medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Comparison between tools for automatic segmentation of white matter hyperintensities of presumed vascular origin in aging: which one, how and why to choose the most suitable for your purpose?

Lucia Torres-Simon[1,2]*; Alberto del Cerro-León[1,2]*; Miguel Yus[3,4]; Ricardo Bruña[1,3,5]; Lidia Gil-Martinez[6]; Alberto Marcos Dolado[3,7,8]; Fernando Maestú[1,2,3]; Juan Arrazola-Garcia[3,4,9]; Pablo Cuesta[1,3,5]

[1] Center of Cognitive and Computational Neuroscience, Universidad Complutense de Madrid (UCM), Madrid, Spain; [2] Department of Experimental Psychology, Cognitive Processes and Speech Therapy, Universidad Complutense de Madrid (UCM), Madrid, Spain; [3] Health Research Institute of the Hospital Clínico San Carlos (IdISSC), Madrid 28040, Spain; [4] Department of Diagnostic Imaging, Hospital Clínico San Carlos, Madrid 28040, Spain; [5] Department of Radiology, Complutense University of Madrid, 28040 Madrid, Spain; [6] Foundation for Biomedical Research at Hospital Clínico San Carlos (FIBHCSC), Hospital Clínico San Carlos, Madrid 28040, Spain; [7] Department of Medicine, School of Medicine, Complutense University of Madrid, Madrid 28040, Spain; [8] Department of Neurology, Hospital Clínico San Carlos, Madrid 28040, Spain; [9] Department of Radiology, Rehabilitation and Radiation Therapy, School of Medicine, Complutense University of Madrid, Madrid 28040, Spain.

* These authors contributed equally to this work

Corresponding autor: Alberto del Cerro-Leon, aldelcer@ucm.es
Facultad de Psicología, Campus de Somosaguas, 28223 Pozuelo de Alarcón

Key words: White matter hyperintensities, automatic detection tools, cerebrovascular disease, aging

Abstract

Cerebrovascular damage consequent to small vessel disease (SVD) is a common companion of healthy and pathological aging. Neuroimaging signatures of SVD are present in virtually all people older than 60 years, and their prevalence increases with age. According to the STRIVE criteria, white matter hyperintensities (WMH) have been associated with different clinical symptoms and constitute a good clinical proxy for SVD. This is important as WMH can be directly assessed via MRI. Currently, the most widely used method to detect and assess the severity of HWH are clinical scales based on visual assessment, but these scales do not offer real quantitative information, making it difficult to assess progression. Quantitative information can be approached through the manual segmentation of physicians, but this process is extremely time-consuming and presents a high inter and intra evaluators variability, which makes its application in routine protocols unfeasible. Therefore, it is imperative to facilitate the use of automatic protocols capable of providing WMH load measurements that are as accurate as possible to those obtained by manual segmentation. In this study, we aim to identify the most accurate software for WMH segmentation, providing not only methodological insights but also usability knowledge that would enlighten the tradeoff between clinical accuracy and real-world implementation. The data set consisted of T1 and Flair images of 45 cognitively healthy older participants (mean age 71 ± 5). The study analyzed WMH segmentations obtained with clinician manual segmentation and four tools included in three of the most widely used neuroimaging toolkits: 1) Lesion Prediction Algorithm (LPA) and Lesion Growth Algorithm (LGA) of the lesion segmentation tool (LST); 2) sequence adaptive multimodal segmentation (SAMSEG); and 3) the brain intensity anomalies classification algorithm (BIANCA). The analysis evaluated the correlations with the Fazekas clinical scale, the influence of the WMH load and evaluated the performance at the individual lesion level. The results showed that the supervised methods (LST-LPA and BIANCA) performed better in all the analyzes and were the only ones capable of consistently capturing small lesions (<26 mm3). However, these tools lose performance when applied to new data. Considering the results (accuracy and ease of use), we concluded that, in the general case, the combined use of the two FSL tools emerged as the best option. We confirmed this conclusion by evaluating WMH segmentation in a dataset of 500 older individuals and found that the LPA results, using LGA to control for divergences, offered valuable and actionable clinical information, both to help clinicians make treatment decisions treatment and to monitor pathological progression.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1 Introduction

White matter hyperintensities (WMH) can be observed in magnetic resonance image (MRI) in a large diversity of both congenital and acquired processes (i.e., autoimmune, infectious, vascular, and toxic-metabolic processes), being small vessel disease (SVD) the most leading underlying cause (Sarbu et al., 2016). These WMH, of presumed vascular origin, are a common accompaniment of ageing, both in dementia and healthy population (Alber et al., 2019; Van Leijsen et al., 2017), being present on images from most individuals over 60 years old (Drebette & Markus, 2010; Kloppenborg et al., 2014). The clinical relevance of WMH has been widely reported in the literature, as it is closely related to the incidence of all types of dementia, especially vascular cognitive impairment (VCI), and to the lifetime risk of stroke (Chutinet & Rost, 2014; Drebette & Markus, 2010). Moreover, WMH total volume in the brain has been associated with severity of cognitive symptoms, progression of disability, and clinical outcome (Bendfeldt et al., 2010; Guerrero et al., 2018). In this way, the presence and progression of WMH in the brain have been linked, with small but robust effects, to cognitive impairment in both the healthy old population (Arvanitakis, Fleischman, et al., 2016; Kloppenborg et al., 2014) and dementia patients (Arvanitakis, Capuano, et al., 2016; Lam et al., 2021; Mortamais et al., 2014; Prins & Scheltens, 2015; Van Den Berg et al., 2018), mainly affecting attentional processes and executive function. Additionally, several studies have depicted the consequences of WMH in mental health disruption (i.e., depression or anxiety) (Anor et al., 2021; Misquitta et al., 2020). Beyond this traits, WMH in aging have been also broadly connected to many other physiopathological signatures; like amyloidosis and tau-pathology (Caballero et al., 2018; Gaubert et al., 2021; Graff-Radford et al., 2019; Soldan et al., 2020); grey matter degeneration (Gaubert et al., 2021; Jang et al., 2017); low cortical thickness (Duering et al., 2012); ventricular dilatation (Bjerke et al., 2014); structural connectivity disruption as assessed with DTI (Veldsman et al., 2020); functional connectivity alterations as measured with fMRI (Schulz et al., 2021) and electrophysiology, specifically using EEG (Quandt et al., 2020; van Straaten et al., 2012, 2015); and genetic risk factors as the ApoE allele ε4 (Silbert et al., 2009). Consequently, an objective and efficient assessment of the presence and progression of WMH is paramount in both the scientific and clinical context to early detection and follow-up of people at risk.

According to the Standards for Reporting Vascular changes on nEuroimaging (STRIVE), WMH are seen as diffuse areas of high signal intensity (hence, "hyperintense") on T2-weighted or fluid-attenuated inversion recovery (FLAIR) MRI (Wardlaw et al., 2013). Thus, MRI is enabled to detect lesions in the white matter, allowing the diagnosis and monitoring of the structural deterioration (Blystad et al., 2016; Garcia-Lorenzo et al., 2011). Nowadays, clinical visual assessment is the most used method to detect and evaluate the severity of WMH, being the Fazekas scale (Fazekas et al., 1987) and the age-related white matter changes scale (ARWMC) (Wahlund et al., 2001) the most recommended visual rating scales (Wahlund et al., 2017). Nevertheless, these scales do not report true quantitative information (e.g., lesion volume or lesions localizations), which makes the evaluation of the WMH progression difficult to quantify when the changes are not abrupt. Exhaustive evaluation for number, volume, location, and distribution of WMH on MRI may provide crucial information on etiology, prognosis, and progression of diseases and, eventually, would be able to help in assessing the effectiveness of potential treatments (Balakrishnan et al., 2021; Qin et al., 2018). In this context, manual segmentation by a clinician is the most accurate method to measure and quantify the lesions in the white matter, and therefore it is considered as the gold standard (Commowick et al., 2018; Heuvel et al., 2006).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Notwithstanding, segmenting the WMH manually is extremely time-consuming, and it comprises high inter- and intra-rater variability, features that make it impractical for clinical and scientific use. For this reason, automatic segmentation tools are increasingly on demand, and different research groups have developed, during the last decade, several automatic WMH segmentation algorithms that vary in requirements, sensitivity, accuracy, runtime, and user accessibility. These WMH automatic segmentation tools can be basically divided according to whether they use supervised or unsupervised methods (García-Lorenzo et al., 2013). On the one hand, unsupervised methods elaborate probabilistic maps to assign a label (white matter, gray matter, cerebrospinal fluid, or WMH) to each voxel (Schmidt et al., 2012). On the other hand, supervised methods require a previously segmented training set to define what is considered as WMH (Griffanti et al., 2016; Schmidt, 2016). Both type of tools has been successfully developed for automatic WMH segmentation but are specially focused in multiple sclerosis patients (Egger et al., 2017; García-Lorenzo et al., 2013). Nevertheless, as WMH signal intensity and spatial distributions vary, displaying unique signatures for each disease (Balakrishnan et al., 2021), it becomes important to study the performance of these tools for WMH of presumed vascular origin. Considering this, a systematic review, covering fully automated methods applied for normal ageing and patients with vascular pathology from 1980 to 2014, can be found in Caliguiri and colleagues (Caligiuri et al., 2015). Extending up to 2016, two reviews were published focused on the discussion of different approaches, both for segmenting WMH and for assessing other neuroimaging markers for small-vessel disease (SVD) (Blair et al., 2017; Wardlaw et al., 2015). Finally, to assess and overview those fully automatic computational methods developed to segment WMH of presumed vascular origin, another review was carried, involving studies from 2015 to 2020 (Balakrishnan et al., 2021). Nevertheless, the heterogeneity in the toolboxes evaluated, the samples size and its characterization, and the methodological perspective of the majority of studies included in these reviews, drive to somehow vague conclusions that are far from being useful for actual clinical practice.

In this concern, the aim of the present piece of work is to develop a comprehensive methodological study, including the three more extended, freely available, and widely used neuroscience packages (SPM12, Freesurfer, and FSL) to facilitate the selection of an algorithm depending on user requirements, to maximize either clinical or research applications. For that purpose, we calculated the WMH volume using four automatic algorithms (Lesion growth algorithm –LGA–, Lesion prediction algorithm –LPA–, Sequence Adaptive Multimodal SEGmentation –SAMSEG– and Brain Intensity AbNormality Classification Algorithm –BIANCA–) in a cohort of 45 cognitively healthy older adults (i.e., over 64 years old) with different loads of WMH of presumed vascular origin, according to the clinical report. In addition to this, all 45 images were manually segmented by neuroradiology experts, and these segmentations were used as gold standard. We analyzed the correlation between the total volume of WMH obtained by each algorithm and both the gold standard and the Fazekas scale. Then, we compared the performance (i.e., sensitivity, precision, and dice score) of each tool compared with the manual segmentation. Additionally, we extend our analysis by evaluating the algorithms' performance depending on the nature of the methodology (supervised vs unsupervised), the amount of WMH, and the individual lesion size. Finally, we included a usability analysis to facilitate the selection of the methods and their applicability. After this assessment, we applicated the algorithms with the best results to a real-world scenario database using an ageing sample of 577 individuals.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2 Materials & Methods

2.1 Participants

The sample of the present study was recruited under the framework of a Spanish-government-funded project (PSI2012-38375-C03-01) focused on research and early detection of dementia. Data were collected between 2013 and 2015 in collaboration with three different clinical centers located in Madrid (Spain): the Neurology Department in "Hospital Universitario Clínico San Carlos," the "Center for Prevention of Cognitive Impairment," and the "Seniors Center of Chamartín District". The sample consisted of 156 participants, age ≥ 50, native Spanish speakers and cognitively healthy, with a score ≥ 26 in the Mini Mental State Examination (MMSE) (Lobo et al., 1979). To assess their general cognitive and functional status, the following set of screening questionnaires were also used: the Global Deterioration Scale (Reisberg et al., 1982), the Geriatric Depression Scale–Short Form (GDS) (Yesavage et al., 1982); the Functional Assessment Questionnaire (FAQ) (Pfeffer et al., 1982); and the questionnaire for Instrumental Activities of Daily Living (Lawton & Brody, 1969). Exclusion criteria included: (1) a mild cognitive impairment (MCI) diagnosis, according to the criteria established by Petersen (Petersen & Negash, 2008); (2) a history of psychiatric or neurological disease; and (3) the use of psychoactive drugs or chronic medication, such as anxiolytics. In addition, we performed, on every participant, a complementary exploration (class II evidence level) to rule out possible causes of cognitive decline, such as B12 vitamin deficit, diabetes mellitus or thyroid problems.

Additionally, all participants underwent an MRI study (T1 and FLAIR sequences). To avoid possible confounding variables, we excluded all the participants with evidence of any brain abnormalities described in the radiologists' report. Out of the 156 initial participants, 33 were excluded due to some degree of brain atrophy, 4 for tumor presence, and 7 for a lack of clinical report. Finally, from the 112 participants, who meet the inclusion and exclusion criteria, we randomly chose 45 according to their corresponding clinical reports: 15 with no alterations in their MRI, 15 with subclinical WMH, and 15 with WMH presumably associated with cerebrovascular disease (CBVD), to ensure that the images of the selected participants included a mixed range of WMH loads, covering a broad spectrum. The demographic and clinical data at baseline evaluation for each participant are described in Table 1.

