1-1-2024

Accuracy of TrUE-Net in comparison to established white matter hyperintensity segmentation methods: An independent validation study

Jeremy F Strain
Washington University School of Medicine in St. Louis

Maryam Rahmani
Washington University School of Medicine in St. Louis

Donna Dierker
Washington University School of Medicine in St. Louis

Christopher Owen
Washington University School of Medicine in St. Louis

Hussain Jafri
Washington University School of Medicine in St. Louis

See next page for additional authors

Recommended Citation

Strain, Jeremy F; Rahmani, Maryam; Dierker, Donna; Owen, Christopher; Jafri, Hussain; Vlassenko, Andrei G; Womack, Kyle; Fripp, Jurgen; Tosun, Duygu; Benzinger, Tammie L S; Weiner, Michael; Masters, Colin; Lee, Jin-Moo; Morris, John C; Goyal, Manu S; and ADOPIC and ADNI Investigators, "Accuracy of TrUE-Net in comparison to established white matter hyperintensity segmentation methods: An independent validation study." NeuroImage. 285, 120494 (2024).
https://digitalcommons.wustl.edu/oa_4/3771

Authors

Jeremy F Strain, Maryam Rahmani, Donna Dierker, Christopher Owen, Hussain Jafri, Andrei G Vlassenko, Kyle Womack, Jurgen Fripp, Duygu Tosun, Tammie L S Benzinger, Michael Weiner, Colin Masters, Jin-Moo Lee, John C Morris, Manu S Goyal, and ADOPIC and ADNI Investigators

Contents lists available at ScienceDirect

NeuroImage

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

Accuracy of TrUE-Net in comparison to established white matter hyperintensity segmentation methods: An independent validation study

Jeremy F. Strain [a,h,1,*], Maryam Rahmani [b,h,1], Donna Dierker [b,h], Christopher Owen [b], Hussain Jafri [b], Andrei G. Vlassenko [b,h], Kyle Womack [a], Jurgen Fripp [f], Duygu Tosun [e], Tammie L.S. Benzinger [b,c], Michael Weiner [e], Colin Masters [g], Jin-Moo Lee [a,b,d], John C. Morris [a,c], Manu S. Goyal [b,h], for the ADOPIC and ADNI Investigators

[a] *Department of Neurology, Washington University School of Medicine, St. Louis, MO, USA*
[b] *Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO, USA*
[c] *Knight Alzheimer Disease Research Center, St. Louis, MO, USA*
[d] *Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA*
[e] *Division of Radiology and Biomedical Imaging, University of California – San Francisco, San Francisco, CA, USA*
[f] *The Australian e-Health Research Centre, CSIRO Health and Biosecurity, Brisbane, QLD, Australia*
[g] *The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, Victoria, Australia*
[h] *Neuroimaging Labs Research Center, Washington University School of Medicine, St. Louis MO, USA*

ARTICLE INFO

Keywords:
WMH
Segmentation Tools
TrUE-Net
Aging
LST

ABSTRACT

White matter hyperintensities (WMH) are nearly ubiquitous in the aging brain, and their topography and overall burden are associated with cognitive decline. Given their numerosity, accurate methods to automatically segment WMH are needed. Recent developments, including the availability of challenge data sets and improved deep learning algorithms, have led to a new promising deep-learning based automated segmentation model called TrUE-Net, which has yet to undergo rigorous independent validation. Here, we compare TrUE-Net to six established automated WMH segmentation tools, including a semi-manual method. We evaluated the techniques at both global and regional level to compare their ability to detect the established relationship between WMH burden and age. We found that TrUE-Net was highly reliable at identifying WMH regions with low false positive rates, when compared to semi-manual segmentation as the reference standard. TrUE-Net performed similarly or favorably when compared to the other automated techniques. Moreover, TrUE-Net was able to detect relationships between WMH and age to a similar degree as the reference standard semi-manual segmentation at both the global and regional level. These results support the use of TrUE-Net for identifying WMH at the global or regional level, including in large, combined datasets.

1. Introduction

Neuroimaging has enabled the quantification of various senescent changes in the aging brain, including those which impact cognition and neurodegenerative diseases. Among the most common changes observed on brain MRI are white matter hyperintensities (WMH), which are hyperintense lesions that are particularly discernible on T2-weighted fluid attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI) sequences (Caligiuri et al., 2015; Wardlaw et al., 2013). WMH presence may arise from various pathologies but have robust associations with age, hypertension and other vascular risk factors (Debette and Markus, 2010; Gouw et al., 2011). The burden of WMH is associated with cognitive decline and vascular dementia, and there is increasing interest in their association with Alzheimer disease and other related neurodegenerative diseases (Haley et al., 2009; Inzitari et al., 2009; Prins and Scheltens, 2015).

Numerous approaches have been introduced for assessing WMH but none is universally applied. Initial methods applied qualitative visual reads by radiologists to grade WMH severity, which remains a common practice, particularly in clinical case studies (Fazekas et al., 1987;

[*] Corresponding author.
E-mail address: strainj@wustl.edu (J.F. Strain).
[1] Contributed equally to the paper.

https://doi.org/10.1016/j.neuroimage.2023.120494
Received 28 June 2023; Received in revised form 23 October 2023; Accepted 9 December 2023
Available online 10 December 2023

Scheltens et al., 1993). However, this method is spatially limited, vulnerable to poor inter-rater reliability, and applies varying criteria among different studies (Mäntylä et al., 1997). Manual segmentation of WMH has been used but is labor-intensive and not feasible for large sample sizes, and is vulnerable to rater bias and sequence variability (Grimaud et al., 1996; Vanderbecq et al., 2020).

Automated WMH segmentation pipelines have evolved as an alternative using a variety of computational techniques, including deep learning (Balakrishnan et al., 2021). Early approaches applied arbitrary thresholds using one or multiple image sequences to capture intensity outliers designated as WMH (Beare et al., 2009). More advanced strategies such as Bayesian regression approaches subsequently emerged (Maillard et al., 2022; DeCarli et al., 2013). Most current popular neuroimaging toolboxes incorporate some combination of such methods including the Lesion Segmentation Toolbox (LST) (Schmidt et al., 2012; Schmidt, 2017) and the MarkVCID pipeline (Maillard et al., 2022).

In 2017, an ambitious WMH segmentation challenge was initiated to assess the capabilities and reliabilities of various WMH segmentation tools. In 2019 the study revealed that deep learning techniques based on convolutional neural networks far surpassed other methods (Kuijf et al., 2019; Maillard et al., 2022; DeCarli et al., 2013). The technique that outperformed the other algorithms, sysu media (SM, Li et al., 2018), exhibited strong reliability across various sequence and scanner types implicating the value of convolutional neural networks for structural imaging. A newer techniques called the Triplanar U-Net ensemble network (TrUE-Net) outperformed several existing pipelines within this class and is of particular interest due to the triplanar model that improves robustness to scanner and acquisition protocol differences, thereby eliminating the need for post-processing measurement harmonization (Sundaresan et al., 2021). However, TrUE-Net has yet to be independently validated. Moreover, the WMH segmentation challenge parameters did not adequately evaluate accuracy regionally nor for biological relevance, which are of critical interest with respect to cognitive decline and neurodegeneration (Biesbroek et al., 2017).

