Title: DeepWML, a deep learning MRI white matter hyperintensity detection applicable to multi-center data

Abstract

Purpose: White matter hyperintense (WMHI) lesions on MR images are an important indication of various types of brain diseases that involve inflammation and blood vessel abnormalities. Automated quantification of the WMHI could be valuable for clinical management of the patients, but existing automated software is often developed for a single type of disease and may not be applicable for clinical scans with thick slices and different scanning protocols.

Methods: We developed and evaluated "DeepWML", a U-net method for fully automated white matter lesion (WML) segmentation of multi-center FLAIR images. We used MRI from 507 patients, including in three distinct WM diseases, obtained in 9 centers, with a wide range of scanners and acquisition protocols. The automated delineation tool was evaluated through quantitative parameters of dice similarity, sensitivity and precision as compared to manual delineation (gold standard).

Results: The overall median dice similarity coefficient was 0.78 (range 0.64~0.86) across the three disease types and multiple centers. The median sensitivity and precision was 0.84 (range 0.67~0.94) and 0.81 (range 0.64~0.92), respectively. The tool's performance increased with larger lesion volumes.

Conclusion: DeepWML was successfully applied to a wide spectrum of MRI data in the three white matter disease types, which is potential to improve practical workflow of white matter lesion delineation.

Keywords: White matter hyperintensity, automated detection and segmentation, multiple sclerosis, multicenter, FLAIR

Introduction

MRI is widely used to detect the white matter hyperintensity (WMHI) lesions in neurological disorders such as multiple sclerosis (MS) [1, 2], neuromyelitis optica spectrum disorders (NMOSD) [3, 4], and cerebral small vessel disease (CSVD) [5, 6]. Accurate identification of WMHI has clinical relevance for diagnosis and predicting prognosis, especially in early disease phases [7, 8].

In the current clinical workflow, though it is not challenging to identify the WM lesions manually, the large amount of images as daily workload is likely to cause decreased efficiency for radiologists. In addition, the delineation of the WM lesions usually relies on manual drawing by experienced radiologists, which is time-consuming. Intra- and inter-rater reproducibility can be compromised by the subjective judgement, the associated workload and the different experience of raters [9, 10]. Various conventional machine learning and deep learning methods have been proposed to automatically detect and segment WM lesions. Dadar *et al.* evaluated the segmentation performance from ten conventional linear and nonlinear classification techniques (naïve Bayes, logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, bagging, and boosting) and observed the superior performance from random forest algorithm [11]. Recently, a few studies reported the white matter hyperintensities (WMHI) segmentation results using convolution neural network [12-16]. Rachmadi *et al.* proposed 2D-CNN scheme to segment WMHI with none or mild vascular pathology, and compared with other 15 types of machine learning segmentation methods [17]. 3D-Unet scheme has also been widely used for automated WM lesion segmentations, for

example, in MS patient studies with T2-FLAIR and MP2RAGE images [18]. As the deep learning technique engendered such applications increasingly, a recent article with CLAIM guidelines has been provided for such studies [19].

To be clinically useful, the segmentation technique needs to provide accurate results under heterogeneous imaging protocols for routinely available FLAIR images (e.g. 2D images with different slice thicknesses), which often calls for multi-center studies[20, 21] . Thus, there is a growing need to develop a fully automated tool that can deal with heterogeneous clinical data across various WM diseases. Unfortunately, previous studies have rarely demonstrated FLAIR based deep learning WMHI lesion segmentation validated on multiple disease types and multicenter data.

In this study, we present a deep learning (DL) based automated WM lesion segmentation tool using a single-modality clinical FLAIR image, that can be applied to data with WMHI lesions from multiple WM disease types (MS, NMOSD and CSVD) and from different clinical centers, with wide range of vendors, image resolutions and scanning slice thicknesses. The tool aims to assist the WM lesion detection and segmentation work for radiologists and clinicians.

Methods

Multi-center dataset

The MRI data from three disease types (MS, NMOSD and CSVD) that include WM lesions were collected from 9 centers (5 centers for MS and NMOSD and 4 centers for CSVD) (Table 1). MS diagnosis was determined according to 2017 McDonald criteria [22]. The NMOSD diagnosis was based on the 2015 International Panel on NMOSD Diagnosis [23] and all patients had antibodies against AQP4 using CBA method. CSVD diagnosis was based on the presence of white matter hyperintensity or more CSVD signs on MRI [24]. Patients who had other abnormalities such as brain tumor on MRI and well-defined macro-vascular stenosis on MRA were excluded.