	N	Age	Sex	Education	MMSE	Fazekas	Volume WMH
Whole sample	45	71 ± 5	30/15	14 ± 6	29 ± 1	0.9 ± 0.8	$(5 \pm 7) \cdot 10^3$
No WMH	15	69 ± 5	11/4	12 ± 5	28 ± 1	0.2 ± 0.4	$(1 \pm 2) \cdot 10^3$
Subclinical WMH	15	71 ± 5	11/4	13 ± 6	29 ± 1	0.7 ± 0.5	$(2 \pm 1) \cdot 10^3$
CBVD WMH	15	72 ± 4	8/7	15 ± 6	28 ± 2	1.8 ± 0.6	$(13 \pm 8) \cdot 10^3$

Table 1. *Demographic and clinical data. Values are presented as mean ± standard deviation. Sex (female/male), age (in years), education (expressed in years of education). MMSE: Mini Mental State Examination. WMH volume (in mm^3) corresponded with the volumes obtained with the radiologist's manual segmentation. Demographics are displayed for the whole sample and for each subsample of interest.*

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.2 MRI acquisition

Each participant was subjected to an MRI assessment that included the acquisition of a 3D T1-weighted and a 3D T2-weighted FLAIR images, obtained with a General Electric 1.5 Tesla magnetic resonance scanner. The images were acquired using a high-resolution antenna and a homogenization PURE filter (Fast Spoiled Gradient Echo sequence). The parameters for the T1 image were repetition time (TR) = 11.2 ms, echo time (TE) = 4.2 ms, inversion time (TI) = 450 ms, Field Of View (FOV) = 25 cm, flip angle (FA) = 12°, 252 coronal slices (in-plain resolution: 256×256), voxel size: 0.98 x 0.98 x 1 mm3 and acquisition time ≃ 8:00 min. The T2-weighted 3D FLAIR images were obtained with the following specifications: TR = 7000 ms, TE = 101 ms, TI = 2112 ms, FOV = 24 cm, 252 sagittal z-axis interpolated slices (in-plain resolution: 256×112), voxel size: 0.94 x 0.94 x 1.6 mm3 and acquisition time ≃ 4:57 min.

2.3 MRI clinical assessment: Fazekas Scores and WMH manual segmentation

The FLAIR images of each participant were co-registered with their respective T1 images to define a single space per participant (Figure 2, 1). Subsequently, the radiology team of the *Hospital Universitario San Carlos* used the FLAIR images to manually segment the WMH of each subject, relying on the T1 image when necessary. All segmentations were performed using FSLeyes (McCarthy, 2018) on the same computer with a monitor resolution of 1920x1080 pixels. Once completed, 2 senior members of the radiology team reviewed the segmentations and applied a common criterion when discrepancies appeared, in order to establish a reliable gold standard. The manually segmented masks were clustered using a three-dimensional 26-connected neighborhood to obtain the individual lesions of each subject. Simultaneously, the clinicians assigned a Fazekas score (0-3) to each participant (Figure 1).

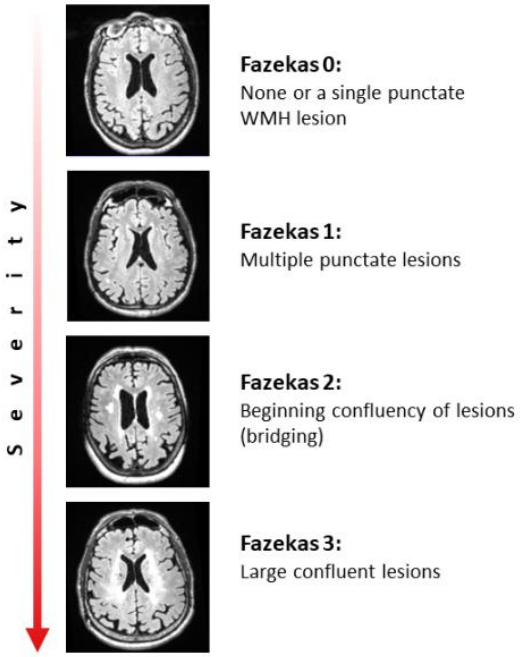


Figure 1. Explanation of the Fazekas scale, from top to bottom the different degrees of severity are detailed from lowest to highest.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.4 Automatic WMH segmentation

The WMH burden for the 45 participants were calculated, using the FLAIR and/or T1 images, by means of 4 automatic segmentation algorithms included in three of the most used neuroscience toolboxes (Figure 2, 2): 1) for the SPM12 toolbox, we employed the algorithm Lesion Segmentation Tool (LST), with two different procedures: Lesion Predictor Algorithm (LST-LPA) and with Lesion Growth Algorithm (LST-LGA); 2) for the FreeSurfer toolbox, we used the Sequence Adaptive Multimodal SEGmentation (SAMSEG) and 3) for FSL, we computed the WMH segmentation through the Brain Intensity AbNormality Classification Algorithm (BIANCA). In all cases, and before starting the assessment of the performance, we analyzed the execution of each tool across their different parameters. For LST-LPA, LST-LGA and SAMSEG, we evaluated their performance at different thresholds (minimum probability criterion to accept a given voxel as lesion), and in the case of BIANCA we tested the algorithm with different training sets (for more details see Supplementary materials 1 and 2). Additionally, as previous studies (Hotz et al., 2021) recommended FreeSurfer's automatic segmentation of WMH over the T1 image, we also include it we Supplementary material 4. Following, we describe each procedure in detail.

SPM12 → Lesion Segmentation Tool (LST)

LST is an open-source toolbox for SPM12 with the capability of segmenting WMHs. We included the two algorithms provided within this toolbox: the Lesion Growth Algorithm (LGA) and the Lesion Prediction Algorithm (LPA).

A. **Lesion Prediction Algorithm (LST-LPA):** LPA automatically segments the WMHs using only a FLAIR image and with no other input from the user. It is a supervised method trained using a logistic regression model from the data of 53 multiple sclerosis patients, whose images are included in the package. The algorithm uses the binary masks and performs a logistic regression (with the spatial and the lesion probability maps as covariates) to compute the voxel-specific changes in lesion probability. When assessing new images, the model uses these parameters to identify the lesions, providing an estimation of the lesion probability for each voxel (for more details see Schmidt et al., 2016. Chapter 6.1). Then, the probability maps were binarized by applying a threshold of 0.5.

B. **Lesion Growth Algorithm (LST-LGA):** LGA-based segmentation requires the use of both T1 and FLAIR images. It is an unsupervised method, which starts by assigning each voxel of the native T1 image to one of three tissue types (grey matter- GM-, white matter – WM- or cerebrospinal fluid -CSF-). Subsequently, the FLAIR hyperintensities outliers of each tissue category are weighed by the spatial probability of being WM, in order to obtain lesion belief maps (BGM, BWM, BCSF). These three lesion belief maps are added together (B), and the BGM is binarized by a pre-chosen threshold (κ), which is used as a seed to elaborate an initial lesion map (Linit). Finally, a topological growth model expands the Linit toward the lesion belief map, generating a probability lesion map in which a threshold can be applied to obtain a binary mask of the WMHs (for more details see Schmidt et al., 2012). For the automatic segmentation of the 45 participants selected by LST-LGA, we set the initial κ threshold to the default value 0.3 and a binarization threshold to 0.5.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

FreeSurfer

C. **FreeSurfer Image Analysis Suite**: FreeSurfer has an automatic structural segmentation tool able to assess those regions that might be including WMH (Fischl, 2012). The method uses Bayesian labeling based on probabilistic local and intensity-related information to tag a voxel as WMH. This information is extracted from a training set of 41 manually segmented images. In our study, the T1 image of each subject was processed using the FreeSurfer version 6.0.1. The WMH masks were reshaped into the original dimensions of the original T1 images.

D. **Sequence Adaptive Multimodal SEGmentation (SAMSEG):** SAMSEG is an open-source software toolbox included in the FreeSurfer neuroimaging analysis package. It is an unsupervised toolbox based on a generative approach, in which a probabilistic forward model is inverted to perform automatic segmentation of multiple anatomical regions from multi-contrast MRI data without any prior assumptions on the specific pulse sequences or the type of scanner applied. In order to identify white matter lesions, SAMSEG augments his generative model adding a binary lesion map and a latent variable h, that constrain the shape of lesions, to a joint likelihood that led the segmentation process (for more details see Cerri et al., 2021). Once completed, a probability map is generated for each segmented anatomical structure, including white matter lesions. Finally, the probability maps are binarized applying a 0.5 threshold.

FMRIB Software Library (FSL)

E. **Brain Intensity AbNormality Classification Algorithm (BIANCA):** BIANCA is an automatic and supervised method for the detection of WMH included in the FSL package. It performs identification of WMHs through the k-nearest neighbor (k-NN) algorithm. However, BIANCA has the ability to use spatial information and dataset-specific training to adapt the segmentation protocol. The algorithm requires a training sample composes by manually segmented lesion masks. Thus, spatial and intensity information are extracted from the training data set and then used to produce a probability lesion map in the new image (for more details see Griffanti et al., 2016). In the present study, we used as training data the manual lesion probability maps for each participant in the original dataset, not including the 45 aforementioned individuals. The parameters used were location of training points = "noborder", 2000 training points, 10000 non lesion points and patch 3D option. The probability maps were binarized by applying a threshold of 0.95.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

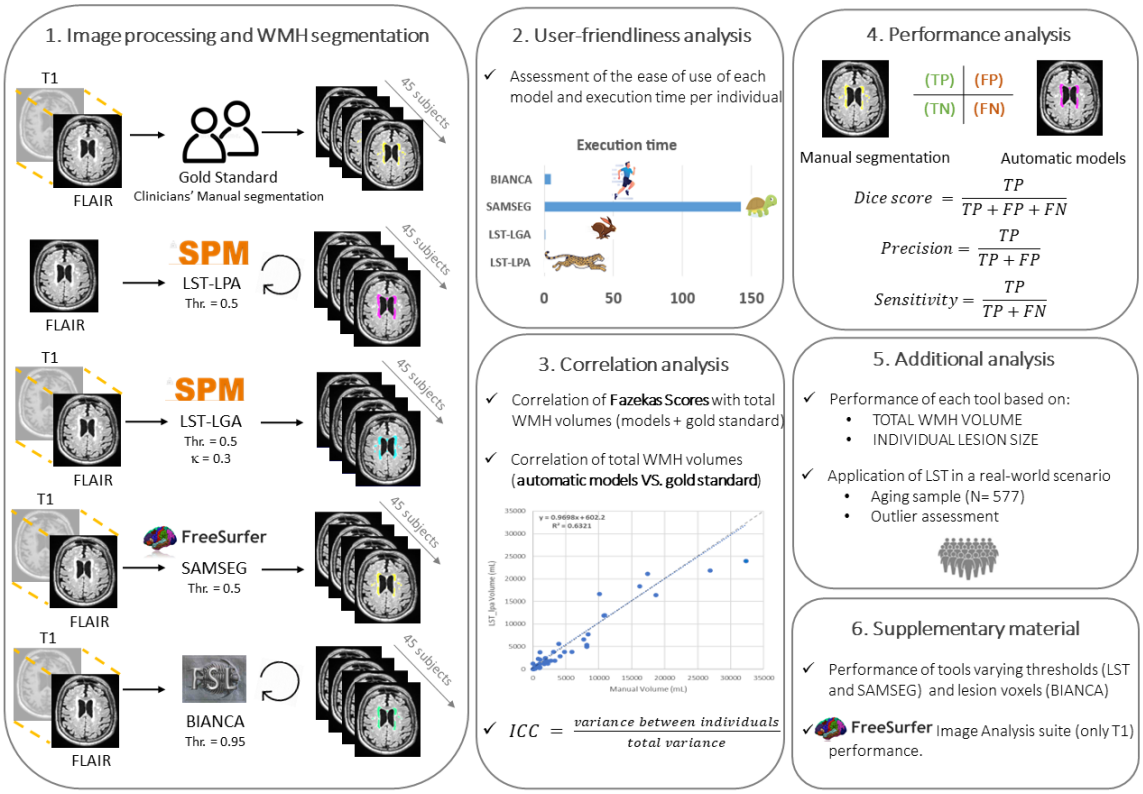


*Figure 2. Methodological pipeline. From left to right: **1:** 3D T1 weighted and 3D FLAIR images co-registration, when required. Optimization of each algorithm performance and computation of the WMH masks of each participant. The circular arrows represent the algorithms that needs a training set. **2:** Assessment of User-friendliness by means of 3 expert neuroscientists. **3:** Battery of comparison based on performance and clinical relevance. **4:** comparison of the WMH masks computed by each algorithm with the corresponding of the gold standard. **5:** Evaluation of the models' performance based on different WMH volumes, when evaluating at the individual lesion level and when testing the algorithm in a large dataset **6**: Supplementary analysis concerning the check about the initial thresholds and lesion voxel for the algorithms analyzed and the results of the FreeSurfer Image analysis performance. The standard WMH segmentation obtained by Freesurfer recon all is described in the supplementary materials.*

2.5 Metrics and Statistical analysis

Algorithms' User-friendliness

To address the user-friendliness (Figure 2, 2) of each algorithm, 3 neuroimaging experts (L.T-S, A.C-L and P.C.) used each of the tools and indicated whether they had encountered any of the following difficulties during their execution: 1) *installation packages* (if the algorithm is included or it requires one or more additional packages); 2) *file preparation* (it makes reference to the images pre-processing required before the segmentation); 3) *brain extraction* (if it is required before segmentation process); 4) *training set* (if it is required for the algorithms and when it is, if this set is included or the user should provide its own); 5) *interface* (if there is an user-friendly interface or the user should work in the terminal environment); 6) *RAM requirements* (amount of available RAM required to run the algorithms); and 7) *execution time* (it was calculated for each algorithm as the corresponding time (in minutes) expended for the computation of the WMH masks, without considering the images or dataset preparations and/or preprocessing).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Correlation analysis

Correlation of WMH volumes and Fazekas Scores: quantitative vs. qualitative approach

Spearman correlations were performed between WMH total volumes (derived from the automatic algorithms and the gold standard) and the Fazekas scores, with the aim to evaluate how near the WMH quantification throughout the segmentation methods (either the automatic ones or the gold standard) were from the most recommended and used visual rating scales in the clinical practice (Figure 2, 3).

Correlations among WMH total volumes

The total lesion volume obtained from each algorithm was compared to the volumes derived from the gold standard, defined by the manual segmentations of the clinicians, through a series of Spearman correlation analysis (Figure 2, 3).

