Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both Multiple Sclerosis and elderly subjects

Philippe Tran, Urielle Thoprakarn, Emmanuelle Gourieux, Clarisse Longo dos Santos, Enrica Cavedo, Nicolas Guizard, François Cotton, Pierre Krolak-Salmon, Christine Delmaire, Damien Heidelberg, et al.

NeuroImage: Clinical 33 (2022) 102940

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both Multiple Sclerosis and elderly subjects

Philippe Tran [a,b,*], Urielle Thoprakarn [a], Emmanuelle Gourieux [i,j], Clarisse Longo dos Santos [a], Enrica Cavedo [a], Nicolas Guizard [a], François Cotton [d,e], Pierre Krolak-Salmon [e,f,g], Christine Delmaire [c], Damien Heidelberg [d], Nadya Pyatigorskaya [h], Sébastian Ströer [h], Didier Dormont [b,h], Jean-Baptiste Martini [a], Marie Chupin [i], Alzheimer's Disease Neuroimaging Initiatives, for the Frontotemporal Lobar Degeneration Neuroimaging Initiative

[a] Qynapse, Paris, France
[b] Equipe-projet ARAMIS, ICM, CNRS UMR 7225, Inserm U1117, Sorbonne Université UMR_S 1127, Centre Inria de Paris, Groupe Hospitalier Pitié-Salpêtrière Charles Foix, Faculté de Médecine Sorbonne Université, Paris, France
[i] CATI, ICM, CNRS UMR 7225, Inserm U1117, Sorbonne Université UMR_S 1127, Paris, France
[j] NeuroSpin, CEA, Saclay, France
[d] Service de Radiologie, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, Pierre-Bénite, France
[e] Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69495, Pierre-Bénite, France
[f] Clinical and Research Memory Centre of Lyon, Hospices Civils de Lyon, Lyon, France
[g] INSERM, U1028, UMR CNRS 5292, Lyon Neuroscience Research Center, Lyon, France
[c] Fondation Adolphe de Rothschild, Paris, France
[h] Department of Neuroradiology, Groupe Hospitalier Pitié-Salpêtrière, AP-HP, Sorbonne Université UMR_S 1127, Paris, France

ARTICLE INFO

Keywords:
White matter hyperintensities
Age-related vascular disorder
3D T2-FLAIR
Automatic segmentation
Multiple sclerosis
MRI

ABSTRACT

Different types of white matter hyperintensities (WMH) can be observed through MRI in the brain and spinal cord, especially Multiple Sclerosis (MS) lesions for patients suffering from MS and age-related WMH for subjects with cognitive disorders and/or elderly people. To better diagnose and monitor the disease progression, the quantitative evaluation of WMH load has proven to be useful for clinical routine and trials. Since manual delineation for WMH segmentation is highly time-consuming and suffers from intra and inter observer variability, several methods have been proposed to automatically segment either MS lesions or age-related WMH, but none is validated on both WMH types. Here, we aim at proposing the White matter Hyperintensities Automatic Segmentation Algorithm adapted to 3D T2-FLAIR datasets (WHASA-3D), a fast and robust automatic segmentation tool designed to be implemented in clinical practice for the detection of both MS lesions and age-related WMH in the brain, using both 3D T1-weighted and T2-FLAIR images. In order to increase its robustness for MS lesions, WHASA-3D expands the original WHASA method, which relies on the coupling of non-linear diffusion framework and watershed parcellation, where regions considered as WMH are selected based on intensity and location characteristics, and finally refined with geodesic dilation. The previous validation was performed on 2D T2-FLAIR and subjects with cognitive disorders and elderly subjects. 60 subjects from a heterogeneous database of dementia patients, multiple sclerosis patients and elderly subjects with multiple MRI scanners and a wide range of lesion loads were used to evaluate WHASA and WHASA-3D through volume and spatial agreement in comparison with consensus reference segmentations. In addition, a direct comparison on the MS database with six available supervised and unsupervised state-of-the-art WMH segmentation methods (LST-LGA and LPA, Lesion-TOADS, lesionBrain, BIANCA and nicMSlesions) with default and optimised settings (when feasible) was conducted. WHASA-3D confirmed an improved performance with respect to WHASA, achieving a better spatial overlap (Dice) (0.67 vs 0.63), a reduced absolute volume error (AVE) (3.11 vs 6.2 mL) and an increased volume

* Corresponding author at: Qynapse SAS, 130 rue de Lourmel, Paris 75015, France. Institut du Cerveau – Paris Brain Institute, Hôpital Pitié, 47 Boulevard de l'Hôpital, 75013, Paris.
E-mail address: ptran@qynapse.com (P. Tran).

https://doi.org/10.1016/j.nicl.2022.102940
Received 28 July 2021; Received in revised form 15 December 2021; Accepted 6 January 2022
Available online 10 January 2022

agreement (intraclass correlation coefficient, ICC) (0.96 vs 0.78). Compared to available state-of-the-art algorithms on the MS database, WHASA-3D outperformed both unsupervised and supervised methods when used with their default settings, showing the highest volume agreement (ICC = 0.95) as well as the highest average Dice (0.58). Optimising and/or retraining LST-LGA, BIANCA and nicMSlesions, using a subset of the MS database as training set, resulted in improved performances on the remaining testing set (average Dice: LST-LGA default/optimized = 0.41/0.51, BIANCA default/optimized = 0.22/0.39, nicMSlesions default/optimized = 0.17/0.63, WHASA-3D = 0.58). Evaluation and comparison results suggest that WHASA-3D is a reliable and easy-to-use method for the automated segmentation of white matter hyperintensities, for both MS lesions and age-related WMH. Further validation on larger datasets would be useful to confirm these first findings.

Nomenclature

AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AVE	Absolute Volume Error
BIANCA	Brain Intensity AbNormality Classification Algorithm
CIS	Clinically Isolated Syndrome
CNN	Convolutional Neural Network
CSF	Cerebrospinal Fluid
FOV	Field Of View
FPR	False Positive Ratio
FSL	FMRIB Software Library
FTD	Frontotemporal Dementia
FTLDNI	Frontotemporal Lobar Degeneration Neuroimaging Initiative
GM	Grey Matter
HC	Healthy Control
ICC	Intraclass Correlation Coefficient
k-NN	k-nearest neighbors
Lesion-TOADS	Lesion-TOpology preserved Anatomical Segmentation
LOP-STAPLE	Logarithmic Optinion Pool – Simultaneous Truth and Performance Level Estimation
LST-LGA	Lesion Segmentation Tool – Lesion Growth Algorithm
LST-LPA	Lesion Segmentation Tool – Lesion Prediction Algorithm
MCI	Mild Cognitive Impairment
MICCAI	Medical Image Computing and Computer-Assisted Intervention
MIPAV	Medical Image Processing, Analysis and Visualization software
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
MSSEG	Multiple Sclerosis lesion Segmentation challenge
PPMS	Primary Progressive Multiple Sclerosis
RRMS	Relapsing Remitting Multiple Sclerosis
SPM12	Statistical Parametric Mapping software v12
SPMS	Secondary Progressive Multiple Sclerosis
T2-FLAIR	T2 Fluid Attenuated Inversion Recovery
TPR	True Positive Ratio
TE	Echo Time
TI	Inversion Time
TR	Repetition Time
WHASA	White matter Hyperintensities Automated Segmentation Algorithm
WHASA-3D	White matter Hyperintensities Automated Segmentation Algorithm for 3D datasets
WM	White Matter
WMH	White Matter Hyperintensities

1. Introduction

White matter hyperintensities (WMH) in the brain and spinal cord are areas with high signal intensities visible on T2-weighted fluid attenuated inversion recovery (T2-FLAIR) MRI sequences, and are common findings in multiple sclerosis patients and elderly people. In MS, those focal areas are designated as "MS lesions" (Filippi et al., 2016; Rovira and León, 2008; Wattjes et al., 2015; Fazekas et al., 1999; Filippi et al., 2019) and for elderly people, WMHs are considered to be a vascular contributor to various disorders such as cognitive decline or dementia (Frey et al., 2019; Kim et al., 2008). Those WMH will then be referred as "age-related WMH" in this study. Consensus guidelines for MRI in MS, such as provided by the French Observatory of MS (OFSEP) (Brisset et al., 2020; Cotton et al., 2015), provide recommendations for imaging techniques to further improve the visualization of lesions (Brisset et al., 2020; Rocca et al., 2013; Simon et al., 2006). An improvement has been reported regarding whole brain lesion detection using 3D T2-FLAIR sequences rather than 2D sequences (Naganawa, 2015), particularly in cortical and infratentorial regions, which are typical locations for MS lesions (Gramsch et al., 2015; Polman et al., 2011).

In MS clinical trials, the assessment of MS lesions volume change is considered a clinically relevant marker of disease progression (Meier et al., 2007) and has been used as an outcome measure, thus considered as a surrogate marker of potential disease-modifying treatments (Mikol et al., 2008; Radue et al., 2012). In the clinical practice, the detection of MS lesions has been included in the MS diagnostic criteria (Polman et al., 2011; Thompson et al., 2018) as they provide useful information for the diagnosis and treatment of MS (Giorgio and De Stefano, 2018). Similarly, the accurate segmentation of age-related WMH could be introduced in clinical practice to support diagnosis, prognosis and treatment monitoring of dementia, as shown in longitudinal studies for Alzheimer's Disease (AD) dementia, cerebral small-vessel disease, frontotemporal dementia and other cognitive disorders (Alber et al., 2019; Frey et al., 2019; Meier et al., 2007; Schmidt et al., 2004; Debette and Markus, 2010). Due to the heterogeneity in WMH appearance, location, size and shape, in addition to anatomical differences between subjects (García-Lorenzo et al., 2013), the identification of WMH on brain MRI in clinical routine is mostly performed with the help of semi-automatic tools, or with visual scales such as Fazekas (Fazekas et al., 1987) by neuroradiologists. Manual outlining of WMH is time-consuming and still suffers significant intra- and inter-rater variability (Commowick et al., 2018; Grimaud et al., 1996; Zijdenbos et al., 2002; Styner et al., 2008), especially since the recent advances in acquisition techniques have enabled a more generalized use of 3D T2-FLAIR imaging with thinner slices. An automatic WMH segmentation method would thus be highly useful in clinical routine and clinical trials. However, such method needs to be reliable, reproducible and efficient to allow processing hundreds or thousands of datasets.

Several automated methods have been described for the delineation of age-related WMH (Caligiuri et al., 2015) and MS lesions (Danelakis et al., 2018; García-Lorenzo et al., 2013) with varying amounts of

Table 1
MR acquisition parameters as given in the DICOM headers.

Database	Cohort	n	Machine (Field strength)	Sequence	TR/TE/TI (ms)	Flip angle	Field of view (FOV, mm)	Voxel size (mm)
MS	LITMS	30	Siemens Magnetom Trio (3T)	2D T1	2000/20/800	120	408×512×152	0.43×0.43×0.82
				3D T2-FLAIR	5000/392/1800	120	192×512×512	0.80×0.47×0.47
Various dementia	ADNI	9	GE Discovery MR750W (3T)	3D T1	7.4/3.1/400	11	196×256×256	1.0×1.0×1.0
				3D T2-FLAIR	4800/116.2/1454	90	218.4×256×256	1.0×1.0×1.02
		1	Philips Ingenia (3T)	3D T1	6.5/2.9/900	9	256×256×211	1.0×1.0×1.0
				3D T2-FLAIR	4800/271/1650	90	192×256×256	1.0××1.0×1.02
	NIFD	15	Siemens TrioTim (3T)	3D T1	2300/3/900	90	160×256×240	1.0×1.0×1.02
				3D T2-FLAIR	6000/388/2100	120	160×250×250	0.98×0.98×1.0
	MEMORA	3	Philips Ingenia (3T)	3D T1	7.2/3.3/None	9	176×256×256	1.0×1.0×1.0
				3D T2-FLAIR	8000/355.5/2400	90	183×240×240	0.83×0.83×1.06
		1	Philips Ingenia (3T)	3D T1	9.4/4.3/None	8	170×250×250	0.74×0.74×0.85
				3D T2-FLAIR	5400/360/1800	90	183×250×250	0.75×0.74×1.04
		1	GE Optima M5450 W (3T)	3D T1	8.8/4.2/None	15	512×512×312	0.5×0.5×0.5
				3D T2-FLAIR	8000/132/2117	90	512×240×512	0.49×0.8×0.49

manual input and/or output postprocessing. They vary greatly in terms of complexity, computational time, required imaging modalities and are generally divided into two groups: supervised and unsupervised methods (García-Lorenzo et al., 2013). Most unsupervised methods rely on clustering techniques to allocate the voxels to different classes (for example tissue – white matter (WM), grey matter (GM), cerebrospinal fluid (CSF) and lesion classes) based on specific features (for example voxel intensity). Those methods are based either on probabilistic models (Jack et al., 2001; Schmidt et al., 2012), thresholding techniques with post-processing (Roura et al., 2015; Schmidt et al., 2012) or both (Samaille et al., 2012). On the other hand, supervised methods rely on a learning step to learn the definition of WMH, thus requiring previously labelled datasets, usually MRI images with manual segmentation. Underlying reported classification methods include k-nearest neighbors (k-NN) (Griffanti et al., 2016; Steenwijk et al., 2013; Fartaria et al., 2016), decision random forests (Geremia et al., 2011), support vector machine (SVM) (Yamamoto et al., 2010), and, more recently, convolutional neural networks (CNNs) (LeCun et al., 2010; Valverde et al., 2019). In order to be used in clinical practice and clinical trials, the performances of WMH segmentation methods, designed either for MS or for dementia/elderly patients, must be validated with respect to a gold standard, such as manual segmentations performed by expert neuroradiologists. The available methods present in the literature are currently validated either on healthy controls and MS patients or vascular and neurodegenerative disease patients (Griffanti et al., 2016; Jain et al., 2015; Samaille et al., 2012; Schmidt et al., 2012; Shiee et al., 2010; Valverde et al., 2019; Weeda et al., 2019; Coupé et al., 2018; Schmidt, 2017). To the best of our knowledge, there is no WMH automated segmentation method that has been validated on both demyelinating and neurodegenerative diseases so far.

In this paper, we aim at proposing the White matter Hyperintensities Automatic Segmentation Algorithm adapted to 3D T2-FLAIR datasets (WHASA-3D), which is a major improvement of the unsupervised method WHASA (Samaille et al., 2012), a fully automatic unsupervised method that relies on non-linear diffusion and watershed-based segmentation followed by intensity and anatomy-based selection, and was up to now only validated on elderly subjects or dementia patients with 2D T2-FLAIR datasets. This study will address the automatic segmentation of WMH in MS patients, healthy controls and patients suffering from neurodegenerative diseases (AD, fronto-temporal dementia (FTD), cognitive impairments) in order to yield a method that is reliable on both age-related WMH and MS lesions. To do so, we will use a database with 60 subjects (healthy controls, MS patients and patients with cognitive disorders) with 3D T2-FLAIR scans from seven centres and with a large lesion load variability, provided with manual segmentations of WMH. We will compare the performance of WHASA-3D with its original version on all 60 subjects, and with other available state-of-the art methods (four unsupervised and two supervised methods) on a subset of 30 MS patients extracted from the full database of 60 subjects.

WHASA-3D is currently included in the medical device QyScore®, a CE-marked and FDA-cleared software, developed by Qynapse (https://www.qynapse.com/), that provides segmentation and volumetric measurements of brain imaging markers.

2. Material and methods

WHASA-3D has been evaluated on a composite database built from several databases. As described below, the database includes various populations, acquired on different scanners and using different imaging protocols, and has been divided into several datasets for the different stages of this work.

2.1. MRI data description

2.1.1. Multiple Sclerosis database

Datasets from a cohort of 30 MS patients were acquired using a 3T Siemens Magnetom Trio MR system at the University Medical Center Ljubljana (Lesjak et al., 2017). Each MR dataset consisted of 2D T1-w, 3D T2-weighted and 3D T2-FLAIR images. T1-w and T2-FLAIR images are used here. They had been interpolated during acquisition, resulting in 0.43×0.43×0.82 mm and 0.80×0.47×0.47 mm apparent resolutions. In order to make these datasets more comparable with those described below, they have been resampled to, respectively, 1×1×0.82 mm and 0.80×1×1mm.

