lesions are hypointense in T1-weigthed MRI, and their intensity is very similar to that of the GM, which can make the performance of these automatic methods variable.

Interestingly, in order to reduce the effect of these hypo-intense T1-weighted MS lesions, lesion filling techniques (Battaglini et al., 2012; Valverde et al., 2014) have already been applied to assess the progression of GM atrophy (Bergsland et al., 2012; Bishop et al., 2017) and have improved the accuracy of tissue volume estimates (Nakamura et al., 2014; Popescu et al., 2014). However, the effects of these lesions on brain structure segmentation methods have not yet been evaluated. One of the largest problems when quantitatively evaluating these methods is that training and testing are difficult due to the limitation of having datasets with both structure ground truth and lesion annotations. Here, we overcome this issue with an approach to synthetically generate MS lesions (from cases with lesion manual annotation) in healthy subjects from whom brain structure ground truth information is available.

In this work, we evaluate the effects of simulated MS lesions on the performance of three well-known automatic brain structure segmentation approaches, each of which follows a different segmentation strategy (González-Villà et al., 2016). FreeSurfer (Fischl et al., 2002) follows a learning-based strategy, FIRST (Patenaude et al., 2011) method uses a deformable approach, and the multiatlas-based segmentation strategy is fused by means of majority voting (Artaechevarría et al., 2009). To the best of our knowledge, this is the first work to study the effect of MS lesions on brain structures segmentation. To evaluate this effect, we generate a set of 100 synthetic MS patients' images with a total of 2174 lesions, using as a base two different databases with structures ground truth (the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database (Landman and Warfield, 2012) (MICCAI'12) and the Internet Brain Segmentation Repository[1] (IBSR18)). The DSC differences in the automatic segmentations between the healthy and the simulated patient images for the three approaches are analyzed separately by brain structure and lesion location.

2. Materials and methods

2.1. Segmentation methods

Three publicly available structure segmentation methods are used in this study. The first of these is the segmentation algorithm (Fischl et al., 2002) included in the well-known software FreeSurfer[2] (Fischl, 2012), which follows a learning-based strategy. The second method is the Bayesian appearance model proposed by Patenaude et al. (2011), which is a deformable strategy and is implemented as part of the FSL[3] package under the name FIRST. The last algorithm follows an atlas-based strategy, more specifically, multi-atlas registration fused by means of the simple and well-known fusion strategy, majority voting. For this last strategy, we follow the procedure described by Artaechevarría et al. (2009). We first perform an affine transformation to align the volumes, followed by a non-rigid B-spline registration using an isotropic grid spacing of 8.0 pixels and mutual information as a similarity metric. As in Artaechevarría et al., Elastix[4] (Klein et al., 2010) is used to perform the registrations. Both FreeSurfer and FIRST are executed with default parameters.

2.2. Synthetic MS patient generation

Currently, there is a lack of public database information regarding MS patients with both brain structures and lesion ground truth. Therefore, to evaluate how WM lesions affect the performance of automatic brain structure segmentation algorithms, we present here our method to generate synthetic lesions from MS patient images to healthy subjects with brain structure ground truth information. In the following sections, we present the steps of this pipeline including lesion dictionary construction, preprocessing and lesion generation.

2.2.1. Lesion dictionary

To build the lesion dictionary, it is necessary to have an MRI dataset from MS patients with a manual lesion annotation (ground truth). This dataset must consist of a set of T1-w volumes and their corresponding lesion delineations.

As MS lesions are usually annotated using either the FLAIR or PD-w sequence and lesions tend to look smaller in T1-w images, we first reduce the lesion masks based on their appearance in the T1-w sequence. To accomplish such a reduction, we perform a tissue segmentation using FSL FAST (Zhang et al., 2001) and discard from the ground truth the voxels classified in the WM class. Once the masks are reduced, we assign an independent label to each lesion in the image with the final objective of evaluating each lesion independently. This is achieved by obtaining the connected components of the lesion mask and considering each mask as a single 3D lesion.

As demonstrated later in this section, we approach strictly WM lesions and WM lesions attached to the lateral ventricles (referred as LV lesions) differently. Therefore, it is important to classify each lesion from the dataset into one of these two groups. To do this, we calculate the 3D Euclidean distance from each lesion contour to the cerebrospinal fluid (CSF). At this point, we have all of the necessary information to build the lesion dictionary, which includes the following information for each lesion in the dataset: the lesion label, the image of procedence, the lesion type (WM or LV) and, for practical issues, the lesion size.

2.2.2. Preprocessing

After construction of the lesion dictionary, and before applying the lesion generation method in situ, some preprocessing steps are required. First, we perform tissue segmentation of the target image, as it is necessary in our lesion generation method to restrict the generated lesion position and avoid overlap with the CSF. Afterwards, we perform both rigid and non-rigid registrations of each MS patient in the dataset to the target image. Both registrations are performed using the original images without any preprocessing (either the MS patient

[1] https://www.nitrc.org/projects/ibsr

[2] http://freesurfer.net/

[3] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

[4] http://elastix.isi.uu.nl/

image or the target image). Therefore, we acquire two new volumes in the target image space − one rigidly registered to the target image and another non-rigidly registered − in addition to the corresponding lesion masks. As lesions can affect the non-rigid registration process, their voxels are masked out in order to achieve a more accurate result. The NiftyReg software[5] is used to perform the registrations.

With all of the selected patients in the target image space, we normalize the source image (registered MS patients) intensities to the target image to create realistic lesions. For this purpose, the skull-stripped images of both the MS registered patients and the target image are used to avoid the influence of non-brain intensities on the histogram matching normalization process. To normalize the images, ITK[6] implementation described by Nyul et al. (2000) is used.

2.2.3. Lesion generation

We finally generate the new MS lesions in the target image as follows:

$$patch_{target} = patch_{lesion} \times mask + patch_{target} \times (1 - mask)$$

where $patch_{target}$ corresponds to the 3D patch in the target image where the lesion is going to be generated, $patch_{lesion}$ is the 3D patch of the registered and normalized patient containing the lesion, and *mask* is the corresponding lesion mask to which we apply a Gaussian filter to make the transition between the healthy tissue and the lesion smoother.

As already mentioned we deal with the lesions separately, generating one lesion at a time and selecting its mask (*mask*) from the corresponding registered image: rigid for WM lesions, since the idea for this type of lesions is to keep the original shape, and non-rigid for LV lesions. These last lesions have a special shape that depends on the morphology of the LV and non-rigid deforming the original lesions allows getting adapted to that structure. Therefore, the final lesion mask is composed by the sum of the independent lesion masks proceeding from the registered images. However, as the original location of the lesion is better captured in the non-rigid registered images, the lesion location in the target image ($patch_{target}$ center), and therefore in the final lesion mask, is always selected based on the non-rigid position of the lesion center.

Fig. 1 shows an example of a generated image applying the proposed methodology. Fig. 1a shows the original MS patient whose lesions have been reproduced, whereas Fig. 1b to d shows the original healthy subject and the lesion generation results.

2.3. Data

To study the performance of the different structure segmentation algorithms, the following two publicly available databases are used: 1) the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database (Landman and Warfield, 2012) (MICCAI'12) and 2) the Internet Brain Segmentation Repository[7] (IBSR18). The first database consists of 15 T1-w MR images for training and 20 for testing as obtained from the OASIS[8] project and labeled by Neuromorphometrics, Inc.[9], which includes labels for the whole brain. The second database consists of 18 T1-w MR images with expert manual segmentations of 43 individual structures provided by the Center for Morphometric Analysis at Massachusetts General Hospital.

The lesion generation method explained in Section 2.2 is used to generate synthetic lesions over the testing subjects. The procedure to select the healthy subjects on which to generate the lesions is as follows. The 38 testing images (20 from MICCAI'12 + 18 from IBSR18) are segmented using the three algorithms presented in Section 2.1. The training cohort from the MICCAI'12 database is used as the atlas for the multi-atlas method. Two subjects from each database are selected based on their Dice similarity coefficients (DSCs) achieved for the evaluated structures. According to this metric, we discard the subjects who represent outliers in any of the analyzed structures in relation to the other subjects' segmentation in the same database. From the remaining subjects, the two that best represent each database are chosen with a DSC for almost all of the structures in the median of all the subjects for each of the analyzed segmentation algorithms. Regarding these criteria, subjects 1004 and 1116 from the MICCAI'12 database and subjects IBSR 08 and IBSR 17 from the IBSR18 database are selected. Some details about the selected images can be found in Table 1.