Interclass Correlation Coefficient (ICC)

As a metric for the reliability of the results, the ICC was computed, by dividing the variance between measurements by the total variance of the measurements. A high ICC indicates high consistency and agreement between the two measurements, while a low ICC indicates variability and heterogeneity in the values (Figure 2, 3).

Algorithms' performance analysis compared with gold standard.

To evaluate the quality of the segmentation carried out by each algorithm, the corresponding binary masks were compared with the gold standard (Figure 2, 4), accounting for matching results (True positives (TP) or True negatives (TN)) and for unmatching ones (False positives (FP) or False Negatives (FN)). TP, TN, FP and FN were extracted for each mask and were used to calculate 3 measures of the reliability of each segmentation: 1) Dice score: degree of similarity between automatic segmentation and the gold standard mask (Dice, 1945); 2) precision: percentage of the segmented lesion that corresponds to an real lesion; and 3) sensitivity: percentage of the lesion present in the gold standard captured by the automatic tool. The mathematical expression per each measure is depicted in the Figure 2, 5. These scores were obtained for participant, obtaining three series of 45 values. The assessment of differences on performance among tools was carried out using a paired t-test. Effect sizes were obtained by means of Hedges-corrected Cohen's D.

Additional analysis

Performance based on WMH volume and individual lesion size.

A series of regression analyzes were performed to observe the effect of total lesion volume on the precision, sensitivity, and Dice score values per each tool (Figure 2, 6). With the aim of evaluating the performance of the tools as a function of individual lesion size, all manual segmentations of the gold standard were clustered using a labeling algorithm based on elements connected in 26 directions in a binary matrix. Only those lesions with a minimum of 4 voxels were considered. Then,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

all lesions of all participants were pooled together creating a data set of 1960 lesions. The analysis consisted of evaluating the sensitivity at each lesion separately. The results were then stratified by lesion volume.

Application of LST algorithms in a real-world scenario and outliers' detection

The performance of the LST-LGA and LST-LPA algorithms was tested in a real-world scenario by assessing the total WMH in a large dataset of 577 older individuals (mean age 68.7 ± 8.6; 384 females/192 males). This large data set comes from different projects developed in the Center for Cognitive and Computational Neuroscience (C^3N) in Madrid. For each participant, we computed the error (squared difference) between the WMH volumes resulted from both algorithms. Subsequently, we calculated the mean and standard deviation of these errors across the whole sample, with an outlier threshold set as 3 standard deviations above the mean. Those cases where the discrepancies between the WMH volumes obtained by the two methods exceeded the threshold were manually reviewed by a neuroimaging expert.

Data availability

The data that support the findings of this study are available from the corresponding author, upon request. All the algorithms used in the present paper are reported in the 'Materials and methods' section.

3 Results

The performance of each WMH segmentation algorithm was evaluated following the structure described in the methods section. This sequence was established to monitor the steps that a researcher must follow to acquire the ability to use the methods, understand the clinical implication of the results obtained, and be able to apply them in a population of older adults.

3.1 User-friendliness analysis

To assess the ease of use and implementation of each tool, 3 neuroimaging experts evaluated the items described in the material and methods section based on their experience (Table 2). In the development of the automatic masks, the procurement and installation of all the tools could be achieved without difficulty. Whilst most tools did not require the installation of additional packages, SAMSEG required the installation of *Tensorflow*, an open-source machine learning library. Regarding the preparation process for each tool, BIANCA was the most demanding, since its operation requires the preparation of an external file to guide the process, a training sample with binary masks of the lesions, and a brain mask. In the experience of the evaluators, ensuring the quality of the brain mask usually required manual intervention. As for the user interface, whereas BIANCA and SAMSEG should be executed from the command line, the LST package has an easy-to-follow user interface, which makes LST-LGA and LST-LPA the most accessible tools. Most of the tools run smoothly under standard computational specifications (in terms of CPU and RAM) with the exception of SAMSEG, whose computations required up to 32 GB of RAM. Finally, in terms of execution times, LST-LPA was the fastest tool (t = 0.4 ± 0.2 min) closely followed by LST-LGA (t =

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

0.6 ± 0.3 min). In a middle range emerged BIANCA (t = 4.8 ± 0.5 min), and the slowest tool was SAMSEG (140 ± 30 min).

Algorithms' User-friendliness

	LST-LPA	LST-LGA	SAMSEG	BIANCA
Installation packages	LST toolbox	LST toolbox	TensorFlow librery	Included in FSL
File preparation	N/A	T1+Flair corregistration	T1+Flair corregistration	T1+Flair Trainning set Master file
Brain Extraction	N/A	N/A	N/A	Required
Training set	Included	N/A	N/A	Own set Required
Interface	SPM User interface	SPM User interface	Linux terminal	Linux terminal
RAM requirements	<8Gb	<8Gb	>16Gb	<8Gb
Execution time (minutes)	0.4 ± 0.2	0.6 ± 0.3	140 ± 30	4.8 ± 0.5

Table 1. Difficulty of using each tool in terms of need to install additional packages, file preparation, brain extraction, manual intervention, need for training sets, user interface, RAM specifications, and runtime performance.

3.1. Correlation analysis

In order to check whether all tools depicted a reasonable relationship with a standard clinical scale of vascular damage, we performed a series of Spearman correlations between the total WMH volume obtained by each tool (and the one corresponding to the gold standard) and the Fazekas scores determined by the clinicians (for more details see **Supplementary material 3**). The results showed that the gold standard showed the higher association (R^2 = 0.84) with the Fazekas scale. Notwithstanding, all measures showed significant correlations (R^2 > 0.78), being the results obtained with BIANCA the ones with the highest correlation (R^2 = 0.80) among the automatic tools (see Table 3).

	LST-LPA	LST-LGA	SAMSEG	BIANCA	Gold Standard
Fazekas	0.78	0.78	0.79	0.80	0.84
Gold Standard	0.62	0.89	0.86	0.91	N/A
ICC	0.91	0.88	0.99	0.93	N/A

Table 2. Fazekas: R^2 values of the Spearman correlation between the total WMH volumes measured by each tool and the one corresponding to the gold standard with the Fazekas scale. Gold Standard: R^2 values of the Spearman correlation between the total WMH volumes measured by each tool with the one corresponding to the gold standard. ICC: intra class correlation coefficient per each tool.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The evaluation of the possible relationships between the total volume of WMH of the gold standard with those obtained through each automatic tool (Figure 3 and Table 3) was carried out by performing a series of Spearman correlation analyses. The results showed that all methods underestimated (slope < 1) the total WMH offered by the gold standard. The WMH volumes obtained with the LST-LPA algorithm were the ones that better matched (depicting a slope of 0.97) the corresponding WMH volumes of the gold standard. The LST-LGA and SAMSEG algorithms obtained good fits, with slopes of 0.89 and 0.86, respectively. Finally, BIANCA obtained the result furthest from the gold standard, with a slope of 0.73. Regarding the adjustment of the data to each model, BIANCA was the tool with the best fit to its linear regression (R^2 = 0.91), followed by LST-LGA (R^2 = 0.89), SAMSEG (R = 0.86) and LST-LPA (R^2 = 0.62).

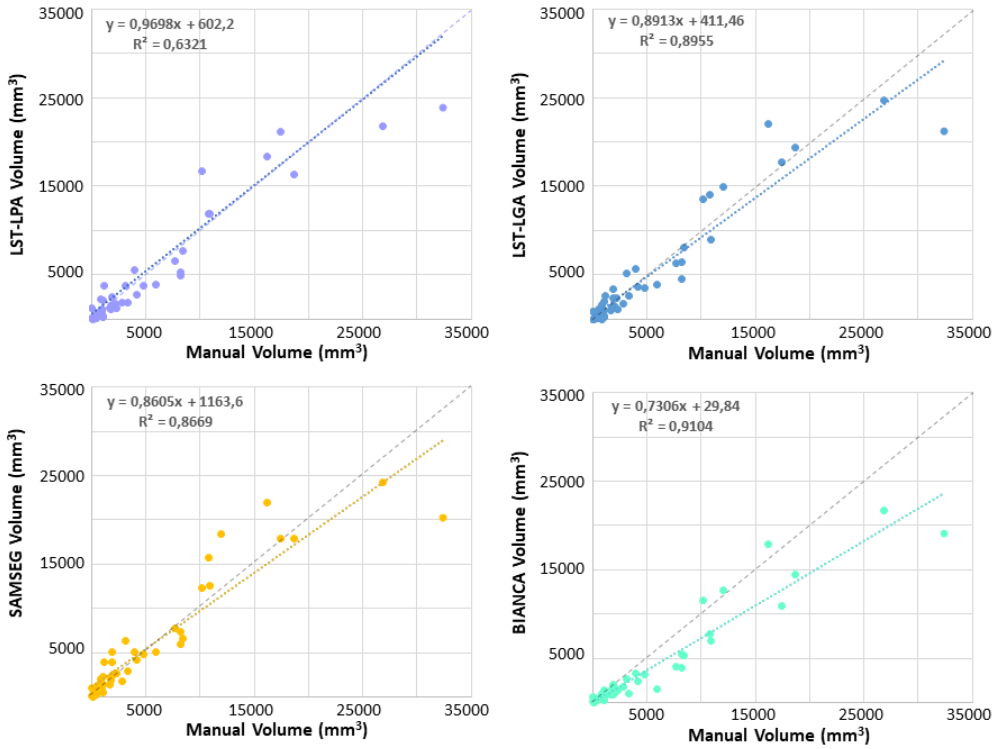


Figure 3. Correlation between the total volume of WMH from the gold standard and those obtained with each automatic segmentation. The graph shows the dispersion of the results, the equation of the trend line and its R2 value. Each colored graph represents an algorithm, and the gray dashed line represents the function y = x (slope of 1).

Finally, to perform an analysis of the consistency of the methods, we calculated the ICC for each tool (see Table 3). Attending the ICC results, all tools achieved a high consistency, being SAMSEG the tool with higher ICC (0.99), followed by BIANCA (0,93), LST-LPA (0,91) and LST-LGA (0.88). The results of FreeSurfer package are available in Supplementary material 4A.

3.2. Performance analysis: mean scores

The performance of each tool was evaluated by comparing their WMH brain masks to the corresponding ones from the gold standard, calculating accuracy, sensitivity, and Dice scores

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(Figure 4). Additionally, the existence of significant differences among methods (see Figure 4, right side) was assessed by means of a paired t-test.

Sensitivity. LST-LPA showed the highest value (mean: 0.5 ± 0.3), followed by BIANCA (0.4 ± 0.2), LST-LGA (0.4 ± 03) and SAMSEG (0.4 ± 0.3). The evaluation of statistical differences between methods showed that all methods differed between each other except for BIANCA and SAMSEG (see Figure 4, top right). The biggest differences were found for the comparisons between the sensitivity scores of LST-LPA and those from the rest of the tools: BIANCA, D = 1.018; LST-LGA, D = 1.284; and SAMSEG, D = 1.115. The comparisons between the sensitivity scores of BIANCA and LST-LGA and between the ones from SAMSEG and those from LST-LGA showed Cohen's d scores 0.427 and 0.534 respectively. No significant results were found for the comparison BIANCA-SAMSEG.

Dice score. LST-LPA showed the highest dice score value (mean: 0.5 ± 0.3) followed by BIANCA (0.5 ± 0.2), LST-LGA (0.3 ± 0.2) and SAMSEG (0.4 ± 0.2). The Dice score of LST-LPA was found to be significantly different (see Figure 4, middle right) to the ones obtained with BIANCA (D = 0.606), with LST-LGA (D = 1.087), and with SAMSEG (D = 1.172). Regarding BIANCA, its dice scores showed significantly higher values than the ones obtained with LST-LGA (D = 0.851) and SAMSEG (D = 0.970). No significant differences were observed between the dice score of LS-LGA and SAMSEG.

Precision. The scores obtained for each tool were: BIANCA (0.6 ± 0.3), LST-LPA (0.6 ± 0.3), LST-LGA (0.3 ± 0.2), and SAMSEG (0.3 ± 0.3). LST-LPA and BIANCA showed significant differences between their precision scores and those obtained with the rest of the groups but between each other (see Figure 4, bottom right). The size effects of the LST-LPA significant comparisons were: LST-LPA vs LST-LGA, D = 0.822; LST-LPA vs SAMSEG, D = 1.181. The size effect of the comparisons of BIANCA with LST-LGA and SAMSEG were D = 0.997 and D = 1.562 respectively. Finally, the comparison between LST-LGA and SAMSEG showed a significant effect size D = 0.369.