2.1.2. Various dementia database

Three different cohorts (including two from publicly available databases), embedding 3D T1-w and 3D T2-FLAIR images, were used and combined to cover a wide range of WMH lesion loads and to have an insight on robustness with respect to MRI scanners and acquisition settings.

ADNI

MRI data from ten subjects (three with Alzheimer's Disease and seven elderly normal controls) were randomly selected from the Alzheimer's Disease Neuroimaging Initiative database (ADNI) (Petersen et al., 2010), a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease. MR images were acquired on 3T GE Discovery MR750W and 3T Philips Ingenia scanners. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

NIFD

Table 2
Demographic information.

Database	Cohort	n	Clinical status	Age range (mean (SD))	Sex proportion (F:M)
MS	LITMS	30	24 RRMS, 2 SPMS, 1 PPMS, 2 CIS, 1 unspecified	25–64 (39.3 (10.1))	23:7
Various dementia	ADNI	10	3 AD, 7 HC	68.3–90.9 (81.7 (6.7))	5:5
	NIFD	15	6 FTD, 2 HC, 7 unspecified	54–83 (67.3 (7.3))	7:8
	MEMORA	5	2 AD, 2 major cognitive disorders, 1 unspecified	76–88 (83 (5))	2:3

*RRMS = Relapsing remitting multiple sclerosis, SPMS = secondary progressive multiple sclerosis, PPMS = primary progressive multiple sclerosis, CIS = Clinically Isolated Syndrome, AD = Alzheimer's Disease, HC = Healthy Control (elderly subjects), FTD = FrontoTemporal Dementia.

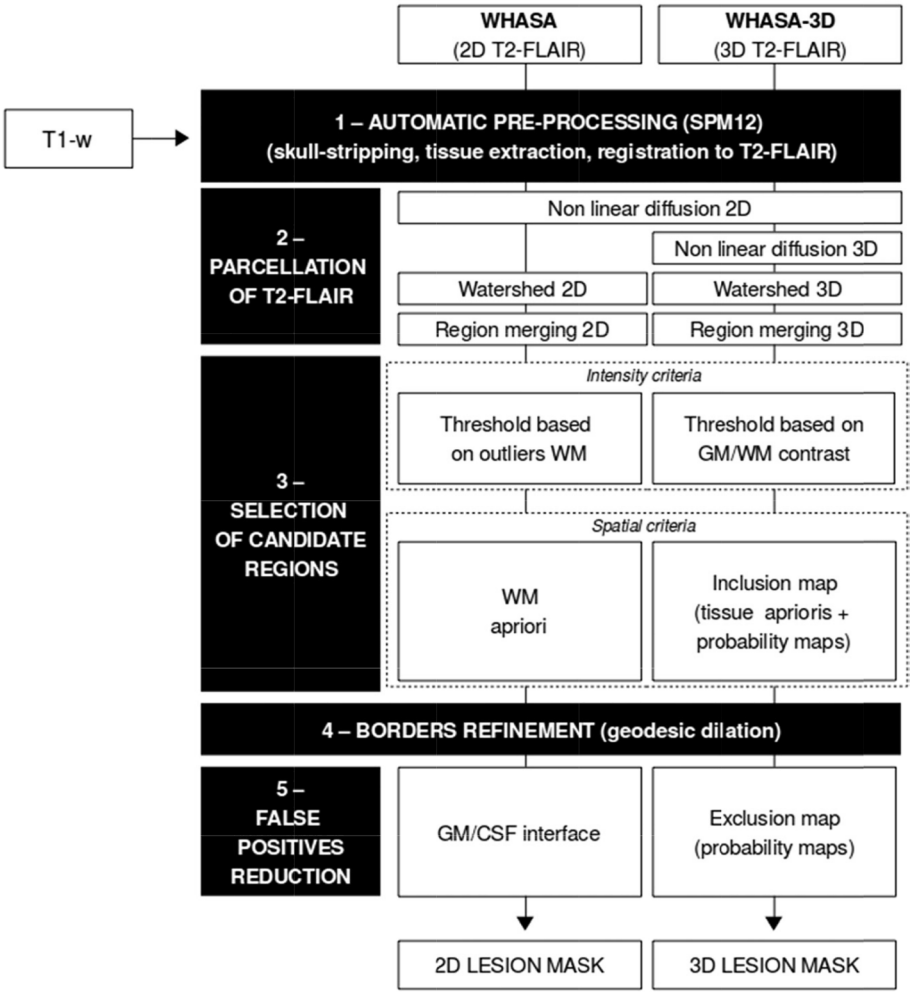

Fig. 1. General overview of WHASA and WHASA-3D for 2D and 3D T2-FLAIR.

MRI data from 15 subjects (six with FTD, two elderly normal controls and seven unspecified diagnostic) were randomly selected from the FrontoTemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI, nicknamed NIFD). MR images were acquired on a 3T Siemens TrioTim. FTLDNI was funded through the National Institute of Aging, and started in 2010. The primary goals of FTLDNI were to identify neuroimaging modalities and methods of analysis for tracking frontotemporal lobar degeneration (FTLD) and to assess the value of imaging versus other biomarkers in diagnostic roles. The Principal Investigator of NIFD was Dr. Howard Rosen, MD at the University of California, San Francisco. The data are the result of collaborative efforts at three sites in North America. For up-to-date information on participation and protocol, please visit http://memory.ucsf.edu/research/studies/nifd.

MEMORA

MRI data from five subjects (four with major cognitive impairment, two with AD, two without diagnosis, and one unspecified diagnosis) were randomly selected from MEMORA, a clinical routine study created to follow patients with cognitive disorders. MRI images were acquired at the Hospices Civils de Lyon centre on a 3T Philips Ingenia and a 3T GE Optima MR450W scanners.

MR parameters are summarized for all three datasets in Table 1 and demographic information in Table 2.

2.1.3. Manual segmentation

The performance of WHASA-3D will be evaluated through systematic comparison with a reference. This ground truth has been defined for all 60 3D T2-FLAIR images as the consensus of three manual segmentations performed by three neuroradiologists, as described below.

Multiple Sclerosis database

Manual lesion segmentations created by three raters were available

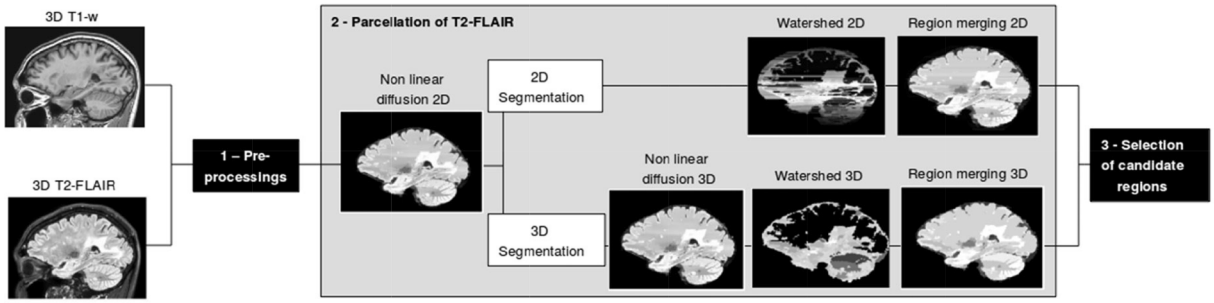

Fig. 2. Parcellation of the T2-FLAIR for 2D and 3D cases to obtain candidate regions.

within the freely available database, as described in (Lesjak et al., 2017): one rater was a second-year radiology intern, while the other two raters were senior radiologists with more than 10 years of experience in assessing MR scans of MS patients. The separate segmentations were carefully revised with the ITK-SNAP software (Yushkevich et al., 2006) by all raters in terms of lesion locations and borders to create the final consensus segmentation.

Various dementia database

Three trained neuroradiologists (with three, six and 18 years of experience in neuroradiology) performed manual segmentation of WMH on 3D T2-FLAIR images. Lesion maps were first generated from T1-w and 3D T2-FLAIR images using LST-LGA v2.0.15 (Lesion Segmentation Tool – Lesion Growth Algorithm), implemented in SPM (Schmidt et al., 2012). The neuroradiologists then corrected each lesion map when necessary, using the manual editing tool ITK-SNAP. The ground truth was then defined as the consensus among those three corrected segmentations, obtained through the LOP-STAPLE algorithm (Akhondi-Asl et al., 2014).

Reference WMH volume

Reference volumes ranged from 0.3 to 68 mL, with a mean of 21 mL and a standard deviation of 15 mL for the whole database (17 $\pm$ 16 mL [0.3–52] for the MS database and 24 $\pm$ 14 mL [0.3–68 mL] for the various dementia database).

2.1.4. Building databases for evaluation and comparison

This section provides more details about the data used for the development and evaluation of WHASA-3D and the comparison to other methods, in order to ensure a fair and unbiased evaluation.

WHASA-3D evaluation

We split our evaluation database (MS and various dementia databases) into a training set and a validation set. Eight subjects (two from each cohort, with a wide range of lesion load and age, MRI systems and pathology) were used to optimize the development of WHASA-3D, while the remaining 52 subjects were used as an independent validation base.

Comparison with state-of-the art methods

The comparison to state-of-the-art methods was focused on the MS database as we wanted to guarantee that our method works properly for MS patients, since most methods had been designed and evaluated for MS patients. The comparison to the methods with their default parameters was conducted on the whole MS database. Since some state-of-the-art methods could be optimized/re-retrained, we then randomly split the MS database into an optimization subset of 10 subjects with a wide range of lesion load for optimization and re-training purposes, while the remaining 20 subjects were used as an independent validation subset. Thus, our MS database was split into three folds of ten subjects, one fold being used for optimization/training and the remaining two folds for validation.

2.2. Methods

This section first gives a brief description of the original automated WMH segmentation method WHASA (Samaille et al., 2012) and details the specific steps of WHASA-3D developed to address the segmentation of 3D T2-FLAIR images. It then introduces the freely available algorithms that will be compared to WHASA and describes the strategy underlying the performance evaluation of the algorithms.

2.2.1. WHASA method

WHASA relies on the coupling of non-linear diffusion and watershed parcellation; regions considered as corresponding to WMH are then selected based on intensity and location characteristics then finally refined with geodesic dilation. Fig. 1 shows the general overview of WHASA and WHASA-3D.

Standard pre-processing steps using SPM12 (Ashburner and Friston, 2005) extract tissue probability maps from the T1-w image, register them to the T2-FLAIR image and correct the T2-FLAIR image for intensity inhomogeneities. Non-linear diffusion then enables to enhance the contrast between hyper-intense areas and surrounding healthy tissue and to reduce the contrast between GM and WM on the inhomogeneity corrected T2-FLAIR image; its combination with the watershed-resulting parcellation yields a piecewise constant image (step "parcellation of T2-FLAIR" on Fig. 2). Candidate lesions are extracted from this piecewise constant image with an automatically computed threshold. Tissue probability maps drive the selection of the relevant candidate lesions according to their location. Finally, a geodesic dilation is then applied in order to refine borders of lesions, with the help of a second lighter non-linear diffusion (diffusion parameter twice smaller) to better take into account large or diffuse WMH (Samaille, 2013).

The original algorithm was designed for 2D T2-FLAIR images with thick slices and several steps were implemented using a 2D slice-by-slice approach to ensure robustness but are not optimal for 3D T2-FLAIR images. In the following subsections we will describe how these steps were redesigned and implemented for 3D T2-FLAIR datasets.

Parcellation of 3D T2-FLAIR

This step aims at parcellating the T2-FLAIR image in homogeneous regions, with alternating iterations of non-linear diffusion (Perona and Malik, 1990) and watershed, followed by a final region merging step, as described for 2D and 3D pipelines in Fig. 2.

For 2D T2-FLAIR images, 2D non-linear diffusion was run, and the diffusion parameter was automatically set as the mean of the intensity gradient on the GM/WM interface obtained from the preprocessing step. A series of 100 iterations with a time-step of 0.1 alternated with a 2D watershed parcellation step until convergence of the whole process, which was reached when two consecutive watershed results were strictly identical. Each region of the final watershed was then labelled with its mean intensity as computed on the T2-FLAIR image. Adjacent regions with close intensity values (mean intensity difference lower than the diffusion parameter) were merged together to reduce the number of regions considered in the candidate region selection step.

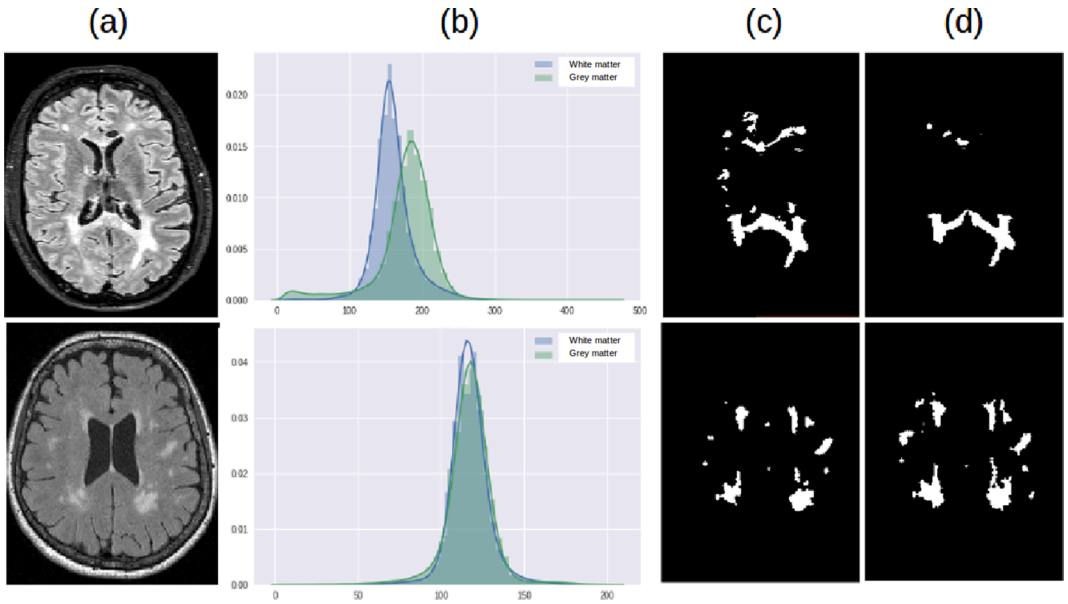

Fig. 3. 3D T2-FLAIR images with two types of GM/WM contrasts, with respective histograms and WMH segmentations. The first row shows high GM/WM contrast, the second row shows low GM/WL contrast. (a) FLAIR (b) Histogram (c) Segmentation with Thr_{WM} (d) Segmentation with Thr_{GM}.

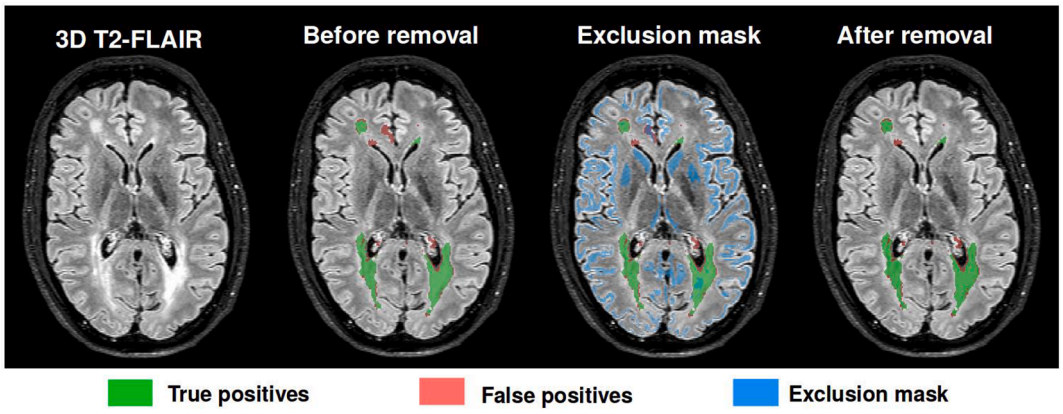

Fig. 4. False positive detection and removal.

