Magnetic Resonance Imaging 76 (2021) 108–115

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

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

Original Contribution

Accuracy and reproducibility of automated white matter hyperintensities segmentation with lesion segmentation tool: A European multi-site 3T study

Federica Ribaldi [a,b,c,am,*], Daniele Altomare [c,am], Jorge Jovicich [d], Clarissa Ferrari [e], Agnese Picco [f], Francesca Benedetta Pizzini [g], Andrea Soricelli [h], Anna Mega [a], Antonio Ferretti [i,j], Antonios Drevelegas [k,l], Beatriz Bosch [m], Bernhard W. Müller [n], Camillo Marra [o], Carlo Cavaliere [h], David Bartrés-Faz [m], Flavio Nobili [p,q], Franco Alessandrini [g], Frederik Barkhof [r,s], Helene Gros-Dagnac [t,u], Jean-Philippe Ranjeva [v], Jens Wiltfang [w], Joost Kuijer [s], Julien Sein [v], Karl-Titus Hoffmann [x], Luca Roccatagliata [q,y], Lucilla Parnetti [z], Magda Tsolaki [aa], Manos Constantinidis [k], Marco Aiello [h], Marco Salvatore [h], Martina Montalti [a], Massimo Caulo [i,j], Mira Didic [ab,ac], Núria Bargallo [ad], Olivier Blin [ae], Paolo M Rossini [af], Peter Schonknecht [ag], Piero Floridi [ah], Pierre Payoux [t], Pieter Jelle Visser [ai], Régis Bordet [aj], Renaud Lopes [aj], Roberto Tarducci [ak], Stephanie Bombois [aj], Tilman Hensch [ag], Ute Fiedler [n], Jill C. Richardson [al], Giovanni B. Frisoni [c,am], Moira Marizzoni [a]

[a] *Laboratory of Alzheimer's Neuroimaging and Alzheimer's Epidemiology, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy*
[b] *Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy*
[c] *Laboratory of Neuroimaging of Aging (LANVIE), University of Geneva, Geneva, Switzerland*
[d] *Center for Mind/Brain Sciences (CIMEC), University of Trento, Rovereto, Italy*
[e] *Unit of Statistics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy*
[f] *Department of Neuroscience, Ophthalmology, Genetics and Mother-Child Health (DINOGMI), University of Genoa, Genoa, Italy*
[g] *Radiology, Dept. of Diagnostic and Public Health, Verona University, Verona, Italy*
[h] *IRCCS SDN, Naples, Italy*
[i] *Department of Neuroscience Imaging and Clinical Sciences, University "G. d'Annunzio" of Chieti, Italy*
[j] *Institute for Advanced Biomedical Technologies (ITAB), University "G. d'Annunzio" of Chieti, Italy*
[k] *Interbalkan Medical Center of Thessaloniki, Thessaloniki, Greece*
[l] *Department of Radiology, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[m] *Department of Psychiatry and Clinical Psychobiology, Universitat de Barcelona and IDIBAPS, Barcelona, Spain*
[n] *LVR-Clinic for Psychiatry and Psychotherapy, Institutes and Clinics of the University Duisburg-Essen, Essen, Germany*
[o] *Center for Neuropsychological Research, Catholic University, Rome, Italy*
[p] *Dept. of Neuroscience (DINOGMI), University of Genoa, Italy*
[q] *IRCCS Ospedale Policlinico San Martino Genova, Italy*
[r] *Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK*
[s] *Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Amsterdam, the Netherlands*
[t] *ToNIC, Toulouse NeuroImaging Center, Université de Toulouse, Inserm, UPS, France*
[u] *Université Toulouse 3 Paul Sabatier, UMR 825 Imagerie Cérébrale et Handicaps Neurologiques, F-31024 Toulouse, France*
[v] *Institut de Neurosciences de la Timone (INT), Aix-Marseille Université, CNRS, UMR 7289, 13005 Marseille, France*
[w] *Department of Psychiatry and Psychotherapy, University Medical Center (UMG), Georg-August University, Göttingen, Germany*
[x] *Department of Neuroradiology, University Hospital Leipzig, Leipzig, Germany*
[y] *Dept. of Health Sciences (DISSAL), University of Genoa, Italy*
[z] *Section of Neurology, Centre for Memory Disturbances, University of Perugia, Perugia, Italy*
[aa] *1st Department of Neurology, Aristotle University of Thessaloniki, Makedonia, Greece*
[ab] *APHM, Timone, Service de Neurologie et Neuropsychologie, APHM Hôpital Timone Adultes, Marseille, France*
[ac] *Aix Marseille Univ, INSERM, INS, Inst Neurosci Syst, Marseille, France*
[ad] *Department of Neuroradiology and Magnetic Resonance Image Core Facility, Hospital Clínic de Barcelona, IDIBAPS, Barcelona, Spain*
[ae] *Aix Marseille University, UMR-INSERM 1106, Service de Pharmacologie Clinique, AP-HM, Marseille, France*
[af] *Dept. Neuroscience & Neurorehabilitation, IRCCS-San Raffaele-Pisana, Rome, Italy*

* Corresponding author at: fribaldi@fatebenefratelli.eu. Laboratory of Alzheimer's Neuroimaging and Alzheimer's Epidemiology, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.

https://doi.org/10.1016/j.mri.2020.11.008

Available online 19 November 2020

[ag] Department of Psychiatry and Psychotherapy, University of Leipzig Medical Center, Leipzig, Germany
[ah] Neuroradiology Unit, Perugia General Hospital, Perugia, Italy
[ai] Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
[aj] Univ. Lille, INSERM, CHU Lille, Lille Neuroscience & Cognition - Degenerative and Vascular Cognitive Disorders-U1172. F-59000 Lille, France
[ak] Medical Physics Unit, Perugia General Hospital, Perugia, Italy
[al] Neurosciences Therapeutic Area, GlaxoSmithKline R&D, Gunnels Wood Road, Stevenage, United Kingdom
[am] Memory Clinic, Geneva University Hospitals, Geneva, Switzerland

ARTICLE INFO

Keywords:
White matter hyperintensities
Automated segmentation algorithms
Lesion segmentation toolbox
Reproducibility
Accuracy