Finally, the performance of the FreeSurfer package was clearly worse than the corresponding to the tools based on FLAIR segmentation (see Supplementary material 4B for more details).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

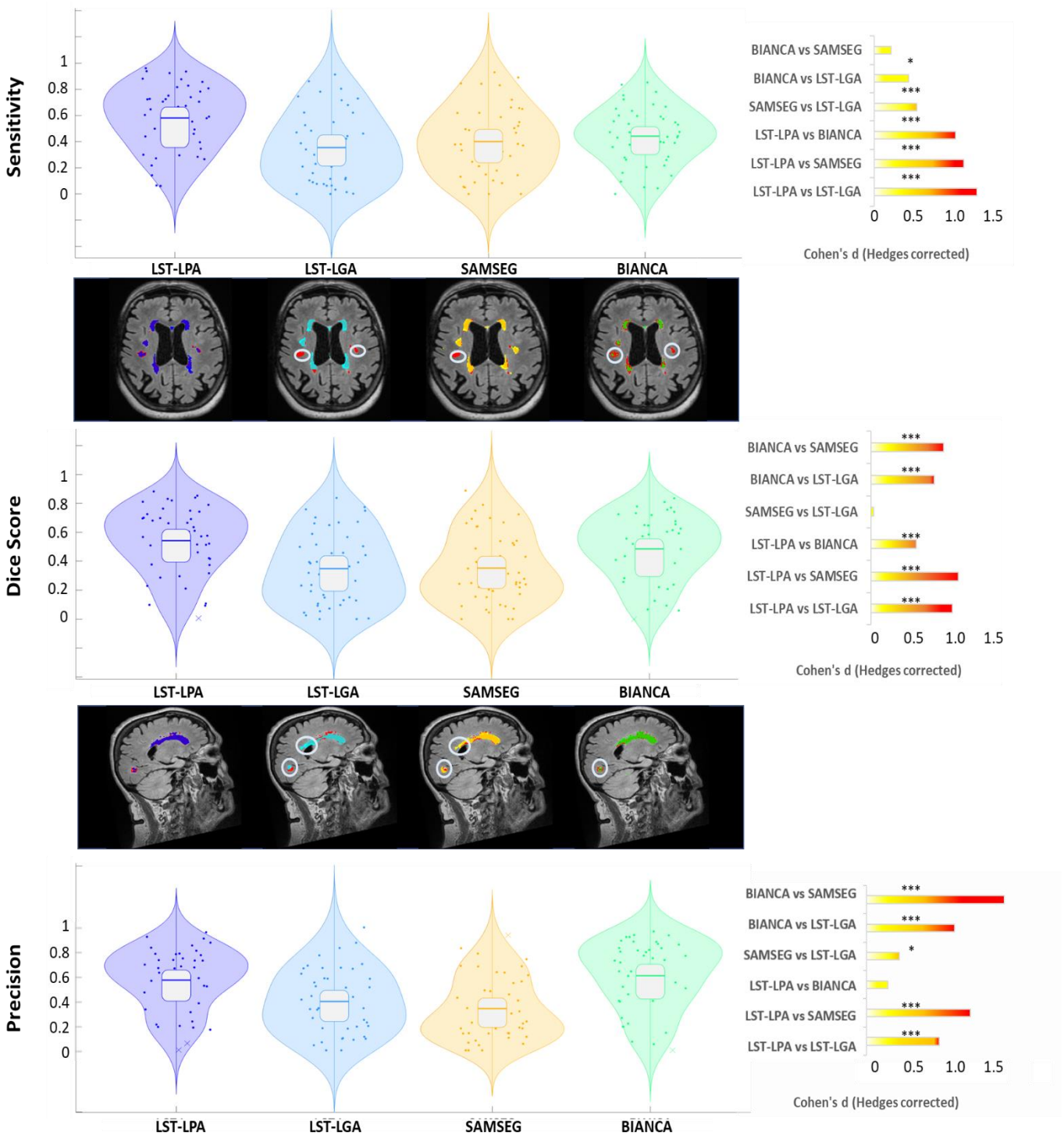


Figure 4. *Lesion segmentation performance in terms of precision, sensitivity, and Dice score for the proposed methods (from left to right: LST-LPA, LST-LGA, SAMSEG and BIANCA). Violin plots depict the distribution of the values obtained in each tool LST-LPA (dark blue), LST-LGA (light blue), SAMSEG (yellow) and BIANCA (green). Inside each violin plot, the corresponding boxplots represent the location of the quartiles. MRI figures show examples of the segmentations carried out by the different tools. The lesions corresponding to the gold standard are shown in red, from LST-LPA are highlighted in dark blue, from LST-LGA are depicted in light blue, from SAMSEG are displayed in yellow and from BIANCA are showed in green. The gray circles on the MRI images mark the main differences between the tools. To the right of each measure, the effect size of the comparisons (paired t.test) between the different tools are represented using the Hedges-corrected Cohen's d (small (yellow), medium (orange) or high (red)) as well as the level of significance obtained (* marks p<0.05 and *** marks p<0.001).*

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

3.3. Performance by total lesion volume

By plotting the Dice score, precision, and sensitivity as a function of the total volume of WMH labeled on the gold standard, the results showed a clear logarithmic influence of WMH loading on the performance of each tool (Figure 5). In addition to the standard R^2 scores, we calculated additional quantitative measures to enhance the differences that the respective logarithmic performances of each method showed relative to each other (Figure 5, left). On the one hand, 95% value (in mm^3) corresponded to the WMH volume that reach the 95% of the performance at 35000 mm^3. On the other hand, ∫f(x)) quantified the area under the fitting curve.

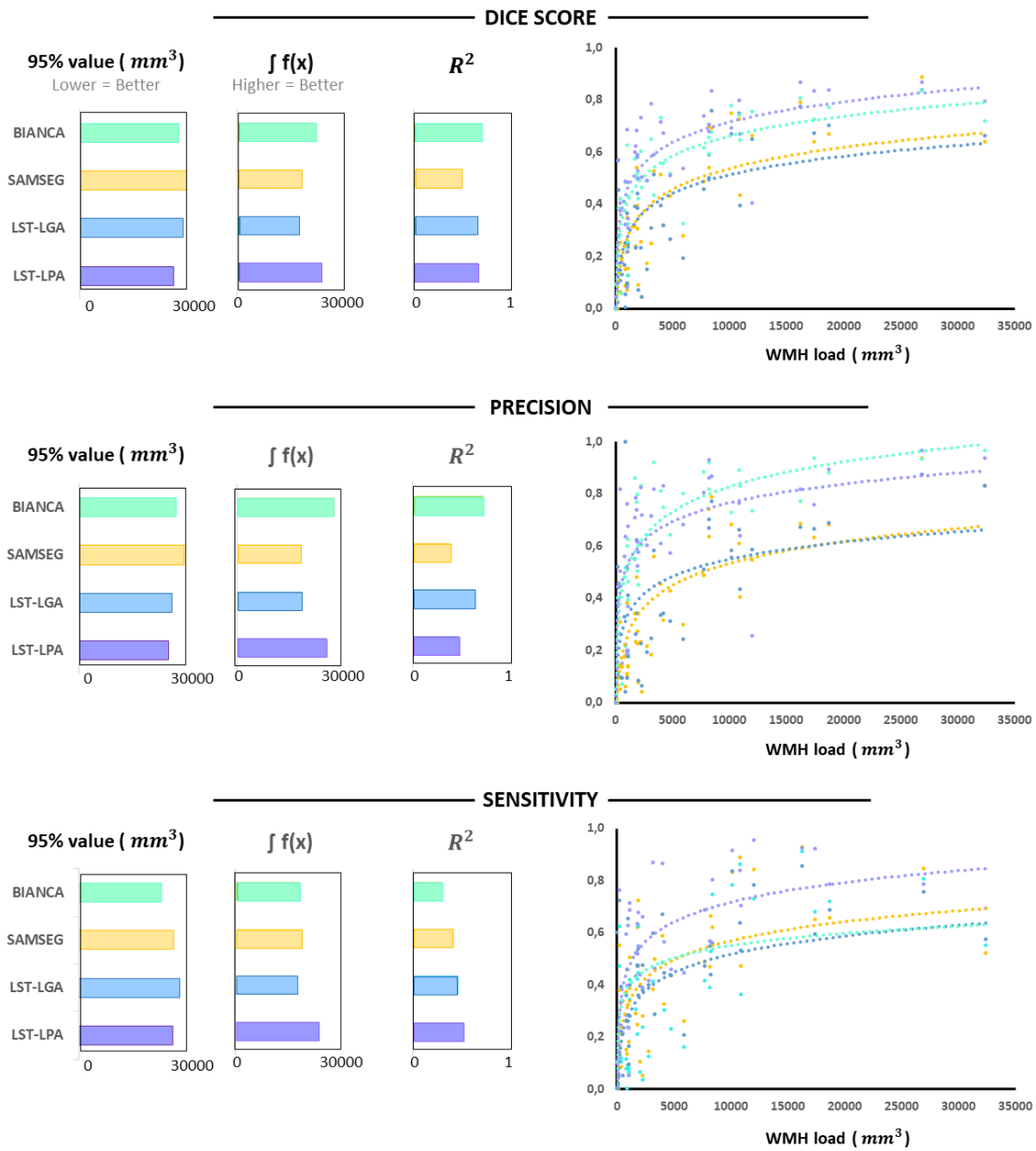


Figure 5. Performance (Dice score, precision, and sensitivity) of each method as a function of the gold standard's total WMH volume. LST-LPA is represented in dark blue, LST-LGA in light blue, SAMSEG in yellow and BIANCA in green. Right, the tendency line is depicted for each tool and measure. Left, representation of the WMH volume that reach the 95% of the performance, the area down the curve and the R^2 value for each tool and measure. The results of BIANCA are represented in green, SAMSEG in yellow, LST-LGA in light blue and LST-LGA in dark blue.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Dice score. LST-LPA results got a good fit to the logarithmic regression (R^2: 0.67) and were the ones that stabilize faster (95% value: 23230 mm^3) and higher ($\int f(x)$ = 26020). The second-best performance was found for BIANCA results (R^2: 0.71, 95% value: 24612 mm^3 and $\int f(x)$: 24108). The results of LST-LGA got intermediate performance (R^2: 0.65, 95% value: 25694 mm^3 and $\int f(x)$: 18870), closely forward by SAMSEG (R^2: 0.50, 95% value: 26406 mm^3 and $\int f(x)$: 19840).

Precision. When assessing precision, the results of BIANCA scored the best (R^2: 0.73, 95% value: 24327 mm^3 and $\int f(x)$: 30180) closely followed by the results of LST-LPA (R^2: 0.48, 95% value: 22433 mm^3 and $\int f(x)$: 27733). SAMSEG (R^2: 0.39, 95% value: 26111 mm^3 and $\int f(x)$: 19751), and LST-LGA (R^2: 0.65, 95% value 23191 mm^3 and $\int f(x)$: 20092) were the tools with lower performance.

Sensibility. The best performance in terms of sensibility was reached by LST-LPA (R^2: 0.53, 95% value: 23371 mm^3 and $\int f(x)$: 26061). After LST-LPA, the results obtained for the resto of the tools were: BIANCA (R^2: 0.3, 95% value: 20678 mm^3 and $\int f(x)$: 19871), SAMSEG (R^2: 0.42, 95% value: 23694 mm^3 and $\int f(x)$: 20849), and LST-LGA (R^2: 0.47, 95% value: 25086 mm^3 and $\int f(x)$: 18981).

Additionally, when assessing Dice Score, precision, and sensibility separately for three WMH volume ranges (0-1000 mm^3, 1000-5000 mm^3 and 5000-35000 mm^3) we found a similar pattern (see Table 4) were LST-LPA obtained the best performance closely followed by SAMSEG. The results obtained for the results of FreeSurfer are available on Supplementary material 4C.

			Dice score	
WMH Total volume	LST-LPA	LST-LGA	SAMSEG	BIANCA
0 - 1000 mm^3	0.30 ± 0.2	0.2 ± 0.2	0.1 ± 0.1	0.3 ± 0.2
1000-5000 mm^3	0.5 ± 0.2	0.3 ± 0.1	0.3 ± 0.2	0.5 ± 0.1
5000-35000 mm^3	0.7 ± 0.1	0.6 ± 0.2	0.6 ± 0.2	0.7 ± 0.1
			Precision	
0 - 1000 mm^3	0.4 ± 0.3	0.3 ± 0.3	0.1 ± 0.1	0.3 ± 0.2
1000-5000 mm^3	0.6 ± 0.2	0.3 ± 0.2	0.3 ± 0.2	0.6 ± 0.2
5000-35000 mm^3	0.8 ± 0.2	0.6 ± 0.2	0.6 ± 0.2	0.8 ± 0.1
			Sensitivity	
0 - 1000 mm^3	0.3 ± 0.2	0.2 ± 0.2	0.2 ± 0.2	0.3 ± 0.2
1000-5000 mm^3	0.5 ± 0.2	0.3 ± 0.2	0.3 ± 0.2	0.4 ± 0.1
5000-35000 mm^3	0.8 ± 0.2	0.6 ± 0.2	0.7 ± 0.2	0.6 ± 0.2

***Table 4**. Mean ± std for Dice score, precision, and sensitivity values for each tool (LST-LPA, LST-LGA, SAMSEG and BIANCA) for three ranges of WMH volumes (0-1000 mm^3, 1000-5000 mm^3 and 5000-35000 mm^3).*

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

3.4. Performance by individual lesion volume

The last analysis consisted of evaluating the sensitivity of each individual lesion outlined in the gold standard, stratifying the analysis based on the volume of the lesion. In the Figure 6 it can be seen the sensibility of each tool for each range of lesions' volumes. The first result was that most lesions of 8 mm^3 or less were not captured by the automatic tools. An important aspect of this analysis was the detection, for each automatic tool, of the point at which individual lesions begin to be captured with some sensitivity. In this regard, BIANCA was the only tool capable of capturing more than the 50% of the lesions from 10 to 13 mm^3, outperforming LST-LPA (whose threshold was 13 to 18 mm^3), and clearly above the results obtained by LST-LGA and SAMSEG (with thresholds of 41-110 and 26-40 mm^3 respectively). It is also worth noting the evolution of the sensitivity levels in each tool, finding a rapid growth in the LST-LPA that reached a median of 0.8 in the most voluminous lesions. On the other hand, BIANCA showed a smoother evolution, reaching a median of 0.6 in the largest lesions. SAMSEG (0.5) and LST-LGA (0.4) were the tools with lower medians on these lesions. The results of FreeSurfer are depicted in Supplementary material 4D.