For 3D T2-FLAIR images, ten iterations of 3D non-linear diffusion have now been added to minimise the between-slice discrepancies yielded by the 2D diffusion process, while still benefiting from its speed and robustness. A 3D watershed ensures to obtain a 3D-consistent piecewise constant image. Adjacent regions of the watershed parcellation are then labelled using their mean intensity on the T2-FLAIR image, then merged with a dedicated iterative 3D process: starting from the region with maximum intensity, neighbouring regions are iteratively merged according to their intensity contrast with the initial region, the merging criterion being the diffusion parameter. This step is crucial to ensure time efficiency for the following steps, as the number of regions generated by the 3D watershed is much larger than for 2D T2-FLAIR images (about 75,000 and 9000 regions generated for 3D and 2D T2-FLAIR respectively).

Selection of candidate regions

Candidate regions still have to be identified as WMH (MS lesions or age-related WMH) through intensity and anatomical rules.

In this step, we select hyperintense regions using an intensity threshold followed by spatial information about regions location, in order to refine the set of candidate lesions. Hyperintense areas corresponding to WMH could be defined as outliers for the WM intensity distribution, as they are mostly found in white matter.

Considering Gaussian distribution, the threshold to detect WMH could thus be defined as follows:

$$Thr_{WM} = \mu_{WM} + 2.698*\sigma_{WM} \quad (1)$$

where μ_{WM} and σ_{WM} are the mean and standard deviation of the WM intensity distribution, computed from the inhomogeneity corrected T2-FLAIR image.

However, depending on acquisition parameters and patients age range, two types of images can be observed among 3D T2-FLAIR images, based on GM/WM contrast characteristics, as shown in Fig. 3b: high GM/WM contrast (first row), with a clear distinction between WM and GM intensity modes, and low GM/WM contrast (second row), with nearly merged WM and GM intensity modes.

For high contrast images, the threshold may result in embedding voxels with intensity belonging to the GM intensity distribution. A threshold using the GM intensity distribution, as introduced in the LST method (Schmidt et al., 2012), may thus be more robust:

$$Thr_{GM} = \mu_{GM} + \sigma_{GM} \quad (2)$$

where μ_{GM} and σ_{GM} are the mean and standard deviation of the GM in-

Table 3
Default settings used for the methods in the comparison experiment.

Methods	WHASA-3D	LST-LGA default		lesionBrain	Lesion-TOADS	LST-LPA		BIANCA default		nicMSlesions default
Version	v1.8	v2.0.15		v1.0	MIPAV v7.2	v2.0.15		FSL 6.0.0		N/A
MRI sequences	T2-FLAIR + T1	T2-FLAIR + T1		T2-FLAIR + T1	T2-FLAIR + T1	T2-FLAIR		T2-FLAIR + T1		T2-FLAIR + T1
Base space	T2-FLAIR	T1		T2-FLAIR	T1	T2-FLAIR		T2-FLAIR		T2-FLAIR
Threshold	/	Initial threshold κ = 0.3 Probability map threshold = 0.5		/	/	Probability map threshold = 0.5		Probability map threshold = 0.9		Probability map threshold = 0.5
Other options	/	/		/	/	/		10,000 non-lesion points; 2000 lesion points; any location of non-lesion points		Batch-size = 20000
Model	/	/		/	/	Pre-trained		Trained with WHASA-3D training set (8 subjects)		Pretrained "baseline_2ch"

tensity distribution, computed from the inhomogeneity corrected T2-FLAIR image. However, for low contrast images, this threshold may result in embedding normal tissue, that would remain below the threshold if the standard deviation of GM is lower than that of WM.

These two intensity behaviours have been confirmed on the training set: Thr_{WM} yields better results on low GM/WM contrast T2-FLAIR images (Fig. 3c), while Thr_{GM} yields better results on high GM/WM contrast T2-FLAIR (Fig. 3d). A contrast-based barycentre was thus introduced between the two thresholds (1) (2) to obtain a generalized threshold robust to GM/WM contrast:

$$Thr_{generalized} = \rho^{*}(Thr_{GM}) + (1-\rho)^{*}(Thr_{WM})$$

with ρ the contrast-based weighting factor, derived from contrast and standard deviation values computed on the tissue probability maps and validated on the training set.

False positive detection and removal

Some false positives, namely voxels mistakenly considered as WMH, remain after the candidate selection and border refinement step; they are often located in the cortical grey matter, even more frequently for 3D T2-FLAIR with high GM/WM contrast, for which cortical folding may result in focal high intensity areas. Although some WMH may truly be located in the cortical grey matter, these are very difficult to distinguish from false positives with only 3D T2-FLAIR image. An automatic post-processing step is thus applied to remove all hyperintense voxels within the cortex from the segmentation mask as illustrated in Fig. 4.

We identify the voxels most likely to belong to GM by creating an exclusion map from the tissue probability maps for WM, GM, CSF previously extracted from the T1-w image at the preprocessing step. A morphological erosion is then applied on this resulting exclusion mask, and the largest connected component is kept as the final exclusion mask to embed only the cortical regions. Candidate lesions are then discarded if they overlap the exclusion mask for more than half of their voxels.

2.2.2. Other methods

In order to evaluate the performance of WHASA-3D, its results were compared with those obtained with state-of-the-art freely available methods on a dedicated dataset. An optimization of the parameters or a model retraining was performed on a optimization dataset for methods that allow it. The state-of-the-art methods, and their re-optimisation step when needed, are described below.

Unsupervised algorithms

LST-LGA

Lesions were segmented by the lesion growth algorithm (Schmidt et al., 2012) as implemented in the LST toolbox version 2.0.15 (www.statistical-modelling.de/lst.html) for SPM. The algorithm first segments the T1-w images into the three main tissue classes (CSF, GM and WM). This information is then combined with the intensities from the co-registered T2-FLAIR in order to compute lesion belief maps. By thresholding these maps with a pre-chosen initial threshold (κ), an initial binary lesion map is obtained which is subsequently grown along voxels that appear hyperintense in the T2-FLAIR image. The result is a lesion probability map. Performance evaluation of LST-LGA was performed in MS patients and healthy subjects (Schmidt et al., 2012).

lesionBrain

lesionBrain 1.0 is an online tool for white matter lesion segmentation (Coupé et al., 2018) and has been integrated into the volBrain platform (https://volbrain.upv.es/). The method first uses the T1-w images to segment several anatomical structures (intracranial cavity, brainstem, cerebellum,lateral ventricles and the brain tissue maps). Lesions are segmented based on a three-stage strategy: multimodal patch-based segmentation, patch-based regularization of the created probability map of lesions and patch-based error correction using an ensemble of shallow neural networks to limit false positives. Its robustness and

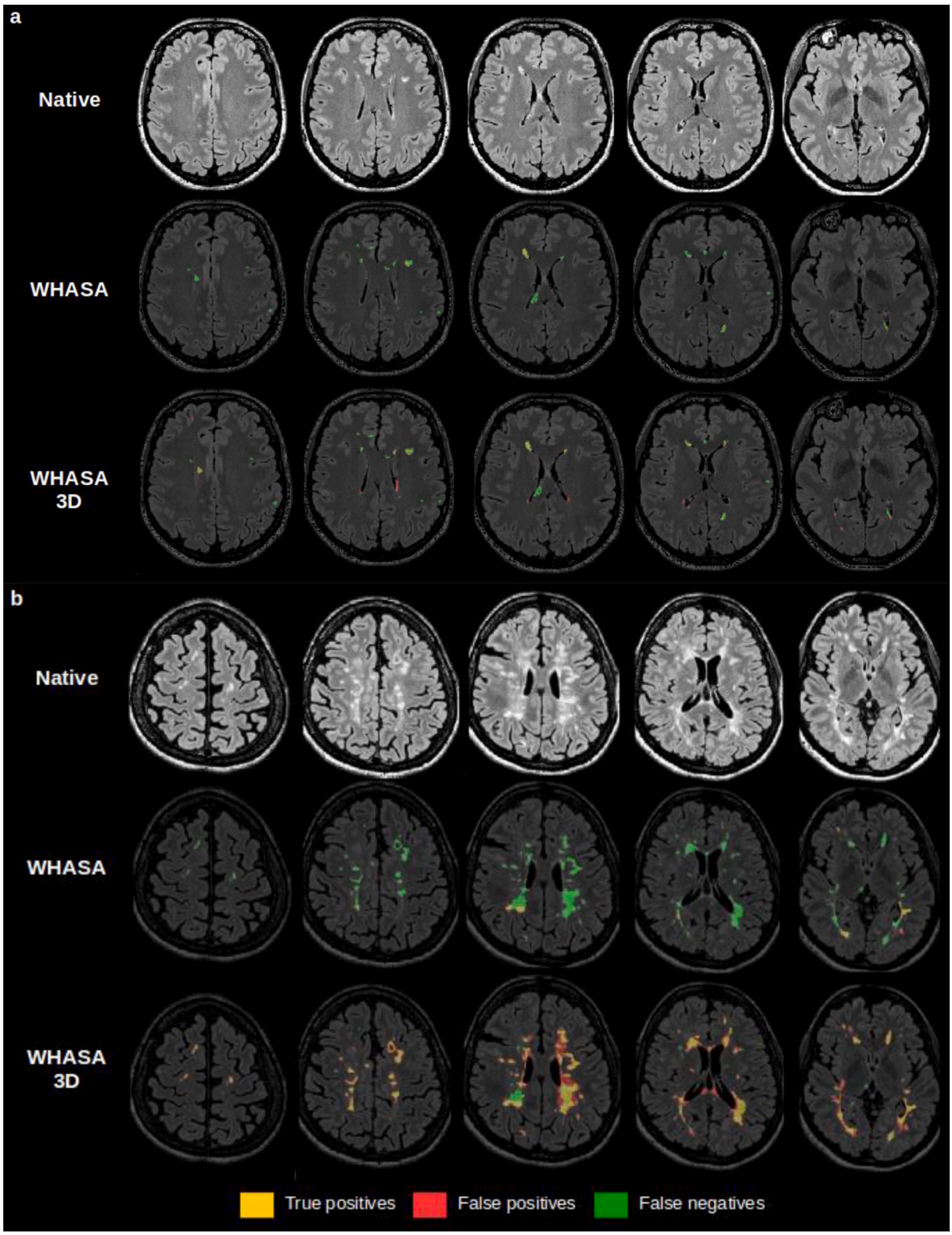

Fig. 5. Illustrations of WHASA and WHASA-3D results on two typical MS subjects with (a) low lesion load (reference volume: 2.51 mL; WHASA volume: 1.30 mL, WHASA-3D volume: 3.71 mL) and (b) high lesion load (reference volume: 50.04 mL, WHASA volume: 7.27 mL, WHASA-3D volume: 37.24 mL).

accuracy have been evaluated on the MSSEG MICCAI Challenge 2016 (Commowick et al., 2018) with 3D T1-w and 3D T2-FLAIR MRI acquired for 15 MS patients (Coupé et al., 2018).

Lesion-TOADS

Lesion-TOpology preserved Anatomical Segmentation (Lesion-TOADS) (Shiee et al., 2010) is a fully automatic method for the segmentation of MS white matter lesions from T1-w and T2-FLAIR images and is available as a plug-in for the MIPAV software (http://mipav.cit.nih.gov/). Lesion-TOADS embeds an iterative algorithm for fuzzy

Table 4
Median (Average ± std [min–max]) for measures of overlap and volumetric agreement with the reference segmentation for WHASA and WHASA-3D.

Database	Metrics median (mean ± std [min–max])	WMH volume	AVE	Dice	F1-score	TPR	FPR	ICC
MS and Various Dementia	Reference	19.9 (21.1 ± 15.7 [0.3–68.0])	N/A	N/A	N/A	N/A	N/A	N/A
	WHASA	16.6 (16.5 ± 13.6 [0.2–58.3])	2.8 (6.2 ± 8.8 [0–42.8])	0.74 (0.63 ± 0.22 [0.13–0.92])	0.39 (0.37 ± 0.14 [0.08–0.70])	0.68 (0.60 ± 0.26 [0.11–0.90])	0.21 (0.23 ± 0.19 [0.01–0.83])	0.78
	WHASA-3D	20.4 (19.7 ± 14.6 [0.5–67.5)	2.0 (3.1 ± 3.2 [0–13.8])	0.76 (0.67 ± 0.20 [0.21–0.91])	0.43 (0.42 ± 0.11 [0.15–0.63])	0.72 (0.67 ± 0.19 [0.26–0.95])	0.22 (0.31 ± 0.23 [0.02–0.83])	0.96
MS	Reference	14.1 (17.4 ± 16.1 [0.3–52.5])	N/A	N/A	N/A	N/A	N/A	N/A
	WHASA	5.7 (8.13 ± 8.55 [0.2–31.4])	3.7 (9.3 ± 11.3 [0.1–42.8])	0.46 (0.50 ± 0.23 [0.13–0.82])	0.29 (0.31 ± 0.13 [0.08–0.53])	0.41 (0.40 ± 0.22 [0.11–0.74])	0.13 (0.22 ± 0.22 [0.01–0.83])	0.61
	WHASA-3D	11.7 (13.9 ± 12.6 [0.5–45.7])	1.9 (3.9 ± 4.1 [0–13.8])	0.66 (0.58 ± 0.22 [0.21–0.86])	0.42 (0.39 ± 0.10 [0.20–0.56])	0.60 (0.55 ± 0.17 [0.26–0.79])	0.23 (0.36 ± 0.27 [0–0.82])	0.95
Various Dementia	Reference	21.0 (24.8 ± 14.5 [0.3–68.0])	N/A	N/A	N/A	N/A	N/A	N/A
	WHASA	23.2 (24.9 ± 12.6 [0.7–58.3])	2.4 (3.2 ± 2.8 [0–11.1])	0.79 (0.77 ± 0.11 [0.41–0.92])	0.45 (0.44 ± 0.12 [0.22–0.70])	0.83 (0.80 ± 0.10 [0.43–0.90])	0.24 (0.25 ± 0.14 [0.06–0.76])	0.95
	WHASA-3D	23.3 (25.4 ± 14.4 [1.0–67.5])	2.2 (2.3 ± 1.8 [0–5.8])	0.79 (0.76 ± 0.14 [0.26–0.91])	0.46 (0.45 ± 0.12 [0.15–0.63])	0.82 (0.80 ± 0.12 [0.32 –0.95])	0.22 (0.26 ± 0.16 [0.12 – 0.83])	0.98

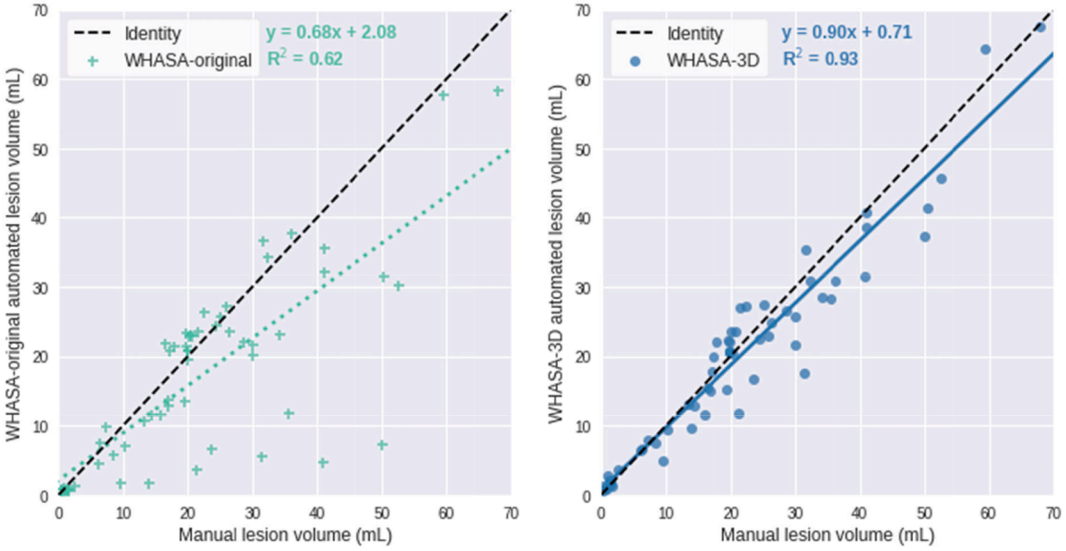

Fig. 6. Scatter plots of manual vs automated lesions volume quantification and linear regression for WHASA (on the left, green crosses) and WHASA-3D (on the right, blue dots) on both MS and Various Dementia database. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

classification of the image intensities, through a combination of topological and statistical atlases. An additional lesion class is added to the brain segmentation, using the same spatial prior as WM; lesions and WM are then separated by selecting the region with the higher membership value. Prior knowledge about areas where false positives commonly appear is used to determine penalty weights based on the distance to these areas. Performance evaluation of Lesion-TOADS was performed in MS patients and simulated images from the Brainweb MS (Shiee et al., 2010).