Accordingly, here we report on an independent assessment of the TrUE-Net algorithm in comparison to sysu media (SM), two techniques from LST, MarkVCID, FreeSurfer (Fischl 2012) and a semi-manual segmentation based on intensity thresholding (MSIT) method, the latter as the reference standard. We demonstrate the performance of these six automated segmentation techniques based on three criteria: 1) consistency in assessing global and regional WMH burden in reference to semi-manual segmentation, 2) accuracy in identifying lesions (DICE, Hausdorff Distance), and 3) detection of biological effects namely the association between WMH and age, at the global and regional level.

2. Methods

2.1. Test data

MRI data from a total of 160 individuals were used to assess the capabilities of the WMH segmentation methods. These data were obtained from the Aging Metabolism & Brain Resilience (AMBR) dataset (Goyal et al., 2023), which included advanced brain MRI and PET, among other assessments in community dwelling adults with or without Alzheimer disease; other known significant neurological illnesses were an exclusion criterion including for example symptomatic stroke. Details have been posted previously described elsewhere (Goyal et al., 2023). The brain MRIs in the AMBR dataset include a high-resolution 3D FLAIR sequence described below. Visual ratings were conducted to provide a clinical metric in terms of WMH burden across our entire cohort. The ratings were classified as none/minimal, mild, moderate, or severe by a neuroradiologist based on the higher Fazekas score for deep versus periventricular regions.

Structural imaging sequences included high-resolution 3D FLAIR (TR/TE = 4800/417 ms, FOV 256 mm, voxel size 1 × 1 × 1 mm, matrix size 256 × 256 × 160 voxels) and T1-weighted rapid gradient-echo (MPRAGE) sequence (TR/TE = 2500.0/(1.81–7.18) ms, FOV 256 mm, voxel size 0.8 × 0.8 × 0.8 mm, matrix size 256 × 256 × 208 voxels, vNav prospective motion corrected). All images were acquired on 3T Trio/Prisma Siemens scanners.

The goal of this WMH technique comparison was not to optimize any of the pipelines but rather perform them as efficiently as possible in order to assess their capability for the general population with unknown cohort size or characteristics. Therefore, no fine-tuning of algorithmic parameters (TrUE-Net and SM) were conducted on our cohort; rather, WMH segmentation parameters were derived from their baseline training dataset. Further details on technique specific adjustments can be found below in the Image Processing section.

2.2. Image processing

A set of standardized preprocessing steps was established to ensure efficient segmentation across all WMH techniques. These steps include: registering the T1-weighted image to the corresponding FLAIR with rigid body registration using FLIRT from FSL, followed by brain extraction and bias field correction using FSL FAST (Jenkinson and Smith, 2001). Several of the applied segmentation techniques incorporated one or more steps of the standardized preprocessing to maximize the number of individuals that passed visual QC for each technique. The full preprocessing pipeline was used for MSIT, MarkVCID and TrUE-Net. The preprocessing for SM only included the co-registration of the T1 to the FLAIR as the remaining preprocessing steps are embedded within the SM package itself. The LST tool box contains all of the above mentioned preprocessing and therefore the raw T1 and FLAIR were used as inputs. No preprocessing or fine tuning was done for FreeSurfer which uses T1-weighted images for WMH segmentation.

Following WMH segmentation, each T1 image in FLAIR space was linearly and nonlinearly registered to the MNI152 template using FLIRT and FNIRT, respectively (Jenkinson and Smith, 2001). The warp parameters created for the T1 were then applied to the binary lesion masks with nearest-neighbor interpolation for further analyses. Finally, the MNI white matter mask was applied to WMH Lesion masks across all techniques to isolate voxels that only resided within the white matter.

2.3. MSIT

The MSIT approach segments WMH by initial intensity thresholding followed by trained manual selection of lesions, seed growing, and quality assessment by a trained neuroradiologist (MG) (Strain et al., 2013). The resulting segmentations were deemed to be the reference standard in this study given this highly rigorous but labor-intensive approach. In short, this method uses an in-house MATLAB code to determine the average intensity and standard deviation at each image slice along the z-plane. A threshold of 1.2 SD was applied as a liberal measurement to have high sensitivity but low specificity. This threshold has shown to be reliable in several prior studies across different sequences and cohorts (Hubbard et al., 2017; Hubbard et al., 2016; Strain et al., 2013). Manual selection of lesions (and as needed tracings) were then performed by identifying true lesion from false positives due to motion, fat signal, ventricles or other sources that would not be considered WMH. To ensure that all WMH clusters were fully represented, the manually selected clusters were treated as seed regions and allowed to expand one voxel in all directions provided the signal intensity was ≥0.5 SD from the slice average intensity.

All WMH binary masks were drawn by the same two raters (MR and CO). Inter- and intra-rater reliability were assessed in a separate cohort of 20 individuals with a range of WMH burden. Each rater was assessed against one another and against an expert with extensive prior experience in WMH segmentation (JS). Each rater was blind to the testing cohort prior to reliability assessment. Following the inter-rater reliability evaluation, the 20 scans were randomized with instructions to identify the WMH a second time with a week between tracings. An average Dice cut-off of 85 % was set as an acceptable reliability rating where >70 % is often considered excellent agreement (Zhang et al., 2007).

2.4. TrUE-Net

We chose not to bias the TrUE-Net algorithm by retraining the model to our data; instead, we utilized the established parameters derived from the MICCAI dataset (Kuijf et al., 2019) and the Neurodegenerative (NDGEN) cohort (Zamboni et al., 2013), to optimize the generalizability of our results. The TrUE-Net segmentation is a convolutional neural net that consists of an ensemble of U-Nets. Briefly, the TrUE-Net algorithm uses a 3-layer deep U-Net consisting of 2D networks for detecting WMH in each plane. Additionally, the TrUE-Net algorithm uses a loss function that considers the anatomical location and distribution of WMHs. To compensate for the bias of larger periventricular WMHs, deep WMH were up-weighted to favor finding small lesions and reducing false negatives. The TrUE-Net algorithm includes a weighted map that increase the likelihood of WMH selection outside PVWMH in order to identify small deep WMH. Further details can be found in the TrUE-Net publication cited in the manuscript (Sundaresan et al., 2021). We selected the recommended threshold of 0.5 from the final probabilistic map to evaluate the model performance.

2.5. LST

The lesion segmentation algorithm toolbox includes two separate algorithms, lesion growth algorithm (LGA, Schmidt et al., 2012) and lesion prediction algorithm (LPA, Schmidt, 2017).

The LGA algorithm needs both T1 and FLAIR images. An initial lesion probability map is produced by combining FLAIR intensities and T1 images segmented into three main tissue classes (CSF, GM and WM). This map is then binarized and lesions were treated as seeds and subsequently grown based on surrounding voxel intensity on FLAIR. The toolbox includes a routine to determine the best initial threshold for the seed map, but in order to not bias the model the default value of 0.3 was used. Masks were then binarized at an a-priori threshold of 0.5.