ABSTRACT

Brain vascular damage accumulate in aging and often manifest as white matter hyperintensities (WMHs) on MRI. Despite increased interest in automated methods to segment WMHs, a gold standard has not been achieved and their longitudinal reproducibility has been poorly investigated. The aim of present work is to evaluate accuracy and reproducibility of two freely available segmentation algorithms. A harmonized MRI protocol was implemented in 3T-scanners across 13 European sites, each scanning five volunteers twice (test-retest) using 2D-FLAIR. Automated segmentation was performed using Lesion segmentation tool algorithms (LST): the Lesion growth algorithm (LGA) in SPM8 and 12 and the Lesion prediction algorithm (LPA). To assess reproducibility, we applied the LST longitudinal pipeline to the LGA and LPA outputs for both the test and retest scans. We evaluated volumetric and spatial accuracy comparing LGA and LPA with manual tracing, and for reproducibility the test versus retest. Median volume difference between automated WMH and manual segmentations (mL) was −0.22 [IQR = 0.50] for LGA-SPM8, −0.12[0.57] for LGA-SPM12, −0.09[0.53] for LPA, while the spatial accuracy (Dice Coefficient) was 0.29[0.31], 0.33[0.26] and 0.41[0.23], respectively. The reproducibility analysis showed a median reproducibility error of 20%[IQR = 41] for LGA-SPM8, 14% [31] for LGA-SPM12 and 10% [27] with the LPA cross-sectional pipeline. Applying the LST longitudinal pipeline, the reproducibility errors were considerably reduced (LGA: 0%[IQR = 0], $p < 0.001$; LPA: 0% [3], $p < 0.001$) compared to those derived using the cross-sectional algorithms. The DC using the longitudinal pipeline was excellent (median = 1) for LGA [IQR = 0] and LPA [0.02]. LST algorithms showed moderate accuracy and good reproducibility. Therefore, it can be used as a reliable cross-sectional and longitudinal tool in multi-site studies.

1. Introduction

White matter hyperintensities (WMHs) are a marker of white matter tissue damage seen as hyperintense signals on T2 and Fluid Attenuated Inversion Recovery (FLAIR) images. They are a potential hallmark of various disorders such as cerebrovascular disease [1], other neurological [2] (e.g. multiple sclerosis, MS, and dementia) [3,4], psychiatric [5] or inflammatory disorders [1,6]. Moreover, WMHs are commonly seen also in cognitively unimpaired people [7], and their prevalence increases with aging [8]. The prevalence of WMHs in community-dwelling elderly is highly variable, ranging from 5.3% to 100% depending on study design, study population, and WMH assessment methods [9–14]. Consistent evidence indicates that the volume of WMHs is positively associated with cognitive decline both in cognitively unimpaired people [15–17] and in patients with cognitive impairment [18,19]. This suggests that WMHs volume is a relevant biomarker and should be taken into account not only for the clinical evaluation of elderly people, but also in research studies [20,21]. Currently, WMHs are often quantified through visual semi-quantitative scales both in clinical and research settings (e.g. the Age-Related White Matter Changes, and Fazekas scales) [22,23]. These scales are relatively quick, but require proper training and show significant inter-rater and intra-rater variability [24]. Similar limitations apply to manual segmentation [25], which are in addition very time-consuming. To overcome these limitations, there is an increasing interest in automated and semi-automated methods allowing reliable and effective WMHs segmentation and quantification (for a review see [26]). We selected an open-source tool, Lesion Segmentation Toolbox (LST) from the Statistical Parametric Mapping (SPM) software package, that segments WMHs using FLAIR and T1 images (http://www.applied-statistics.de/lst.html). We choose LST because: (i) it does not require training; (ii) is freely available; (iii) is fully automated; (iv) includes also a longitudinal pipeline, in addition to cross-sectional algorithms, that could be useful for monitoring the WMHs evolution over time [27]. LST was originally developed for MS lesion segmentation [28], but it was used to segment WMHs also in other diseases such as diabetes mellitus [29]. To the best of our knowledge, accuracy and test–retest reproducibility of WMH measurements in multi-site studies of elderly subjects have been poorly investigated. The aim of this study was to assess i) the accuracy of the two LST cross-sectional algorithms versus manual segmentation performed by an expert rater (as a standard of truth), and ii) the test-retest reproducibility of the two cross-sectional algorithms and the longitudinal pipeline of LST. We also tested whether accuracy and reproducibility were affected by MRI scanner or site effects.

2. Materials and methods

For the present study, we analyzed data from the PharmaCog project [30]. Participants, study design and further details have been exhaustively described in previous studies [31–35], but are briefly summarized below.

2.1. Participants

Thirteen 3T MRI sites across Italy (Verona, Genoa, Rome, Perugia and Naples), Spain (Barcelona), France (Marseille, Lille, and Toulouse), Germany (Essen, Leipzig), Greece (Thessaloniki) and the Netherlands (Amsterdam) provided imaging data. Each site enrolled four-to-five cognitively unimpaired elderly (age range: 50–78 years, 60% of females) scanned twice as follow: (i) baseline (test); (ii) after 7–32 days (retest) (median interval: 13.5 days, IQR = 15). This short test–retest interval minimizes potential biological changes, allowing to address the reproducibility of MRI assessment tools. All participants had no history of psychiatric, neurological or systemic disease, were Caucasian and provided written informed consent following procedures approved by the local institutional review board of the institution where scanning was performed. Detailed inclusion and exclusion criteria were described elsewhere [33].

2.2. MRI acquisition

The 13 MRI sites used different MRI scanners (Siemens, GE, Philips) and

only vendor-provided sequences. For each participant, axial 2D structural FLAIR images were obtained in different sessions two weeks apart for the test-retest evaluations. The acquisition parameters in each session followed mostly the harmonization suggestions from the ADNI-2 protocol (http://adni.loni.usc.edu/methods/documents/mri-protocols). For all sites the following basic FLAIR parameters were maintained: voxel 0.9 × 0.9 × 4 mm3; inversion flip angle 1500, no fat suppression, full k space, acceleration factor in the range of 1.5–2 was used where possible. Parameters that change across sites are reported in Table 1. Note that some parameters vary considerably across vendors due to differences in sequence implementations and definitions. The test–retest raw data from this study will be made available on request.

2.3. MRI manual segmentation

A rater with expertise in lesion segmentation performed the 2D manual segmentation of WMH only on the test FLAIR images using FSLview version 5.0.3 blinded to the results of the automated segmentation. The process of manual tracing resulted in the definition of binary masks, considered as a standard of truth. For each subject, WMHs volumes (expressed in mL) were calculated automatically using FSL (*fslstats* of FSLUTILS).

2.4. MRI automated segmentation: LST

We have processed the test and retest images using the two LST algorithms, and their outputs were further processed using the LST longitudinal pipeline for the reproducibility assessment. These tools are described below.