Five MS databases including MICCAI'08[10], MICCAI'16[11], ISBI'15[12] and two in-house databases are included in our lesion dictionary, with a total of 140 patients and 4291 lesions. The lesions from each patient are simulated in the selected healthy subjects, and once completed, a second selection is performed to obtain the final images included in our study. Twenty-five MS patient images are selected in such a way that lesions are represented in all of the analyzed structures, and different patient volumes and lesion numbers are achieved. The simulations of these 25 patients in the four selected subjects are chosen as the cohort on which to perform our experiments, which includes a total of 100 simulated images. Details of the original 25 MS patients are shown in Table 2. The original cohort has a total of 1429 WM lesions, but for practical issues only lesions larger than 27 mm^3 are simulated.

The lesions of the same MS patients are replicated on the four selected healthy subjects, leading to simulated MS patients with lesion loads ranging from 0.44 to 59.93 ml and a number of lesions per patient ranging from 1 to 62. The total number of generated lesions over the 100 synthetic images is 2174.

As stated previously, the 25 MS patients are selected in such a way that there are lesions represented in all of the analyzed structures. As a result, the generated cohort contains 27 images with lesions in the left thalamus, 35 with lesions in the right thalamus, 83 in the left caudate, 81 in the right caudate, 25 in the left putamen, 30 in the right putamen, 4 in the left pallidum, 5 in the right pallidum, 41 in the left hippocampus, 64 in the right hippocampus, 27 in the left amygdala, 36 in the right amygdala, 28 in the left accumbens, 23 in the right accumbens and 57 in the brainstem.

Notice that registration inaccuracies affect the lesion generation procedure in different ways. First of all, when we perform registration to move the MS patient image to the healthy image space, if we are moving to a lower resolution space and the lesions are small and only visible in a low number of slices it may happen that these lesions disappear from the registered image. Moreover, registration can make the lesion position displace to an upper/lower slice or even change the lesion position, which for some small lesions could result in being or not overlaid on the same structure. Furthermore, even though lesions are masked out to perform the non-rigid registration, their morphology (shape and size) may be affected differently from one healthy to another. Moreover, as we set the restriction that only lesions above 27 mm^3 are simulated, it could happen that the same lesion is simulated in one healthy but not in the others.

2.4. Evaluation

Images from both healthy subjects and their corresponding simulated MS patients are segmented using the three segmentation strategies presented in Section 2.1. Since FIRST only provides segmentation results for the subcortical structures and the brainstem, the perfor-

[5] http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg
[6] https://itk.org/
[7] https://www.nitrc.org/projects/ibsr
[8] http://www.oasis-brains.org/
[9] http://neuromorphometrics.com/
[10] http://www.ia.unc.edu/MSseg/
[11] https://portal.fli-iam.irisa.fr/msseg-challenge/overview
[12] http://iacl.ece.jhu.edu/index.php/MSChallenge

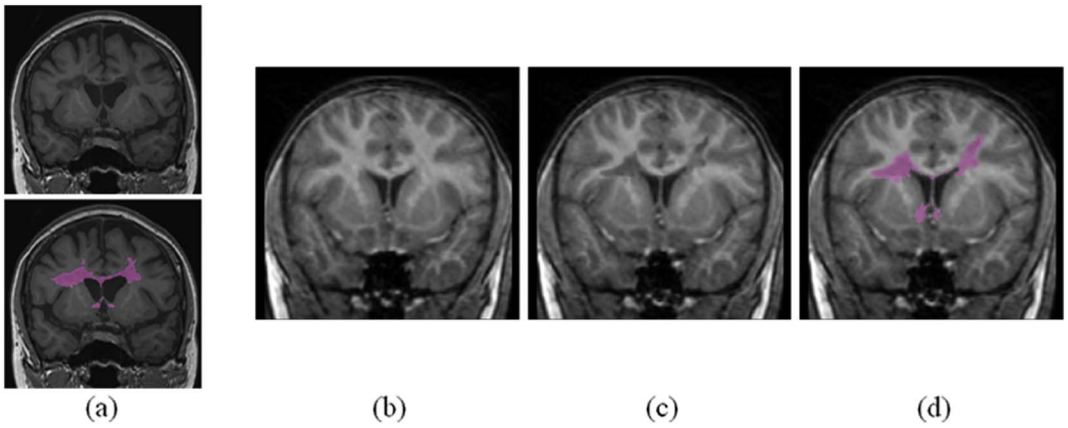

Fig. 1. Example of MS lesions generation. (a) Original MS patient image and corresponding lesion mask (patient 01016SACH form the MICCAI'16 Challenge database) (b) Healthy image (subject IBSR 17 from the IBSR18 database) (c) Generated image (d) Lesion mask.

Table 1
Properties of the four selected healthy subjects.

Name	Database	Age	Scanner	Volume (mm)	Voxel (mm)
1004	MICCAI' 12 (Landman and Warfield, 2012)	23	Siemens (1.5 T)	256 × 256 × 256	1 × 1 × 1
1116	MICCAI' 12 (Landman and Warfield, 2012)	61	Siemens (1.5 T)	256 × 256 × 256	1 × 1 × 1
IBSR_08	IBSR[a]	60	Siemens (1.5 T)	256 × 256 × 128	1 × 1 × 1.5
IBSR_17	IBSR[a]	8	GE (1.5 T)	256 × 256 × 128	0.84 × 0.84 × 1.5

[a] https://www.nitrc.org/projects/ibsr

Table 2
Properties of the twenty five selected MS patients. Lesion volumes and number of lesions are calculated from the reduced masks based on lesion appearance in the T1-w sequence.

Database	No.	Scanner	Volume (mm)	Voxel (mm)	Lesion vol (ml)	No. lesions	Lesion size (mm^3)
MICCAI'08[a]	4	Siemens Allegra (3 T)	512 × 512 × 512	0.5 × 0.5 × 0.5	6.28–14.39	51–125	$0.13–0.41 \times 10^4$
MICCAI'16[b]	5	Siemens Verio (3 T)	176 × 256 × 256	1 × 1 × 1	0.89–59.79	7–69	$1.00–3.28 \times 10^4$
MICCAI'16[b]	3	Siemens Aera (1.5 T)	256 × 256 × 176	1.08 × 1.08 × 0.9	1.43–35.26	19–65	$1.05–1.65 \times 10^4$
MICCAI'16[b]	3	Philips Ingenia (3 T)	2: 200 × 336 × 336 1: 210 × 336 × 336	0.85 × 0.74 × 0.74	4.78–26.60	51–133	$0.47–1.54 \times 10^4$
ISBI'15[c]	2	Philips (3 T)	256 × 256 × 120	0.82 × 0.82 × 1.17	12.68–25.67	42–45	$0.80–1.39 \times 10^4$
IN-HOUSE 1	7	Siemens Trio Tim (3 T)	6: 128 × 240 × 256 1: 128 × 232 × 256	1.2 × 1 × 1	1.30–14.25	20–99	$1.20–0.21 \times 10^4$
IN-HOUSE 2	1	Siemens Simphony Quantum (1.5 T)	192 × 256 × 46	0.98 × 0.98 × 3	31.60	9	$48.64–2.30 \times 10^4$

[a] http://www.ia.unc.edu/MSseg/
[b] https://portal.fli-iam.irisa.fr/msseg-challenge/overview
[c] http://iacl.ece.jhu.edu/index.php/MSChallenge

mance of the three algorithms and the effects of the lesions are only evaluated on this subset of brain structures. All of the structures are evaluated separately for the left and right hemispheres, except the brainstem, which is a unique structure.

The Dice similarity coefficient (DSC) (Dice, 1945) and the volume differences between healthy controls and simulated patients are used as the main metrics for evaluation, and other measures such as false positive Dice (over-segmentation) or false negative Dice (under-segmentation) are also analyzed (Babalola et al., 2009) but finally omitted in this paper for the sake of simplicity.