LPA uses parameters from a high dimensional logistic regression model to estimate the probability of each voxel being a lesion, with lesion belief maps based on FLAIR intensity and tissue classification incorporated as a covariate; a spatial covariate is also included. The established logistic regression was trained on binary lesion maps of multiple sclerosis patients We did not train the model in our data and used the established training parameters. A recommended threshold of 0.5 was then used to create the binary mask from the lesion probability maps.

2.6. Freesurfer

FreeSurfer is a widely used software for brain surface mapping and morphometry and uses T1-weighted images to identify white matter hypointense lesions. The software uses a spatial intensity gradient across all tissue classes to identify WM hypointensities.

2.7. Sysu media (SM)

This algorithm ranked 1st in the evaluations done on the WMH Segmentation Challenge, MICCAI 2017. To avoid biasing the results we used the parameters set by the MICCAI training dataset and did not retrain the model on our data. Sysu Media is an ensemble of 3 fully convolutional U-net like structures. Original training model also included data augmentation and a loss function. A final post processing removes any lesions detected in the first and last 1/8 of the slices in order to exclude any unreasonable lesion detections. The final lesion probability mask is then binarized at the recommended 0.5 to produce a WMH lesion map.

2.8. MarkVCID

The established MarkVCID WMH segmentation method utilizes a Bayesian regression by creating a four-tissue segmentation based on the multimodal T1-weighted and FLAIR images. Each voxel is converted to a probability value that pertains to the likelihood of being a WMH lesion. Standard cut-offs were applied to create a binary mask as previously defined (Maillard et al., 2022; DeCarli et al., 2013).

2.9. Statistics

Global WMH volumes were computed as the sum of all voxels designated as lesion in cubic millimeters for each WMH segmentation technique. Total WMH volumes followed a log-normal distribution, and were thus log10 transformation prior to further analysis. Regional WMH volumes were calculated by overlaying the binary lesion masks onto 18 predefined WM tracts in MNI space (Strain et al., 2017). All voxels that resided within the WM tract masks were added together to represent the total lesion burden within the corresponding tract.

The WMH techniques were evaluated with three different criteria: 1) consistency in assessing global and regional WMH burden, 2) accuracy in identifying lesions, and 3) detection of biological effects—namely the association between WMH and age. 1) Pearson correlations were computed between the semi-manually segmented WMH volumes and each automated WMH technique to assess WMH segmentation consistency globally. 2) We calculated Dice coefficients, false positive and false negative rates of voxels and clusters using the Bianca_overlap_measures function within FSL to assess accuracy for identifying lesions (Udupa et al., 2006). In addition, we calculated the average Hausdorff distance which evaluates the lesion boundaries between the reference standard (MSIT) and the respective WMH segmentation method (Aydin et al., 2021; Taha and Hanbury, 2015). As the focus of this paper is on TrUE-Net we evaluated the overlap and distance measures computed for TrUE-Net compared to every other technique using a Wilcoxon signed-rank test. 3) Pearson correlations were conducted for each technique comparing WMH volumes (global and regional) to age, as the most readily available and highly established correlate of WMH burden. In a subsequent analysis we evaluated the reliability of each technique for quantifying WMH burden in young and elderly individuals with a cut-off of 50 to define the two groups. As WMH can differ across gender similar analyses were performed on this demographic variable as well. The correlation coefficients for each WMH technique with age and MSIT were then compared. To ascertain the strength of the correlations between the three techniques we calculated the test of the difference between two dependent correlations (Lee and Preacher, 2013).

3. Results

3.1. Demographics

Among the 160 individuals, 14 had cognitive impairment/dementia (Clinical Dementia Rating®, CDR®>0), including 11 who had AD pathological levels of PET amyloid and 3 who were PET amyloid negative. Mean age was 63.7 ± 13.7 years; 93 of the participants were female and 67 were male. Participants included 53 participants without significant WMH according to neuroradiology assessment (mean age 51.9 ± 13.8 years), 78 had mild and 26 had moderate WMH burden (mean ages of 67.4 ± 9.3 and 74.6 ± 7.4 years, respectively). Another 7 participants had severe WMH burden (mean age of 74.1 ± 8.8). Details of demographic characteristics subdivided by qualitative WMH burden are available in Table S1.

3.2. MSIT ratings

For the MSIT technique, both segmentation raters achieved an average DICE coefficient that exceeded our established cutoff of 85 %

with the expert rater (90 % and 88 %) and between each other (87 %). The median DICE for each inter-rater reliability assessment was 94 %. The average intra-rater reliability agreement exceeded 90 % for each rater. These DICE percentages are similar to prior publications that have used similar semi-manual tracing techniques (Hubbard et al., 2016, 2017).

3.3. Method time

On average (based on work time divided by number of scans), the MSIT technique per subject took approximately 25 min of manual effort to complete, though this varied significantly according to the degree of WMH burden. The MarkVCID technique took approximately 120 min of computing time, whereas TrUE-Net and both LST techniques needed approximately 2 min to complete, and computing time for SM was 3 min. Each of these times exclude preprocessing time of about 15 min per scan for MarkVCID and TrUE-Net and 7 min for SM; reorientation of T1 and FLAIR scans for the LST algorithms took less than a minute per scan.

3.4. WMH frequency maps

All WMH frequency maps were created in template space for each of the WMH segmentation strategies (Fig. S1). In general, each segmentation map revealed higher frequency of WMH in similar regions, particularly along the anterior and posterior horns of the lateral ventricles. Visually, the MSIT technique produced the highest WMH frequency but the LPA had the largest distribution. TrUE-Net and SM reported the lowest WMH burden compared to all the other methods. Anatomically, all techniques showed modest WM coverage but the LST methods and MarkVCID technique labeled regions of the septum, and the hippocampal commissure as WMH (Fig. S1).

3.5. Correlation among WMH segmentation methods

All WMH segmentation methods strongly correlated with MSIT for log-transformed global WMH volumes across participants with correlation coefficients ranging from 0.45 (SM) to 0.96 (TrUE-Net) (Fig. 1). Although all techniques associated MSIT, the correlation coefficient for TrUE-Net was significantly stronger than all other techniques ($p < 0.001$ all). Individuals that contained fewer WMH revealed more variability across most of the WMH techniques.

3.6. Reliability

Most techniques revealed low Dice values (Table 1) compared to the MSIT output (Fig. 2). In comparison to TrUE-Net the LST and Freesurfer methods had significantly higher DICE but the SM had significantly lower reliability (Table 1). Similar results were observed for the Matthew correlation coefficient (MCC) metric (Fig. S2) with the LPA showing greater reliability with MSIT. However, TrUE-Net revealed significantly better Hausorff distance values than all other techniques except for LPA (Fig. 3). See Table 1 for statistical comparisons.