1. Lesion Growth Algorithm (LGA): LGA is implemented both in SPM8 and SPM12. The algorithm first segments the T1 images into the three main tissue classes (CSF, GM and WM). This information is then combined with the coregistered FLAIR intensities in order to calculate 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 FLAIR image. The result is a lesion probability map. We used LGA with an optimized parameter (Kappa = 0.25) set by visual inspection of the segmentations resulting from different test parameters [36].
2. Lesion Prediction Algorithm (LPA): LPA is implemented only in SPM12. 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 as for the lesion growth algorithm [28] was used 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. The algorithm requires only a FLAIR image, however T1 might improve WMH segmentation. We used LPA using both T1 and FLAIR. No parameters needed to be set [37].

1) Longitudinal Pipeline: A longitudinal LST pipeline is implemented only for SPM12. Segmented lesion maps of test and retests were compared using the longitudinal pipeline implemented in the LST toolbox. This pipeline consists of the following steps: first, lesion maps and FLAIR images are coregistered to the images of the first time point; then, relative differences of FLAIR intensities are calculated along all voxels that were segmented as lesions in at least one time point; finally, significant increase and decrease of lesion voxels are identified if their differences exceed or fall below a certain threshold that is obtained by analyzing healthy white matter. As a final result, lesion change labels are produced for all consecutive time points. In these images the three possible cases decrease, no change and increase are labeled by the numbers 1, 2, and 3, respectively. Both the LGA and LPA cross-sectional outputs from SPM12 were further processed using this pipeline [27].

Since LGA and LPA output are on T1-space, we linearly registered

Table 1
Summary of demographic, MRI system and 2D FLAIR acquisition differences across MRI sites (largely based on ADNI-2).

Site (location)	3T MRI Scanner	Sequence parameters 2D FLAIR				Acquisition matrix	Voxel (read x phase x slice mm3)	Subjects' age, median (IQR)	Test-Retest Days interval, median (IQR)	Gender, (female /N)
		TR (ms)	TE (ms)	TI (ms)	FA (°)					
1 (Verona)	Siemens Allegra	9760	86	2500	150	256 × 256		68 (7)	7 (14)	2/5
2 (Barcelona)	Siemens TrioTim	9000						73 (1)	12 (2)	4/4
3 (Leipzig)	Siemens TrioTim		90					62 (4)	14 (1)	3/5
4 (Marseille)	Siemens Verio							65 (11)	14 (28)	4/5
5 (Essen)	Siemens Skyra		91					52 (3)	9 (6)	2/5
6 (Naples)	Siemens Biograph mMR		90					58 (2)	7 (27)	2/5
7 (Genoa)	GE HDxt	11,000	147	2250			0.9 × 0.9 × 4.0	58 (3)	14 (10)	2/4
8 (Thessaloniki)	GE HDxt	8002	126					56 (11)	32 (13)	2/4
9 (Amsterdam)	Discovery MR750							63 (10)	7 (7)	3/5
10 (Lille)	Philips Achieva	9000	90	2500		256 × 237		66 (2)	8 (15)	3/5
11 (Toulouse)	Philips Achieva							60 (4)	19 (9)	3/5
12 (Chieti)	Philips Achieva	11,000				256 × 211		69 (4)	9 (1)	4/5
13 (Perugia)	Philips Achieva	9000						60 (12)	7 (3)	2/3

Abbreviations: TR, repetition time; TE, echo time; TI, inversion time; FA, flip angle.

them to the FLAIR-space (FSL-FLIRT, 6 DOF and trilinear interpolation) where the manual segmentations were performed.

2.5. Accuracy analysis

First, a threshold of 0.5 was applied to the LGA and LPA lesion maps in order to create binary masks. After that, we assessed the accuracy of the cross-sectional algorithms (i.e. LGA and LPA) vs manual WMHs segmentation in terms of: i) volumetric accuracy (differences between volumes of the manual and automated segmentations), and ii) spatial accuracy (Dice coefficient, DC). DC was calculated with the following formula [38,39]:

$$DC = 2\frac{|Automated\ segmentation \cap Manual\ segmentation|}{|Automated\ segmentation| + |Manual\ segmentation|}$$

2.6. Reproducibility analysis

We assessed the test-retest reproducibility of LST in terms of: i) volumetric reproducibility, using the reproducibility error (ε) and the intraclass correlation coefficient (ICC), and ii) spatial reproducibility using the DC. Reproducibility error was calculated with following formula [31]:

$$\varepsilon = 100\frac{|Retes\ volume - Test\ volume|}{(Retest\ volume + Test\ volume)/2}$$

2.7. Statistical analysis

One-way Kruskal–Wallis test was used to test for MRI site and scanner effects on the participants' distribution of age, volumes, accuracy and reproducibility measures (significance threshold set at $p < 0.05$). If significant, we used post-hoc pairwise comparisons using Dunn's all-pairs test. For the spatial accuracy analyses, Spearman rank correlation between DCs and manual WMHs volumes was performed to evaluate if the association between WMH volumes and spatial overlap was significant. An independent 2-group Mann-Whitney *U* Test was used to assess the DCs differences between the low WMH volume group (≤5 mL) compared to the group with medium to high WMH volume (>5 mL). All analyses were performed using R, version 3.5.2 (R Foundation for statistical computing, https://www.r-project.org/).

3. Results

3.1. Participants' features across MRI sites

Table 1 shows participants' demographic features across MRI sites. Age was similar across sites, except for the participants of site 5 (Essen) who were younger (age: 52 years, IQR = 3) than those of site 2 (Barcelona, age: 73 years, IQR = 1, $p = 0.004$) and site 12 (Chieti, age: 69 years, IQR = 4, $p = 0.022$). No differences across sites were observed in gender distribution, and time interval between test and retest scans (Table 1). The median WMH volume measured by manual rater was 0.54 mL (IQR = 1.58), and no differences across sites or scanners are observed for this measure.

3.2. Accuracy results

Of the 60 subjects, two were excluded from accuracy analysis due to lack of lesions detected by the expert ($n = 1$), or for low signal-to-noise ratio (n = 1). Visual inspection of the WMHs segmentation showed different segmentation quality across MRI scanners, in particular the visual quality assessment shows high performance for LPA segmentations (Fig. 1).

3.2.1. Volumetric accuracy

Fig. 2 shows median volume differences between manual and automated segmentations. Volumetric accuracy of LPA SPM12 (volume difference: -0.09, IQR = 0.53) seemed numerically better than that of LGA SPM8 (−0.22, IQR = 0.50, $p = 0.024$), but was not statistically different than that of LGA SPM12 (−0.12, IQR = 0.57, $p = 0.084$) (Fig. 2).