Statistical analysis is performed using the Matlab software package[13]. After performing Lilliefors test on our data we see that we cannot assume normality, which restricts us to the set of statistical tests for non-normal variables. Differences in the performance of the three analyzed methods when segmenting the healthy subjects of both databases are analyzed using pairwise Wilcoxon rank sum tests. Moreover, the Pearson's linear correlation coefficient is used to compute the correlation between the total lesion volume and the changes in DSC and in structure volume. We also compare the methods robustness with respect to each other when simulated lesions are introduced. In order to rank the methods on both IBSR and MICCAI'12 databases, significant pairwise method permutation tests of the absolute DSC differences and the absolute structure volume differences are performed. Furthermore, we test for significant differences in the robustness of the three strategies when the lesions are overlaid in the structure of analysis and when they are not. To perform such analysis, series of permutation tests on the absolute structure volume differences with respect to the healthy controls are performed. For our experiments, we have adapted the implementation provided by Klein et al. (2009). For all the permutation tests performed in our experiments, we set the number of comparisons between each pair of methods to $N = 1000$. In all the analysis, we consider data significant at p-values < 0.05.

3. Results

In this section, we analyze the behavior of the automatic brain structure segmentation methods presented in Section 2.1. First, we analyze how these methods behave when the healthy subjects from the

[13] http://es.mathworks.com/products/matlab

two databases with ground truth available (IBSR18 and MICCAI'12) are segmented. Then, we compare the automatic segmentations obtained with the three different approaches for both the generated patients and the healthy subject images and perform an analysis of how the simulated WM lesions affect the performance of each software method based on structure and lesion location.

Table 3
Healthies DSC. Structure acronyms are: left thalamus (L.Tha), right thalamus (R.Tha), left caudate (L.Cau), right caudate (R.Cau), left putamen (L.Put), right putamen (R.Put), left pallidum (L.Pal), right pallidum (R.Pal), left hippocampus (L.Hip), right hippocampus (R.Hip), left amygdala (L.Amy), right amygdala (R.Amy), left accumbens (L.Acc), right accumbens (R.Acc) and brainstem (BS). The table shows the DSC values (mean ± std) for the MICCAI'12 (M) and IBSR18 (I) databases, separated by segmentation strategy (FreeSurfer, FIRST and majority voting (M.V.)). Highlighted areas show the best segmentation strategy results for a given structure and database. Statistically significant ($p \leq 0.05$) better method performance is shown in bold. Notice that MV strategy is highly dependent on the training set.

Structure	DB	FreeSurfer	FIRST	M. V.
L.Tha	I	81.53 ± 5.59	**89.34 ± 1.69**	81.68 ± 2.94
	M	83.01 ± 1.77	88.92 ± 1.72	87.73 ± 3.46
R.Tha	I	86.36 ± 2.23	**88.46 ± 1.20**	74.70 ± 4.99
	M	84.88 ± 2.07	89.02 ± 1.83	86.53 ± 4.51
L.Cau	I	79.61 ± 4.96	78.27 ± 4.39	67.39 ± 6.15
	M	80.83 ± 7.89	79.72 ± 11.66	76.64 ± 8.92
R.Cau	I	80.92 ± 4.84	**87.04 ± 2.75**	60.32 ± 6.22
	M	80.11 ± 4.16	**83.66 ± 4.57**	75.82 ± 8.77
L.Put	I	78.88 ± 3.81	**86.88 ± 2.01**	60.25 ± 7.51
	M	77.13 ± 3.86	85.98 ± 7.95	86.24 ± 4.13
R.Put	I	82.92 ± 3.10	**88.05 ± 1.05**	74.29 ± 5.59
	M	79.87 ± 2.62	87.59 ± 6.00	87.75 ± 4.37
L.Pal	I	63.17 ± 17.05	**81.05 ± 3.33**	51.73 ± 10.45
	M	69.25 ± 18.93	81.49 ± 6.04	84.23 ± 2.79
R.Pal	I	77.44 ± 3.23	**80.89 3.70**	68.43 ± 6.34
	M	79.15 ± 8.53	79.93 ± 8.80	**85.23 ± 5.31**
L.Hip	I	76.00 ± 3.58	**80.64 ± 2.31**	62.42 ± 4.82
	M	78.35 ± 5.37	80.85 ± 1.40	78.49 ± 4.48
R.Hip	I	76.66 ± 6.03	**81.68 ± 2.26**	61.13 ± 4.84
	M	79.44 ± 2.54	80.97 ± 2.16	79.32 ± 3.54
L.Amy	I	66.06 ± 6.94	**74.18 ± 6.35**	55.77 ± 7.91
	M	58.47 ± 6.41	72.13 ± 5.44	71.24 ± 7.57
R.Amy	I	69.05 ± 6.73	**75.72 ± 6.18**	46.62 ± 6.19
	M	57.58 ± 7.59	70.67 ± 5.25	73.44 ± 7.14
L.Acc	I	60.42 ± 7.08	**68.40 ± 9.77**	52.44 ± 11.13
	M	62.98 ± 5.51	69.94 ± 8.12	71.03 ± 7.96
R.Acc	I	57.36 ± 7.41	**70.27 ± 7.62**	50.35 ± 9.33
	M	44.26 ± 6.46	67.77 ± 8.87	68.79 ± 10.50
BS	I	84.12 ± 1.96	82.50 ± 2.69	80.79 ± 2.05
	M	85.67 ± 1.96	83.34 ± 1.66	**91.64 ± 1.57**

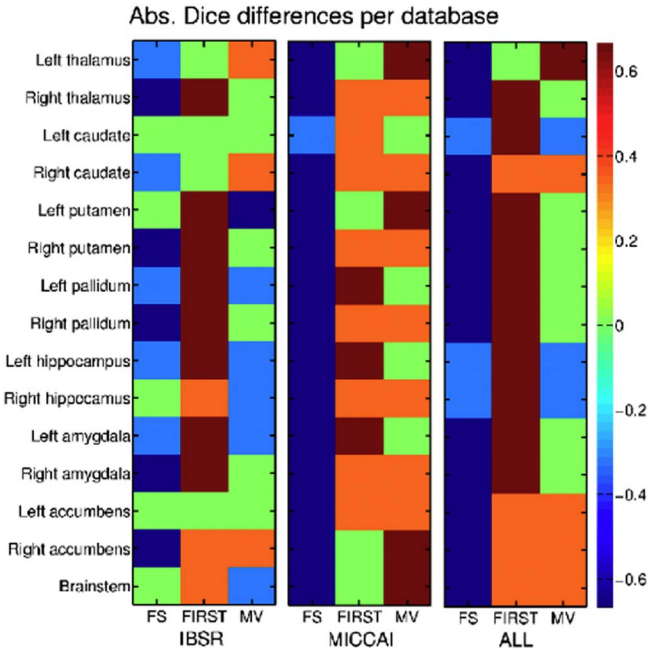

Fig. 2. Ranking of the segmentation methods separated per database and brain structure obtained from the permutation tests. Color scale that reflects the relative robustness of the segmentation methods when simulated lesions are introduced (with red indicating higher consistency with relation to the healthy segmentation). Each colored square represents the average score for a given method and structure, averaged over 100 segmentations. The scores are values indicating the pairwise robustness of the method relative to each of the other methods, according to DSC differences (generated patient − healthy). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Database performance

Table 3 shows the DSC results of the three analyzed segmentation strategies on the healthy subjects from both databases (20 from MICCAI'12 + 18 from IBSR18). As shown in this table, FreeSurfer provides similar results for both databases; however, we can highlight some structures such as the amygdalas ($p \leq 0.001$), the right accumbens ($p < 0.001$), the right thalamus ($p < 0.05$) and the right putamen ($p < 0.01$) that achieve better segmentation results, on average, for the healthy subjects from the IBSR18 database. On the other hand, better results are obtained for the left pallidum ($p < 0.05$), the right pallidum ($p < 0.01$), the left hippocampus ($p = 0.01$) and the brainstem ($p < 0.05$) when the healthy subjects from the MICCAI'12 are segmented.