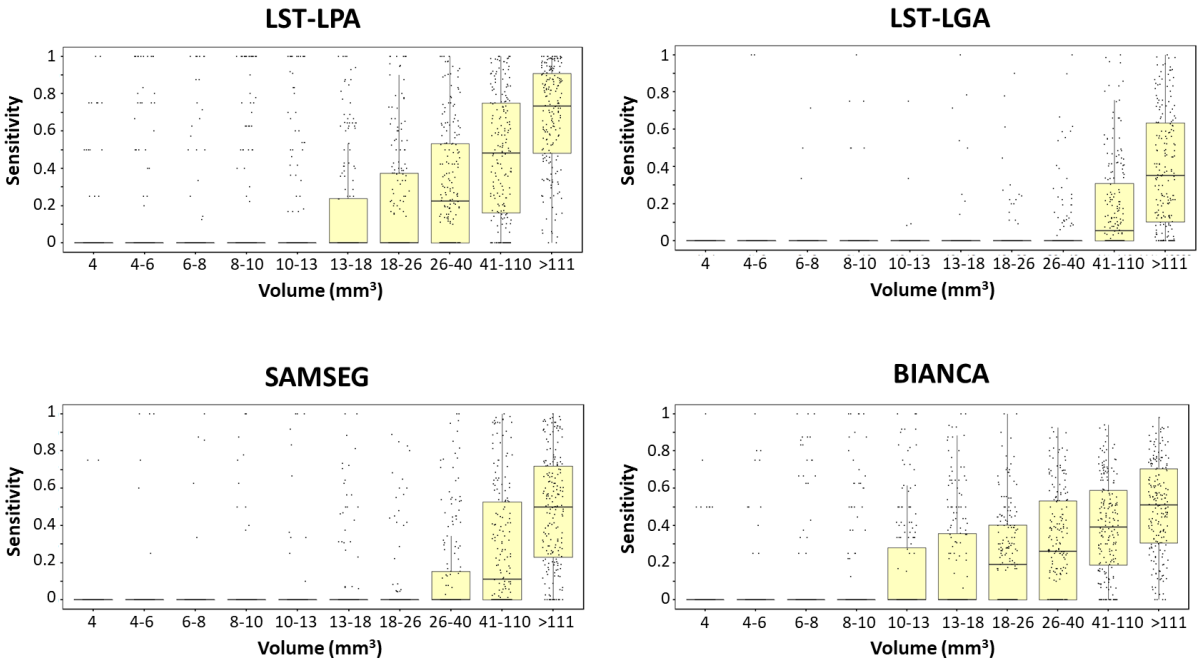


Figure 6. Sensitivity of each tool when evaluating individual lesions stratified by volume ranges. Each volume range contained the same number of lesions (196). For each tool and volume range, the boxplot indicates the median and the second and third quartiles. The whiskers represent the first and fourth quartiles, ignoring outlier individuals. The dots represent the distribution of sensitivity obtained for each individual lesion.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

3.5. Real world scenario and Outlier evaluation

In view of the results obtained, we opted to process a sample with 577 old individuals from our laboratory (C^3N) with both LST algorithms. LST tools were selected by their performance, especially in the case of LST-LPA, the fast processing, and the ease of use. The analysis consisted of evaluating the consistency of the differences between LST-LPA and LST-LGA. The results showed that LST-LPA malfunctioned in 9 patients, while LST-LGA did so in 2. Importantly, the results showed that the combination of both LST tools provided a robust estimate of WMH volumes, as only those participants with large differences between the results of both tools needed to be visually verified. Figure 7 shows the difference between the results of both tools for everyone in the sample and MR images for those cases with poor WMH segmentation.

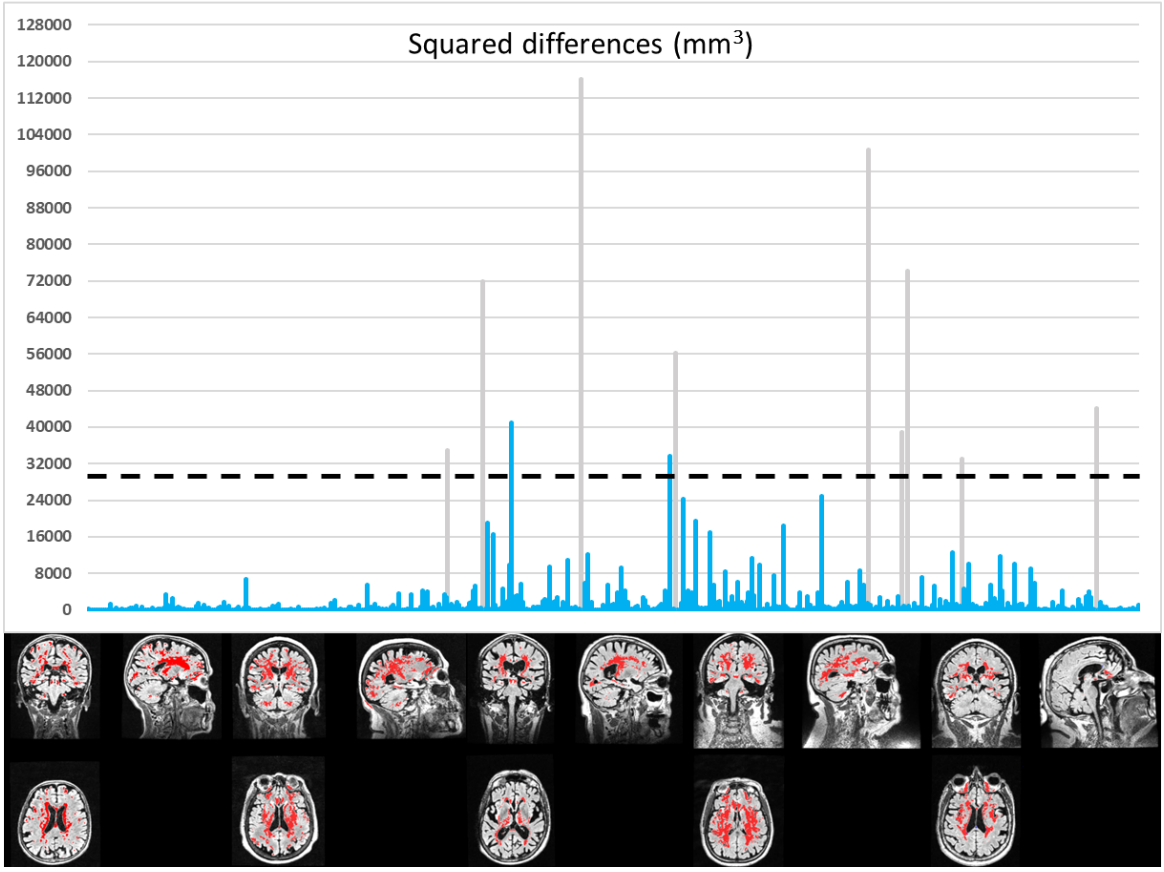


Figure 7. Representation of the squared differences between LST-LPA and LST-LGA along 577 subjects. The dash line marks the outlier threshold of three standard deviation above the mean. Blue bars represent differences in subjects below the outlier line or with good quality segmentations in the LST-LPA tool. Black bars denote outliers with poor WMH segmentation in the LST-LPA tool. Examples of bad LST-LPA executions are shown at the bottom of the figure.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4 Discussion

In this piece of work, we aimed to assist clinicians and researchers in the process of selecting the most appropriate methodology for the evaluation and follow-up of WMH of suspected vascular origin, based on their specific conditions and requirements. For this purpose, we carefully evaluated the performance of 4 freely available algorithms (LST-LPA, LST-LGA, SAMSEG and BIANCA), designed specifically for WMH segmentation in MRI, in a sample of 45 cognitively healthy elders with a broad distribution of WMH load classified according to the clinical reports (no WMH, subclinical WMH, and CBVD WMH). These algorithms are included in three of the neuroimaging toolboxes most widely used in both clinical and research settings (FSL, FreeSurfer and SPM).

The performance of each tool was evaluated by calculating the correlation between its results and both the Fazekas clinical scale and manual segmentations performed by radiologists. In addition, we carried out analyzes based on WMH loads, studied the sensitivity of each method at the lesion level, and ended with a real application of two automated tools to a large data set consisting of 577 older individuals. The results showed that all the metrics achieved acceptable results, useful for clinical practice, since they showed a high correlation with the Fazekas clinical scale and with the gold standard. Overall, the best algorithms in terms of performance were LST-LPA and BIANCA, but taking all the analysis into account, we decided to dig deeper into the large data set with both LST tools, which we found to be complementary.

4.1 Clinical relevance: where the automatic results related to the clinical status?

The total volume of WMH obtained with all the tools showed high correlations with the Fazekas scores (Table 3), showing values close to those obtained by the gold standard. In addition, the correlations between the total WMH volumes computed by means of the automatic tools and the ones obtained from the gold standard were consistent but marked by a general tendency by the automated tools to underestimate the volume of the lesion (Figure 2). These two results highlight the clinical relevance of automated tools, as they have been shown to serve as a proxy for the Fazekas scale and manual segmentation. We combined the Fazekas volume and total WMV correlations because together they are a good indicator of vascular damage, since the strong ceiling and floor effect of the Fazekas scale is offset by the discriminatory capacity of the total WMH volume. The correlation analyses between automatic and manual segmentations have been previously replicated by other studies carried out with LST-LPA (Heinen et al., 2019; Tran et al., 2022), LST-LGA (Heinen et al., 2019; Schmidt et al., 2012; Tran et al., 2022) and BIANCA (Griffanti et al., 2016; Hotz et al., 2021; Ling et al., 2018). Regarding SAMSEG algorithm, the correlation analysis with manually segmented volume was not assessed by the original article of Cerri and collaborators, and to the best of our knowledge it has not been depicted in any study. Although the results in the literature are heterogeneous, most studies report a slight tendency for these tools to provide smaller lesions volume when compared to manual segmentation (Griffanti et al., 2016; Heinen et al., 2019; Hotz et al., 2021; Ling et al., 2018). As can be seen in Figure 3, LST-LPA was the tool with the closest correlation with manual segmentation, closely followed by LST-LGA and SAMSEG. However, in other study the WMH measured through LST-LGA has been even better than LST-LPA (Tran et al., 2022). Finally, BIANCA presented the furthest correlation from the clinical measurements, but it is the one that best fits the regression, which may reflect a systematic problem of the tool in the treatment of false negatives, as can be seen in the sensitivity measurements (Figure 4 and 5).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4.2 General performance comparison: total volume matters

The stability of the results was assessed by means of ICC, showing that all algorithms presented a high degree of stability. SAMSEG was the tool with the highest ICC, followed by BIANCA and both LST algorithms. In this sense, the ICC results previously reported for LST-LPA, LST-LGA, and BIANCA are very heterogeneous, finding results similar to ours in some cases but also different in others (Griffanti et al., 2016; Heinen et al., 2019; Hotz et al., 2021; Tran et al., 2022).

In other vein, we evaluated the relationship between WMH volumetric masks obtained from automated methods and manual segmentations performed by radiologists by calculating precision, sensitivity, and Dice scores. In this analysis, LST-LPA emerged as the most sensitive algorithm and with the greatest degree of similarity (Dice score) compared to the manual segmentation. Regarding precision, LST-LPA and BIANCA showed similar performance, followed by SAMSEG and LST-LGA (Figure 4). Overall, the supervised methods (LST-LPA and BIANCA) offered higher execution values than the unsupervised ones (LST-LGA and SAMSEG). This phenomenon probably arises from the fact that the algorithms used by the supervised methods, such as logistic regressions in LST-LPA (Schmidt, 2016) or K nearest neighbors in BIANCA (Griffanti et al., 2016), imply a higher degree of complexity than the clustering algorithms used in the unsupervised methods. Furthermore, we found that the performance of the algorithms increased logarithmically in sensitivity, accuracy, and Dice score as lesion burden expanded (Figure 5). Our results agree with previous literature (Commowick et al., 2018; Griffanti et al., 2016; Hotz et al., 2021), that stated the same influence of lesion load and lesion size on performance, and the same logarithmic behavior indicating that the effectiveness of these tools fails in those individuals with very low total lesion load, but rapidly increases within lesion volume. In this work we try to go further by evaluating the stabilization rate of the results as a function of the volume of WMH, properties that have not been dealt with in depth in the literature (Table 4 and Figure 5). Our results pointed out that, of the supervised methods, LST-LPA presented higher values of improvement depending on the WMH volume than BIANCA in Dice score and sensitivity, and very similar results in precision. In terms of growth velocity, LST-LPA reached 95% of the cut-off value in our sample faster than BIANCA in precision and Dice score. However, despite performing better than BIANCA in sensitivity, this tool presented a faster sensitivity evolution. This effect could be due to, as will be detailed later, the fact that the BIANCA algorithm is capable of segmenting smaller lesions than LST-LPA, but with lower sensitivity. The unsupervised methods showed a similar behavior, but while SAMSEG presented a better evolution in sensitivity, LST-LGA stood out in precision. In speed, both tools showed a slow evolution in precision and Dice score, but LST-LGA presented a fast evolution in sensitivity.

According to previous literature, we obtained similar results for the supervised tools (Griffanti et al., 2016; Heinen et al., 2019; Hotz et al., 2021; Khademi et al., 2021; Tran et al., 2022; Vanderbecq et al., 2020) and SAMSEG (Cerri et al., 2021), but, in our case, LST-LGA showed a lower performance than previous literature (Heinen et al., 2019; Schmidt et al., 2012; Tran et al., 2022; Vanderbecq et al., 2020). Taking into account the variability of the execution depending on the total volume of the lesion, other researchers have observed the same evolution but with a milder evolution pattern, mainly due to a more efficient execution at lower loads of vascular damage (Griffanti et al., 2016; Heinen et al., 2019; Schmidt et al., 2012). The discrepancy could lie in the difference between images acquired with MR machines with 3T and 1.5T magnetic fields, the latter being the most used in the clinical setting but with poorer resolution, especially for small lesions (Cerri et al., 2021). This effect may be especially compelling for the LST-LGA tool, as it relies on hypointensities on the

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

T1 image to initialize lesions that will subsequently grow on the FLAIR image (Schmidt et al., 2012). As mentioned by the authors of LST-LGA, hypointensities in T1 images are sensitive to field strength, and may not exist for certain lesions on low-resolution scans (Schmidt et al., 2012). Another possibility could be a different etiology of the lesions (Balakrishnan et al., 2021) since most of these studies were carried out in patients with multiple sclerosis.