The data of the current study is from prospective studies, and the data partition was at image level. The consent forms were signed by patients and the data were anonymized. Routine clinical FLAIR images were acquired in axial orientation on 3.0T or 1.5T scanners from multiple vendors (Philips Achieva, GE Discovery MR750/Signa HDxt/Optima MR360 and Siemens Skyra/TrioTim). Each patient's FLAIR contained 17-30 slices to cover the whole brain. After data acquisition, one patient with MS was excluded due to poor image quality. One patient with NMOSD was excluded due to the history of brain trauma. Nine patients with CSVD were excluded due to >50% intra-cranial macro-vascular stenosis on MRA, and one CSVD patient was excluded due to incidental mengioma Fig. S1). The final dataset consisting of 507 patients with 10753 image slices took part in this multi-center study, including 135 MS patients (82 women; mean age (SD) 37.2 (12.7) years), 74 NMOSD patients (62 women; 39.5 (13.5) years), and 298 CSVD patients (177 women; 43.0 (16.1) years). Original data had a wide range of in-plane resolution from 0.4102 × 0.4102 mm^2/pixel to 0.9375 × 0.9375 mm^2/pixel, with slice thickness ranging from 3 mm to 8 mm. Details of the data parameters and distribution are shown in Table 1. As pre-processing step, all FLAIR image slices

were intensity-normalized and resampled to a matrix of 256 x 256 before the network training.

Manual labeling of the WM lesions was performed using the software 3D Slicer [25]. Two experienced radiologists (Y.D with 12 years' experience and W.G with 10 years' experience) firstly completed a training session to reach a consensus on the evaluation of the imaging findings. After the training, they performed the manual labelling independently, and then Dice of the lesion masks from the two radiologists was calculated. The labels with relatively poor consistency (Dice < 0.85) need to be re-labelled to reach a good consensus (Dice $\geq$ 0.85) as ground truth for further analysis.

Deep learning neural network

We employed a 2-D Unet strategy [26] with network architecture shown in Figure 1. It consisted of a contracting path and an expanding path with skip connections between them. The operation block in the contracting path consisted of two blocks of 3x3 convolutions with a rectified linear unit (ReLU) (blue arrows in Fig.1). A 2x2 max pooling (red arrow in Fig.1) was performed at the end of each block to down-sample feature maps. In the expanding path, each operation block started with a 2x2 deconvolution for up-sampling (green arrow in Fig.1), followed by two blocks of 3x3 convolutions with ReLU. Skip connections (dashed line in Fig.1) carried the features from contracting path to the expanding path. The final layer was a fully convolutional layer with 1x1 kernel (orange arrow in Fig.1) which translated 64-channel

features to a single channel feature map. The output logits were compressed to range 0-1 to predict the final lesion activation map.

Network Training and Testing

The FLAIR images from all the centers were used for training and testing sessions. In the training and validation steps, 60% and 20% of the image dataset were randomly selected respectively, from each center and each disease type. This is to include data samples from all centers and all three disease types. This led to a total of 8640 image slices included in the training and validation procedure. The remaining 20% of the dataset was used as test set. The U-net was trained with a loss function of combined binary cross-entropy loss and Dice coefficient loss [27]. Training was performed for 200 epochs using Adam optimizer [28] with a starting learning rate of e^{-4}. Data augmentation was performed including random horizontal flip and rotation (-10 to 10 degrees). The model was implemented using the Python Pytorch 1.4 framework [29], with GPU NVIDIA GTX 1080 Ti*2 processor.

Performance evaluation

To quantitatively evaluate the performance of the networks, we calculated three evaluation metrics -- the Dice similarity coefficient (DSC), the true positive rate (i.e. Sensitivity),

and the Precision -- on each image in the test set. The median of the three metrics was calculated along each disease type and each clinical center, respectively.