No site effect was observed for volumetric differences, while a scanner effect was observed only for LGA SPM12 (−0.44, IQR = 1.11 for GE vs −0.01, IQR = 0.70 for Siemens, $p = 0.010$) and LPA SPM12 (−0.33, IQR = 1.23 for GE vs 0.01, IQR = 1.00 for Philips, $p = 0.003$).

3.2.2. Spatial accuracy

Fig. 3 shows the spatial accuracy between manual and automated segmentations for each subject, expressed by the DC coefficient. Subject were ordered based on the manual segmentation WMH volumes, showing a clear trend for worse performance at lower volumes. Median DC was 0.41 for LPA (from 0.34, IQR = 0.21, in Philips to 0.43, IQR = 0.34, in Siemens), 0.29 for LGA SPM8 (from 0.25, IQR = 0.27 in Philips to 0.42, IQR = 0.34 in Siemens), and 0.33 for LGA SPM12 (from 0.31, IQR = 0.35 in GE to 0.40, IQR = 0.28 in Siemens). No statistically significant differences were observed among the algorithms ($p > 0.05$). Moreover, as expected, the DC coefficient increased with increasing WMHs volume, independently of the considered algorithm (LGA SPM8: rho = 0.70, $p < 0.001$; LGA SPM12: rho = 0.62, $p < 0.001$; LPA SPM12: rho = 0.62, $p < 0.001$) (Fig. 3). Table 2 reports the DC of the two algorithms divided by lesion volume in low (≤5 mL) and medium to high (>5 mL). Indeed, the DC is higher for the group with higher WMH, irrespective of the algorithm ($p < 0.005$). No site or scanner effects were observed on these measures.

3.3. Reproducibility results

Of the 60 subjects, three were excluded from reproducibility analysis due to lack of lesions detected by the expert ($n = 1$, the same subjects excluded from the accuracy analysis), or for low signal-to-noise ratio (n = 1, the same subjects excluded from the accuracy analysis), or because the longitudinal segmentation failed (n = 1).

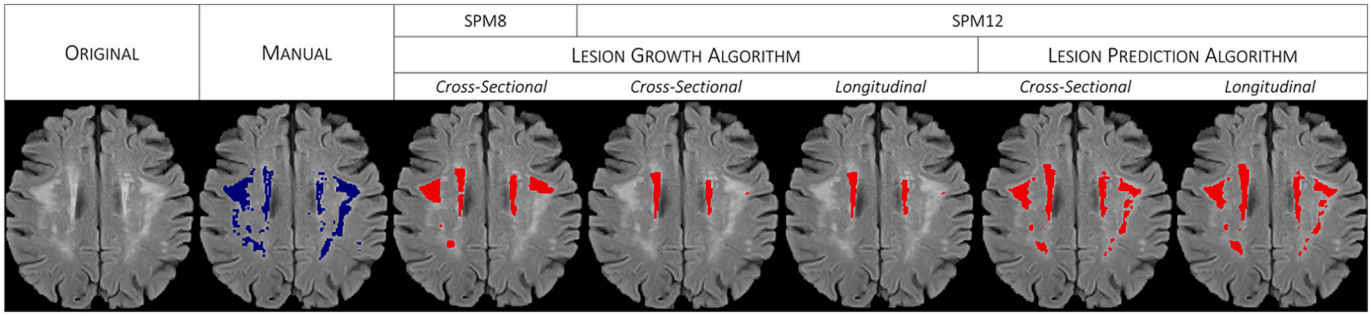

Fig. 1. Manual and automated WMHs segmentations overlaid on sample subject 2D FLAIR scan. Abbreviations: SPM, statistical parametric mapping.

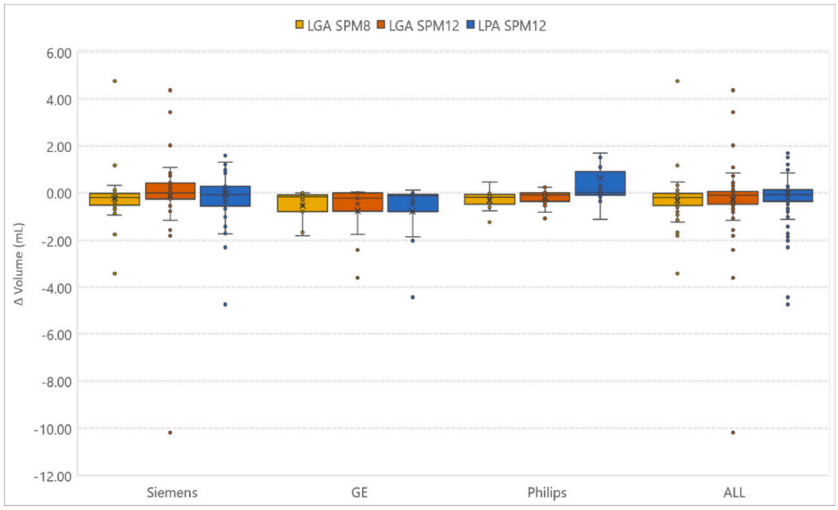

Fig. 2. Volumetric Accuracy: comparison between automated and manual segmentations.Boxplot represents lesion volumes differences between automated and manual segmentations, separated for scanner type and in the whole group. The overall comparison showed that the difference between manual and LPA SPM12 volumes was lower than that between manual and LGA SPM8 ($p = 0.024$), while not statistically different than that between manual and LGA SPM12 ($p = 0.084$), meaning that the volume accuracy is better for LPA SPM12 compared to LGA SPM8.
Abbreviations: LGA, lesion growth algorithm; SPM, statistical parametric mapping; LPA, lesion prediction algorithm; Cross, cross-sectional.

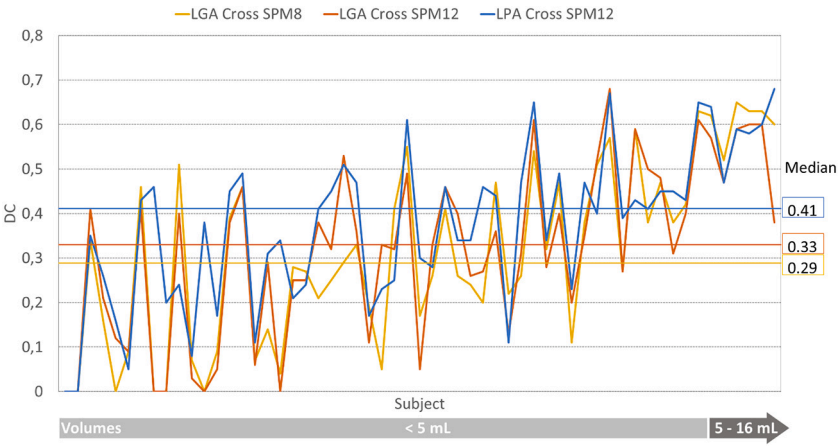

Fig. 3. Spatial accuracy is expressed by Dice coefficients for each subject ordered by lesion volumes. Median DC was 0.41 for LPA, 0.29 for LGA on SPM8 and 0.33 for LGA SPM12. DC coefficients increased with increasing WMHs volume, independently of the considered algorithm (LGA SPM8: rho=0.70, p>0.001; LGA SPM12: rho=0.62, p>0.001; LPA SPM12: rho=0.62, p>0.001). Abbreviations: DC, dice coefficient; LGA, lesion growth algorithm; SPM, statistical parametric mapping; LPA, lesion prediction algorithm; Cross, cross-sectional. Spatial Accuracy: comparison between automated and manual segmentations.