4.3 Individual lesion size bias

The detection of small lesions is of paramount interest in the clinical and scientific fields, especially in studies focused on healthy aging or premorbid stages of pathological aging, where vascular damage is usually mild, or in longitudinal studies intended to follow the evolution of the disease from the growth of the subtle lesion. For instance, the WMH location and pattern of distribution (i.e., extent or symmetry) in the brain have been also associated with different underlying pathologies (Balakrishnan et al., 2021; Dadar et al., 2022) and with the degree of impairment for different cognitive functions (Jiménez-Balado et al., 2022). Therefore, in studies like ours, this factor is key, because most automatic segmentation tools have been optimized by maximizing Dice's sensitivity, precision, and score values. However, despite the widespread validity of these methods, based on these global measures, performance on smaller lesions does not contribute much to their values (Balakrishnan et al., 2021; Ghafoorian et al., 2016).

Individual lesion volume and definition of lesion borders often differ depending on the etiology of WMH (Balakrishnan et al., 2021). In this sense, since some metrics are sensitive to the surface-volume ratio of segmented structures, it may be important to consider the size of the lesion to correctly understand the results and evaluate the performance of the tools (Commowick et al., 2018; García-Lorenzo et al., 2013). In this regard, one of the difficulties that arise when segmenting small lesions is that they tend to have a different range of intensities due to the partial volume effect that occurs when a lesion is located in the limits of several tissues, making its neighbors not homogeneous (Balakrishnan et al., 2021; García-Lorenzo et al., 2013; Ghafoorian et al., 2016; Lesjak et al., 2018; Park et al., 2018). In addition, these lesions usually appear in blob structures and with a more variable location than the large lesions, which usually appear around the periventricular area (Ghafoorian et al., 2016). This heterogeneity in small lesions results in a nonlinear problem that is difficult to solve for automatic algorithms, and that directly affects its performance (Commowick et al., 2018). To assess this phenomenon, we measured the sensitivity to each lesion individually. Considering our results, the size of the individual lesions directly influences the quality of the segmentation. However, each tool has performance evolution and a cutoff point where its segmentations begin to be reliable. BIANCA was the only tool capable of consistently segmenting lesions between 10-13 mm^3, agreeing with results previously reported (Park et al., 2018). LST-LPA presented its sensitivity threshold in lesions between 13-18 mm^3, however, as lesion size increases, its performance improves more rapidly and up to values higher than those obtained in BIANCA. The unsupervised methods were only able to consistently segment lesions larger than 26 mm^3, with SAMSEG being slightly more sensitive than LST-LGA. These execution problems could have had their origin in the difficulty of these tools to define the borders of the lesion. Therefore, smaller lesions, which have a higher surface-to-volume ratio, would be more affected. Summarizing these results, it should be noted that, despite the good result of BIANCA in the detection of small lesions, no algorithm was able to consistently segment lesions smaller than 10 mm^3, leaving a wide spectrum of lesions not covered by any of the methods discussed here. However, several promising machine learning strategies for segmenting small lesions have been developed in recent years, such

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

as the use of multiscale deep features (Park et al., 2018) or two different machine learning systems for small and large lesions (Ghafoorian et al., 2016). Despite these good results, it would be necessary to carry out external validations of these methods and test their effectiveness in larger and more varied samples.

4.4 Comparison between algorithms: Which algorithm should be used?

As reported in previous sections, supervised methods generally obtain lesion segmentations closer to the gold standard, with high sensitivity to smaller lesions. At the same time, these algorithms can better delineate the edges of the lesion, avoiding the occurrence of false positives and obtaining better accuracy scores. These effects are due to a larger number of possible characteristics to identify a lesion , while unsupervised methods work with a limited selection of pre-defined parameters (García-Lorenzo et al., 2013).

Attending to unsupervised methods, SAMSEG and LST-LGA showed total lesion values close to the gold standard, marked by a slight tendency to underestimate total lesion volume. LST-LGA showed higher precision but lower sensitivity than SAMSEG, mainly driven by a lesion sensitivity threshold of 26 mm^3 in SAMSEG and 41 mm^3 in LST-LGA. The advantages of LST are an intuitive user interface, low average computational requirements, and very short execution times. However, SAMSEG provides more information than only WMH, as volumetric values of different cortical and subcortical regions that may be of interest in the academic field.

Regarding the supervised methods, both BIANCA and LST-LPA achieve high-quality segmentations with excellent sensitivity, precision, and Dice score parameters. LST-LPA outperforms BIANCA in sensitivity and Dice score, achieving a total lesion volume closer to the gold standard and a more rapid evolution as a function of volume. Nevertheless, BIANCA can detect low-volume lesions better than LST-LPA, although the latter greatly improves its sensitivity as lesion volume increases. Methodologically, the main difference between both algorithms is how they are prepared; while BIANCA was trained on the gold standard sample, LST-LPA features an algorithm pre-trained on samples from multiple sclerosis patients, although the original authors claim that they are in the process of releasing a version capable of adapting the learning process to each sample (Schmidt, 2016). Regarding preprocessing, BIANCA usually failed when defining the exclusion masks, which led to the appearance of multiple false positives that would be difficult to control in large databases. All these features, added to its ease of use, low computational requirements, and short execution times, make LST-LPA one of the best options available.

In recent years, computational advances have enabled the development of convolutional neural networks (CNNs) which have been shown, in some validations, to perform better than LST-LPA with the additional advantage of being less influenced by vascular load (Khademi et al., 2021). On the other hand, effective methods have been proposed to improve the segmentation of these tools, such as taking into account the spatial variability of intensities along a tissue using region-specific multivariate Gaussian distributions (Harmouche et al., 2006). Taken together, these results highlight the existence of complex and non-linear features in the lesion voxels that could be taken into account to improve the performance of automated tools and successfully capture the variability of vascular lesions (intensity, shape, size and texture) (Khademi et al., 2021; Tran et al., 2022). Nevertheless other studies have indicated that, despite their high specificity, CNN-based tools perform similarly to simpler tools such as LST-LPA and BIANCA (Balakrishnan et al., 2021).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Taking all the results into account, we decided to automatically segment the WMH of a larger sample of older adults with both LST tools, since its combination supposes a great quality/cost balance (i.e., computed with the same toolbox, with low computational requirements, and with short execution time). By checking the difference between the total WMH volumes provided by both tools, we were able to detect 11 outliers out of 577 participants. The visual inspection of these outliers showed that in 9 of them LST-LPA performed poorly, providing a high number of false positive around several brain structures (Figure 7, bottom). The remaining 2 cases were images where LST-LGA malfunctioned. The advantage of LST-LGA could be based on the non-use of training samples, which always depends on manual segmentations that have a high intra-evaluator variability (García-Lorenzo et al., 2013). Indeed, it is not uncommon to observe a better performance of LST-LGA than LST-LPA or BIANCA once applied to large samples that present a wide variability (Hotz et al., 2021; Tran et al., 2022). In this regard, Vanderbecq and colleagues evaluated routine clinical and research images with artifacts in order to assess the performance of different algorithms. Their conclusion was that, despite the good results in image research, as the tools became more complex they presented more problems with routine images, being the most detrimental those based on CNN and followed by supervised methods with simpler strategies such as LST-LPA or BIANCA (Vanderbecq et al., 2020). This drawback implies that, when dealing with large databases, the required review and quality control takes additional time and effort. On the other hand, the unsupervised tools, albeit obtaining more modest values, present a stable quality level throughout the measurements. In our research, we found that the combined use of both LST tools adequately tackled these issues. On the one hand, we obtained reliable results in most cases with LST-LPA. On the other hand, we were able to control the appearance of errors in the LST-LPA segmentation thanks to its comparison with the results from LST-LGA. Therefore, we believe that both tools constitute a useful proxy for WMH segmentation, providing synergistic information, with an extraordinarily efficient process in terms of ease of use and processing time. We would only recommend BINCA in those cases in which there is a special interest in studying small lesions, since BIANCA outperformed LST-LPA in these cases (10-13 mm^3 vs 13-18 mm^3).

4.5 Limitations and Future research

Despite the useful findings of this study, it is important to describe the methodological limitations we faced. One concern was the use of manual segmentations as the gold standard. As other authors have previously pointed out, the definition of WMH, although widely accepted, is open to a large number of interpretations (García-Lorenzo et al., 2013; Vanderbecq et al., 2020). This lack of precision in the definition leads to a high inter-rater variability, which makes it difficult to differentiate the errors of the automatic tools from those associated with differences in criteria between clinicians. On the other hand, although the use of 1.5T scanners brings us closer to clinical practice, it also makes it difficult to compare with other studies in the field, where most use higher resolution, 3T images. Taken together, these characteristics difficult the comparison with previous studies, and may be the cause of some of the discrepancies found.

As mentioned above, many automatic segmentation methods with different approaches have appeared in recent years. In this scenario, the combination and comparison of methods is the way to develop a high-precision tool capable of overcoming the limitations we see today. However, many of them either are not available or their code is not open, which makes it difficult to compare the results obtained in different validations. Furthermore, this lack of transparency does not allow for a combination of methods to refine and standardize the measurements. For this reason, only

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

tools publicly available and widely used in the clinical and scientific fields were used in this work. Finally, to combine different efforts, it is essential to clarify terms such as WMH and its limits, and to develop a standardized consensus on the recommendations for the segmentation of this type of lesions.

5 Conclusion

As mentioned in many previous studies, a fully automated segmentation method suitable for clinical use is not yet available. Therefore, when choosing a WMH segmentation method, we must take into account the metrics that interest us, as well as the characteristics of our sample. On one hand, for the calculation of the total WMH volume, the unsupervised tools offer results that are close to manual segmentations, reaching high reliability. On the other hand, for more precise analyses of lesion location or in samples with abundant small lesions, supervised methods offer a clear advantage. However, it is advisable to carry out a quality control of the images obtained to detect possible outliers in the segmentation. In our experience, the LST package includes two promising and complementary methods that are widely available, easy to use, and have short execution times. Combining the results of LST-LPA (supervised) and LST-LGA (unsupervised) is a feasible procedure with the ability to be used as a useful proxy for WMH segmentation.

6 Founding

Lucia Torres-Simon and Alberto del Cerro-León acknowledge the financial support of predoctoral researchers grants from Universidad Complutense de Madrid (CT42/18-CT43/18) & (CT58/21-CT59/21) respectively, that were co-funded by Santander bank.

7 Declaration of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Compliance with ethical standards

The Hospital Universitario San Carlos Ethics Committee (Madrid) approved the study, and all participants signed a written informed consent prior to participation.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

9 Bibliography

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., … Hainsworth, A. H. (2019). White matter hyperintensities in vascular contributions to cognitive impairment and dementia (VCID): Knowledge gaps and opportunities. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, *5*, 107–117. https://doi.org/10.1016/j.trci.2019.02.001

Anor, C. J., Dadar, M., Collins, D. L., & Tartaglia, M. C. (2021). The Longitudinal Assessment of Neuropsychiatric Symptoms in Mild Cognitive Impairment and Alzheimer's Disease and Their Association With White Matter Hyperintensities in the National Alzheimer's Coordinating Center's Uniform Data Set. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(1), 70–78. https://doi.org/10.1016/j.bpsc.2020.03.006

Arvanitakis, Z., Capuano, A. W., Leurgans, S. E., Bennett, D. A., & Schneider, J. A. (2016). Relation of cerebral vessel disease to Alzheimer's disease dementia and cognitive function in elderly people: a cross-sectional study. *The Lancet Neurology*, *15*(9), 934–943. https://doi.org/10.1016/S1474-4422(16)30029-1

Arvanitakis, Z., Fleischman, D. A., Arfanakis, K., Leurgans, S. E., Barnes, L. L., & Bennett, D. A. (2016). Association of white matter hyperintensities and gray matter volume with cognition in older individuals without cognitive impairment. *Brain Structure and Function*, *221*(4), 2135–2146. https://doi.org/10.1007/s00429-015-1034-7

Balakrishnan, R., Valdés Hernández, M. del C., & Farrall, A. J. (2021). Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data – A systematic review. *Computerized Medical Imaging and Graphics*, *88*(December 2020). https://doi.org/10.1016/j.compmedimag.2021.101867

Bendfeldt, K., Blumhagen, J. O., Egger, H., Loetscher, P., Denier, N., Kuster, P., Traud, S., Mueller-Lenke, N., Naegelin, Y., Gass, A., Hirsch, J., Kappos, L., Nichols, T. E., Radue, E. W., & Borgwardt, S. J. (2010). Spatiotemporal distribution pattern of white matter lesion volumes and their association with regional grey matter volume reductions in relapsing-remitting multiple sclerosis. *Human Brain Mapping*, *31*(10), 1542–1555. https://doi.org/10.1002/hbm.20951

Bjerke, M., Jonsson, M., Nordlund, A., Eckerström, C., Blennow, K., Zetterberg, H., Pantoni, L., Inzitari, D., Schmidt, R., & Wallin, A. (2014). Cerebrovascular Biomarker Profile Is Related to White Matter Disease and Ventricular Dilation in a LADIS Substudy. *Dementia and Geriatric Cognitive Disorders Extra*, *4*(3), 385–394. https://doi.org/10.1159/000366119

Blair, G. W., Hernandez, M. V., Thrippleton, M. J., Doubal, F. N., & Wardlaw, J. M. (2017). Advanced Neuroimaging of Cerebral Small Vessel Disease. *Current Treatment Options in Cardiovascular Medicine*, *19*(7). https://doi.org/10.1007/s11936-017-0555-1

Blystad, I., Håkansson, I., Tisell, A., Ernerudh, J., Smedby, Ö., Lundberg, P., & Larsson, E.-M. (2016). Quantitative MRI for Analysis of Active Multiple Sclerosis Lesions without Gadolinium-Based Contrast Agent. *American Journal of Neuroradiology*, *37*(1), 94–100. https://doi.org/10.3174/ajnr.A4501

Caballero, M. Á. A., Suárez-Calvet, M., Duering, M., Franzmeier, N., Benzinger, T., Fagan, A. M., Bateman, R. J., Jack, C. R., Levin, J., Dichgans, M., Jucker, M., Karch, C., Masters, C. L., Morris, J. C., Weiner, M., Rossor, M., Fox, N. C., Lee, J. H., Salloway, S., … Ewers, M. (2018). White matter diffusion alterations precede symptom onset in autosomal dominant Alzheimer's