With MSIT as the reference standard, TrUE-Net demonstrated very few false positive voxels (Fig. 4) and clusters (Fig. S3) but relatively high number of false negative voxels (Fig. S4) and clusters (Fig. S5). In contrast, LPA had a relatively lower false negative rate compared to all other techniques but also had a high false positive rate. When analyzed within WM tracts of interest, TrUE-Net and LPA demonstrated similar amount of WMH burden across all WM tracts. Across all techniques higher variability was visually observed in the forceps major, forceps minor, and fronto-occipital fasciculus tracts (Fig. S6). Similarly, as observed with the global metrics TrUE-Net consistently underreported the amount of WMH within each track in reference to MSIT. However greater subject variability was visually observed for other techniques. SM and LGA methods did not perform as well when evaluating WMH burden within WM tracks suggesting a lack of focal reliability.

3.7. Correlation with demographic variables

All techniques demonstrated strong positive correlations between global WMH burden and age (Fig. S8) the highest being for MSIT ($r = 0.71$, Fig. S7). The TrUE-Net ($r = 0.7$, respectively) correlation coefficient was significantly higher compared to the other automated WMH techniques (MarkVCID $p = 0.031$; FreeSurfer $p = 0.029$; LPA $p = 0.028$;

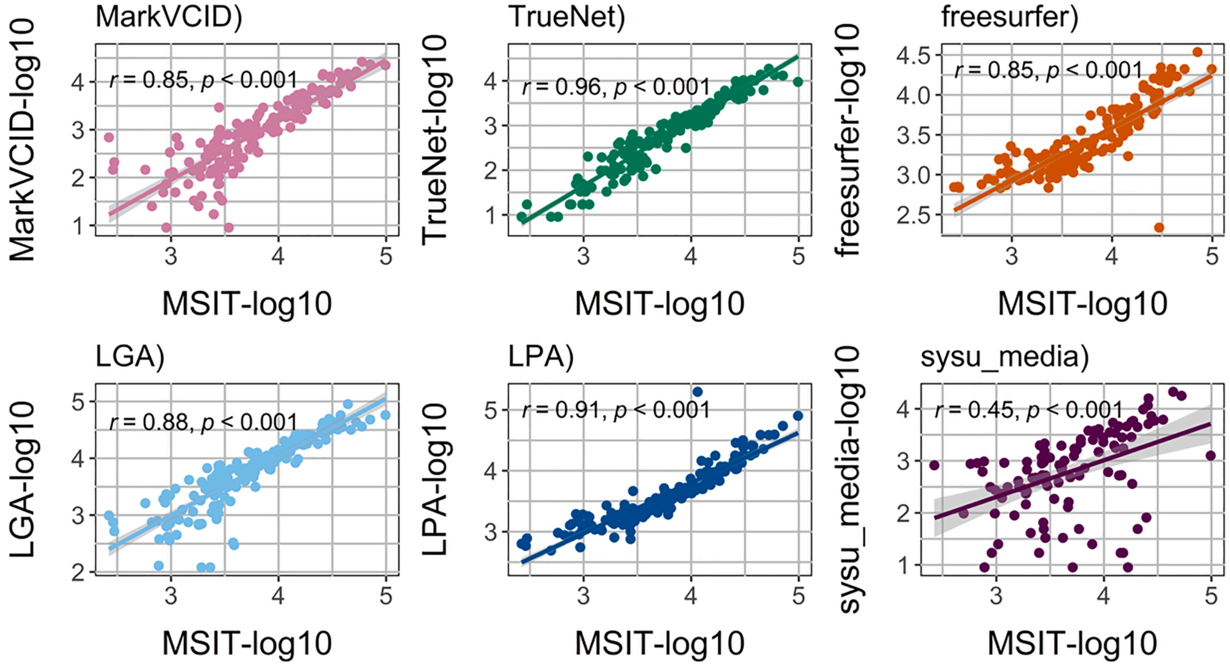

Fig. 1. Global WMH volume for each participant was computed as the summation of all segmented voxels (mm^3) and was log10 transformed for normalization. Pearson correlation was calculated between each WMH segmentation technique with MSIT as the ground truth. All techniques showed a high correlation with MSiT, and increased variability in lower lesions burdens. (Lines represent linear correlation with shaded bars reflecting standard errors).

Table 1
Performance metrics for different segmentation tools, data are reported as median and 1st and 3rd quartiles Correlations are reported as 95 % confidence intervals.

	Median volume (ml)	Processing failure	DSC	MCC	Balanced Average Hausdorff distance	False-positive voxels	False-positive clusters	False-negative voxels	False-negative-clusters	Correlation with age	Correlation with MSIT
True-Net	0.51	5	0.19 (0.1–0.29)	0.30 (0.18–0.41)	1.64 (0.98–3.4)	0.0002 (0.0–0.01)	0.00 (0.0–0.07)	0.89 (0.83–0.94)	0.61 (0.50–0.71)	0.7 (0.60_0.77)	0.96 0.94_0.97)
MarkVCID	0.86	12	0.20 (0.09–0.32)	0.32 (0.23–0.41)	2.25* (1.33–4.0)	0.11* (0.3–0.36)	0.25* (0.13–0.4)	0.88 (0.8–0.95)	0.80* (0.66–0.87)	0.57* (0.45_0.67)	0.85* (0.8–0.9)
LST-LGA	6.55	3	0.41† (0.23–0.50)	0.42† (0.24–0.37)	3.21* (1.49–6.0)	0.6* (0.51–0.64)	0.41* (0.25–0.45)	0.56† (0.95–1.0)	0.7* (0.6–0.82)	0.7 (0.62–0.77)	0.89* (0.83_0.91)
LST-LPA	2.77	0	0.44† (0.35–0.54)	0.46† (0.36–0.57)	1.51† (0.88–2.05)	0.39* (0.22–0.41)	0.62* (0.48–0.76)	0.63† (0.55–0.69)	0.50† (0.30–0.62)	0.58* (0.47–0.67)	0.94* (0.91–0.95)
sysu_ media	0.81	50	0.15* (0.04–0.31)	0.21* (0.09–0.39)	3.9* (1.53–8.5)	0.26* (0.0.4–0.68)	0.59* (0.3–0.76)	0.90 (0.8–0.98)	0.62 (0.4–0.84)	0.24* (0.05 – 0.41)	0.45* (0.26_0.58)
freesurfer	1.92	5	0.30† (0.22–0.42)	0.36† (0.24–0.47)	1.98* (1.38–3.0)	0.42* (0.23–0.68)	0.67* (0.53–0.80)	0.79† (0.68–0.83)	0.68* (0.5–0.8)	0.57* (0.49–0.66)	0.85* (0.78_0.88)

MCC, Matthews correlation coefficient.
* Indicates significant advantage of true-net compared to tested method.
† Indicates significant advantage of tested method compared to true net.
p-values are computed using Wilcoxon ranked test for non correlation metrics, *P* values <0.05 were considered significant.
Correlation metrics were compared using the difference between two dependent correlations.
Distance values are measured in voxels.