Smaller differences between the segmentation results in both databases are obtained for the deformable strategy. In this method we can highlight three structures on which this difference is statistically significant: the left caudate ($p < 0.05$), the right caudate ($p < 0.01$) and the right amygdala ($p < 0.01$).

Regarding the majority voting strategy, we observe from the table that it provides higher DSC values for MICCAI'12 than for IBSR18 in all of the analyzed structures ($p \leq 0.001$). This difference arise because the atlases used proceed from the training cohort of the MICCAI'12 database, and thus, their similarity in the scanner acquisition configuration and the rank of intensities allow better registration results when segmenting the MICCAI'12 subjects. These results indicate that this strategy strongly depends on the training dataset.

In a database-specific analysis, we observe from Table 3 that for the MICCAI'12 the differences between the segmentation performance provided by FIRST and majority voting are mostly not significant. However, we can highlight three structures on which these two strategies differentiate, which are the right caudate ($p < 0.01$), the right pallidum ($p < 0.001$) and the brainstem ($p < 0.001$). On the

Table 4
Permutation tests average ranking based on the method robustness when lesions are introduced. Ranks after conducting permutation tests between absolute DSC differences (1 − |DSC generated − DSC healthy|) of the generated MS patient images and their corresponding healthy controls (averaged across structures) for each pair of methods, then calculating the percentage of p-values less or equal to 0.05 (of 1000 tests). Members within ranks 1, 2 and 3 have means lying within one, two and three standard deviations of the highest mean, respectively. μ = mean; σ = standard deviation.

	Dice differences					
	IBSR	μ ± σ	MICCAI	μ ± σ	ALL	μ ± σ
Rank 1	FIRST	0.42 ± 0.29	MV FIRST	0.33 ± 0.25 0.31 ± 0.23	FIRST	0.53 ± 0.21
Rank 2	MV	−0.09 ± 0.29				
Rank 3	FS	−0.33 ± 0.28	FS	−0.64 ± 0.09	MV FS	0.07 ± 0.29 −0.60 ± 0.14

other hand, the results obtained for the IBSR18 database show that for most of the structures, the strategy that provides the best accuracy is FIRST ($p < 0.05$).

3.2. Lesion effects per segmentation strategy

The effect of the generated lesions on the three segmentation strategies is analyzed here separately by method performance (DSC) and structure volume.

3.2.1. Dice difference

Fig. 2 presents the robustness of each method relative to the other two (relative robustness) when lesions are introduced as a color-coded table, separated for structure and database. We compare here the three segmentation strategies based on how consistent their segmentations are when lesions are introduced, compared to those obtained for the corresponding healthy subject. This figure shows that for the IBSR18 database, FIRST provides the most robust results for almost all the structures, however majority voting is more consistent when segmenting the left thalamus and the right caudate. On the other hand, FreeSurfer achieves more unstable results than the other two methods for seven of the analyzed structures, but it is more consistent than majority voting when segmenting the left putamen, the right hippocampus and the brainstem. Regarding the MICCAI'12 database, FreeSurfer shows to be the least robust strategy for all the analyzed structures whereas the other two methods have a similar behavior. Combining both databases FIRST achieves the most consistent results for ten of the analyzed structures whereas the segmentations obtained with FreeSurfer are the ones that seem more affected by the lesions presence. Analyzing the overall DSC differences averaged across all the structures, we see that for the IBSR18 database, FIRST achieves the most consistent results (0.56 ± 0.95), followed by majority voting (1.10 ± 0.94) and FreeSurfer (1.57 ± 1.83). On the other hand, for the subjects of the MICCAI'12 database, the most robust results are obtained when segmenting with majority voting (0.49 ± 0.44), followed by FIRST (0.53 ± 0.67). However, for this database FreeSurfer seems to be highly affected by the presence of lesions (4.05 ± 1.24). As stated in Section 3.1 we observe again here that majority voting is highly dependent on the training set, achieving more robust segmentation results for the MICCAI'12 database. Table 4 presents the ranking of methods separated by database, according to the percentage of permutation tests whose p-values are less or equal to 0.05. Members within ranks 1, 2 and 3 have means lying within one, two and three standard deviations of the highest mean, respectively.

Fig. 3 shows the boxplots of the differences in terms of DSC between healthy subjects and the corresponding generated MS patients (patient − healthy) as separated by segmentation strategy. Green boxplots show the differences for the patients who do not have lesions overlaid on the analyzed structure but do in other parts of the brain. On the other hand, the red boxplots show the differences when these patients have at least one lesion overlaid on that structure, independently of the patient having lesions in other parts of the brain. This figure shows that lesions have an unpredictable effect on the segmentation performance, since in some cases, it might help to improve the segmentation performance, whereas in other cases, it might produce worse overall segmentation results.

By analyzing each strategy individually, we can see that FreeSurfer does not show a clear difference in its performance when lesions are overlaid in a particular structure or not (red vs green). As shown in Fig. 3a, the trends for several structures look similar when lesions are present (in red) or absent (in green), such as in the right thalamus, both putamens, the right hippocampus, the left accumbens and the brainstem. Furthermore, we observe a trend in some of the analyzed structures towards improvement in their segmentation performance when lesions are introduced (anywhere in the brain), as seen in the right caudate, both putamens, the right hippocampus both amygdalas and the brainstem.

As shown in Fig. 3b, FIRST seems to be quite robust when lesions are present, showing DSC differences for the non-lesioned structures (in green) that range from −0.36 ± 0.39 (right pallidum) to 0.44 ± 0.94 (right amygdala), whereas the differences for the structures with lesions (in red) achieve values from −2.40 ± 5.54 (right accumbens) to 0.19 ± 0.62 (brainstem). In this strategy, the standard deviations are below 1.50 for all of the structures except the right caudate (3.71) and both accumbens (2.53 left; 5.54 right), but the three of them when lesions are overlaid in these structures (in red), showing that the segmentations provided by this method are consistent, particularly when the lesions in the structure are not present.

In the multi-atlas strategy, the differences between the healthy subjects and the simulated patients are small as shown in Fig. 3c. These differences ranged from −1.24 ± 1.53 (left amygdala) to 0.59 ± 1.45 (right accumbens) for the structures without lesions (in green) and from −0.90 ± 1.16 (left putamen) to 0.59 ± 1.64 (right amygdala) for the structures affected by lesions (in red). As can be deduced from these numbers there is not a clear difference in the method performance when the lesions are overlaid or not, as better segmentation results are achieved for non-lesioned structures in only half of the analyzed cases (8 over 15). In general, majority voting seems to underperform, on average, when segmenting the simulated patients compared with the healthy subjects for almost all of the structures, independent of whether the lesions are overlaid or not.

On the other hand, in analysis of each structure, we observe that independently of the segmentation strategy, the structure that shows more variability when the lesions are introduced is the nucleus accumbens (1.32 ± 4.00 with FreeSurfer, −1.12 ± 2.53 with FIRST and −0.24 ± 0.81 with majority voting for the left hemisphere, and 9.65 ± 9.87, −2.40 ± 5.54 and −0.48 ± 2.67 for the right hemisphere), whereas those for which the segmentation is more robust are the brainstem (1.03 ± 1.31, 0.19 ± 0.62, and −0.20 ± 0.38) and the thalamus (0.74 ± 0.89, 0.18 ± 0.53 and 0.03 ± 0.35 for the left hemisphere, and −0.48 ± 1.08, −0.04 ± 0.22 and −0.25 ± 0.56 for the right one). On the other hand, the structure for which we have