Supervised algorithms

LST-LPA

Lesions were segmented by the lesion prediction algorithm (Schmidt, 2017) as implemented in the LST toolbox version 2.0.15 (www.statistical-modelling.de/lst.html) for SPM. This algorithm consists of a binary classifier in the form of a logistic regression model trained on the data of 53 MS patients with severe lesion patterns. Data were obtained at the Department of Neurology, Technische Universität München, Munich, Germany. As covariates for this model, a similar lesion belief map was used, as for the lesion growth algorithm (Schmidt et al., 2012), as well as a spatial covariate that takes into account voxel specific changes in lesion probability. Parameters of this model fit are used to segment lesions in new images by providing an estimate for the lesion probability for each voxel. A pre-trained model is provided, however to date, no solution to re-train this model is yet available. Performance evaluation of LST-LPA was undertaken in MS patients (Schmidt et al., 2017).

nicMSlesions

nicMSlesions is a deep learning based method (Valverde et al., 2019), designed to automatically segment MS lesions from several brain MRI sequences, and validated in MS patients. Only T1-w and T2-FLAIR images are mandatory. The method is based on a cascade of two convolutional neural networks (CNN), the first being trained to be more sensitive to candidate lesion voxels, and the second being trained to reduce the number of false positives. A pre-trained model called

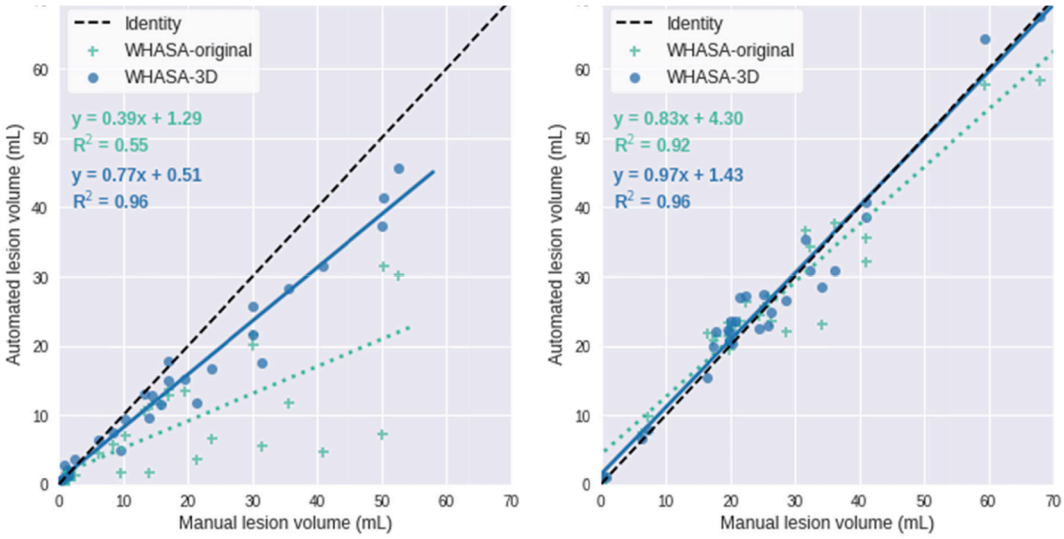

Fig. 7. Scatter plots of manual vs automated lesions volume quantification and linear regression for WHASA (green cross and dotted lines) and WHASA-3D (blue dots and straight lines) on the MS database (on the left) and the Various Dementia database (on the right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

"baseline_2ch" is therefore provided and the output is a lesion probability map. It was fully trained on two public MS lesion datasets (MSSEG MICCAI challenges 2008 and 2016) and evaluated with a private MS dataset and the ISBI2015 challenge dataset, using spatial and volumetric agreement (Valverde et al., 2019).

BIANCA

Brain Intensity AbNormality Classification Algorithm (BIANCA) (Griffanti et al., 2016) is a fully automated supervised method for WMH segmentation embedded in the FSL toolbox. The algorithm is based on the k-nearest neighbor framework (k-NN) and classifies the voxels based on their intensity and spatial features. BIANCA is flexible in terms of MRI modalities to use (either T1-w and T2-FLAIR or T2-FLAIR only) and offers several options (spatial weighting, local spatial intensity averaging, choice of the number and location of the training points). The output image is a probability map. The method has been validated on a cohort of neurodegenerative and vascular patients with manual segmentations. The training dataset consisted of a combination of those two datasets to train and optimize the model parameters, with a leave-one-out cross validation. Once optimized, BIANCA was used to segment WMH on the remaining subjects of the two cohorts, to be further evaluated with spatial and volumetric agreement.

2.2.3. Settings

Default settings

Default settings for the above methods are summarized in Table 3. There was no parameter to tune for WHASA-3D, lesionBrain and LesionTOADS. The remaining methods allow the user to tune some parameters to possibly improve the resulting segmentation. Because the output is a probability map for LST-LGA, LST-LPA and nicMSlesions, a default threshold has been set to 0.5, in order to obtain binary segmentations as recommended in the official LST website[1]. LST-LGA has an additional initial threshold, set to 0.3 by default. Please note that no pre-trained model has been provided with BIANCA, we thus trained BIANCA with the optimal configuration described in Griffanti et al. 2016 on the same 8 subjects training database as used for WHASA-3D development. For nicMSlesions, we used the pretrained model provided by the method called "baseline_2ch".

Optimized settings

Optimization was performed on LST-LGA, BIANCA and nicMSlesions, based on the highest average Dice score in comparison to expert manual segmentation on the optimization subset of 10 subjects from the MS database.

For LST-LPA and LST-LGA, the default probability threshold was kept at 0.5. The optimization of the initial threshold κ is detailed in Supplementary Table 1 showing an optimal threshold κ of 0.05. BIANCA has many possible configurations since the method offers the possibility to tune many parameters: number of lesion points, non-lesion points, location of non-lesion points, probability threshold… Every combination of these options is reported in Supplementary Table 2 and the optimal configuration reached was as follows: 2000 lesion points, 2000 non-lesion points and "any" location of the non-lesion training points, no spatial weighting and no 3D patch used. Finally, in order to obtain binary masks from probability maps, optimal thresholds have also been determined for nicMSlesions and BIANCA, with values of 0.6 and 0.75 respectively (Supplementary Table 3).

2.2.4. Evaluation

The performances of WHASA-3D and the other methods were evaluated by comparing segmentation results with reference segmentations at the voxel level, through volume and spatial agreement. In addition, evaluation was also considered at the WMH level, to assess the performance at the lesion level, as the counting task, which is a crucial component of MS diagnosis (Commowick et al., 2018).

Volume agreement

Total WMH volume gives an overall indication of the performance of the method and was evaluated using intra-class correlation coefficient (ICC) and absolute volume error (AVE, mL) between the automatic and reference segmentations. The ICC was derived from a two-way mixed model with absolute agreement definition. The relative volume difference is classically used for this type of evaluation, but would emphasize too much small differences for small lesion loads, and thus make it difficult to compare differences between small and large lesion loads. The absolute volume error was used instead, and computed as follows: $AVE = |V_R - V_A|$ with V_R the reference volume and V_A the automatic segmentation volume. The result is thus given in mL, the optimal value being 0 mL.

Spatial agreement

Total WMH volume gives no indication about spatial agreement. The automatic segmentation could have the same volume as the reference segmentation without any common voxel. The spatial agreement between reference and automatic segmentations is evaluated based on the

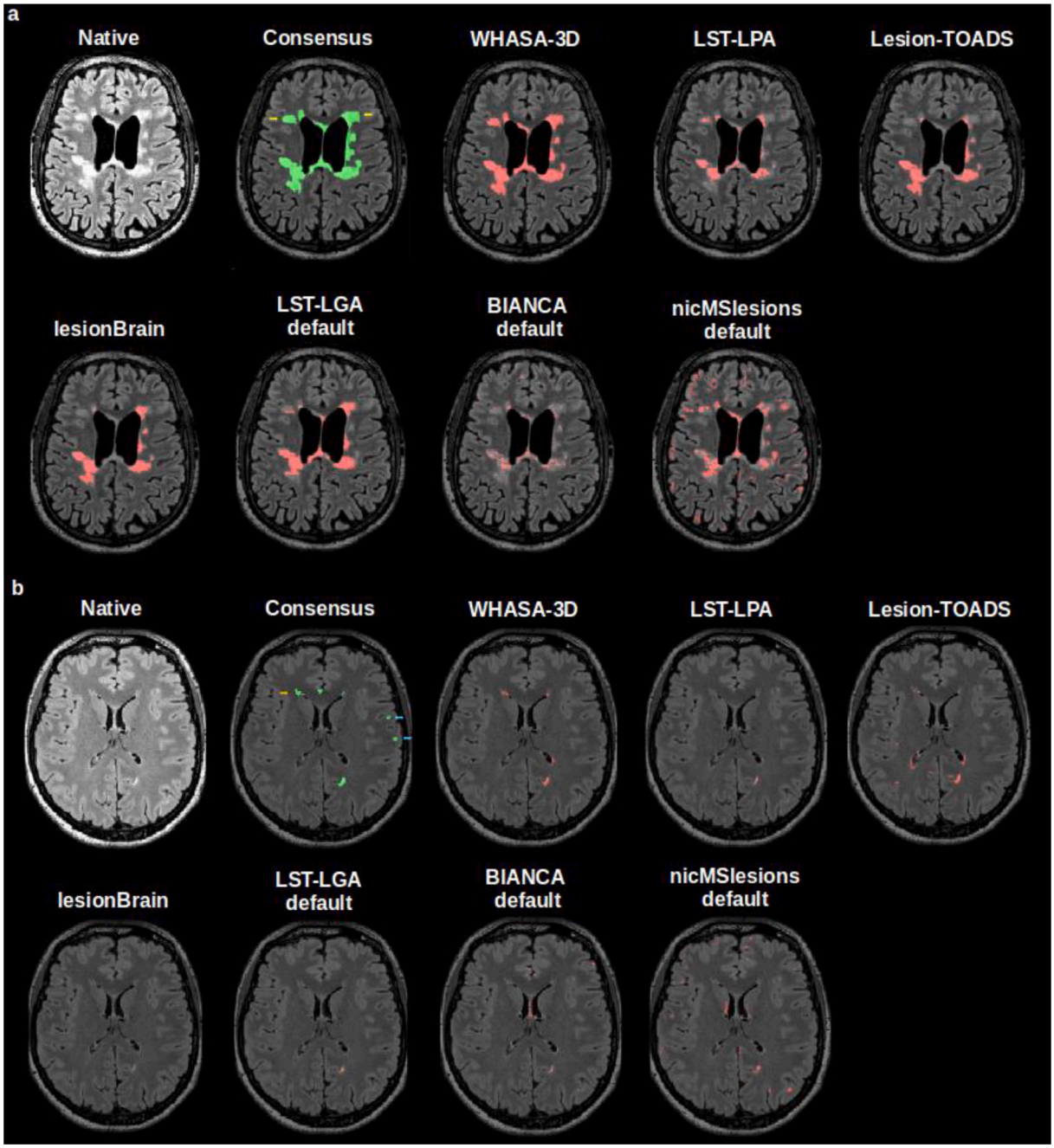

Fig. 8. S database 3D T2-FLAIR images and superposed segmentations from the consensus reference segmentation and all methods with their default settings on subjects with the (a) highest and (b) lowest Dice (0.86 and 0.21 resp.) for WHASA-3D in comparison to the reference segmentation. Yellow arrows shows WMH that are correctly detected by WHASA-3D but either missed or underestimated by other methods, and blue arrows shows WMH missed by all methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

volume of true positives ($V_{TP} = V_{AR}$), false positives ($V_{FP} = V_A - V_{AR}$) and false negative ($V_{FN} = V_R - V_{AR}$), computed at the voxel level. Three indices ranging between 0 and 1 are then used at the voxel level: the Dice similarity index (Dice) (Dice, 1945) (perfect agreement: 1), the false positive ratio (FPR) (perfect agreement: 0) and true positive ratio (TPR) (perfect agreement: 1) defined as follows:

$$\text{Dice} = \frac{2*V_{TP}}{V_{FP} + V_{FN} + 2*V_{TP}}$$

$$\text{FPR} = \frac{V_{FP}}{V_A}$$

$$\text{TPR} = \frac{V_{TP}}{V_R}$$

WMH agreement

Evaluation of WMH detection relies on determining how many WMH have been correctly or incorrectly detected. The WMH agreement relies on identifying individual WMH in the reference and automatic segmentation, based on the number of WMH in the reference (L_R), the number of WMH in the automatic segmentation (L_A), the number of WMH in the reference correctly detected by the segmentation ($L_{TP(R)}$) and the number of WMH in the segmentation for which there is a WMH in the reference ($L_{TP(A)}$). From the number of WMH in each segmentation