Moreover, we partitioned the test set by different lesion volumes. For each 2D image, the lesion volume (LV) was calculated as the manually labeled lesion area multiplied by its slice thickness. LV groups were partitioned as: G1) LV < 0.2ml; G2) 0.2ml ≤ LV < 0.7ml; G3) 0.7ml ≤ LV < 2ml; G4) 2ml ≤ LV < 5ml; G5) LV ≥ 5ml. The three metrics within each LV group were compared.

Statistical analysis

To explore the performance of our tool in the multicenter multi-diseased dataset, statistical analysis was applied on the three evaluation metrics using python scipy package (https://www.scipy.org). We employed the Shapiro-Wilk test to check data normality, and the Kruskal-Wallis test to evaluate the performance difference across disease types and groups based on lesion volume ranges. The Mann-Whitney tests were further performed as pairwise comparisons for those tests with significance (p-value < 0.05), and Bonferroni correction was performed.

Results

Figure 2 shows representative cases with WM lesions from MS (Fig.2-left panel), NMOSD (Fig.2-middle panel) and CSVD (Fig.2-right panel), with manual labeling (overlay in red) and automated segmentations (overlay in green) for comparison. Four typical cases in each disease type were randomly selected. Our tool could automatically segment the WM lesions in manner consistent with the manual labels, regardless of the variation in lesion patterns, locations, disease types and imaging parameters. The processing time for automated segmentation on each slice was within 0.3s. For a patient data with 17-30 slices in our dataset, the total time for a whole brain lesion segmentation was about 5-9s.

Table 2 lists the quantitative results for each imaging center. The median DSC for the testing dataset had an average value of 0.78, with 0.80 for MS lesions, 0.77 for NMOSD lesions, and 0.78 for CSVD lesions. The median sensitivity and precision were 0.84 and 0.81, respectively. Figure 3 further details the distributions of the three performance metrics for each imaging center (Fig.3A-3C) and for each disease (Fig.3D-3F). The Kruskal-Wallis test showed that there was no significant difference in DSC (p-value = 0.09) (Fig.3D) or sensitivity (p-value = 0.54) (Fig.3E) among three disease types. However, the segmentation tool behaved differently among disease types in precision (p-value=0.009); in particular, the Mann–Whitney test demonstrated that the segmentation precision for MS lesions performed significantly better than for NMOSD and CSVD (p-value after Bonferroni correction << 0.05) (Fig.3F).

Figure 4 shows the segmentation performance for different WM lesion volumes including all three disease types of lesions. The Kruskal-Wallis test showed a significantly higher performance (DSC, sensitivity and precision) as the lesion volume increased (p-value < .001 for

DSC; p-value = 0.03 for Sensitivity; p-value < .001 for Precision). As indicated in Fig.4A (DSC), significant DSC improvements were found among the LV groups of G3 versus G1 ($p = 0.00025$), G4 versus G3 ($p = 2.4e^{-5}$), and G5 versus G4 ($p = 6.5e^{-6}$). For Sensitivity (Fig.4B), there was significant improvement of G5 versus G4 ($p = 0.0024$). For Precision, there was a significant improvement of G3 versus G1 ($p = 0.0002$) and G4 versus G3 ($p = 0.00011$).

Discussion

We present DeepWML, an automated WM lesion delineation tool using DL network that is largely agnostic to disease and scanner. The network was trained based on a large amount of data with a wide range of imaging conditions, disease types and lesion sizes, therefore, can be applied to a wide spectrum of MRI data for automated WM lesion segmentation in three main WM diseases (MS, NMOSD and CSVD).

In this study, we have employed three evaluation metrics (DSC, sensitivity and precision) to quantify the segmentation performance. The DSC depicts the overlap between the manual labeling and the automated result, and the median DSC reaches 0.78 for the overall multi-center dataset. This is comparable to the results reported from alternative approaches [15].