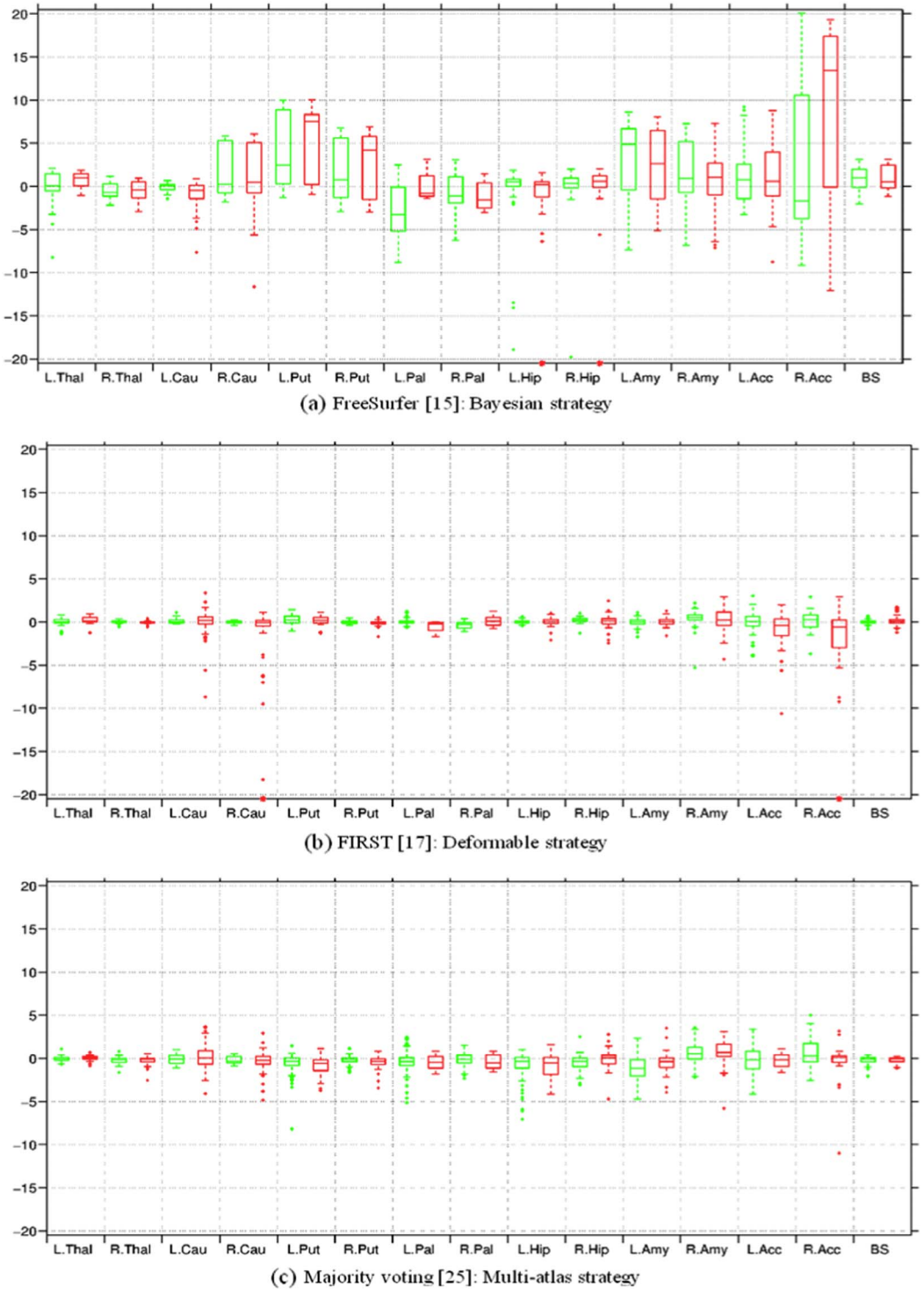

Fig. 3. DSC differences (generated patients - healthy) for the 100 generated images using various publicly available software. The green boxplots show the differences when there were MS lesions generated on the brain but not on top of that particular structure. On the other hand, red boxplots stand for lesioned structures. The red asterisks on top of the x axis mean that there are more outliers below the − 20%. Acronyms from left to right are: left thalamus (L.Thal), right thalamus (R.Thal), left caudate (L.Cau), right caudate (R.Cau), left putamen (L.Put), right putamen (R.Put), left pallidum (L.Pal), right pallidum (R.Pal), left hippocampus (L.Hip), right hippocampus (R.Hip), left amygdala (L.Amy), right amygdale (R.Amy), left accumbens (L.Acc), right accumbens (R.Acc) and brainstem (BS). Notice that the red boxplots for both pallidums and the green boxplots for both caudates contain < 20 cases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

encountered the largest number of outliers is the caudate nucleus, which is also the structure on which we find the largest number of lesions, affected in 83 images (left caudate) and 81 images (right caudate), respectively.

We also analyze the extent to which total lesion volume affects each of the segmented structures by computing the Pearson linear correlation coefficient between differences in DSC and lesion volume. Lesion volume do not correlate with the DSC differences found for FreeSurfer in any of the analyzed structures, except for the left caudate ($r = 0.52$, $p < 0.001$). A similar behavior is found for FIRST, where a significant correlation is seen for both caudates ($r = 0.58$, $p < 0.001$ and $r = 0.51$, $p < 0.001$) and the left hippocampus ($r = 0.47$, $p < 0.001$), while obtaining a moderate correlation for the right accumbens ($r = 0.38$,

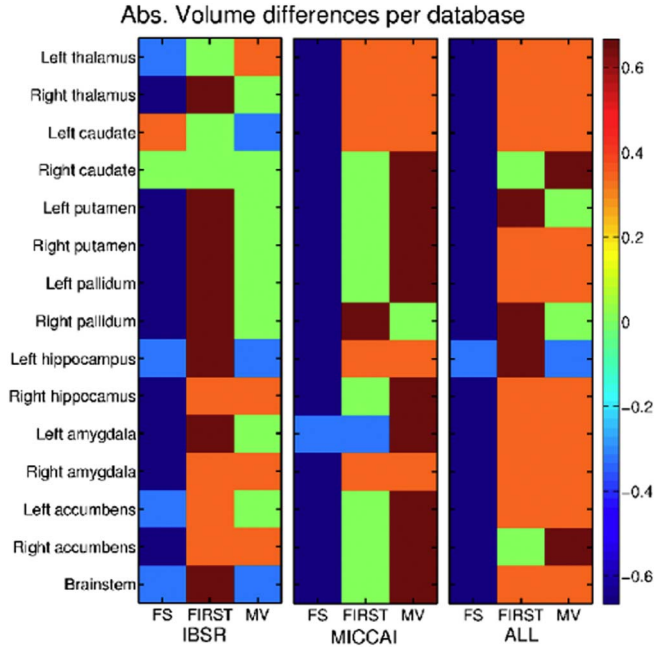

Fig. 4. Ranking of the segmentation methods separated per database and brain structure obtained from the permutation tests. Color scale that reflects the relative robustness of the segmentation methods when simulated lesions are introduced (with red indicating higher consistency with relation to the healthy segmentation). Each colored square represents the average score for a given method and structure, averaged over 100 segmentations. The scores are values indicating the pairwise robustness of the method relative to each of the other methods, according to volume differences (generated patient − healthy). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

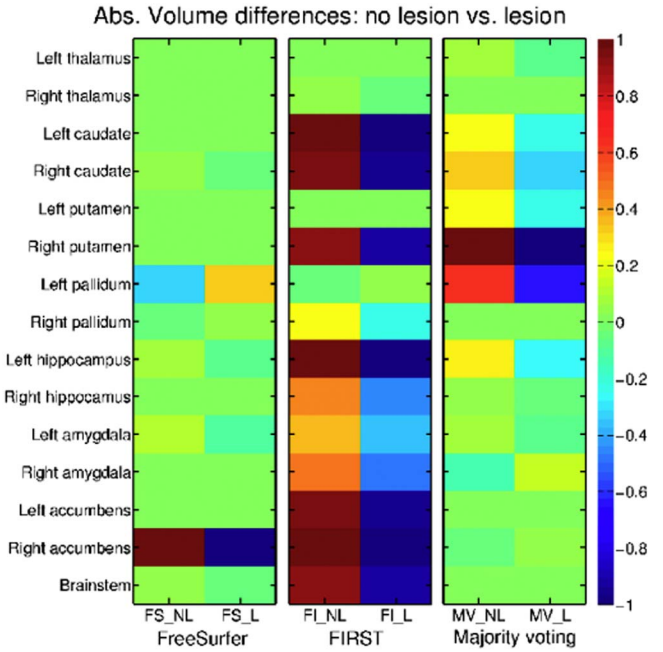

Fig. 5. Result of the permutation tests. Relative volume consistency of the segmented structures with the three segmentation strategies when lesions are overlaid or not on the structure compared to the healthy controls. Color scale displays the relative robustness for each method (with red indicating higher consistency with relation to the healthy segmentation). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$p < 0.001$). Stronger correlations are found for the majority voting strategy for the left thalamus ($r = 0.51$, $p < 0.001$), both caudate ($r = 0.73$, $p < 0.001$ and $r = 0.45$, $p < 0.001$), the right putamen ($r = 0.43$, $p < 0.001$) and the left pallidum ($r = 0.42$, $p < 0.001$) whereas a moderate correlation is found for the left putamen ($r = 0.38$, $p < 0.001$), the right pallidum ($r = 0.35$, $p < 0.001$) and both hippocampus ($r = 0.35$, $p < 0.001$ and $r = 0.34$, $p = 0.001$).