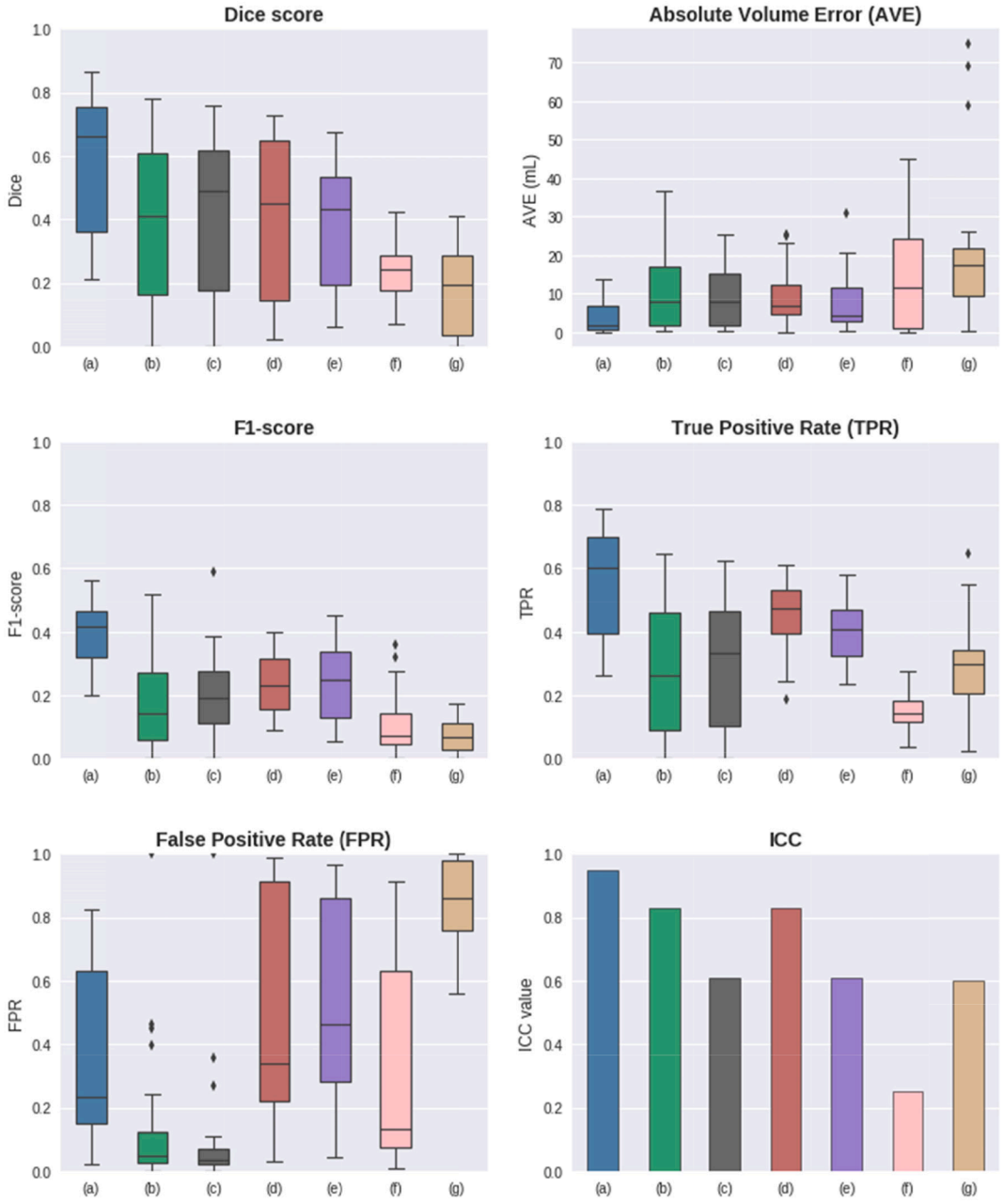

Fig. 9. Box-and-whisker plots (median, interquartile range and extrema) showing Dice, F1-score, Absolute Volume Error, True Positive Rate and False Positive Rate in comparison to the manual lesion segmentation on the MS database for WHASA-3D (a), LST-LGA default (b), LST-LPA (c), lesionBrain (d), Lesion-TOADS (e), BIANCA default (f) and nicMSlesions default (g).

(reference and automatic) and the numbers computed above ($L_{TP(R)}$ and $L_{TP(A)}$), the F1-score is computed, which accounts for the sensitivity (i.e the proportion of detected WMH in the reference) and the positive predictive value (i.e the proportion of true positive WMH inside the automatic segmentation) (Commowick et al., 2018). F1-score ranges from 0 to 1 and gives a global idea of the detection performance (perfect detection: 1).

$$F1-score = \frac{2*\left(\frac{L_{TP(R)}}{L_R}\right)*\left(\frac{L_{TP(A)}}{L_A}\right)}{\left(\frac{L_{TP(R)}}{L_R}\right)+\left(\frac{L_{TP(A)}}{L_A}\right)}$$

Statistics

Statistical analysis was performed using the Scipy version 1.2.1 Python library. For the comparison of WHASA-3D with WHASA, a non-parametric Wilcoxon test was used for the volumetric (absolute volume error) and spatial agreement (dice score) with respect to the manual segmentation, and results were considered statistically significant upon p-value < 0.05. Regarding the comparison of WHASA-3D with multiple methods, a non-parametric Friedman test of differences among repeated measures and post-hoc analyses with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.05/n$ with n the number of comparisons made for the volumetric and spatial agreement.

3. Results

WHASA-3D was first qualitatively and quantitatively evaluated and compared with WHASA on the MS and Various dementia databases. Its performance was then compared with the other methods mentioned

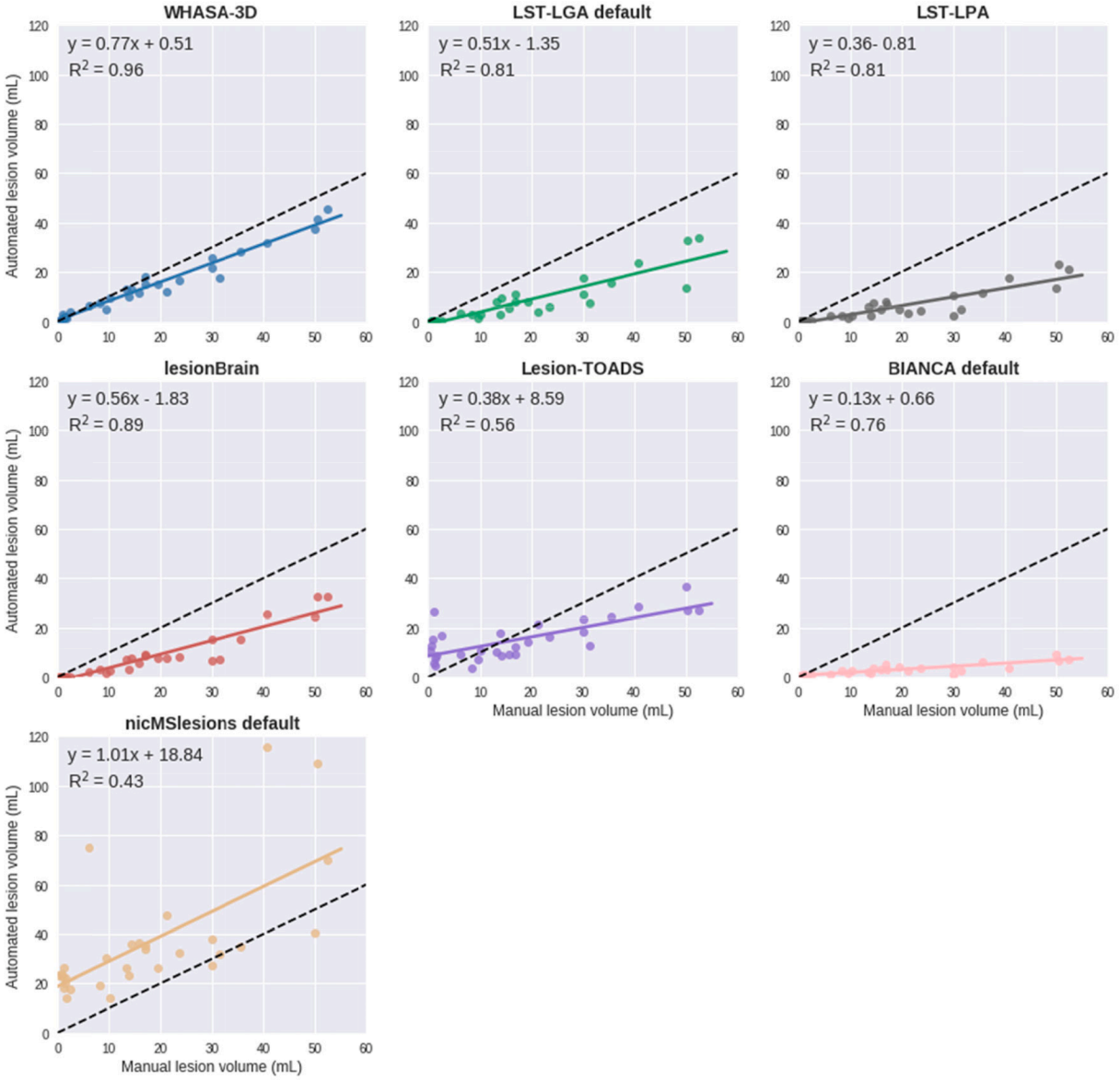

Fig. 10. Scatter plots of manual vs automated lesion volume and linear regression for methods with default parameters on the MS database. Identity is represented as a dotted line.

above only on the MS database.

3.1. Comparison of WHASA-3D with WHASA

WHASA-3D showed consistent behaviour on the MS database, while WHASA exhibited insufficient WMH segmentation as illustrated in Fig. 5. Comparison results of quantitative metrics for WHASA and WHASA-3D with reference segmentation in MS and Various dementia databases are presented in Table 4**.** Most metrics showed an improvement for WHASA-3D compared to WHASA: the average Dice score has increased from 0.63 to 0.67, the F1-score from 0.37 to 0.42 and the absolute volume error has decreased from 6.2 to 3.1 mL; the ICC value has increased from 0.78 to 0.96, and TPR has also increased from 0.60 to 0.67 demonstrating a better correlation of WHASA-3D with the experts' reference volumes compared to WHASA. TPR, resp. FPR, has increased from 0.60 to 0.67, resp. from 0.23 to 0.31, with WHASA-3D, which meant a better detection of WMH but also a higher risk of detecting false positives. Regression analysis between manual and automated lesion volume, as illustrated in Fig. 6, showed increased correlation (R^2 from 0.62 to 0.93) and a better regression slope (from 0.68 to 0.90) using WHASA-3D on the combination of MS and Various dementia databases.

The improvement was highly prominent on the MS database ($R^2_{WHASA} = 0.55$ and $R^2_{WHASA\text{-}3D} = 0.96$) and less visible on the database including various dementia types, but WHASA-3D also performs better ($R^2_{WHASA} = 0.92$ and $R^2_{WHASA\text{-}3D} = 0.96$) as shown in Fig. 7.

Statisical analysis showed no significant difference on Dice score between WHASA and WHASA-3D, either in MS or the Various Dementia database despite a global reduction in average dice scores between methods. However, we report a significant difference for the absolute volume error for the whole database and the MS database (MS and Various dementia, $p = 3.93E\text{-}5$; MS database, $p = 1.97E\text{-}5$), but not for the Various dementia database.

3.2. Comparison of WHASA-3D with other lesion segmentation methods

In this section, we will present a comparison of WHASA-3D with other methods freely available in the literature and described in the methods section, based on results obtained on the MS database as it appeared to be the most challenging for WHASA-3D. To ensure fair comparison, methods will be run with their default settings and with

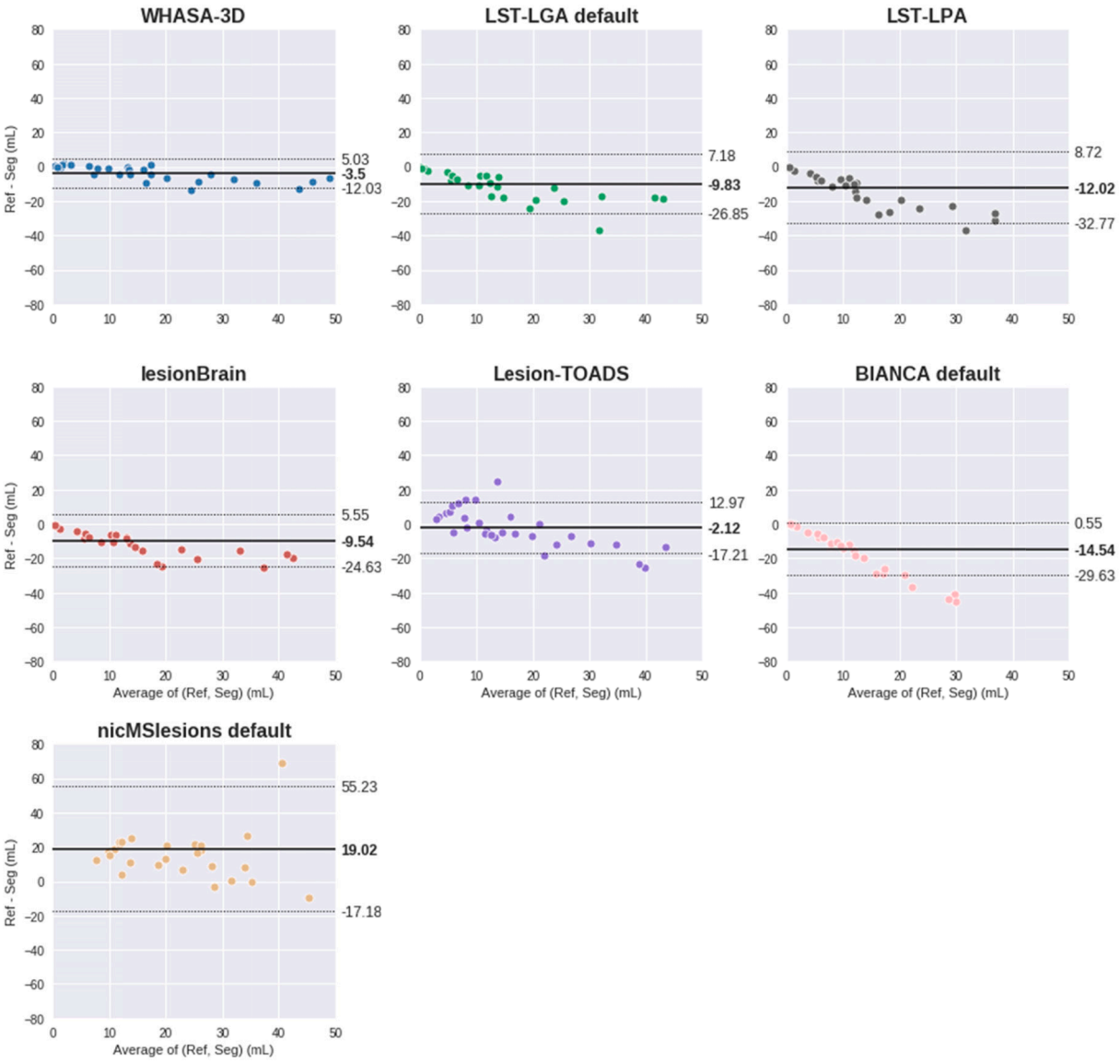

Fig. 11. Bland-Altman plots for methods with default parameters on the MS database. Mean bias (straight line) and 95% limits of agreements (dotted lines) are also displayed for each method.

parameter optimization and/or model re-training whenever possible.

3.2.1. *Default settings*

Results of WMH segmentations with all methods at default settings are displayed on Fig. 8 on two subjects with the highest and lowest Dice obtained by WHASA-3D. WHASA-3D, LST-LGA, LST-LPA, Lesion-TOADS and lesionBrain showed consistent segmentation results in comparison to the consensus reference segmentation, whereas BIANCA default and nicMSlesions default revealed large under or over-segmentation. The box-and-whisker plots for each volume and spatial agreement metric are displayed in Fig. 9. WHASA-3D showed the highest volume agreement ($ICC_{WHASA\text{-}3D} = 0.95$) as well as the highest average Dice, F1-score and TPR on this database. All methods except nicMSlesions default and Lesion-TOADS showed a lower average FPR than WHASA-3D, but all had a lower average TPR. Regarding LST algorithms, LST-LPA performed better on this dataset than LST-LGA with default parameters, except for volume agreement ($ICC_{LST\text{-}LPA} = 0.61$ and $ICC_{LST\text{-}LGA\ default} = 0.81$). The two supervised methods, BIANCA and nicMSlesions, when used with their default settings, showed poor performances in terms of Dice, F1-score and AVE. All measures are reported in the Supplementary Table S4.

Volume consistency with manual segmentation for each automated method (Fig. 10) showed that volumes obtained with WHASA-3D are the most consistent with manual segmentation. Bland-and-Altman plots (Fig. 11), show an underestimation of lesion volume for all methods except nicMSlesions, and a narrower interval between limits of agreement for WHASA-3D compared to the other methods on this dataset.

There was a statistically significant difference in spatial (Friedman test, p-value = 3E-22) and volumetric agreement (Friedman test, p-value = 1E-11) between WHASA-3D and the other methods in their default settings. Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.0071$. After Bonferroni correction, WHASA-3D outperfom all the method considered, with a significant difference between WHASA-3D and the other methods for the spatial and volumetric agreement. All p-values for volumetric and spatial agreement are reported in the Supplementary Table S5.