SM $p < 0.001$) except for LGA, and showed a similar association when compared to the reference standard MSIT ($p = 0.431$). Regional WM track associations with age revealed similar findings for TrUE-Net, LPA, MSIT, and FreeSurfer. However, MarkVCID had weaker associations in the cingulum and LGA and SM showed weaker relationships among several WM tracks (Fig. 5). Dichotomizing the individuals into young (<50) and elderly (≥50) revealed similar findings of reliability as observed across the entire cohort. All techniques performed better in reference to MSIT values in the elderly cohort compared to the younger cohort (Figs. S9–S11). The LPA technique performed the best in terms of overall Dice values but TrUE-Net had the lowest false positive rate for both elderly and young individuals. Performing similar analyses across gender revealed comparable reliability for most techniques except for the sysu media that was less reliably for males (Figs. S12–S14).

4. Discussion

The aim of our study was to evaluate the consistency, accuracy, and strength of the TrUE-Net method compared to other established WMH techniques along with a semi-manual reference standard and a biological standard. We specifically avoided retraining any of the algorithms to better test their "out-of-box" efficacy. In this context, TrUE-Net correlated best with the reference standard MSIT, produced minimal false positives albeit with greater false negatives, and detected high associations between WMH burden and age. Moreover, TrUE-Net was computationally efficient. However, the potentially resistance of TrUE-Net to different FLAIR acquisition parameters is a particularly valuable feature, likely resulting from its training using a wide variety of FLAIR sequences and tri-planar decomposition of the data. Altogether, these results suggest that TrUE-Net is highly suited for WMH segmentation in large and small data sets.

A "gold standard" for WMH segmentation is not well-defined (Vanderbecq et al., 2020), and it is difficult to determine where to mark a boundary around WMH. This is in part due to the partial volume effects occurring at the irregular margins of WMH, which might themselves be smaller than the resolution of the MRI sequence, and in som cases might also be due to a gradient of signal change at the boundary. Nonetheless, manual segmentation is often used as a reference standard and has been shown to be superior to automated segmentation across different performance parameters (Commowick et al., 2018). Indeed, our semi-manual segmentation method (MSIT) also found the strongest WMH to age associations. However, manual segmentation is highly labor intensive, dependent like other methods on arbitrary threshold selection, and vulnerable to inter-study bias.

In our comparison to reference standard MSIT WMH volume estimates, we observed relatively low Dice similarity coefficients for all WMH segmentation methods except notably for LPA. The Dice coefficient is a common metric for defining similarity between two segmentation techniques but has been subjected to increased scrutiny in the field. The Dice coefficient has a known bias towards true lesion voxels and is not adquately by correct identification of non-lesion voxels (Taha and Hanbury, 2015). The percentage of false positives was dramatically reduced between TrUE-Net and MSIT compared to the other WMH segmentation techniques, suggesting that the low Dice coefficient for TrUE-Net resulted from a high false negative rate, i.e., TrUE-Net identified less WMH as compared to MSIT. Based on our findings we encourage future studies to evaluate the FDR or FNR directly in addition to DICE or measures that incorporate both (e.g., Hausdorff distance and MCC).

Indeed, our data shows that only focusing on DICE performance may unjustly favor techniques that sacrifice sensitivity over specificity. This confound was observed in the LPA mask that yielded a significantly higher DICE than TrUE-Net in association with MSIT. However, TrUE-Net far surpassed LPA in terms of false positives and specificity. Moreover, this increased specificity of TrUE-Net is further reflected in a stronger biological signal in terms of the relationship between WMH and

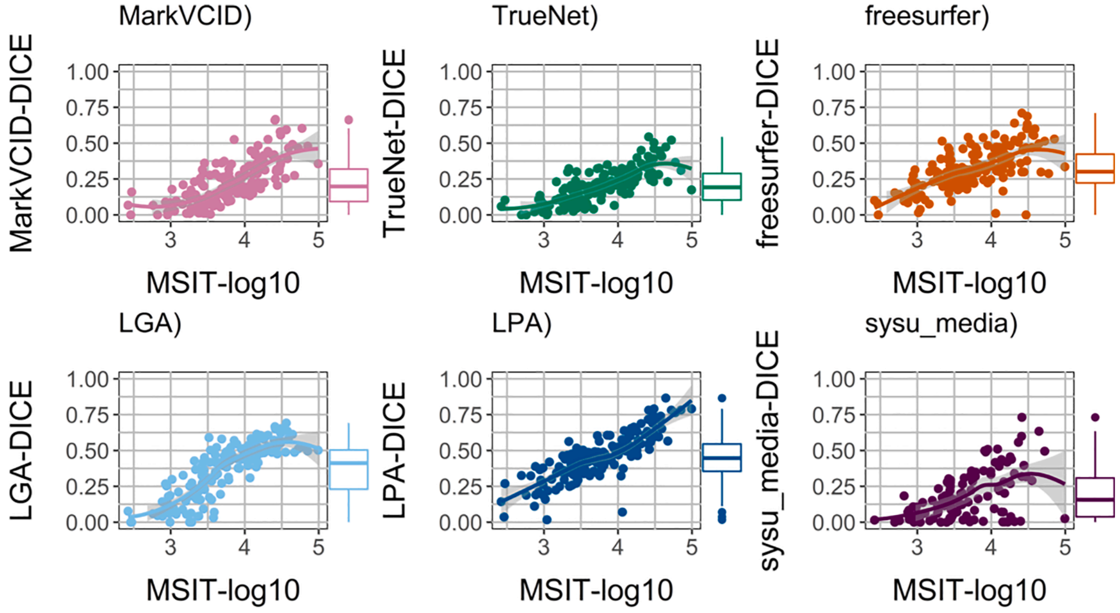

Fig. 2. Dice similarity coefficient between the automated techniques and MSIT as the ground truth. Box plots represent the median and quartile ranges for each WMH segmentation association with MSIT. (Lines represent a Loess function with shaded bars reflecting standard errors).