3.2.2. Volume difference

Fig. 4 shows the robustness of each method relative to the other two (relative robustness) in terms of volume change, when lesions are introduced to the healthy control. The procedure is the same as in Section 3.2.1. From this figure we observe that in the IBSR database, FIRST provides the most robust results for eight of the analyzed structures whereas its consistency is comparable to majority voting in four other structures. Regarding the MICCAI'12 database, the majority voting strategy seems to provide more robust results that the other two methods in nine of the fifteen structures whereas it achieves more unstable results than FIRST only for the right pallidum. For this database, majority voting is more robust than FreeSurfer for all the analyzed structures. Combining both databases we observe that FIRST and majority voting are comparable in terms of structure volume change consistency, however FreeSurfer seems to be the least robust against lesions achieving only similar results to majority voting for the left hippocampus. Analyzing the overall volume difference, we observe the same behavior as in Section 3.2.1. Again, for the IBSR18 database FIRST provides the most robust results ($53.88 \pm 76.37\ mm^3$), followed by majority voting (74.82 ± 64.16) and FreeSurfer (120.83 ± 134.98). Moreover, for the MICCAI'12 database, majority voting is the most consistent method (38.97 ± 38.64), followed by FIRST (68.26 ± 88.41) and FreeSurfer (352.07 ± 116.65). Table 5 presents the general ranking of the methods based on the permutation tests performed.

Fig. 5 presents the robustness in terms of structure volume changes for each method comparing the cases on which lesions are overlaid in the structure and they are not (the same case as in Fig. 3 for red and green boxplots, in addition, the corresponding volume boxplots are provided as Supplementary material). We observe that FIRST tends to be significantly more robust when lesions are not overlaid in the analyzed structure, whereas FreeSurfer is equally affected wherever the lesions are, except for the right accumbens, on which the segmentation result appears more consistent when lesions are not overlaid. Regarding

Table 5
Permutation tests average ranking based on the method robustness when lesions are introduced. Ranks after conducting permutation tests between absolute volume differences of the generated MS patient images and their corresponding healthy controls (averaged across structures) for each pair of methods, then calculating the percentage of p-values less or equal to 0.05 (of 1000 tests). Members within ranks 1, 2 and 3 have means lying within one, two and three standard deviations of the highest mean, respectively. μ = mean; σ = standard deviation.

	Volume differences					
	IBSR	μ ± σ	MICCAI	μ ± σ	ALL	μ ± σ
Rank 1	FIRST	0.44 ± 0.27	MV	0.51 ± 0.21	FIRST	0.36 ± 0.20
					MV	0.29 ± 0.25
Rank 2	MV	0.02 ± 0.23	FIRST	0.13 ± 0.25		
Rank 3	FS	−0.47 ± 0.3	FS	−0.64 ± 0.09	FS	−0.64 ± 0.09

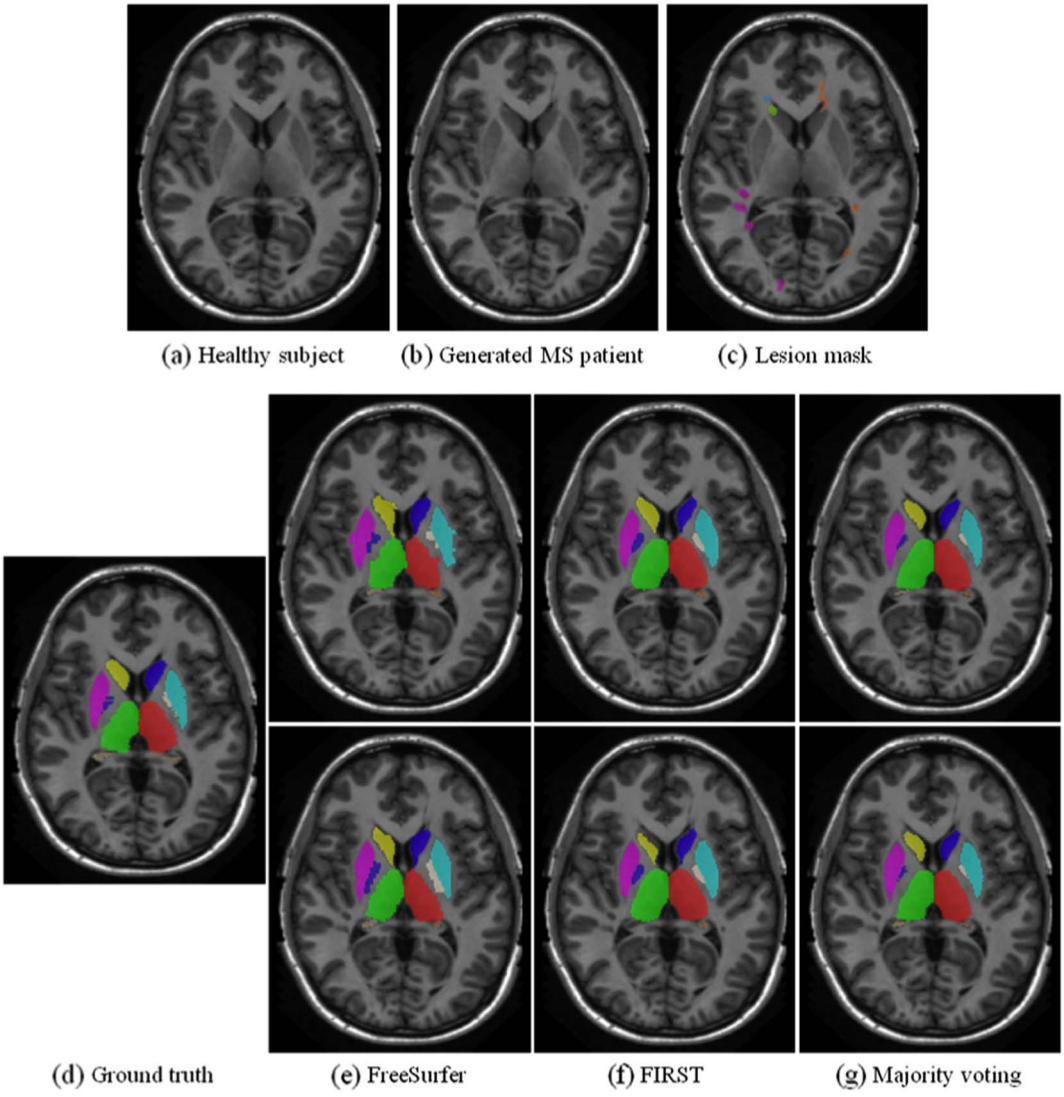

Fig. 6. Automatic brain structures segmentation. Figures (a)–(c) show the original healthy subject (1004 from MICCAI'12 database), the generated MS patient and the corresponding lesion mask. Figures (d)-(g) show the structures ground truth and automatic segmentation of both images (without lesions in the top row and with lesions bottom row) for the three segmentation strategies. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

the majority voting strategy, it seems there is a trend to be slightly more unstable when lesions are overlaid on the analyzed structure, however this difference is not conclusive.