3.2.2. *Optimized settings*

With the optimized settings, we observed largely improved

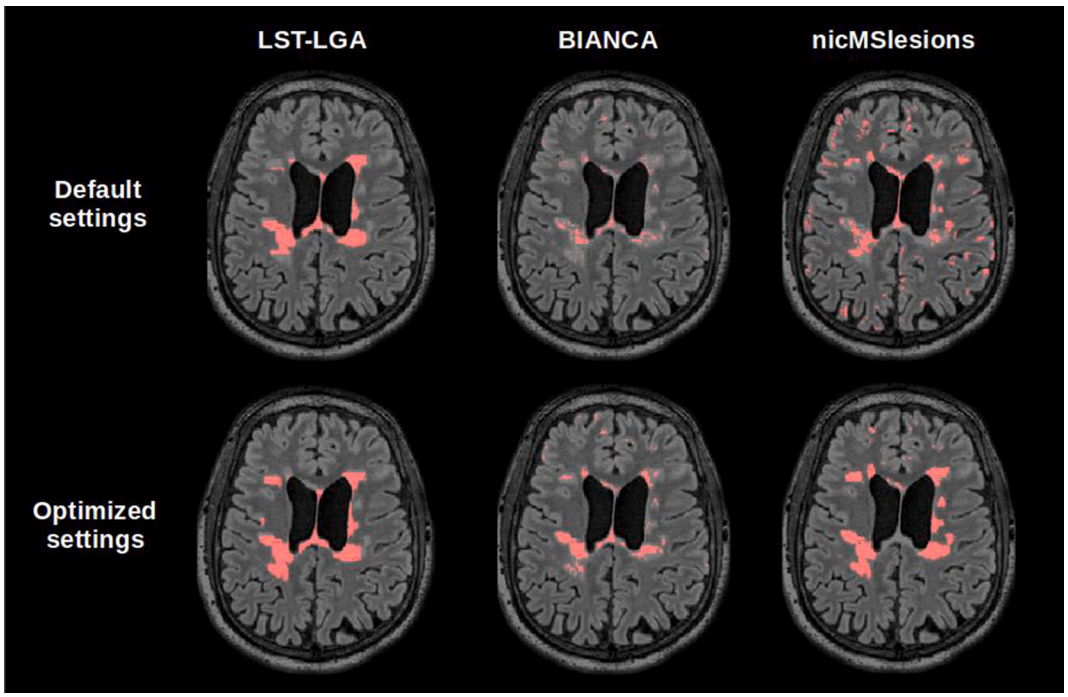

Fig. 12. 3D T2-FLAIR image and superposed segmentations from LST-LGA, BIANCA and nicMSlesions methods with default (first row) and optimized settings (second row).

segmentation performances for LST-LGA, BIANCA and nicMSlesions compared to the large under or over-segmentation with their default settings, as displayed in Fig. 12**.** Results of the comparisons between WHASA-3D and the three methods for which optimisation could be undertaken are reported in Fig. 13 for the validation subset of the MS database. Performance in terms of overlap and volume agreement after optimization are both revealed by the Average Dice (LST-LGA default/optimized = 0.41/0.51; BIANCA default/optimized = 0.22/0.39; nicMSlesions default/optimized = 0.17/0.63) and the ICC (LST-LGA default/optimized = 0.86/0.95; BIANCA default/optimized = 0.23/0.71; nicMSlesions default/optimized = 0.61/0.88). The volume consistency between automated and manual volumes is also displayed with scatter plots and Bland-Altman plots (Figs. 14 and 15**)**. The highest volume agreement is obtained by WHASA 3D and LST-LGA (ICC 0.97 and 0.95), and the best spatial agreement by WHASA 3D and nicMSlesions optimized (Average Dice 0.58 and 0.63, TPR 0.56 and 0.59). All measures are reported in the Supplementary Table S6**.**

There was a statistically significant difference in spatial (Friedman test, p-value = 3E-07) and volumetric agreement (Friedman test, p-value = 8E-05) between WHASA-3D and the other methods in their optimized settings. Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.0125$. For spatial agreement, all comparison between WHASA-3D and the other methods were significantly different at the exception of nicMSlesions. p-values for volumetric and spatial agreement are reported in the Supplementary Table S7.

3.3. Processing time

In order to consider using methods in clinical routine, results have to be delivered in a short time. All methods were run on a computer with an Intel CPU 3.50 GHz (8 cores) processor and 16go RAM. Table 5 shows the computational time for each method. It greatly varies depending on the underlying framework. Unlike unsupervised methods, supervised methods that needs to be trained before-hand (BIANCA, nicMSlesions) require training time, that ranges from few minutes for BIANCA to up to 15 h for the deep-learning-based method nicMSlesions. Note that training has to be performed only once for a given type of sequence. Among the unsupervised methods, LST-LGA remains the fastest.

4. Discussion

We have presented here WHASA-3D, an extension of WHASA (Samaille et al., 2012) dedicated to the automatic segmentation of age-related WMH and MS lesions from 3D T2-FLAIR images in a multicenter and multi-disease framework. Validation of WHASA-3D was undertaken on a database with 60 subjects, built from four different cohorts, with subjects acquired on seven MRI scanners, displaying a wide range of lesion load and including 30 patients with age-related WMH (elderly subjects and various dementia) and 30 patients with MS lesions. WHASA-3D outperformed WHASA when evaluated in comparison with consensus manual segmentation masks in terms of overlap and volume agreement. We also compared WHASA-3D with three unsupervised methods and three supervised methods with default and optimized settings when recommended. When default “pre-trained” parameters were used, WHASA-3D showed the best volume and spatial agreement with the highest ICC and Dice, followed by LST-LGA, lesionBrain and Lesion-TOADS. After retraining the methods that could be retrained on a separate subset, nicMSlesions performances improved dramatically (average Dice and F1-score raised from 0.17 to 0.63 and from 0.06 to 0.56) showing the best performance. However, nicMSlesions outperformed WHASA 3D exclusively in the Dice and F1-score, while WHASA 3D still showed better performances for ICC.

The GM/WM contrast greatly varies between subjects and between MRI protocols (Gabr et al., 2017), and this variability has to be taken into account when developing segmentation tools, in order to be able to detect lesions with all types of contrasts. In addition, a large variability of WMH lesion characteristics can also be observed: the most common WMHs are age-related WMH and MS lesions; MS lesions show different shape, contrast and distribution compared to age-related WMH (Caligiuri et al., 2015; Schmidt et al., 2012). We therefore developed the algorithm using a training database embedding eight subjects selected to be as representative as possible of the variations of WMHs visibility, by ensuring variability in the following criteria: scanners, MRI protocols,

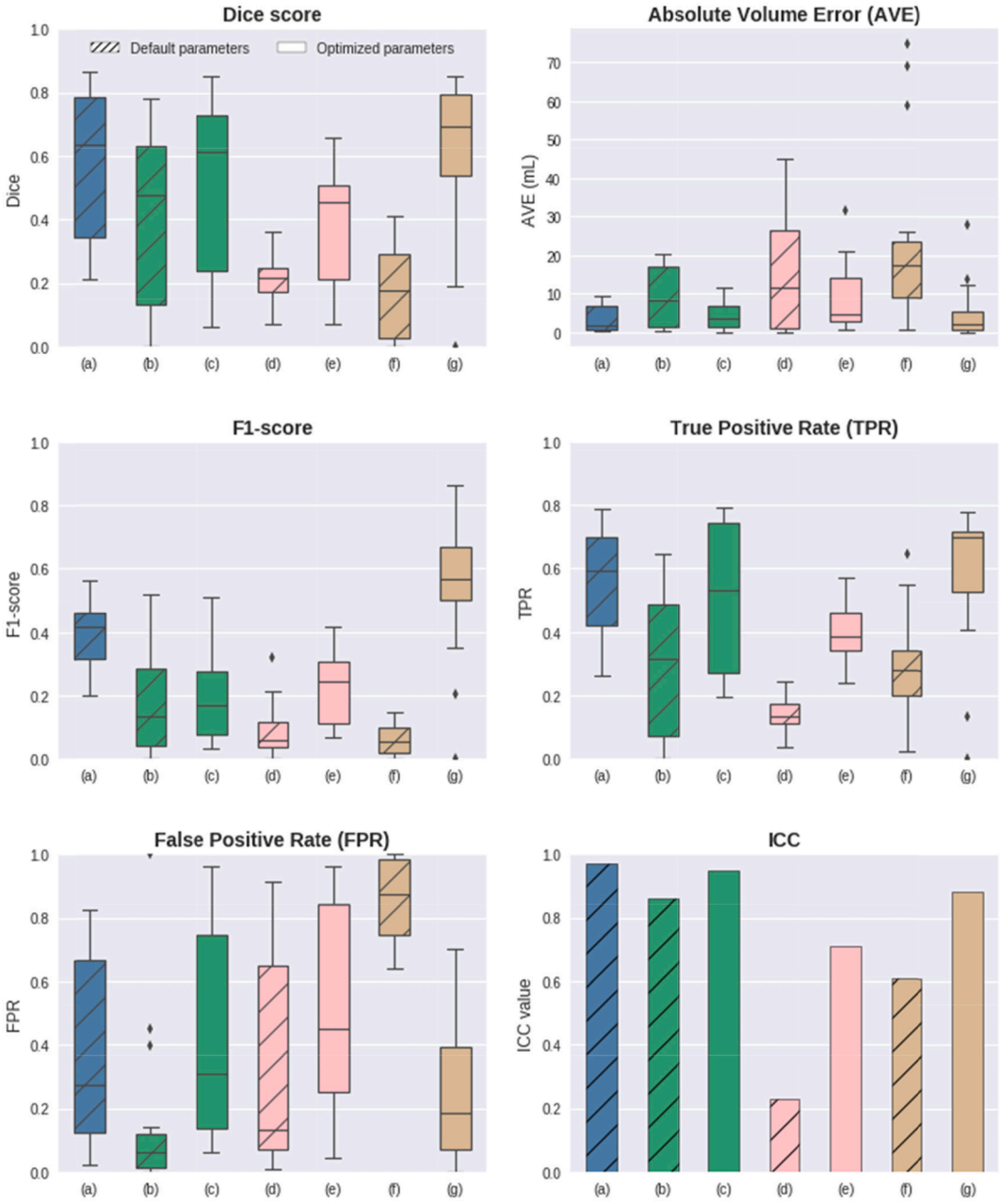

Fig. 13. Box-and-whiskers plot (median, interquartile range and extrema) showing Dice, F1-score, Absolute Volume Error, True Positive Rate and False Positive Rate in comparison to the manual lesion segmentation on the validation subset of the MS database for WHASA-3D (a), LST-LGA default (b), LST-LGA optimized (c), BIANCA default (d), BIANCA optimized (e), nicMSlesions default (f) and nicMSlesions optimized (g).

diagnosis (AD, FTD, HC, MS, or patients with cognitive disorders), and lesion load. This allowed settings that can perform consistently on a wide range of acquisition types, as was confirmed with the evaluation results. An important validation of WHASA-3D was indeed conducted on a database built from four different cohorts, displaying a wide range of WMH and including the same different diagnosis as in the training database. Besides, to highlight the adaptation of WHASA to MS lesions specificities, we divided this validation dataset equally into "Various dementia" and "MS" databases with 30 subjects each.

WHASA-3D was thus designed not only to ensure proper 3D segmentation but also to be able to segment datasets with various contrast and lesion characteristics. On the "Various dementia" database, WHASA-3D showed an average Dice score of 0.76, compared to 0.77 for WHASA-original, which indicates good performance, compared to the originally proposed version validated on patients with this type of lesion (Samaille et al., 2012). Regarding the "MS" database, intraclass correlation (ICC) increased greatly between WHASA and WHASA-3D, pointing towards a greater correlation with the consensus reference, confirmed with the decreased volume error and the better linear regression (Fig. 7) between automated and manual segmentation volume. On this database, the compromise between sensitivity and specificity measures, indicated through TPR and FPR, is shifted towards higher TPR rather than lower FPR, from WHASA to WHASA-3D. It shows, in fact, a better ability to detect every lesion, in line with the fact that it is a crucial component of MS diagnosis according to McDonald criteria (Thompson et al., 2018). Part of the increased false positive regions is due to partial volume effect around MS lesions; lesion edges are usually not clearly defined and even experts are often unsure of how to delineate border, most of all when dirty WM is involved (Lesjak et al., 2017; Seewann et al., 2009). This dirty aspect can also influence the

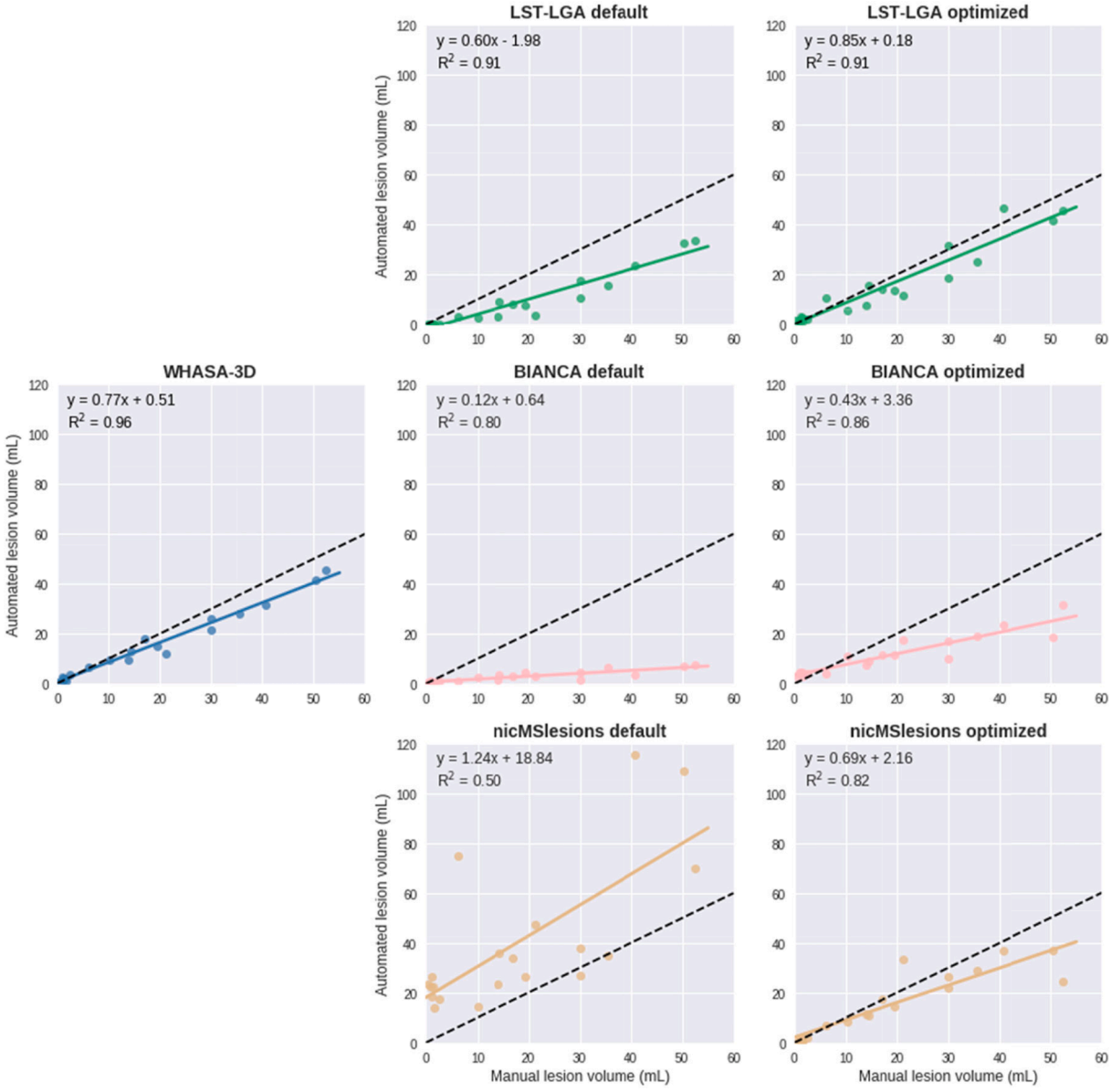

Fig. 14. Scatter plots of the manual vs automated lesions volume quantification on the validation subset of the MS database obtained from WHASA-3D, LST-LGA, lesionBrain, Lesion-TOADS, BIANCA and nicMSlesions optimized and linear regression for each method. Identity is represented as a dotted line.

estimation of the generalized threshold implemented in WHASA-3D: the method tends to segment the hyperintense dirty WM around lesions as well, even though they appear less intense than focal lesions, thus resulting in a disagreement with the consensus segmentation, though these areas are in fact uncertain. It may thus be of interest to distinguish both lesion types, in order to better characterise lesions (Dadar et al., 2021; Seewann et al., 2009). False positives regions can also be found in tight cortical folding patterns where GM can appear hyperintense in 3D T2-FLAIR images. Most false positives areas of this type are excluded with the exclusion mask, but a few still remain due to the high contrast between GM and WM. Such an exclusion mask may result in erroneously removing cortical WMH that can be found in those areas. However, removing cortical false positives is a complex task, as it requires distinguishing them from cortical WMH, which are very relevant in the diagnosis of MS (Thompson et al., 2018). This is also the case for infratentorial lesions, as tissue segmentations are less precise in infratentorial area, and infratentorial lesions can also be falsely removed by the exclusion mask. Additional work is planned to improve the segmentation of cortical WMH, with the help of specific sequences like double inversion recovery sequences (DIR), which better reveal cortical WMH compared to the use of FLAIR sequences.