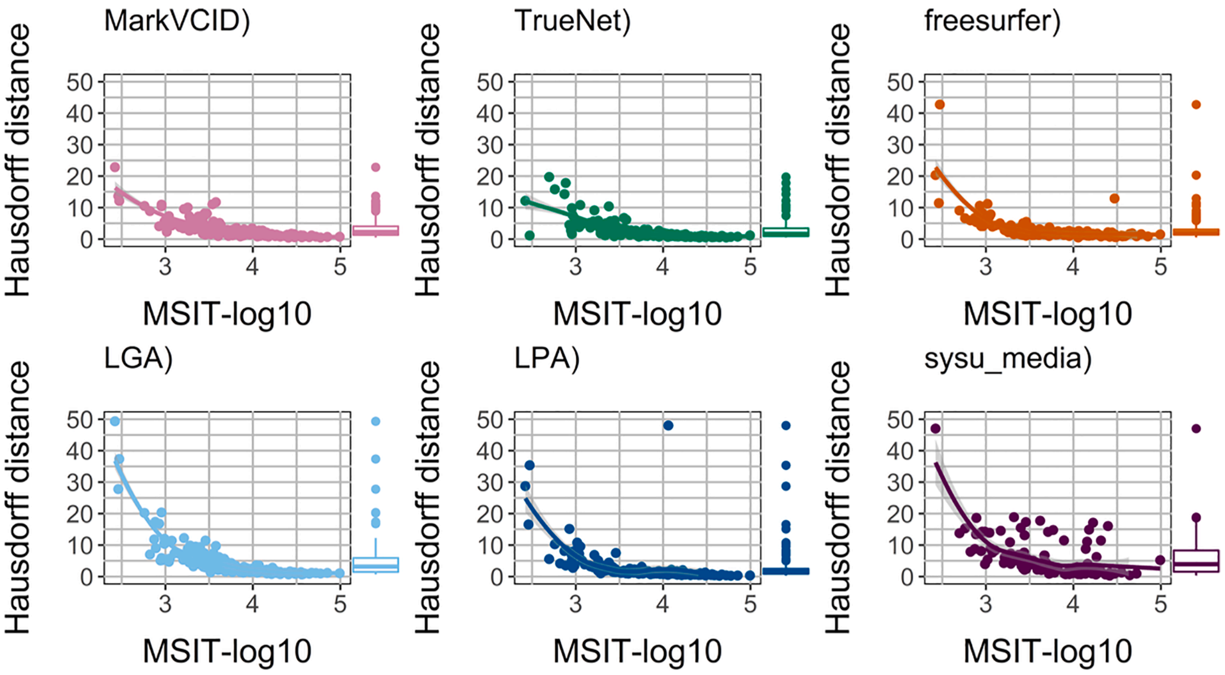

Fig. 3. Average Hausdorff distance values for each WMH segmentation technique. Box plots represent the median and quartile ranges for each WMH segmentation association with MSIT (Lines represent a Loess function with shaded bars reflecting standard errors).

age, which was significantly weaker with LPA. This does not suggest that methods that weight specificity over sensitivity will always produce stronger results because MSIT associated with age just as strongly as TrUE-Net. Accordingly, the sacrifice that LPA makes to increase the sensitivity and thereby inflate reliability metrics like DICE may actually produce noisier outcomes and potentially reduce biological relevance. Accordingly, TrUE-Net may have innate advantages on identifying biologically relevant global or regional WMH values compared to other techniques that sacrifice specificity over sensitivity.

WMH severity is commonly calculated as a single global value but recent studies have suggested that the spatial pattern of WMH can be disease specific (McAleese et al., 2021). The WM tract analysis was thus conducted to assess the reliability of the WMH segmentation techniques within different regions. Tract size is not a factor in this analysis as the goal was to evaluate the pattern across tracts among WMH segmentation methods and not the level of burden within an individual tract. Interestingly, our findings suggest some variability among the different tracts in comparing TrUE-Net with MSIT but Bayesian methods in particular were far less reliable at regional burden across all tracks as opposed to the global burden. This suggests that differences in regional performance is indeed a concern when comparing WMH segmentation methods and should also be considered when evaluating segmentation

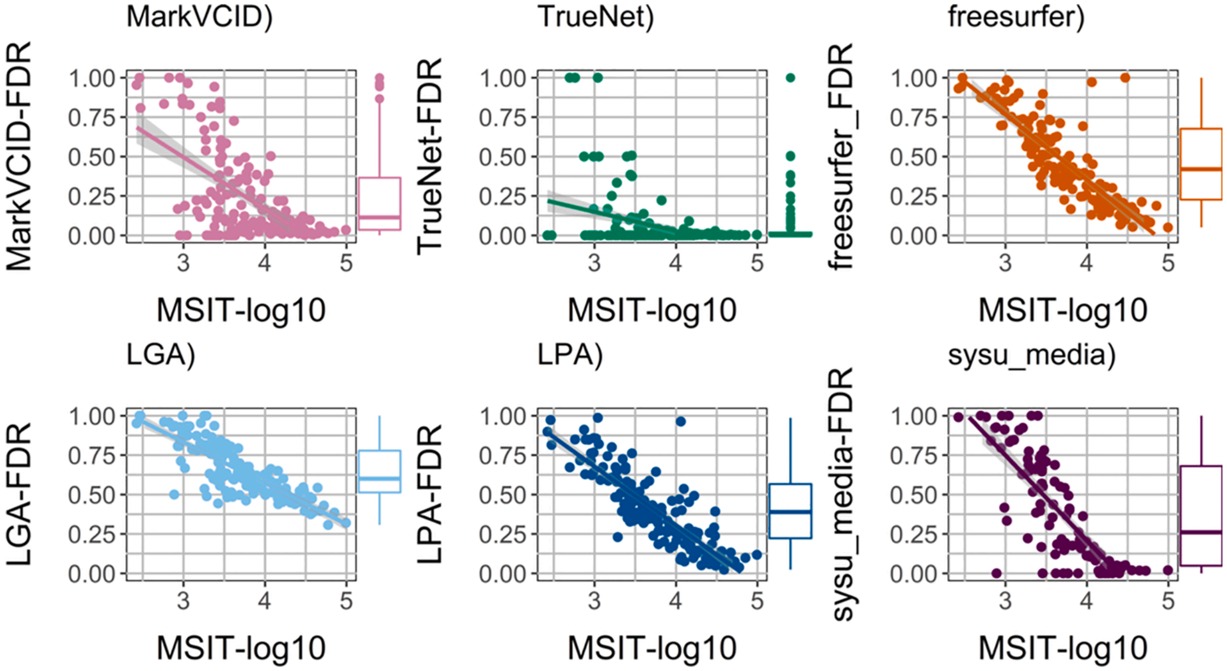

Fig. 4. Overlap measures for each WMH technique were computed as the number of voxels incorrectly labeled as lesion (false positive) compared to MSIT divided by the total number of WMH voxels. All WMH values were log10 transformed (mm^3). (Lines represent a linear function with shaded bars reflecting standard errors). TrUE-Net showed consistently lower false-positive voxels compared to all other techniques.