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

disease. *Brain*, *141*(10), 3065–3080. https://doi.org/10.1093/brain/awy229

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

Cerri, S., Puonti, O., Meier, D. S., Wuerfel, J., Mühlau, M., Siebner, H. R., & Van Leemput, K. (2021). A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. *NeuroImage*, *225*, 117471. https://doi.org/10.1016/j.neuroimage.2020.117471

Chutinet, A., & Rost, N. S. (2014). White matter disease as a biomarker for long-term cerebrovascular disease and dementia topical collection on cerebrovascular disease and stroke. *Current Treatment Options in Cardiovascular Medicine*, *16*(3). https://doi.org/10.1007/s11936-013-0292-z

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., … Barillot, C. (2018). Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Scientific Reports*, *8*(1), 13650. https://doi.org/10.1038/s41598-018-31911-7

Dadar, M., Mahmoud, S., Zhernovaia, M., Camicioli, R., Maranzano, J., & Duchesne, S. (2022). White matter hyperintensity distribution differences in aging and neurodegenerative disease cohorts. *NeuroImage: Clinical*, *36*(September). https://doi.org/10.1016/j.nicl.2022.103204

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

Drebette, S., & Markus, H. S. (2010). The clinical importance or WMH on brain MR systematic review and meta-analysis. *BMJ Open*, *314*. https://doi.org/https://doi.org/10.1136/bmj.c3666

Duering, M., Righart, R., Csanadi, E., Jouvent, E., Herve, D., Chabriat, H., & Dichgans, M. (2012). Incident subcortical infarcts induce focal thinning in connected cortical regions. *Neurology*, *79*(20), 2025–2028. https://doi.org/10.1212/WNL.0b013e3182749f39

Egger, C., Opfer, R., Wang, C., Kepp, T., Sormani, M. P., Spies, L., Barnett, M., & Schippling, S. (2017). MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? *NeuroImage: Clinical*, *13*, 264–270. https://doi.org/10.1016/j.nicl.2016.11.020

Fazekas, F., Chawluk, J. B., & Alavi, A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Neuroradiology*, *8*(3), 421–426.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

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. *Medical Image Analysis*, *17*(1), 1–18. https://doi.org/10.1016/j.media.2012.09.004

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Garcia-Lorenzo, D., Prima, S., Arnold, D. L., Collins, D. L., & Barillot, C. (2011). Trimmed-Likelihood Estimation for Focal Lesions and Tissue Segmentation in Multisequence MRI for Multiple Sclerosis. *IEEE Transactions on Medical Imaging*, *30*(8), 1455–1467. https://doi.org/10.1109/TMI.2011.2114671

Gaubert, M., Lange, C., Garnier-Crussard, A., Köbe, T., Bougacha, S., Gonneaud, J., de Flores, R., Tomadesso, C., Mézenge, F., Landeau, B., de la Sayette, V., Chételat, G., & Wirth, M. (2021). Topographic patterns of white matter hyperintensities are associated with multimodal neuroimaging biomarkers of Alzheimer's disease. *Alzheimer's Research and Therapy*, *13*(1), 1–11. https://doi.org/10.1186/s13195-020-00759-3

Ghafoorian, M., Karssemeijer, N., Van Uden, I. W. M., De Leeuw, F. E., Heskes, T., Marchiori, E., & Platel, B. (2016). Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Medical Physics*, *43*(12), 6246–6458. https://doi.org/10.1118/1.4966029

Graff-Radford, J., Arenaza-Urquijo, E. M., Knopman, D. S., Schwarz, C. G., Brown, R. D., Rabinstein, A. A., Gunter, J. L., Senjem, M. L., Przybelski, S. A., Lesnick, T., Ward, C., Mielke, M. M., Lowe, V. J., Petersen, R. C., Kremers, W. K., Kantarci, K., Jack, C. R., & Vemuri, P. (2019). White matter hyperintensities: Relationship to amyloid and tau burden. *Brain*, *142*(8), 2483–2491. https://doi.org/10.1093/brain/awz162

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. https://doi.org/10.1016/j.neuroimage.2016.07.018

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M. C., Dickie, D. A., Wardlaw, J., & Rueckert, D. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, *17*(December 2017), 918–934. https://doi.org/10.1016/j.nicl.2017.12.022

Harmouche, R., Collins, L., Arnold, D., Francis, S., & Arbel, T. (2006). Bayesian MS lesion classification modeling regional and local spatial information. *Proceedings - International Conference on Pattern Recognition*, *3*, 984–987. https://doi.org/10.1109/ICPR.2006.318

Heinen, R., Steenwijk, M. D., Barkhof, F., Biesbroek, J. M., van der Flier, W. M., Kuijf, H. J., Prins, N. D., Vrenken, H., Biessels, G. J., de Bresser, J., van den Berg, E., Biessels, G. J., Boomsma, J. M. F., Exalto, L. G., Ferro, D. A., Frijns, C. J. M., Groeneveld, O. N., Heinen, R., van Kalsbeek, N. M., … Vriens, E. (2019). Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-52966-0

Heuvel, D. M. J. van den, Dam, V. H. ten, Craen, A. J. M. de, Admiraal-Behloul, F., Es, A. C. G. M. van, Palm, W. M., Spilt, A., Bollen, E. L. E. M., Blauw, G. J., Launer, L., Westendorp, R. G. J., & Buchem, M. A. van. (2006). Measuring Longitudinal White Matter Changes : *Ajnr. American Journal Of Neuroradiology*, 875–878.

Hotz, I., Deschwanden, P. F., Liem, F., Mérillat, S., Malagurski, B., Kollias, S., & Jäncke, L. (2021). Performance of three freely available methods for extracting white matter hyperintensities: FreeSurfer , UBO Detector, and BIANCA . *Human Brain Mapping*, *January*, 1–20. https://doi.org/10.1002/hbm.25739

Jang, H., Kwon, H., Yang, J. J., Hong, J., Kim, Y., Kim, K. W., Lee, J. S., Jang, Y. K., Kim, S. T., Lee, K. H.,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Lee, J. H., Na, D. L., Seo, S. W., Kim, H. J., & Lee, J. M. (2017). Correlations between Gray Matter and White Matter Degeneration in Pure Alzheimer's Disease, Pure Subcortical Vascular Dementia, and Mixed Dementia. *Scientific Reports*, *7*(1), 1–9. https://doi.org/10.1038/s41598-017-10074-x

Jiménez-Balado, J., Corlier, F., Habeck, C., Stern, Y., & Eich, T. (2022). Effects of white matter hyperintensities distribution and clustering on late-life cognitive impairment. *Scientific Reports*, *12*(1), 1–13. https://doi.org/10.1038/s41598-022-06019-8

Khademi, A., Gibicar, A., Arezza, G., DiGregorio, J., Tyrrell, P. N., & Moody, A. R. (2021). Segmentation of white matter lesions in multicentre FLAIR MRI. *Neuroimage: Reports*, *1*(4), 100044. https://doi.org/10.1016/j.ynirp.2021.100044

Kloppenborg, R. P., Nederkoorn, P. J., Geerlings, M. I., & Van Den Berg, E. (2014). Presence and progression of white matter hyperintensities and cognition: A meta-analysis. *Neurology*, *82*(23), 2127–2138. https://doi.org/10.1212/WNL.0000000000000505

Lam, S., Lipton, R. B., Harvey, D. J., Zammit, A. R., & Ezzati, A. (2021). White matter hyperintensities and cognition across different Alzheimer's biomarker profiles. *Journal of the American Geriatrics Society*, *69*(7), 1906–1915. https://doi.org/10.1111/jgs.17173

Lawton, M. P., & Brody, E. M. (1969). Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living. *The Gerontologist*, *9*(3 Part 1), 179–186. https://doi.org/10.1093/geront/9.3_Part_1.179

Lesjak, Ž., Galimzianova, A., Koren, A., Lukin, M., Pernuš, F., Likar, B., & Špiclin, Ž. (2018). A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. *Neuroinformatics*, *16*(1), 51–63. https://doi.org/10.1007/s12021-017-9348-7

Ling, Y., Jouvent, E., Cousyn, L., Chabriat, H., & De Guio, F. (2018). Validation and Optimization of BIANCA for the Segmentation of Extensive White Matter Hyperintensities. *Neuroinformatics*, *16*(2), 269–281. https://doi.org/10.1007/s12021-018-9372-2

Lobo, A., Ezquerra, J., Gómez Burgada, F., Sala, J. M., & Seva Díaz, A. (1979). [Cognocitive mini-test (a simple practical test to detect intellectual changes in medical patients)]. *Actas Luso-Espanolas de Neurologia, Psiquiatria y Ciencias Afines*, *7*(3), 189–202. http://www.ncbi.nlm.nih.gov/pubmed/474231

Misquitta, K., Dadar, M., Louis Collins, D., & Tartaglia, M. C. (2020). White matter hyperintensities and neuropsychiatric symptoms in mild cognitive impairment and Alzheimer's disease. *NeuroImage: Clinical*, *28*(May), 102367. https://doi.org/10.1016/j.nicl.2020.102367

Mortamais, M., Artero, S., & Ritchie, K. (2014). White Matter Hyperintensities as Early and Independent Predictors of Alzheimer's Disease Risk. *Journal of Alzheimer's Disease*, *42*, S393–S400. https://doi.org/10.3233/JAD-141473

Park, B. yong, Lee, M. J., Lee, S. hak, Cha, J., Chung, C. S., Kim, S. T., & Park, H. (2018). DEWS (DEep White matter hyperintensity Segmentation framework): A fully automated pipeline for detecting small deep white matter hyperintensities in migraineurs. *NeuroImage: Clinical*, *18*(February), 638–647. https://doi.org/10.1016/j.nicl.2018.02.033

Petersen, R. C., & Negash, S. (2008). Mild Cognitive Impairment: An Overview. *CNS Spectrums*, *13*(1), 45–53. https://doi.org/10.1017/S1092852900016151

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Pfeffer, R. I., Kurosaki, T. T., Harrah, C. H., Chance, J. M., & Filos, S. (1982). Measurement of Functional Activities in Older Adults in the Community. *Journal of Gerontology*, *37*(3), 323–329. https://doi.org/10.1093/geronj/37.3.323

Prins, N. D., & Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: An update. *Nature Reviews Neurology*, *11*(3), 157–165. https://doi.org/10.1038/nrneurol.2015.10

Qin, C., Guerrero, R., Bowles, C., Chen, L., Dickie, D. A., Valdes-Hernandez, M. del C., Wardlaw, J., & Rueckert, D. (2018). A large margin algorithm for automated segmentation of white matter hyperintensity. *Pattern Recognition*, *77*, 150–159. https://doi.org/10.1016/j.patcog.2017.12.016

Quandt, F., Fischer, F., Schröder, J., Heinze, M., Lettow, I., Frey, B. M., Kessner, S. S., Schulz, M., Higgen, F. L., Cheng, B., Gerloff, C., & Thomalla, G. (2020). Higher white matter hyperintensity lesion load is associated with reduced long-range functional connectivity. *Brain Communications*, *2*(2), 1–12. https://doi.org/10.1093/braincomms/fcaa111

Reisberg, B., Ferris, S., De Leon, M., & Crook, T. (1982). Lent of Primary. *Gerontologist*, *139*(9), 1136–1139.

Sarbu, N., Shih, R. Y., Jones, R. V., Horkayne-Szakaly, I., Oleaga, L., & Smirniotopoulos, J. G. (2016). White matter diseases with radiologic-pathologic correlation. *Radiographics*, *36*(5), 1426–1447. https://doi.org/10.1148/rg.2016160031

Schmidt, P. (2016). *Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. Dissertation, LMU München: Faculty of Mathematics, Computer Science and Statistics*. *November*, Chapter 6.1. https://edoc.ub.uni-muenchen.de/20373/

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. https://doi.org/10.1016/j.neuroimage.2011.11.032

Schulz, M., Malherbe, C., Cheng, B., Thomalla, G., & Schlemm, E. (2021). Functional connectivity changes in cerebral small vessel disease - a systematic review of the resting-state MRI literature. *BMC Medicine*, *19*(1), 1–29. https://doi.org/10.1186/s12916-021-01962-1

Silbert, L. C., Howieson, D. B., Dodge, H., & Kaye, J. A. (2009). Cognitive impairment risk: White matter hyperintensity progression matters. *Neurology*, *73*(2), 120–125. https://doi.org/10.1212/WNL.0b013e3181ad53fd

Soldan, A., Pettigrew, C., Zhu, Y., Wang, M. C., Moghekar, A., Gottesman, R. F., Singh, B., Martinez, O., Fletcher, E., Decarli, C., & Albert, M. (2020). White matter hyperintensities and CSF Alzheimer disease biomarkers in preclinical Alzheimer disease. *Neurology*, *94*(9), e950–e960. https://doi.org/10.1212/WNL.0000000000008864

Tran, P., Thoprakarn, U., Gourieux, E., dos Santos, C. L., Cavedo, E., Guizard, N., Cotton, F., Krolak-Salmon, P., Delmaire, C., Heidelberg, D., Pyatigorskaya, N., Ströer, S., Dormont, D., Martini, J. B., & Chupin, M. (2022). Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both Multiple Sclerosis and elderly subjects. *NeuroImage: Clinical*, *33*. https://doi.org/10.1016/j.nicl.2022.102940

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Van Den Berg, E., Geerlings, M. I., Biessels, G. J., Nederkoorn, P. J., & Kloppenborg, R. P. (2018). White Matter Hyperintensities and Cognition in Mild Cognitive Impairment and Alzheimer's Disease: A Domain-Specific Meta-Analysis. *Journal of Alzheimer's Disease*, *63*(2), 515–527. https://doi.org/10.3233/JAD-170573