Table 2
Summary of the accuracy and reproducibility measure (expressed in DC) grouped by low and medium to high volume.

Comparison		Low (N = 54)	Medium to high (N = 6)
	Volume	≤5 mL	>5 mL
Automated vs manual test	DC Cross LGA SPM8	0.27 [0.29]	0.62 [0.02]
	DC Cross LGA SPM12	0.32 [0.28]	0.58 [0.10]
	DC Cross LPA SPM12	0.38 [0.22]	0.60 [0.05]
Automated test vs retest	DC Cross LGA SPM8	0.61 [0.31]	0.83 [0.06]
	DC Cross LGA SPM12	0.64 [0.29]	0.80 [0.02]
	DC Long LGA SPM12	1 [0.0]	1 [0.01]
	DC Cross LPA SPM12	0.65 [0.15]	0.81 [0.03]
	DC Long LPA SPM12	1 [0.03]	1 [0.01]

Abbreviations: DC, dice coefficient; LGA, lesion growth algorithm; SPM, statistical parametric mapping; LPA, lesion prediction algorithm; Long, longitudinal.

3.3.1. *Volumetric reproducibility*

Using the cross-sectional algorithms, we observed a median reproducibility error of 10% (IQR = 27) using LPA, 14% (IQR = 31) for LGA SPM12 and 20% (IQR = 41) for LGA SPM8.

Applying the LST longitudinal pipeline to LGA and LPA, the reproducibility errors were considerably reduced (LGA: 0%, IQR = 0, $p < 0.001$; LPA: 0%, IQR = 3, $p < 0.001$) compared to those observed using the cross-sectional algorithms only (Fig. 4). We observed a scanner effect only on the reproducibility error of SPM12 LGA (9% for Siemens vs 15% for GE, $p = 0.029$; and 30% for Philips, $p = 0.003$). Moreover, we found no site effect. We observed an excellent test-retest volumetric agreement using both cross-sectional algorithms and applying the longitudinal pipeline (ICC = 1).

3.3.2. *Spatial reproducibility*

The comparison between test-retest cross-sectional algorithms showed a similar ($p = 0.975$) DC for SPM8 LGA (0.65, IQR = 0.26), SPM12 LGA (0.67, IQR = 0.31) and LPA (0.66, IQR = 0.17). For cross-sectional algorithms the DC is higher in the group with higher WMH volume (Table 2, $p < 0.005$). The DC for both LGA and LPA with longitudinal processing was very high (LGA: median = 1, IQR = 0 vs LPA: median = 1, IQR = 0.02; $p = 0.04$). We observed no site or scanner effect on this measure (Fig. 5).

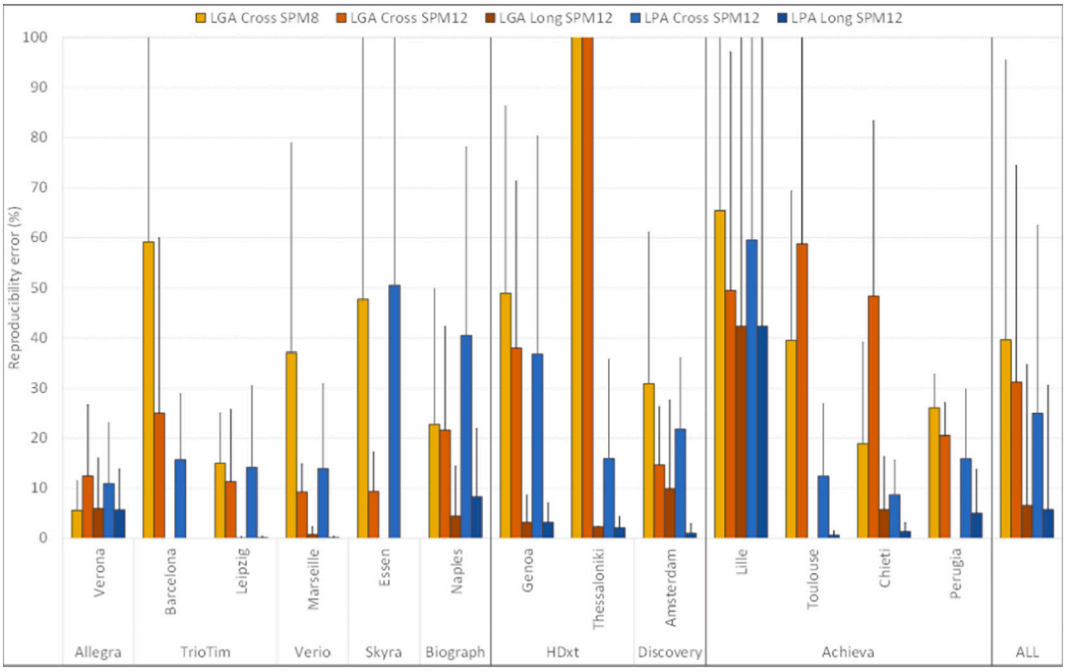

Fig. 4. Bars represent reproducibility error (%) = |Vretest-Vtest|/(Vretest+Vtest)/2. We assessed the volumetric reproducibility of the whole WMHs segmentation algorithms and pipelines for each site and grouped for scanner type. We observed a scanner effect only on the reproducibility error of SPM12 LGA (9% in Siemens vs 15% in GE, p=0.029; and 30% in Philips, p=0.003). Abbreviations: LGA, lesion growth algorithm; SPM, statistical parametric mapping; LPA, lesion prediction algorithm; Cross, cross-sectional; Long, longitudinal. Volumetric reproducibility: comparison between test and retest automated WMHs segmentations.