In order to have an estimate of a consistent aim for the best performance results were assessed by comparing them to available state-of-the-art methods. As stated in previous work (Caligiuri et al., 2015), several automated segmentation methods have been developed for MS lesions detection, similarly to methods focused on WMH segmentation, but the techniques trained in MS patients perform only moderately well when applied to elderly patients. On the other hand, automated segmentation methods developed for WMH segmentation might perform poorly when applied to MS patients. This is partly due to the aspect of white matter hyperintensities. In MS, lesions are usually focal with clear edges while WMH in the elderly or dementia population have a more diffuse pattern. Automatic methods have to take into account the type of WMH to process, because they may share the same characteristics (high intensities compared to the normal appearing white matter in FLAIR sequences) but have very distinct features (edges, localization). We therefore focused the comparison study on the MS database, since such data had not been used previously for the WHASA validation (Samaille

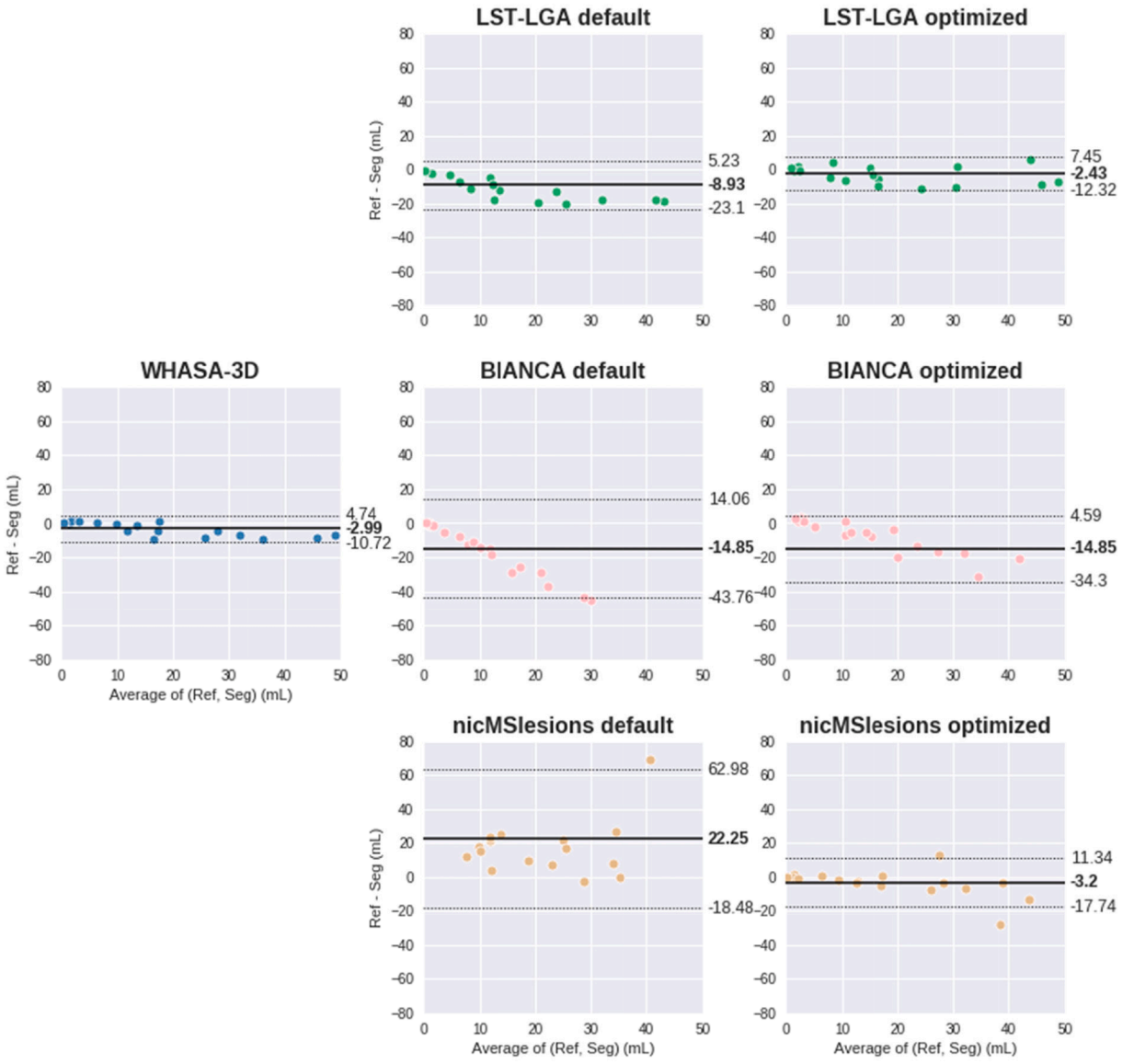

Fig. 15. Bland-Altman plots on the validation subset of the MS database for each optimized method and their default parameters. Mean bias (straight line) and 95% limits of agreements (dotted lines) are also displayed for each method.

Table 5
Computational time of the different methods per subject.

Type	Methods	Training time (approx.)	Preprocessing time (approx.)	Segmentation time (approx.)
Unsupervised	WHASA-3D	None	10 min	10 min
	LST-LGA	None	3 min	2 min
	LesionBrain	None	N/A	Results available after 30 min
	Lesion-TOADS	None	N/A	45 min
Supervised	LST-LPA	No retraining possible	3 min	2 min
	BIANCA	5 min	1–2 h (FSL)	2 min
	nicMSlesions	15 h	15 min	5 min

et al., 2012), and to guarantee that our method worked properly for MS patients. The three unsupervised methods (LST-LGA, lesionBrain, Lesion-TOADS) and the three supervised methods (LST-LPA, BIANCA, and the deep-learning based method nicMSlesions) were compared with default settings in order to give an insight of the feasibility of using the same set of parameters for all datasets (new or returning patient), for a convenient use in clinical routine (Commowick et al., 2018). While unsupervised methods are designed to adapt well to new datasets, most supervised methods are made available with a pre-trained model obtained on a specific dataset, in addition to default parameters settings. We therefore used pretrained models provided with nicMSlesions and LST-LPA methods (Schmidt et al., 2012; Valverde et al., 2019); and as for BIANCA, no pretrained model was available, and it was thus trained on the same 8 subjects training database as used for WHASA-3D development, but the optimal set of parameters reported in (Griffanti et al., 2016) was used. As could be expected, unsupervised methods mostly outperformed supervised methods used with their default configuration regarding segmentation accuracy. Among all methods, WHASA-3D shows the best volume and spatial agreement with the highest ICC and Dice, followed by LST-LGA, lesionBrain and Lesion-TOADS. These methods had been specifically designed and validated for MS subjects (Schmidt et al., 2012; Shiee et al., 2010; Coupé et al., 2018). Regarding supervised methods, although better results were reported in recent

WMH or MS lesion segmentation challenges after retraining on specific training datasets (Commowick et al., 2018; Kuijf et al., 2019), these methods may show generalisation issues when faced with subjects from new centers or with unseen pathological characteristics. This is true for both LST algorithms, where LST-LGA (unsupervised) performed better than LST-LPA (supervised), with an ICC of 0.83 compared to 0.61. It is worth mentioning that the retraining of nicMSlesions using only one manual delineated subject as input data is possible, but only if input lesion volumes in the given training data are sufficient enough to retrain the last layers of the network (Weeda et al., 2019).

We also compared the performance of three methods (LST-LGA, BIANCA and nicMSlesions) after dedicated retraining or parameter optimisation, in order to ensure fair comparison and assess the impact of optimization and re-training on the segmentation performance. In fact, even though it allows to optimise the final segmentation on a given type of data, it may be difficult to apply on larger multicentre studies. Please note that no optimization step was done for WHASA-3D for the specific dataset, as variability was already taken into account in the automatic contrast-adapted intensity parameters.After optimisation and retraining for the three methods, results were improved both in terms of overlap and volume agreement, nicMSlesions showing the larger improvement (average Dice and F1-score raised from 0.17 to 0.63 and from 0.06 to 0.56), thus outperforming all other methods. Results were also improved for LST-LGA even though the optimized threshold k for LST-LGA, reported in the Supplementary Table S1, corresponded to the lower limit of the search range, suggesting a sub-optimal behaviour for this dataset. Deep-learning based methods have been proven very efficient in segmentation tasks (García-Lorenzo et al., 2013) but may require retraining to adapt to new datasets, that is likely to involve high computational power to run the training step on a specific hardware GPU, while most algorithms can run on regular computer CPU (Kuijf et al., 2019). Here, the retraining of nicMSlesions took 15 h to re-train the full 11-layer cascaded CNN on the optimization subset.

While the comparison study presented in this paper allows to evaluate of the performance on subjects with wide range of lesion load and different clinical stages, the MS database used contained data from one center only, acquired on a single MRI system (Siemens Magnetom Trio) (Lesjak et al., 2017). Ensuring a consistent performance on all data type would require a multi-centered dataset, representative of the acquisition variability with different MRI acquisition protocols and MRI systems. To overcome this issue, an initiative has been proposed to standardize MRI sequences for MS (Arevalo et al., 2019; Brisset et al., 2020), but no open-access database of MRI images is yet available (Marek et al., 2011; Wyman et al., 2013). In addition, although accuracy and robustness across different scanners and acquisitions is the most widely performed type of validation, clinicians are also very concerned with reproducibility of measures over time and between MRI systems (García-Lorenzo et al., 2013). An automated method is considered reproducible and consistent if it shows low volume difference and high spatial agreement between the scan and the rescan in dedicated experiments (Fartaria et al., 2019; Jain et al., 2015; Weeda et al., 2019). This was not yet evaluated for WHASA-3D as no such dataset was available but will be undertaken in the future to ensure that differences in segmentation result from pathological changes rather than from changes related to acquisition and segmentation.

Currently, T1-weighted and T2-FLAIR images are mandatory as inputs, in order to automatically segment WMH for the WHASA methods. In fact, it needs a reliable estimate of the grey matter/white matter interface, that is obtained from the tissue segmentation from SPM12. An additional T1 is therefore necessary to generate those segmentations, even though good quality 3D FLAIR images may be sufficient to derive this segmentation. Future work is planned to create a new version of WHASA, without the need of T1-weighted images.

5. Conclusion

The proposed automated white matter lesion segmentation algorithm WHASA-3D has proven to be a reliable extension for MS patients of the original method WHASA. WHASA-3D automatically segments age-related WMH and MS lesions from 3D T2-FLAIR and T1 images in multi-centered datasets with a processing time of twenty minutes per subject. Evaluation was performed on 60 patients, acquired on different MRI scanners displaying various diagnose and a wide range of lesion load, by computing volume and spatial agreement measures as compared to expert manual segmentations. For MS lesions, performances have been further compared with six other methods (three unsupervised and three supervised), with their default settings to recreate the use in clinical routine, and after optimization when available, to illustrate the maximum potential of methods. Better results have been observed in the default settings for WHASA-3D over all methods, and the method still shows among the best volumetric and spatial agreement after optimization and retraining of methods that could be optimized. This suggests that WHASA-3D is a fast, reliable and easy-to-use method with no optimisation or retraining needed for the automated segmentation of MS lesions and age-related WMH. Nevertheless, further validation on larger datasets and reproducibility studies are needed to fully validate our method.

Disclosures

Marie Chupin, Didier Dormont, Christine Delmaire, Sébastian Ströer and Emmanuelle Gourieux have nothing to disclose. Philippe Tran, Clarisse Longo dos Santos, and Enrica Cavedo are employees of Qynapse. Urielle Thoprakarn and Jean-Baptiste Martini are not currently employees of Qynapse, but this work has been performed during their previous position at Qynapse. François Cotton and Pierre Krolak-Salmon have received personal compensation for consulting services from Qynapse. Damien Heidelberg and Nadya Pyatigorskaya have received personal fees for their time spent on this project.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data collection and sharing for this project was funded by the Frontotemporal Lobar Degeneration Neuroimaging Initiative (National Institutes of Health Grant R01 AG032306). The study is coordinated through the University of California, San Francisco, Memory and Aging Center. FTLDNI data are disseminated by the Laboratory for Neuro

Imaging at the University of Southern California.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2022.102940.

References

Akhondi-Asl, A., Hoyte, L., Lockhart, M.E., Warfield, S.K., 2014. A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights. IEEE Trans Med Imaging. 33 (10), 1997–2009.

Alber, J., Alladi, S., Bae, H.-J., Barton, D.A., Beckett, L.A., Bell, J.M., Berman, S.E., Biessels, G.J., Black, S.E., Bos, I., Bowman, G.L., Brai, E., Brickman, A.M., Callahan, B.L., Corriveau, R.A., Fossati, S., Gottesman, R.F., Gustafson, D.R., Hachinski, V., Hayden, K.M., Helman, A.M., Hughes, T.M., Isaacs, J.D., Jefferson, A. L., Johnson, S.C., Kapasi, A., Kern, S., Kwon, J.C., Kukolja, J., Lee, A., Lockhart, S.N., Murray, A., Osborn, K.E., Power, M.C., Price, B.R., Rhodius-Meester, H.F.M., Rondeau, J.A., Rosen, A.C., Rosene, D.L., Schneider, J.A., Scholtzova, H., Shaaban, C.E., Silva, N.C.B.S., Snyder, H.M., Swardfager, W., Troen, A.M., Veluw, S. J., Vemuri, P., Wallin, A., Wellington, C., Wilcock, D.M., Xie, S.X., Hainsworth, A.H., 2019. White matter hyperintensities in vascular contributions to cognitive impairment and dementia (VCID): Knowledge gaps and opportunities. Alzheimers Dement (N Y) 5 (1), 107–117. https://doi.org/10.1016/j.trci.2019.02.001.

Arevalo, O., Riascos, R., Rabiei, P., Kamali, A., Nelson, F., 2019. Standardizing Magnetic Resonance Imaging Protocols, Requisitions, and Reports in Multiple Sclerosis: An Update for Radiologist Based on 2017 Magnetic Resonance Imaging in Multiple Sclerosis and 2018 Consortium of Multiple Sclerosis Centers Consensus Guidelines. J Comput Assist Tomogr. 43 (1), 1–12.

Ashburner, J., Friston, K.J., 2005. Unified Segmentation. Neuroimage 26 (3), 839–851.