Van Leijsen, E. M. C., Van Uden, I. W. M., Ghafoorian, M., Bergkamp, M. I., Lohner, V., Kooijmans, E. C. M., Van Der Holst, H. M., Tuladhar, A. M., Norris, D. G., Van DIjk, E. J., Rutten-Jacobs, L. C. A., Platel, B., Klijn, C. J. M., & De Leeuw, F. E. (2017). Nonlinear temporal dynamics of cerebral small vessel disease. *Neurology*, *89*(15), 1569–1577. https://doi.org/10.1212/WNL.0000000000004490

van Straaten, E. C. W., de Haan, W., de Waal, H., Scheltens, P., van der Flier, W. M., Barkhof, F., Koene, T., & Stam, C. J. (2012). Disturbed oscillatory brain dynamics in subcortical ischemic vascular dementia. *BMC Neuroscience*, *13*(1). https://doi.org/10.1186/1471-2202-13-85

van Straaten, E. C. W., den Haan, J., de Waal, H., van der Flier, W. M., Barkhof, F., Prins, N. D., & Stam, C. J. (2015). Disturbed phase relations in white matter hyperintensity based vascular dementia: An EEG directed connectivity study. *Clinical Neurophysiology*, *126*(3), 497–504. https://doi.org/10.1016/j.clinph.2014.05.018

Vanderbecq, Q., Xu, E., Ströer, S., Couvy-Duchesne, B., Diaz Melo, M., Dormont, D., & Colliot, O. (2020). Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. *NeuroImage: Clinical*, *27*(July), 102357. https://doi.org/10.1016/j.nicl.2020.102357

Veldsman, M., Tai, X. Y., Nichols, T., Smith, S., Peixoto, J., Manohar, S., & Husain, M. (2020). Cerebrovascular risk factors impact frontoparietal network integrity and executive function in healthy ageing. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-18201-5

Wahlund, L. O., Barkhof, F., Fazekas, F., Bronge, L., Augustin, M., Sjögren, M., Wallin, A., Ader, H., Leys, D., Pantoni, L., Pasquier, F., Erkinjuntti, T., & Scheltens, P. (2001). A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke*, *32*(6), 1318–1322. https://doi.org/10.1161/01.STR.32.6.1318

Wahlund, Lars Olof, Westman, E., van Westen, D., Wallin, A., Shams, S., Cavallin, L., & Larsson, E. M. (2017). Imaging biomarkers of dementia: recommended visual rating scales with teaching cases. *Insights into Imaging*, *8*(1), 79–90. https://doi.org/10.1007/s13244-016-0521-6

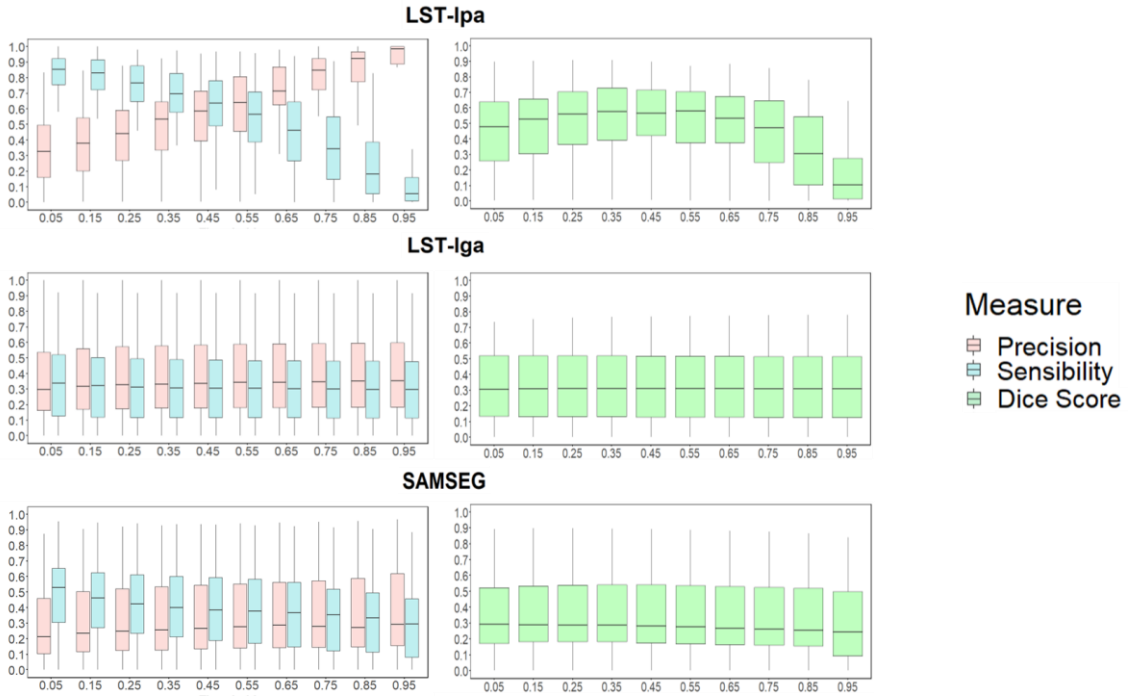
Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R. I., O'Brien, J. T., Barkhof, F., Benavente, O. R., Black, S. E., Brayne, C., Breteler, M., Chabriat, H., DeCarli, C., de Leeuw, F. E., Doubal, F., Duering, M., Fox, N. C., … Dichgans, M. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, *12*(8), 822–838. https://doi.org/10.1016/S1474-4422(13)70124-8

Wardlaw, J. M., Valdés Hernández, M. C., & Muñoz-Maniega, S. (2015). What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. *Journal of the American Heart Association*, *4*(6), 001140. https://doi.org/10.1161/JAHA.114.001140

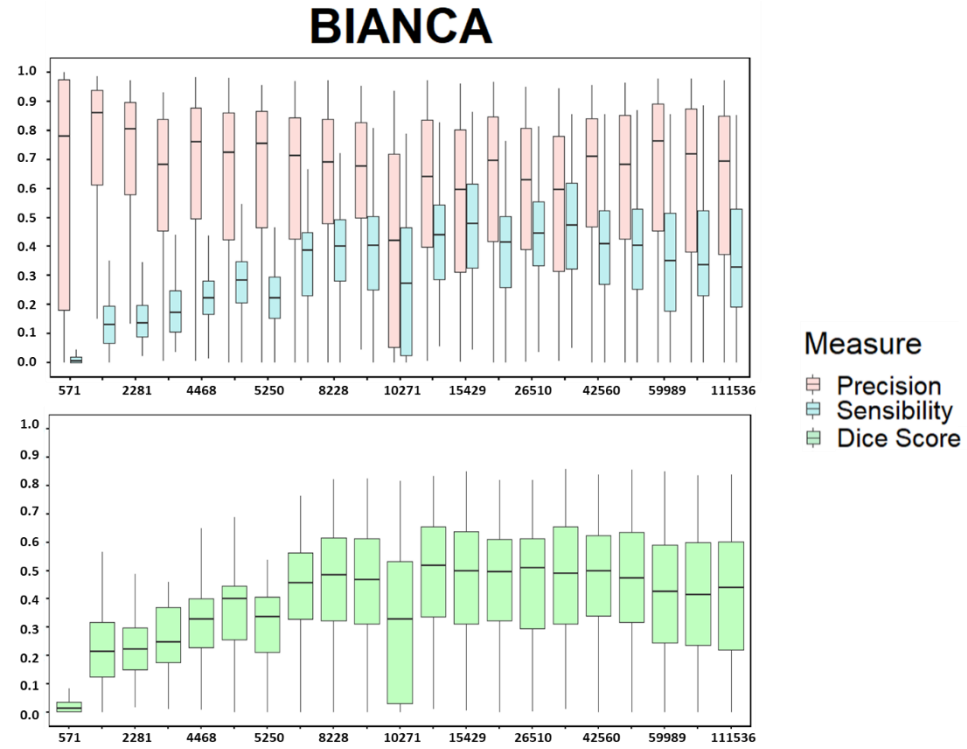
Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research*, *17*(1), 37–49. https://doi.org/10.1016/0022-3956(82)90033-4

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

7. Supplementary material



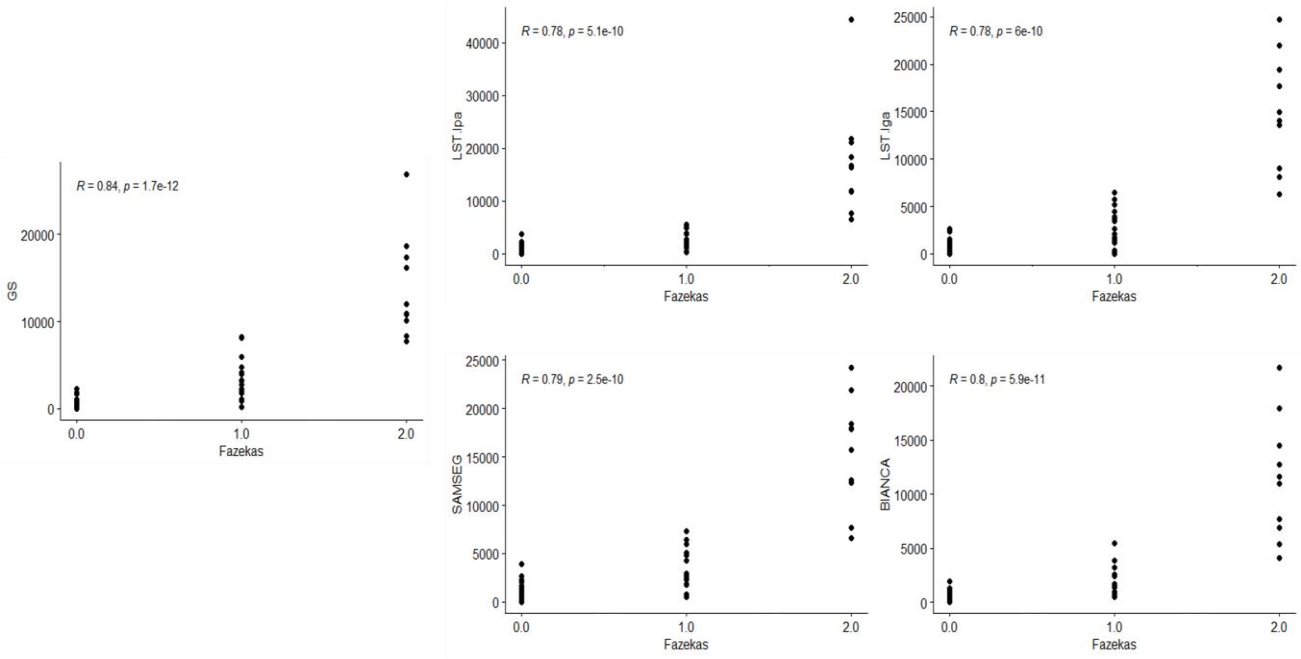
Supplementary 1. Lesion segmentation performance in terms of precision, sensitivity and Dice score for the proposed methods by threshold (from up to down: LST-LPA, LST-LGA y SAMSEG). Blue boxplots represent sensitivity distributions, red boxplots represent precision distributions and green boxplots represent Dice score distributions.



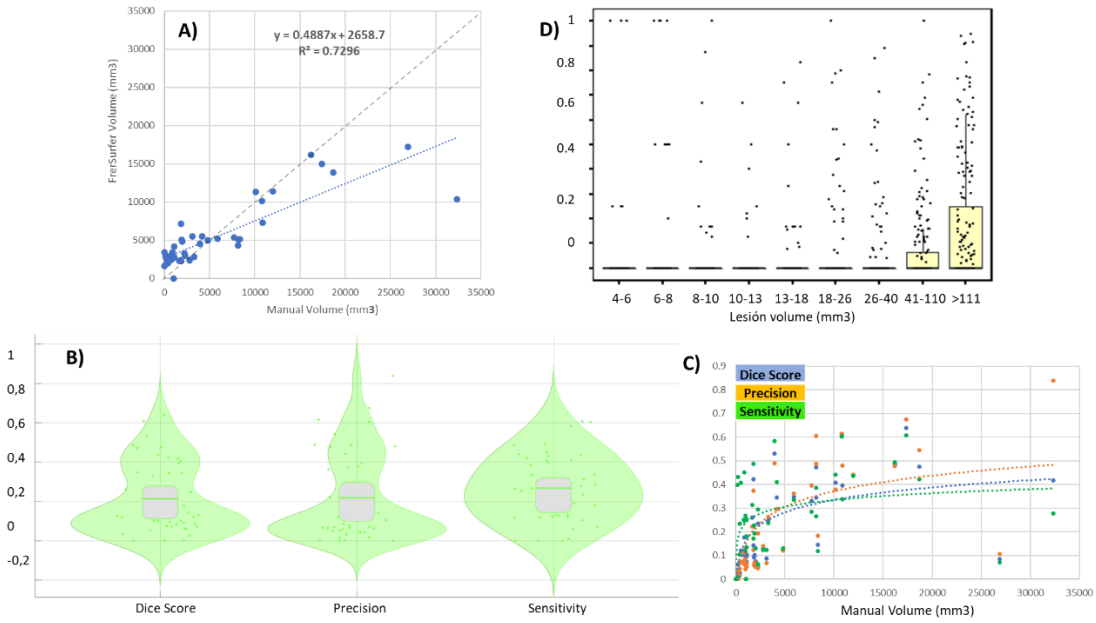
Supplementary 2. Lesion segmentation performance in terms of precision, sensitivity, and Dice score for BIANCA by the lesion voxels employed in the training. Blue boxplots represent sensitivity distributions, red boxplots represent precision distributions and green boxplots represent Dice score distributions.

Supplementary 2. Lesion segmentation performance in terms of precision, sensitivity, and Dice score for BIANCA by the lesion voxels employed in the training. Blue boxplots represent sensitivity distributions, red boxplots represent precision distributions and green boxplots represent Dice score distributions.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.30.23287946; this version posted April 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



Supplementary 3. Volume-Fazekas correlations. The volume values for each automatic tool and the Gold Standard are plotted as a function of the Fazekas scale assigned by the clinicians. the spearman correlation r-value is shown in each graph.



Supplementary 4. Freesurfer performance evaluation. A) Correlation with Gold standard volume; B) Violin plot of Dice score, precision and sensitivity. C) Influence of WMH load on performance. D) Sensitivity at the individual lesion level.