The relation between the observed change in structure volume and the total lesion volume introduced in the simulated images is also analyzed for the three evaluated strategies. Significant correlations are not found for FreeSurfer in any of the analyzed structures whereas the volume differences found in FIRST seem to significantly correlate with the total lesion volume in both caudates ($r = 0.63$, $p < 0.001$ and $r = 0.61$, $p < 0.001$), the right putamen ($r = 0.46$, $p < 0.001$), the left hippocampus ($r = 0.43$, $p < 0.001$) and the right accumbens ($r = 0.45$, $p < 0.001$). A similar behavior than the one seen in the DSC change is seen here for the majority voting strategy. For this method, correlation between structure volume changes and total lesion volume is seen in a higher number of structures than in the other two strategies: both thalamus ($r = 0.45$, $p < 0.001$ and $r = 0.48$, $p < 0.001$), both caudates ($r = 0.60$, $p < 0.001$ and $r = 0.56$, $p < 0.001$), both putamens ($r = 0.45$, $p < 0.001$ and $r = 0.45$, $p < 0.001$) and both pallidums ($r = 0.43$, $p < 0.001$ and $r = 0.45$, $p < 0.001$). A trend towards moderate correlation is also observed in the left hippocampus ($r = 0.37$, $p < 0.001$) and the brainstem ($r = 0.32$, $p = 0.001$).

3.3. Qualitative results

Fig. 6 shows a qualitative example of the segmentation results obtained with the three methods. Fig. 6a to c shows a central slice of the healthy subject and its corresponding simulated MS patient and lesion mask. The automatic segmentation obtained with the analyzed strategies for both a healthy (top row) and simulated MS patient (bottom row) is shown in Fig. 6e to g, whereas the structure ground truth is shown in Fig. 6d. As seen in Fig. 6e, FreeSurfer improves its segmentation performance for the right caudate (in yellow) and both putamens (in pink and cyan) when the lesions are added, thus reducing the number of false positives obtained for the healthy images. In this case, the lesions may have modified the global intensity distribution, making the Gaussian chosen to represent the structure intensity more precise and improving the segmentation performance. However, as we can observe, the performance for both pallidums (in blue and white) decreases when lesions are present, increasing the number of false positives. On the other hand, as shown in the FIRST segmentations in Fig. 6f, we can see that opposite to FreeSurfer, local lesions interfere with the segmentation performance. This can be seen for the right caudate (in yellow) where the green lesion shown in Fig. 6c is constraining the deformation performed by the method to obtain the

final segmentation. The results obtained for the rest of the structures are similar for both images, since the generated lesions are far from the structures of interest. Finally, for the majority voting strategy, no changes are visually appreciated for this particular slice in Fig. 6g. As shown, the differences in the segmentation performance of the healthy subject and the generated MS patients tend to be small for this method. Furthermore, the lesions shown in this slice do not necessarily affect the performance of the structures shown here but may interfere in other parts of the brain, since, as stated before, in this strategy the segmentation performance oscillates independently of the lesion location.

4. Discussion

FreeSurfer is the most affected method when MS lesions are present. Despite being a method that deals with WM hypo-intensities (such as MS lesions), its segmentation performance significantly varies when MS lesions are introduced. This may be because this method is based on a Bayesian strategy that tries to infer the most likely segmentation given the image intensities and prior information in the form of an atlas, and thus, adding MS lesions may affect this method in two different ways. First, the registration of the atlas priors can be affected by the lesions, as seen in the multi-atlas strategy. Moreover, the incorporation of the lesions modifies the image intensities and consequently the intensity distribution of each structure, which is modeled as a Gaussian, may be affected and produce a different segmentation result.

The segmentation performance of the majority voting strategy is also affected when MS lesions are present. Similar to that of FreeSurfer, this method performance oscillates independently of the lesion location. As the atlas registration is performed globally, not only local lesions but also lesions in other parts of the brain have an effect on the registration result. On the other hand, although FIRST also performs registration to align the image and the model, it provides the most robust results. In this case, the method performs a local registration for which it uses a subcortical mask that determines whether a voxel is included or not in the calculation of the similarity function, which allows the registration to concentrate only on the subcortical alignment, and therefore, if the lesions are located outside the mask, they do not interfere with the registration result.

Regarding the effect of the lesion location, we have observed that FreeSurfer and majority voting are, in general, equally affected when the lesions are overlaid on the structure of analysis or placed in other parts of the brain. As these methods provide segmentation for the whole brain, the most logical approach is to perform a unique and global registration instead of trying to maximize the similarity for each structure independently by means of local registrations, as FIRST does with the group of subcortical structures. However, this global registration, despite being quicker, allows registration errors produced by brain irregularities such as WM lesions or tumors to be propagated to other parts of the brain and consequently affect the segmentation performance independent of the lesion location. On the other hand, the analyzed results show that in the case of FIRST, the lesion location has a direct effect on the method performance. In this case, the lesions that are overlaid on the structures worsen the segmentation result compared with that of the non-damaged structures, whereas the effects of lesions in other parts of the brain are inappreciable. It should come as no surprise that since FIRST is a local deformable strategy and segments each structure independently, the shape and intensities of lesions far from the structure of interest do not interfere with the deformation process.

As for the method robustness, FIRST provides the most consistent segmentations. Since this method is based on a deformable strategy constrained by shape and intensity, deformations that exceed the average geometric variation of the structure are avoided, and therefore, only specific lesions attached to the structure may cause the method performance oscillation. Furthermore, and contrary to the other two analyzed strategies, FIRST only segments the subcortical structures and works only with a small region of the brain instead of the whole volume.

Regarding the analyzed brain structures, in terms of DSC, the accumbens is the most affected by the presence of lesions, whereas the thalamus and the brainstem provide the most robust results when lesions are present. However, this behavior may be closely related to the fact that the accumbens is the smallest structure, whereas the brainstem and the thalamus are the largest ones. Thus, small changes in the segmentation result can imply large differences in the DSC when the structure volume is small and can have an insignificant effect when the structure volume is large. On the other hand, the structure in which we find the largest number of lesions is the caudate nucleus, which makes sense since the vast majority of MS patients have at least one ovoid periventricular lesion that augments the probability of this structure of being affected. The large number of outliers found for this structure when lesions are overlaid can be explained by the number of cases. Although we see that around the 50% of the analyzed cases do not show an excessive DSC difference with respect to the healthy image, the segmentation result may have more chances of being affected due to the variability in lesions location and intensity (sometimes similar to that of the caudate) found in the different images. Furthermore, as the caudate nucleus is a small structure, a larger effect is seen in the DSC when a small change in the segmentation result is produced.

In recent years, several brain structure segmentation methods following different segmentation strategies (González-Villà et al., 2016) have been described. Despite the large number of studies available in the literature, these approaches have been tested with images of non-lesioned brains, and therefore, how WM lesions affect their performance has not been evaluated. Performing such analysis with real images is not trivial, since there is no publicly available database with both structure ground truth and MS lesion annotation. Because of that, in an attempt to evaluate this effect, we generated a set of synthetic images as done in (Battaglini et al., 2012; Chard et al., 2010; Gelineau-Morel et al., 2012; Nakamura and Fisher, 2009) to evaluate the effect of the lesions on tissue segmentation. The effect of such lesions on automatic tissue segmentation strategies has been widely evaluated, and also strategies to reduce this effect, such as masking out the lesions or fill them before segmentation, have been proposed (Battaglini et al., 2012; Chard et al., 2010; Valverde et al., 2015).

In spite of the effect of MS lesions has been evaluated on several tissue segmentation methods, as far as we know only Gelineau-Morel et al. (2012) have evaluated how these lesions affect the segmentation of the subcortical brain structures. In their work, they use FIRST as the baseline method for segmentation, concluding that WM lesions led to an artificial decrease in all the DGM structures volume except the hippocampus. Corroborating that statement, a negative correlation between the lesion volume and the total DGM volume was found in our experiments when segmenting with FIRST. Despite the experiments performed here are not the same and they dealt with right and left hemispheres combined, our results are consistent with their findings, except for the left thalamus (for which we saw an average volume increase compared to the healthy controls) and the left hippocampus (which experimented an average volume loss). Furthermore, when analyzing the DGM structures individually we found that their volume differences (patient - healthy) did not have the same direction of correlation with the lesion volume, which is also consistent with the trend seen in their work (Gelineau-Morel et al., 2012).