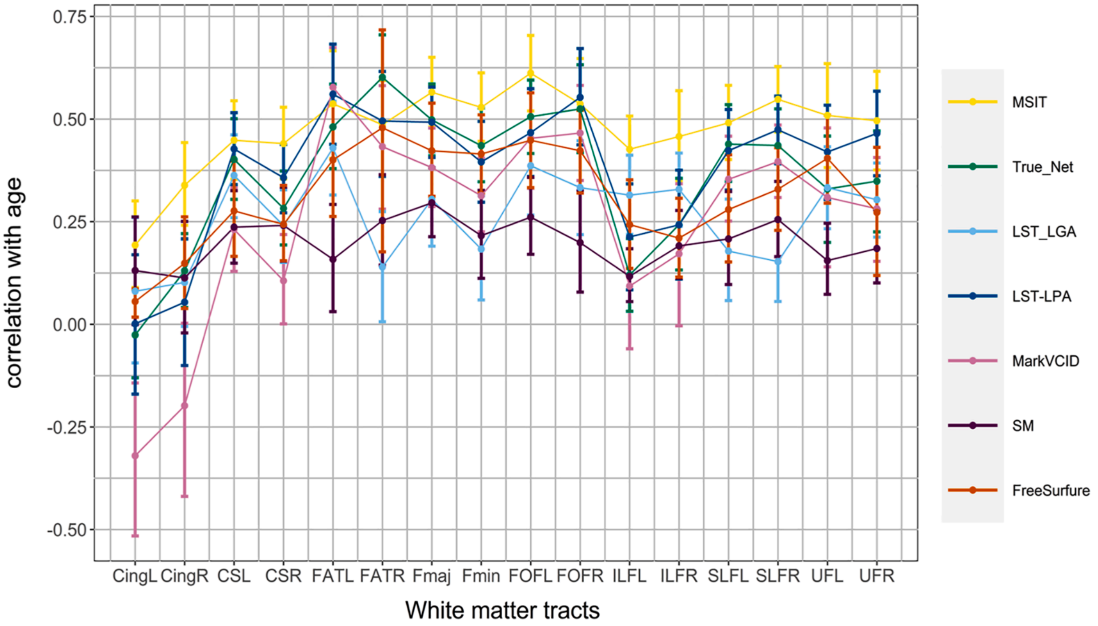

Fig. 5. Regional age correlation with WMH volumes and bootstrapped 95 % confidence intervals were computed within individual WM tracts. (Dots represent bootstrapped Pearson correlation values, and the bars show the 0.025, and 0.975 bootstrapped confidence intervals. Each color represents different segmentation techniques.).

performance in the future.

It is important to restate that we did not train TrUE-Net or SM on our data set, which would almost certainly improve their performance on the metrics assessed in this study. This was done deliberately to assess the generalizability of these techniques for any dataset despite cohort size. This innately provides more stringent criteria on the machine learning techniques over others that incorporate subject specific information for producing WMH segmentation maps. Nevertheless, TrUE-Net performed relatively well compared to the other techniques.

There are some limitations to our study. The AMBR dataset includes sequences not routinely performed and/or available in all clinical settings. The AMBR data set is comprised of a large data set with varying WMH severity levels but the number of severe WMH cases are limited and therefore our findings may not generalize to more severe degrees of WMH burden. We were unable to preprocess every technique with the exact same tools, which may have introduced additional bias. However, the decision to alter the preprocessing was to maximize the number of individuals for each technique that would pass our quality control checks for evaluation. In order to avoid suboptimal results from specific techniques due to preprocessing requirements, we adapted the order or inclusion of the brain extraction and bias field correction. Although manual segmentations are typically used as reference standards for estimating segmentation reliability this can still introduce bias. To address these concerns, we used age as an unbiased measure of a known biological correlate of WMH burden. Although age is not the only known biological source it is a consistent finding among prior work and based on our results we believe it was an appropriate unbiased reference metric for segmentation technique evaluation. We did not determine the effects of other demographic or biological factors aside from age, as these are being reserved for more detailed analyses in the future. Furthermore, the association with age could have been affected by other factors and further validation studies are required. Finally, this study only investigated cross-sectional data; future longitudinal studies are underway.

5. Conclusion

Compared to a semi-manual reference standard segmentation technique, TrUE-Net demonstrates high accuracy and reliability in identifying WMH and is also computationally efficient. Though currently conservative in defining WMH boundaries, TrUE-Net is well suited to large and potentially mixed datasets for estimating WMH burden at the global and regional level.

Code and data availability statement

The Aging Metabolism & Brain Resilience (AMBR) (PI: Manu Goyal, Andre Vlassenko) and the Knight Alzheimer Disease Research Center (ADRC) datasets were acquired at Washington University in St. Louis (PI: John Morris). These data may be made available to qualified investigators by a data use agreement and reasonable written request to the corresponding author (Manu Goyal) and Knight ADRC steering committee. Image processing used openly available tools for TrUE-Net (https://git.fmrib.ox.ac.uk/vaanathi/truenet). Further R scripts for algebraic and statistical analyses may be available upon request to Jeremy F. Strain.

CRediT authorship contribution statement

Jeremy F. Strain: Conceptualization, Formal analysis, Visualization, Writing – original draft. **Maryam Rahmani:** Data curation, Formal analysis, Conceptualization, Investigation, Visualization, Writing – original draft. **Donna Dierker:** Data curation, Validation, Writing – review & editing. **Christopher Owen:** Data curation, Validation, Writing – review & editing. **Hussain Jafri:** Conceptualization, Writing – review & editing. **Andrei G. Vlassenko:** Conceptualization, Project administration, Funding acquisition, Writing – review & editing. **Kyle Womack:** Conceptualization, Writing – review & editing. **Jurgen Fripp:** Writing – review & editing. **Duygu Tosun:** Writing – review & editing. **Tammie L.S. Benzinger:** Funding acquisition, Writing – review & editing. **Michael Weiner:** Writing – review & editing. **Colin Masters:** Writing – review & editing. **Jin-Moo Lee:** Writing – review & editing. **John C. Morris:** Conceptualization, Project administration, Writing – review & editing. **Manu S. Goyal:** Conceptualization, Project administration, Supervision, Validation, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgments

Funding for this research was provided by the Barnes-Jewish Hospital Foundation (JCM), the James S. McDonnell Foundation, the McDonnell Center for Systems Neuroscience, and the NIH/NIA (R01AG058676, R01AG053503, R01AG057536, P30AG066444, P01AG026276, P01AG003991 and RF1AG073210). Some of the MRI sequences used to produce the AMBR dataset were obtained from the Massachusetts General Hospital. The authors thank Vaanathi Sundaresan and Hongwei Li for their assistance with operating TrUE-Net and Sysu Media on our data. We are particularly grateful to our research participants and their families for their altruism. We also acknowledge the directors and staff of the Neuroimaging Labs Research Center, Center for Clinical Imaging Research, and Knight Alzheimer Disease Research Center for making this research possible.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120494.

References

Aydin, O.U., Taha, A.A., Hilbert, A., Khalil, A.A., Galinovic, I., Fiebach, J.B., Frey, D., Madai, V.I., 2021. On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking. Eur. Radiol. Exp. 5, 4.

Balakrishnan, R., Valdés Hernández, M.D.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. Comput. Med. Imaging Graph 88, 101867.

Beare, R., Srikanth, V., Chen, J., Phan, T.G., Stapleton, J., Lipshut, R., Reutens, D., 2009. Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities. Neuroimage 47, 199–203.

Biesbroek, J.M., Weaver, N.A., Biessels, G.J., 2017. Lesion location and cognitive impact of cerebral small vessel disease. Clin. Sci. 131, 715–728.

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, 261–276.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Ameli, R., Ferre, J.C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., Mahbod, A., Wang, C., McKinley, R., Wagner, F., Muschelli, J., Sweeney, E., Roura, E., Llado, X., Santos, M.M., Santos, W.P., Silva-Filho, A.G., Tomas-Fernandez, X., Urien, H., Bloch, I., Valverde, S., Cabezas, M., Vera-Olmos, F.J., Malpica, N., Guttmann, C., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S.K., Cotton, F., Barillot, C., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci. Rep. 8, 13650.