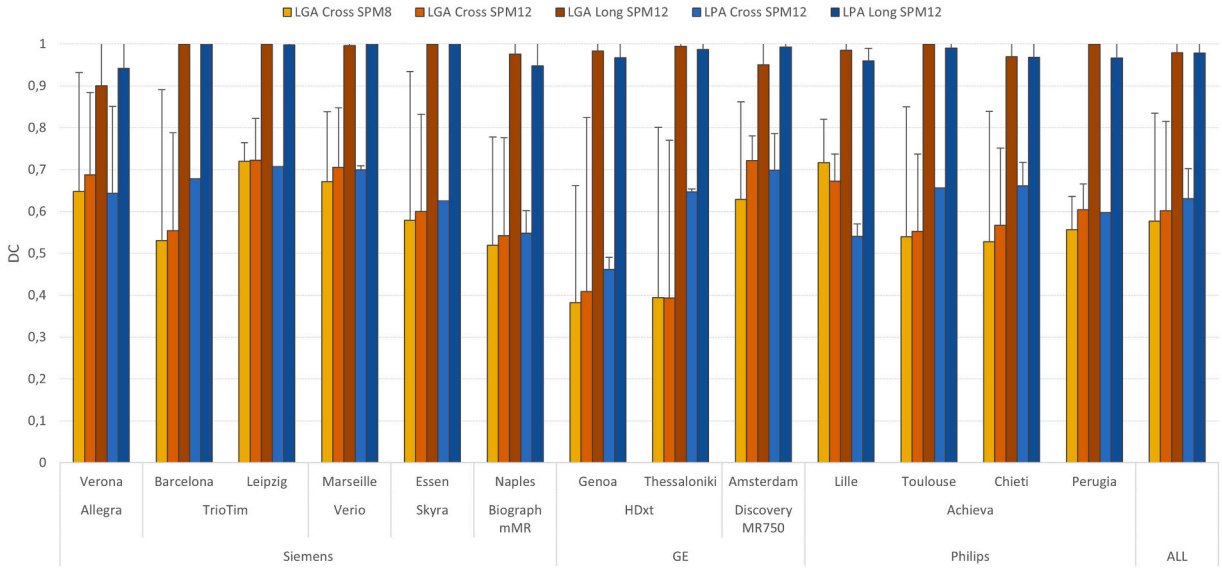

Fig. 5. Spatial reproducibility: comparison between test and retest automated WMHs segmentations.
Bars represent Dice Coefficients between test and retest automated segmentations.
We assessed the spatial reproducibility of the cross-sectional and longitudinal pipelines of WMHs segmentation algorithms for each site and grouped for scanner type. No site or scanner effect were observed.Abbreviations: DC, dice coefficient; LGA, lesion growth algorithm; SPM, statistical parametric mapping; LPA, lesion prediction algorithm; Long, longitudinal.

4. Discussion

In this study, we evaluated the accuracy and reproducibility of LST in a population of healthy elderly subjects scanned twice in a 3T MRI multi-site cohort, showing a good performance of its cross-sectional algorithms and longitudinal pipeline in terms of volumes accuracy and reproducibility. In particular, our main results are summarized as follows: (i) LPA and LGA show a good volumetric accuracy, but LPA performed overall better than LGA; (ii) the LGA and LPA's spatial accuracy increases with the amount of WMHs; (iii) volumetric reproducibility reveals that LST longitudinal pipeline steeply reduces the reproducibility error; (iv) spatial reproducibility of the longitudinal pipeline applied to LGA and LPA outputs was optimal.

Compared to De Sitter and colleagues [40] in a cohort of 52 MS

patients (mean WMH volume = 4.85 mL), we found a slightly better volumetric accuracy comparing both LPA (mean volume difference = 0.45 mL) and LGA (mean volume difference = 2.88 mL) using SPM12, or other tools that they have tested such as Cascade [41,42] (mean volume difference = 0.67 mL), Lesion-Topology preserving Anatomical Segmentation (Lesion-TOADS) [43] (mean volume difference = 2.18 mL) or using k-Nearest Neighbor with Tissue Type Priors (kNN-TTP) [44] (mean volume difference = −1.46 mL). Similar results have been reported by Egger and colleagues [45] for LGA SPM8 (median volume difference = 0.68 mL), LGA SPM12 (median volume difference = 0.93 mL), LPA SPM12 (median volume difference = 0.85 mL).

On the other hand, the spatial accuracy we observed using both LGA and LPA (DC range = 0.29–0.41) was lower compared to previous studies using LST algorithms or other supervised/unsupervised or automated methods (DC range = 0.75–0.84, [44] mean WMH volume = 16.33 mL, [46–48]), except for De Sitter and colleagues that showed comparable results (DC range = 0.23–0.44) [40]. Conversely, studies on healthy subjects showed more variability, e.g. DC = 0.47 in Ong et al. [49] (mean WMH volume = 5.182.603 mm^3))), DC between 0.63 and 0.75 in Manjon et al. [50], or DC = 0.77 in Wang et al. [48,51] (mean WMH volume = 20.43 mL). The lower spatial accuracy found in our work might be due to the following reasons: (i) our sample consists of healthy participants with overall small WMHs volumes, which correlate with a lower DC (as represented in Fig. 3); (ii) we acquired only 2D FLAIR images, so we could expect a higher accuracy using 3D FLAIR; (iii) given the multi-site nature of this study, we expected a considerable heterogeneity of the algorithms' performance across sites and scanners, which is a non-systematic bias for both manual and automated segmentations. Indeed, a high variability was observed for all quantification methods. Nevertheless, the multi-site nature of this study is a strength of this study, making the research setting closer to the real clinical setting and improving the generalization of our results.

As far as we know, only a few studies have investigated the reproducibility of WMHs segmentation tools. The reproducibility of a FMRIB's tool to segments automatically WMHs, Brain Intensity AbNormality Classification Algorithm (BIANCA), was tested in a sample of 20 subjects scanned twice on the same scanner, revealing a very similar reproducibility error compared to our results for both LGA and LPA on SPM12 (reproducibility error mean = 10%) [48]. However, the longitudinal pipeline of LST allows to reduce the reproducibility error approximately to 0%.

A quick and reliable WMHs quantification across different sites and scanners is needed both in clinical practice, in order to improve the diagnostic workup and track the disease progression of elderly people with suspected neurodegenerative diseases, and in research settings to improve the selection of the population of interest. Further studies on larger datasets are needed to confirm the accuracy and reproducibility of LST and to provide normative data on WMHs. Moreover, studies on the cause of the low accuracy, like lesions location, type or shape will enhance our knowledge of the WMHs and will help the algorithm developers.

5. Limitation

Marizzoni et al. (2015, 32] and Jovicich et al. (2013, 2014) [31,33] have already discussed some limitations of the study design, but some of the issues are addressed here for completeness.