Brisset, J.-C., Kremer, S., Hannoun, S., Bonneville, F., Durand-Dubief, F., Tourdias, T., Barillot, C., Guttmann, C., Vukusic, S., Dousset, V., Cotton, F., Ameli, R., Anxionnat, R., Audoin, B., Attye, A., Bannier, E., Barillot, C., Ben Salem, D., Boncoeur-Martel, M.-P., Bonhomme, G., Bonneville, F., Boutet, C., Brisset, J.C., Cervenanski, F., Claise, B., Commowick, O., Constans, J.-M., Cotton, F., Dardel, P., Desal, H., Dousset, V., Durand-Dubief, F., Ferre, J.-C., Gaultier, A., Gerardin, E., Glattard, T., Grand, S., Grenier, T., Guillevin, R., Guttmann, C., Krainik, A., Kremer, S., Lion, S., Champfleur, N.M.D., Mondot, L., Outteryck, O., Pyatigorskaya, N., Pruvo, J.-P., Rabaste, S., Ranjeva, J.-P., Roch, J.-A., Sadik, J.-C., Sappey-Marinier, D., Savatovsky, J., Stankoff, B., Tanguy, J.-Y., Tourbah, A., Tourdias, T., Brochet, B., Casey, R., Cotton, F., De Sèze, J., Douek, P., Guillemin, F., Laplaud, D., Lebrun-Frenay, C., Mansuy, L., Moreau, T., Olaiz, J., Pelletier, J., Rigaud-Bully, C., Stankoff, B., Vukusic, S., Debouverie, M., Edan, G., Ciron, J., Lubetzki, C., Vermersch, P., Labauge, P., Defer, G., Berger, E., Clavelou, P., Gout, O., Thouvenot, E., Heinzlef, O., Al-Khedr, A., Bourre, B., Casez, O., Cabre, P., Montcuquet, A., Créange, A., Camdessanché, J.-P., Bakchine, S., Maurousset, A., Patry, I., De Broucker, T., Pottier, C., Neau, J.-P., Labeyrie, C., Nifle, C., 2020. New OFSEP recommendations for MRI assessment of multiple sclerosis patients: special consideration for gadolinium deposition and frequent acquisitions. Journal of Neuroradiology 47 (4), 250–258.

Caligiuri, M.E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., Cherubini, A., 2015. Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review. Neuroinformatics 13 (3), 261–276. https://doi.org/10.1007/s12021-015-9260-y.

Cotton, F., Kremer, S., Hannoun, S., Vukusic, S., Dousset, V., 2015. OFSEP, a nationwide cohort of people with multiple sclerosis: Consensus minimal MRI protocol. Journal of Neuroradiology 42 (3), 133–140.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Améli, R., Ferré, J.C., Kerbrat, A., Tourdias, 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.J., Malpica, N., Guttmann, C., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S.K., Cotton, F., Barillot, C., 2018 Sep 12. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Sci Rep. 8 (1), 13650. https://doi.org/10.1038/s41598-018-31911-7. PMID: 30209345; PMCID: PMC6135867.

Dadar, M., Narayanan, S., Arnold, D.L., Collins, D.L., Maranzano, J., 2021. Conversion of diffusely abnormal white matter to focal lesions is linked to progression in secondary progressive multiple sclerosis. Mult Scler 27 (2), 208–219.

Debette S, Markus HS. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ*. 2010;341: c3666. Published 2010 Jul 26. doi:10.1136/bmj.c3666.

Danelakis, et al., 2018. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. Computerized Medical Imaging and Graphics. https://doi.org/10.1016/j.compmedimag.2018.10.002.

Dice, L., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26 (3), 297–302. https://doi.org/10.2307/1932409.

Fartaria, M. J., Bonnier, G., Roche, A., Kober, T., Meuli, R., Rotzinger, D., Frackowiak, R., Schluep, M., Du Pasquier, R., Thiran, J.-P., Krueger, G., Bach Cuadra, M., and Granziera, C. (2016). Automated detection of white matter and cortical lesions in early stages of multiple sclerosis. Journal of Magnetic Resonance Imaging, 43: 1445–1454/.

Fartaria, M.J., Sati, P., Todea, A., Radue, E.-W., Rahmanzadeh, R., O'Brien, K., Reich, D. S., Bach Cuadra, M., Kober, T., Granziera, C., 2019. Automated Detection and Segmentation of Multiple Sclerosis Lesions Using Ultra–High-Field MP2RAGE: Investigative Radiology 54 (6), 356–364.

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

Fazekas, F., Chawluk, J.B., Alavi, A., Hurtig, H.I., Zimmerman, R.A., 1987. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. AJR Am. J. Roentgenol. 149, 351–356.

Filippi, M., et al., 2016. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. Lancet Neurol. 15, 292–303.

Filippi M, Preziosa P, Banwell BL, et al. Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain*. 2019;142(7):1858-1875. doi:10.1093/brain/awz144.

Frey, B.M., Petersen, M., Mayer, C., Schulz, M., Cheng, B., Thomalla, G., 2019;10:238.. Characterization of White Matter Hyperintensities in Large-Scale MRI-Studies. Front Neurol. https://doi.org/10.3389/fneur.2019.00238. Published 2019 Mar 26.

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

Gabr, R.E., Pednekar, A.S., Govindarajan, K.A., Sun, X., Riascos, R.F., Ramírez, M.G., Hasan, K.M., Lincoln, J.A., Nelson, F., Wolinsky, J.S., Narayana, P.A., 2017 Aug. Patient-specific 3D FLAIR for enhanced visualization of brain white matter lesions in multiple sclerosis. J Magn Reson Imaging. 46 (2), 557–564. https://doi.org/10.1002/jmri.25557. Epub 2016 Nov 21 PubMed PMID: 27869333.

Gramsch, C., Nensa, F., Kastrup, O., Maderwald, S., Deuschl, C., Ringelstein, A., Schelhorn, J., Forsting, M., Schlamann, M., 2015 May. Diagnostic value of 3D fluid attenuated inversion recovery sequence in multiple sclerosis. Acta Radiol. 56 (5), 622–627. https://doi.org/10.1177/0284185114534413. Epub 2014 May 27 PubMed PMID: 24867222.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., Jenkinson, M., 2016. BIANCA (Brain intensity abnormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. Neuroimage. 141, 191–205.

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

Geremia, et al., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage. https://doi.org/10.1016/j.neuroimage.2011.03.080.

Giorgio, A., De Stefano, N., 2018. Effective Utilization of MRI in the Diagnosis and Management of Multiple Sclerosis. Neurol Clin. 36 (1), 27–34. https://doi.org/10.1016/j.ncl.2017.08.013.

Jack, C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., Erickson, B.J., 2001. FLAIR histogram segmentation for measurement of leukoaraiosis volume. J Magn Reson Imaging 14 (6), 668–676.

Jain, M. Sima, Diana, Ribbens, Annemie, Melissa, Cambron, et al., 2015. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. Neuroimage: Clinical. https://doi.org/10.1016/j.nicl.2015.05.003.

Kim, K.W., MacFall, J.R., Payne, M.E., 2008. Classification of white matter lesions on magnetic resonance imaging in elderly persons. Biol Psychiatry. 64 (4), 273–280. https://doi.org/10.1016/j.biopsych.2008.03.024.

Kuijf, H.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Llado, X., Biesbroek, J.M., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S. H., Pezold, S., Puybareau, E., De Bresser, J., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Heinen, R., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., 2019 Nov. Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge. IEEE Trans Med Imaging. 38 (11), 2556–2568. https://doi.org/10.1109/TMI.2019.2905770. Epub 2019 Mar 19 PubMed PMID: 30908194.

LeCun, Yann, Kavukcuoglu, Koray, Farabet, Clement, et al., 2010. Convolutional networks and applications in vision. IEEE Xplore. https://doi.org/10.1109/ISCAS.2010.5537907.

Lesjak, Ž, Galimzianova, A., Koren, A., et al., 2017. A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. Neuroinformatics. https://doi.org/10.1007/s12021-017-9348-7.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., Poewe, W., Mollenhauer, B., Klinik, P.-E., Sherer, T., Frasier, M., Meunier, C., Rudolph, A., Casaceli, C., Seibyl, J., Mendick, S., Schuff, N., Zhang, Y., Toga, A., Crawford, K., Ansbach, A., De Blasio, P., Piovella, M., Trojanowski, J., Shaw, L., Singleton, A., Hawkins, K., Eberling, J., Brooks, D., Russell, D., Leary, L., Factor, S., Sommerfeld, B., Hogarth, P., Pighetti, E., Williams, K., Standaert, D., Guthrie, S., Hauser, R., Delgado, H., Jankovic, J., Hunter, C., Stern, M., Tran, B., Leverenz, J., Baca, M., Frank, S., Thomas, C.-A., Richard, I., Deeley, C., Rees, L., Sprenger, F., Lang, E., Shill, H., Obradov, S., Fernandez, H., Winters, A., Berg, D., Gauss, K., Galasko, D., Fontaine, D., Mari, Z., Gerstenhaber, M., Brooks, D., Malloy, S., Barone, P., Longo, K., Comery, T., Ravina, B., Grachev, I., Gallagher, K., Collins, M., Widnell, K.L., Ostrowizki, S., Fontoura, P., Ho, T., Luthman, J., Brug, M.V.D., Reith, A.D., Taylor, P., 2011. The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol. 95 (4), 629–635. https://doi.org/10.1016/j.pneurobio.2011.09.005.

Meier, D.S., Weiner, H.L., Guttmann, C.R.G., 2007. Time-series modeling of multiple sclerosis disease activity: A promising window on disease progression and repair potential? Neurotherapeutics 4 (3), 485–498.
Mikol, D.D., Barkhof, F., Chang, P., et al., 2008. Comparison of subcutaneous interferon beta-1a with glatiramer acetate in patients with relapsing multiple sclerosis (the REbif vs Glatiramer Acetate in Relapsing MS Disease [REGARD] study): A multicentre, randomized, parallel, open-label trial. Lancet Neurol 7, 903–914.
Perona, Malik, 1990. Scale-space and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.56205.
Petersen, R.C., Aisen, P.S., Beckett, L.A., et al., 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI). Neurology 74, 201–209.
Pierrick Coupé, Thomas Tourdias, Pierre Linck, Jose Romero, Jose Manjon. LesionBrain: An Online Tool for White Matter Lesion Segmentation. Lecture Notes in Computer Science, Springer, 2018,pp.95 – 103. 10.1007/978-3-030-00500-9_11. hal-01918438.
Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F.D., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A.J., Waubant, E., Weinshenker, B., Wolinsky, J.S., 2011. Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. Ann. Neurol. 69, 292–302.
Radue, E.W., O'Connor, P., Polman, C.H., et al., 2012. Impact of fingolimod therapy on magnetic resonance imaging outcomes in patients with multiple sclerosis. Arch Neurol 69, 1259–1269.
Rocca, M.A., Anzalone, N., Falini, A., Filippi, M., 2013. Contribution of magnetic resonance imaging to the diagnosis and monitoring of multiple sclerosisDiagnosi precoce e monitoraggio nella sclerosi multipla: il contributo della risonanza magnetica. Radiol Med 118 (2), 251–264.
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.
Rovira, À., León, A., 2008. MR in the diagnosis and monitoring of multiple sclerosis: an overview. Eur. J. Radiol. 67 (3), 409–414.
Samaille, T., Fillon, L., Cuingnet, R., Jouvent, E., Chabriat, H., Dormont, D., Colliot, O., Chupin, M., Herholz, K., 2012. Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation. PLoS One. 7 (11), e48953. https://doi.org/10.1371/journal.pone.004895310.1371/journal.pone.0048953.g00110.1371/journal.pone.0048953.g00210.1371/journal.pone.0048953.g00310.1371/journal.pone.0048953.g00410.1371/journal.pone.0048953.g00510.1371/journal.pone.0048953.g00610.1371/journal.pone.0048953.g00710.1371/journal.pone.0048953.g00810.1371/journal.pone.0048953.t00110.1371/journal.pone.0048953.t00210.1371/journal.pone.0048953.t00310.1371/journal.pone.0048953.s00110.1371/journal.pone.0048953.s00210.1371/journal.pone.0048953.s00310.1371/journal.pone.0048953.s00410.1371/journal.pone.0048953.s005.
Samaille, T (2013). Segmentation automatique des anomalies de la substance blanche du sujet âgé [PhD thesis, Pierre and Marie Curie University].
Schmidt, R., Scheltens, P.h., Erkinjuntti, T., Pantoni, L., Markus, H.S., Wallin, A., Barkhof, F., Fazekas, F., 2004. White matter lesion progression: a surrogate endpoint for trials in cerebral small-vessel disease. Neurology. 63 (1), 139–144.
Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Mühlau, M., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. Neuroimage. 59 (4), 3774–3783.
Schmidt P. LMU München; 2017. Bayesian Inference for Structured Additive Regression Models for Large-Scale Problems with Applications to Medical Imaging.
Seewann, A., Vrenken, H., van der Valk, P., Blezer, E.L., Knol, D.L., Castelijns, J.A., Polman, C.H., Pouwels, P.J., Barkhof, F., Geurts, J.J., 2009. Diffusely abnormal white matter in chronic multiple sclerosis: imaging and histopathologic analysis. Archives of neurology 66 (5), 601–609. https://doi.org/10.1001/archneurol.2009.57.
Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage. 49 (2), 1524–1535.
Simon, J.H., Li, D., Traboulsee, A., Coyle, P.K., et al., 2006. Standardized MR imaging protocol for multiple sclerosis: Consortium of MS centers consensus guidelines. Am J Neuroradiol 27, 455–461.
Naganawa, S., 2015. The Technical and Clinical Features of 3D-FLAIR in Neuroimaging. MRMS 14 (2), 93–106.
Steenwijk, M.D., Pouwels, P.J.W., Daams, M., van Dalen, J.W., Caan, M.W., a., Richard, E., Barkhof, F., and Vrenken, H, 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage: Clinical 3, 462–469.
Styner M., Lee J., Chin B. 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. MIDAS Journal. 2008:1–6.
Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S., Fujihara, K., Galetta, S.L., Hartung, H.P., Kappos, L., Lublin, F.D., Marrie, R.A., Miller, A.E., Miller, D.H., Montalban, X., Mowry, E.M., Sorensen, P.S., Tintoré, M., Traboulsee, A.L., Trojano, M., Uitdehaag, B.M.J., Vukusic, S., Waubant, E., Weinshenker, B.G., Reingold, S.C., Cohen, J.A., 2018. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol. 17 (2), 162–173. https://doi.org/10.1016/S1474-4422(17)30470-2.
Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Salvi, J., Oliver, A., Lladó, X., 2019. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. NeuroImage: Clinical 21, 101638. https://doi.org/10.1016/j.nicl.2018.101638.
Wattjes, M.P., Rovira, À., Miller, D., 2015. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis – establishing disease prognosis and monitoring patients. Nat. Rev. Neurol. 11 (10), 597–606.
Weeda, M.M., Brouwer, I., de Vos, M.L., de Vries, M.S., Barkhof, F., Pouwels, P.J.W., Vrenken, H., 2019. Comparing lesion segmentation methods in multiple sclerosis: Input from one manually delineated subject is sufficient for accurate lesion segmentation. NeuroImage: Clinical 24, 102074. https://doi.org/10.1016/j.nicl.2019.102074.
Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., Hill, D., Killiany, R.J., Pachai, C., Schwarz, A.J., Schuff, N., Senjem, M.L., Suhy, J., Thompson, P.M., Weiner, M., Jack, C.R., 2013. Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimer's & Dementia 9 (3), 332–337.
Yamamoto, D., Arimura, H., Kakeda, S., Magome, T., Yamashita, Y., Toyofuku, F., Ohki, M., Higashida, Y., Korogi, Y., 2010 Jul. Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. Comput Med Imaging Graph. 34 (5), 404–413. https://doi.org/10.1016/j.compmedimag.2010.02.001. Epub 2010 Feb 26 PubMed PMID: 20189353.
Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006 Jul 1. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage. 31 (3), 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015. Epub 2006 Mar 20 PMID: 16545965.
Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans. Med. Imaging. 21 (10), 1280–1291.