There are a number of limitations in this work that have to be considered. Due to the lack of databases with both annotated MS lesions and structures ground truth, this study has been conducted using simulated images instead of real ones. Furthermore, our claims about the evaluated methods performance have to be prudent, given the fact that only four healthy images with structures ground truth have been used as the basis of our generation method. Moreover, as the simulated

lesions follow the spatial distribution found in real cases, the number of cases per structure with lesions overlaid is not the same for all the analyzed structures.

In summary, in this study we have presented an analysis about the effect of simulated MS lesions on three well-known automatic brain structure segmentation methods (FreeSurfer, FIRST and multi-atlas fused by majority voting), based on different segmentation strategies. We have demonstrated that there exists a direct effect of MS lesions on the performance of automatic brain structure segmentation methods. FreeSurfer seems to be the most affected algorithm whereas FIRST has shown to be the most robust against lesions. The lesions location does not seem to have a direct effect on the global strategies (FreeSurfer and majority voting) whereas in FIRST, which is a local strategy, the segmentation performance of a brain structure is more affected when there are lesions either overlaid or close to it. To the best of our knowledge, how different segmentation strategies to automatically segment the brain structures are affected by MS lesions had not been evaluated until now. This study addresses an important problem of the automatic segmentation methods of the DGM structures, which is related to MS lesion interference for the optimal segmentation. Investigating the influence of lesions on the segmentation on other diseases is indeed an important aspect of future research for the community.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2017.05.003.

Acknowledgements

S. González-Villà holds a UdG-BRGR2015 grant from the University of Girona. This work has been partially supported by “La Fundació la Marató de TV3” Ref. 201425 30, by Retos de Investigación TIN2014-55710-R and TIN2015-73563-JIN, and by MPC UdG 2016/022 grant.

References

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. Imaging 28 (8), 1266–1277.

Audoin, B., Zaaraoui, W., Reuter, F., Rico, A., Malikova, I., Confort-Gouny, S., Cozzone, P.J., Pelletier, J., Ranjeva, J.-P., 2010. Atrophy mainly affects the limbic system and the deep grey matter at the first stage of multiple sclerosis. J. Neurol. Neurosurg. Psychiatry 81 (6), 690–695.

Babalola, K.O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., Rueckert, D., 2009. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. NeuroImage 47 (4), 1435–1447.

Battaglini, M., Jenkinson, M., De Stefano, N., 2012. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Hum. Brain Mapp. 33 (9), 2062–2071.

Bergsland, N., Horakova, D., Dwyer, M., Dolezal, O., Seidl, Z., Vaneckova, M., Krasensky, J., Havrdova, E., Zivadinov, R., 2012. Subcortical and cortical gray matter atrophy in a large sample of patients with clinically isolated syndrome and early relapsing-remitting multiple sclerosis. Am. J. Neuroradiol. 33 (8), 1573–1578.

Bishop, C.A., Newbould, R.D., Lee, J.S., Honeyfield, L., Quest, R., Colasanti, A., Ali, R., Mattoscio, M., Cortese, A., Nicholas, R., Matthews, P.M., Muraro, P.A., Waldman, A.D., 2017. Analysis of ageing-associated grey matter volume in patients with multiple sclerosis shows excess atrophy in subcortical regions. Neuroimage Clin. 13, 9–15.

Calabrese, M., Rinaldi, F., Mattisi, I., Bernardi, V., Favaretto, A., Perini, P., Gallo, P., 2011. The predictive value of gray matter atrophy in clinically isolated syndromes. Neurology 77 (3), 257–263.

Chard, D.T., Jackson, J.S., Miller, D.H., Wheeler-Kingshott, C.A., 2010. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. J. Magn. Reson. Imaging 32 (1), 223–228.

Debernard, L., Melzer, T.R., Alla, S., Eagle, J., Stockum, S.V., Graham, C., Osborne, J.R., Dalrymple-Alford, J.C., Miller, D.H., Mason, D.F., 2015. Deep grey matter MRI abnormalities and cognitive function in relapsing-remitting multiple sclerosis. Psychiatry Res. Neuroimaging 234 (3), 352–361.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Fischl, B., 2012. FreeSurfer. NeuroImage 62 (2), 774–781.

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.

Gelineau-Morel, R., Tomassini, V., Jenkinson, M., Johansen-Berg, H., Matthews, P.M., Palace, J., 2012. The effect of hypointense white matter lesions on automated gray matter segmentation in multiple sclerosis. Hum. Brain Mapp. 33 (12), 2802–2814.

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.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33 (1), 115–126.

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.

Iglesias, J., Sabuncu, M., Van Leemput, K., 2012. A generative model for probabilistic label fusion of multimodal data. Multimodal Brain Image Anal. 7509, 115–133.

Jacobsen, C., Hagemeier, J., Myhr, K.-M., Nyland, H., Lode, K., Bergsland, N., Ramasamy, D.P., Dalaker, T.O., Larsen, J.P., Farbu, E., et al., 2014. Brain atrophy and disability progression in multiple sclerosis patients: a 10-year follow-up study. J. Neurol. Neurosurg. Psychiatry 85, 1109–1115.

Kazi, A.Z., Joshi, P.C., Kelkar, A.B., Mahajan, M.S., Ghawate, A.S., 2013. MRI evaluation of pathologies affecting the corpus callosum: a pictorial essay. Indian J Radiol Imaging 23 (4), 321–332.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46 (3), 786–802.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. Elastix: A toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29 (1), 196–205.

Landman, B., Warfield, S., 2012. MICCAI 2012 workshop on multi-atlas labeling. In: Landman, B.A., Warfield, S.K. (Eds.), MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling. CreateSpace Independent Publishing Platform, Nice, France.

Lee, J.H., Ryan, J., Andreescu, C., Aizenstein, H., Lim, H.K., 2015. Brainstem morphological changes in Alzheimers disease. Neuroreport 26 (7), 411–415.

Mak, E., Bergsland, N., Dwyer, M., Zivadinov, R., Kandiah, N., 2014. Subcortical atrophy is associated with cognitive impairment in mild Parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis. Am. J. Neuroradiol. 35 (12), 2257–2264.

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.

Nakamura, K., Fisher, E., 2009. Segmentation of brain magnetic resonance images for measurement of gray matter atrophy in multiple sclerosis patients. NeuroImage 44 (3), 769–776.

Nakamura, K., Guizard, N., Fonov, V.S., Narayanan, S., Collins, D.L., Arnold, D.L., 2014. Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. Neuroimage Clin. 4, 10–17.

Nyul, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging 19 (2), 143–150.

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 Clin. 4, 366–373.

Schoonheim, M.M., Popescu, V., Lopes, F.C.R., Wiebenga, O.T., Vrenken, H., Douw, L., Polman, C.H., Geurts, J.J., Barkhof, F., 2012. Subcortical atrophy and cognition sex effects in multiple sclerosis. Neurology 79 (17), 1754–1761.

Štecková, T., Hluštík, P., Sládková, V., Odstrčil, F., Mareš, J., Kaňovský, P., 2014. Thalamic atrophy and cognitive impairment in clinically isolated syndrome and multiple sclerosis. J. Neurol. Sci. 342 (1), 62–68.

Valverde, S., Oliver, A., Lladó, X., 2014. A white matter lesion-filling approach to improve brain tissue volume measurements. Neuroimage Clin. 6, 86–92.

Valverde, S., Oliver, A., Díez, Y., Cabezas, M., Vilanova, J., Ramió-Torrentà, L., Rovira, À., Lladó, X., 2015. Evaluating the effects of white matter multiple sclerosis lesions on the volume estimation of 6 brain tissue segmentation methods. Am. J. Neuroradiol. 36 (6), 1109–1115.

Wang, H., Yushkevich, P., 2013. Multi-atlas segmentation with joint label fusion and corrective learning - An open source implementation. Front. Neuroinform. 7 (27).

Weisenfeld, N.I., Warfield, S.K., 2011. Learning Likelihoods for Labeling (L3): A General Multi-Classifier Segmentation Algorithm. Springer, Toronto, Canada, pp. 322–329.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57.