First, the number of participants included in this study is small (n = 60), and each site contributed with a different number of participants (from 4 to 5). We have grouped the MRI sites with the same scanner in order to increase the number of subjects per group. Moreover, the median WMHs volume of our sample was low (0.54 mL, IQR = 1.58). Therefore, our sample might be little representative of a healthy subject's population. We only had 2D FLAIR (instead of 3D sequences), and their lower resolution might explain the low spatial agreement.

Lastly, we did not use a training dataset to set the threshold (see method, LGA) but only visual inspection to be coherent with clinical practice.

6. Conclusion

LST is a free, easy-to-use and quick automated method allowing to accurately and reliably assess WMHs volume, even at multiple time points. We suggest the use of this tool in observational longitudinal research studies as a reliable tool to quantify WMHs overtime.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mri.2020.11.008.

Acknowledgements

We wish to thank all the co-Investigators from our Study Group (WP5-PharmaCog/European Alzheimer's Disease Neuroimaging Initiative) (for the complete list see Supplementary Table 1). FB is supported by the NIHR biomedical research center at UCLH.

References

[1] Rosenberg GA. Inflammation and white matter damage in vascular cognitive impairment. Stroke 2009. https://doi.org/10.1161/STROKEAHA.108.533133.
[2] Mortamais M, Artero S, Ritchie K. Cerebral white matter hyperintensities in the prediction of cognitive decline and incident dementia. Int Rev Psychiatry 2013. https://doi.org/10.3109/09540261.2013.838151.
[3] Trip SA, Miller DH. Imaging in multiple sclerosis. J Neurol Neurosurg & amp Amp Psychiatry 2005. https://doi.org/10.1136/jnnp.2005.073213. 76:iii11 LP-iii18.
[4] Prins ND, Scheltens P. White matter hyperintensities, cognitive impairment and dementia: an update. Nat Rev Neurol 2015. https://doi.org/10.1038/nrneurol.2015.10.
[5] Gunde E, Blagdon R, Hajek T. White matter hyperintensities from medical comorbidities to bipolar disorders and back. Ann Med 2011. https://doi.org/10.3109/07853890.2011.595733.
[6] Geissler A, Andus T, Roth M, Kullmann F, Caesar I, Held P, et al. Focal white-matter lesions in brain of patients with inflammatory bowel disease. Lancet (London, England) 1995;345:897–8.
[7] Garde E, Mortensen EL, Krabbe K, Rostrup E, Larsson HBW. Relation between age-related decline in intelligence and cerebral white-matter hyperintensities in healthy octogenarians: a longitudinal study. Lancet 2000. https://doi.org/10.1016/S0140-6736(00)02604-0.
[8] Morris Z, Whiteley WN, Longstreth WT, Weber F, Lee YC, Tsushima Y, et al. Incidental findings on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 2009. https://doi.org/10.1136/bmj.b3016.
[9] Hopkins RO, Beck CJ, Burnett DL, Weaver LK, Victoroff J, Bigler ED. Prevalence of white matter hyperintensities in a young healthy population. J Neuroimaging 2006;16:243–51. https://doi.org/10.1111/j.1552-6569.2006.00047.x.
[10] Breteler MM, van Swieten JC, Bots ML, Grobbee DE, Claus JJ, van den Hout JH, et al. Cerebral white matter lesions, vascular risk factors, and cognitive function in a population-based study: the Rotterdam study. Neurology 1994;44:1246–52. https://doi.org/10.1212/WNL.44.7.1246.
[11] de Leeuw FE, de Groot JC, Achten E, Oudkerk M, Ramos LM, Heijboer R, et al. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. The Rotterdam scan study. J Neurol Neurosurg Psychiatry 2001;70:9–14. https://doi.org/10.1136/jnnp.70.1.9.
[12] Launer LJ, Berger K, Breteler MM, Dufouil C, Fuhrer R, Giampaoli S, et al. Regional variability in the prevalence of cerebral white matter lesions: an MRI study in 9 European countries (CASCADE). Neuroepidemiology 2006;26:23–9. https://doi.org/10.1159/000089233.
[13] Wen W, Sachdev P. The topography of white matter hyperintensities on brain MRI in healthy 60- to 64-year-old individuals. Neuroimage 2004;22:144–54. https://doi.org/10.1016/j.neuroimage.2003.12.027.
[14] Kim KW, MacFall JR, Payne ME. Classification of white matter lesions on magnetic resonance imaging in elderly persons. Biol Psychiatry 2008;64:273–80. https://doi.org/10.1016/j.biopsych.2008.03.024.
[15] Debette S, Beiser A, Decarli C, Au R, Himali JJ, Kelly-Hayes M, et al. Association of MRI markers of vascular brain injury with incident stroke, mild cognitive impairment, dementia, and mortality: the Framingham offspring study. Stroke 2010. https://doi.org/10.1161/STROKEAHA.109.570044.
[16] Kuller LH, Lopez OL, Newman A, Beauchamp NJ, Burke G, Dulberg C, et al. Risk factors for dementia in the cardiovascular health cognition study. Neuroepidemiology 2003. https://doi.org/10.1159/000067109.
[17] Prins ND, Van Dijk EJ, Den Heijer T, Vermeer SE, Koudstaal PJ, Oudkerk M, et al. Cerebral white matter lesions and the risk of dementia. Arch Neurol 2004. https://doi.org/10.1001/archneur.61.10.1531.
[18] Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol 2014;13:614–29.
[19] Tosto G, Zimmerman ME, Carmichael OT, Brickman AM. Predicting aggressive decline in mild cognitive impairment: the importance of white matter