Debette, S., Markus, H.S., 2010. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. Bmj 341, c3666.

DeCarli, C., Maillard, P., Fletcher, E., 2013. Four Tissue Segmentation in ADNI II. Department of Neurology and Center for Neuroscience, University of California, Davis.

Fazekas, F., Chawluk, J.B., Alavi, A., Hurtig, H.I., Zimmerman, R.A., 1987. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. Am. J. Neuroradiol. 8, 421–426.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774–781.

Gouw, A.A., Seewann, A., van der Flier, W.M., Barkhof, F., Rozemuller, A.M., Scheltens, P., Geurts, J.J., 2011. Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations. J. Neurol. Neurosurg. Psychiatry 82, 126–135.

Goyal, M.S., Blazey, T., Metcalf, N.V., McAvoy, M.P., Strain, J.F., Rahmani, M., Durbin, T.J., Xiong, C., Benzinger, T.L.-S., Morris, J.C., Raichle, M.E., Vlassenko, A. G., 2023. Brain aerobic glycolysis and resilience in Alzheimer disease. Proc. Natl. Acad. Sci. 120, e2212256120.

Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G.J., Plummer, D.L., Tofts, P.S., McDonald, W.I., Miller, D.H., 1996. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn. Reson. Imaging 14, 495–505.

Haley, A.P, Hoth, K.F., Gunstad, J., Paul, R.H., Jefferson, A.L., Tate, D.F., Ono, M., Jerskey, B.A., Poppas, A., Sweet, L.H., 2009. Subjective cognitive complaints relate to white matter hyperintensities and future cognitive decline in patients with cardiovascular disease. Am. J. Geriatr. Psychiatry 17, 976–985.

Hubbard, N.A., Turner, M., Hutchison, J.L., Ouyang, A., Strain, J., Oasay, L., Sundaram, S., Davis, S., Remington, G., Brigante, R., 2016. Multiple sclerosis-related white matter microstructural change alters the BOLD hemodynamic response. J. Cereb. Blood Flow Metab. 36, 1872–1884.

Hubbard, N.A., Turner, M.P., Ouyang, M., Himes, L., Thomas, B.P., Hutchison, J.L., Faghihahmadabadi, S., Davis, S.L., Strain, J.F., Spence, J., 2017. Calibrated imaging reveals altered grey matter metabolism related to white matter microstructure and symptom severity in multiple sclerosis. Hum. Brain Mapp. 38, 5375–5390.

Inzitari, D., Pracucci, G., Poggesi, A., Carlucci, G., Barkhof, F., Chabriat, H., Erkinjuntti, T., Fazekas, F., Ferro, J., Xe, M.H., Langhorne, P., O'Brien, J., Scheltens, P., Visser, M.C., Wahlund, L.-O., Waldemar, G., Wallin, A., Pantoni, L., 2009. Changes in white matter as determinant of global functional decline in older independent outpatients: three year follow-up of LADIS (leukoaraiosis and disability) study cohort. BMJ 339, 279–282.

Jenkinson, M., Smith, S., 2001. Global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Llado, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S.H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE Trans. Med. Imaging 38, 2556–2568.

Lee, I.A., and K.J. Preacher. 2013. Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software].

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., Menze, B., 2018. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. Neuroimage 183, 650–665.

Maillard, P., Lu, H., Arfanakis, K., Gold, B.T., Bauer, C.E., Zachariou, V., Stables, L., Wang, D.J.J., Jann, K., Seshadri, S., 2022. Instrumental validation of free water, peak-width of skeletonized mean diffusivity, and white matter hyperintensities: MarkVCID neuroimaging kits. Alzheimer's Dementia 14, e12261.

Mäntylä, R., Erkinjuntti, T., Salonen, O., Aronen, H.J., Peltonen, T., Pohjasvaara, T., Standertskjöld-Nordenstam, C.-G., 1997. Variable agreement between visual rating scales for white matter hyperintensities on MRI: comparison of 13 rating scales in a poststroke cohort. Stroke 28, 1614–1623.

McAleese, K.E., Miah, M., Graham, S., Hadfield, G.M., Walker, L., Johnson, M., Colloby, S.J., Thomas, A.J., DeCarli, C., Koss, D., Attems, J., 2021. Frontal white matter lesions in Alzheimer's disease are associated with both small vessel disease and AD-associated cortical pathology. Acta Neuropathol. 142, 937–950.

Prins, N.D., Scheltens, P., 2015. White matter hyperintensities, cognitive impairment and dementia: an update. Nat. Rev. Neurol. 11, 157–165.

Scheltens, P., Barkhof, F., Leys, D., Pruvo, J.P., Nauta, J.J.P., Vermersch, P., Steinling, M., Valk, J., 1993. A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. J. Neurol. Sci. 114, 7–12.

Schmidt, P. 2017. 'Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging', lmu.

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, 3774–3783.

Strain, J., Didehbani, N., Cullum, C.M., Mansinghani, S., Conover, H., Kraut, M.A., Hart, J., Womack, K.B., 2013. Depressive symptoms and white matter dysfunction in retired NFL players with concussion history. Neurology 81, 25.

Strain, J.F., Didehbani, N., Spence, J., Conover, H., Bartz, E.K., Mansinghani, S., Jeroudi, M.K., Rao, N.K., Fields, L.M., Kraut, M.A., 2017. White matter changes and confrontation naming in retired aging National Football League athletes. J. Neurotrauma 34, 372–379.

Sundaresan, V., Zamboni, G., Rothwell, P.M., Jenkinson, M., Griffanti, L., 2021. Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. Med. Image Anal. 73, 102184.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15, 29.

Udupa, J.K., LeBlanc, V.R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L.M., Hirsch, B. E., Woodburn, J., 2006. A framework for evaluating image segmentation algorithms. Comput. Med. Imaging Graph. 30, 75–87.

Vanderbecq, Q., Xu, E., Stroer, S., Couvy-Duchesne, B., Diaz Melo, M., Dormont, D., Colliot, O., Initiative Alzheimer's Disease Neuroimaging, 2020. Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. Neuroimage Clin. 27, 102357.

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., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., Oostenbrugge, R., Pantoni, L., Speck, O., Stephan, B.C., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P.B., Dichgans, M., 2013. 'Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration'. Lancet Neurol. 12, 822–838.

Zamboni, G., Wilcock, G.K., Douaud, G., Drazich, E., McCulloch, E., Filippini, N., Tracey, I., Brooks, J.C.W., Smith, S.M., Jenkinson, M., Mackay, C.E., 2013. Resting functional connectivity reveals residual functional activity in Alzheimer's disease. Biol. Psychiatry 74, 375–383.

Zhang, L., Dean, D., Liu, J.Z., Sahgal, V., Wang, X., Yue, G.H., 2007. Quantifying degeneration of white matter in normal aging using fractal dimension. Neurobiol. Aging 28, 1543–1555.