hyperintensities. JAMA Neurol 2014. https://doi.org/10.1001/jamaneurol.2014.667.
[20] Carmichael O, Schwarz C, Drucker D. Longitudinal changes in white matter disease and cognition in the first year of the Alzheimer Disease. Arch Neurol 2010;67(11):1370–8. https://doi.org/10.1001/archneurol.2010.284.
[21] Schmid R, Roob G, Kapeller P, Schmidt H, Berghold A, Lechner A, et al. Longitudinal change of white matter abnormalities. J Neural Transm Suppl 2000;59:9–14.
[22] Wahlund LO, Barkhof F, Fazekas F, Bronge L, Augustin M, Sjogren M, et al. A new rating scale for age-related white matter changes applicable to MRI and CT. Stroke 2001;32:1318–22. https://doi.org/10.1161/01.STR.32.6.1318.
[23] Fazekas F, Chawluk JB, Alavi A. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. Am J Neuroradiol 1987;8:421–6. https://doi.org/10.2214/ajr.149.2.351.
[24] Grimaud J, Lai M, Thorpe J, Adeleine P, Wang L, Barker GJ, et al. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn Reson Imaging 1996;14:495–505. https://doi.org/10.1016/0730-725X(96)00018-5.
[25] Ashton EA, Takahashi C, Berg MJ, Goodman A, Totterman S, Ekholm S. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. J Magn Reson Imaging 2003. https://doi.org/10.1002/jmri.10258.
[26] Caligiuri ME, Perrotta P, Augimeri A, Rocca F, Quattrone A, Cherubini A. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. Neuroinformatics 2015;13:261–76. https://doi.org/10.1007/s12021-015-9260-y.
[27] Schmidt P, Pongratz V, Küster P, Meier D, Wuerfel J, Lukas C, et al. Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. NeuroImage Clin 2019. https://doi.org/10.1016/j.nicl.2019.101849.
[28] Schmidt P, Gaser C, Arsic M, Buck D, Förschler A, Berthele A, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. Neuroimage 2012;59:3774–83.
[29] Maldjian JA, Whitlow CT, Saha BN, Kota G, Vandergriff C, Davenport EM, et al. Automated white matter total lesion volume segmentation in diabetes. Am J Neuroradiol 2013. https://doi.org/10.3174/ajnr.A3590.
[30] Galluzzi S, Marizzoni M, Babiloni C, Albani D, Antelmi L, Bagnoli C, et al. Clinical and biomarker profiling of prodromal Alzheimer's disease in workpackage 5 of the innovative medicines initiative PharmaCog project: a "European ADNI study.". J Intern Med 2016;279:576–91.
[31] Jovicich J, Marizzoni M, Sala-Llonch R, Bosch B, Bartr??s-Faz D, Arnold J, et al.. Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. Neuroimage 2013;83:472–84.
[32] Marizzoni M, Antelmi L, Bosch B, Bartr??s-Faz D, M??ller BW, Wiltfang J, et al.. Longitudinal reproducibility of automatically segmented hippocampal subfields: a multisite European 3T study on healthy elderly. Hum Brain Mapp 2015;36:3516–27. https://doi.org/10.1002/hbm.22859.
[33] Jovicich J, Marizzoni M, Bosch B, Bartr??s-Faz D, Arnold J, Benninghoff J, et al.. Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects. Neuroimage 2014;101:390–403. https://doi.org/10.1016/j.neuroimage.2014.06.075.
[34] Jovicich J, Minati L, Marizzoni M, Marchitelli R, Sala-Llonch R, Bartr??s-Faz D, et al.. Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: a multicentric resting-state fMRI study. Neuroimage 2016;124:442–54. https://doi.org/10.1016/j.neuroimage.2015.07.010.
[35] Marchitelli R, Minati L, Marizzoni M, Bosch B, Bartr??s-Faz D, M??ller BW, et al.. Test-retest reliability of the default mode network in a multi-centric fMRI study of healthy elderly: effects of data-driven physiological noise correction techniques. Hum Brain Mapp 2016;37:2114–32. https://doi.org/10.1002/hbm.23157.
[36] Schmidt P, Gaser C, Arsic M, Buck D, Förschler A, Berthele A, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. Neuroimage 2012. https://doi.org/10.1016/j.neuroimage.2011.11.032.
[37] Schmidt P. Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. 2017.
[38] Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 1994. https://doi.org/10.1109/42.363096.
[39] Dice LR. Measures of the amount of ecologic association between species. Ecology 2006. https://doi.org/10.2307/1932409.
[40] de Sitter A, Steenwijk MD, Ruet A, Versteeg A, Liu Y, van Schijndel RA, et al. Performance of five research-domain automated WM lesion segmentation methods in a multi-center MS study. Neuroimage 2017;163:106–14. https://doi.org/10.1016/j.neuroimage.2017.09.011.
[41] Damangir S, Manzouri A, Oppedal K, Carlsson S, Firbank MJ, Sonnesyn H, et al. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. J Neurol Sci 2012. https://doi.org/10.1016/j.jns.2012.07.064.
[42] Damangir S, Westman E, Simmons A, Vrenken H, Wahlund L-O, Spulber G. Reproducible segmentation of white matter hyperintensities using a new statistical definition. Magn Reson Mater Phys Biol Med 2016. https://doi.org/10.1007/s10334-016-0599-3.
[43] Shiee N, Bazin PL, Ozturk A, Reich DS, Calabresi PA, Pham DL. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage 2010. https://doi.org/10.1016/j.neuroimage.2009.09.005.
[44] Steenwijk MD, Pouwels PJW, Daams M, Van Dalen JW, Caan MWA, Richard E, et al. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage Clin 2013;3:462–9. https://doi.org/10.1016/j.nicl.2013.10.003.
[45] Egger C, Opfer R, Wang C, Kepp T, Sormani MP, Spies L, et al. MRI FLAIR lesion segmentation in multiple sclerosis: does automated segmentation hold up with manual annotation? NeuroImage Clin 2017;13:264–70. https://doi.org/10.1016/j.nicl.2016.11.020.
[46] Admiraal-Behloul F, Van Den Heuvel DMJ, Olofsen H, Van Osch MJP, Van Der Grond J, Van Buchem MA, et al. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. Neuroimage 2005;28:607–17. https://doi.org/10.1016/j.neuroimage.2005.06.061.
[47] Anbeek P, Vincken KL, Van Osch MJP, Bisschops RHC, Van Der Grond J. Probabilistic segmentation of white matter lesions in MR imaging. Neuroimage 2004. https://doi.org/10.1016/j.neuroimage.2003.10.012.
[48] Griffanti L, Zamboni G, Khan A, Li L, Bonifacio G, Sundaresan V, et al. BIANCA (brain intensity AbNormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. Neuroimage 2016;141:191–205. https://doi.org/10.1016/j.neuroimage.2016.07.018.
[49] Ong KH, Ramachandram D, Mandava R, Shuaib IL. Automatic white matter lesion segmentation using an adaptive outlier detection method. Magn Reson Imaging 2012. https://doi.org/10.1016/j.mri.2012.01.007.
[50] Manjón JV, Coupé P, Raniga P, Xia Y, Fripp J, Salvado O. HIST: HyperIntensity Segmentation Tool. Cham: Springer; 2016. p. 92–9. https://doi.org/10.1007/978-3-319-47118-1_12.
[51] Wang Y, Catindig JA, Hilal S, Soon HW, Ting E, Wong TY, et al. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. Neuroimage 2012. https://doi.org/10.1016/j.neuroimage.2012.02.034.