Moreover, a good median sensitivity (84%) ensures that most of the lesion pixels can be correctly identified; a good precision (81%) signifies that there are few false positive pixels. Importantly, the computation time per segmentation slice is within 0.3s (5-9s for each patient), which compares favorably with other DL algorithms. The few failure cases with discordant

segmentation result against the manual labels (shown in Fig. S2) were analyzed, and there are the following three causes of the detection failure. Firstly, some extremely small WM lesions which occupy less than 10 pixels (Fig.S2-Panel A) are difficult to be detected by the current DeepWML model, probably due to small portion of training data with extremely small lesion size. Secondly, Some WMHIs are contaminated with the normal tissue around the ventricle area (Fig.S2 – Panel B), because they share the high intensity feature. A third type of failure cases is due to the subtle intensity difference in WM lesion against their surrounding tissues (Fig.S2 – Panel C). For the performance with different lesion volumes, our study reaches median dice of 0.87 for lesion volume > 5ml, and 0.71 for lesion volume < 0.2ml which is good performance for such small lesions. Small lesions below the in-plane resolution are confusing for both the automatic tool and manual labeling. For the spatial accuracy expressed by Dice coefficient, our study showed a trend of increased median Dice along with the lesion volume increase (Fig.4-left). This result is highly consistent with the finding from previous work [30], in which “a clear trend for worse performance at lower volumes” was stated. A larger dataset is recommended to further confirm this finding in the subgroup analysis.Our algorithms showed several advantages to be applied in clinical practice. First, many previous studies used multiple MR image contrasts to enrich the input information in order to achieve better segmentation performance on MS patients data [16, 31, 32]. However, the multi-contrast MR data requires additional scan time and usually need co-registration to integrate information from multiple channels. The input of our tool only requires routine 2D FLAIR images and can be applied for images from multiple scanners with different vendors, with a wide range of in-plane resolution, slice thickness, and other imaging parameters. Second, the applicability of the trained U-net

tool is not limited to the WM lesion of one specific disease type. Our results showed no significant difference in the overlap ratio between automated and manual results (DSC) from MS, NMOSD and CSVD, implying that it is not required to have specified diagnostic information before commencing the WM lesion segmentation. Lastly, the tool is not dependent on skull-stripping of the images, which makes the pre-processing easier. The abovementioned features indicate great potential for our tool to be used in common clinical scenarios.

There are several limitations for our tool. First, the proposed network is based on 2D images, as well as the evaluation of the lesions on 2D slice level. It is possible that expansion to a 3D network may achieve similar or better performance. One reason that we stick to 2D network is that there is large variation of the slice thickness in our clinical routine 2D FLAIR images (3 mm ~ 8 mm). This leads to very limited number of slices for certain patients, and the 2D strategy may fit this situation. Second, the trained network was tested on the three common white matter disease types (MS, NMOSD and CSVD), while the performance of our tool in less common types of white matter diseases may need further validation. Lastly, the tool performance with small lesion volumes (e.g. less than 0.2 ml) needs to be improved.

Conclusion

In summary, we developed DeepWML, a DL-based automated WM lesion segmentation tool for WM lesion identification and delineation with satisfied performance in three main WM diseases. The robustness of the tool and the computation efficiency would allow integration into the clinical routine workflow.

Reference

1. Rovira, A. and A. León. (2008). *MR in the diagnosis and monitoring of multiple sclerosis: an overview.* European journal of radiology. **67**(3): p. 409-414.
2. Wattjes, M.P., À. Rovira, D. Miller, T.A. Yousry, M.P. Sormani, M.P. de Stefano, et al. (2015). *Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis--establishing disease prognosis and monitoring patients.* Nature reviews. Neurology. **11**(10): p. 597-606.
3. Asgari, N., H.P.B. Skejoe, S.T. Lillevang, T. Steenstrup, E. Stenager, and K.O. Kyvik. (2013). *Modifications of longitudinally extensive transverse myelitis and brainstem lesions in the course of neuromyelitis optica (NMO): a population-based, descriptive study.* BMC neurology. **13**: p. 33-33.
4. Dutra, B.G., A.J. da Rocha, R.H. Nunes, and A.C.M.J. Maia. (2018). *Neuromyelitis Optica Spectrum Disorders: Spectrum of MR Imaging Findings and Their Differential Diagnosis.* Radiographics : a review publication of the Radiological Society of North America, Inc. **38**(1): p. 169-193.
5. Zhang, X., Y. Tang, Y. Xie, C. Ding, J. Xiao, X. Jiang, et al. (2017). *Total magnetic resonance imaging burden of cerebral small-vessel disease is associated with post-stroke depression in patients with acute lacunar stroke.* European journal of neurology. **24**(2): p. 374-380.
6. Li, G., C. Zhu, J. Li, X. Wang, Q. Zhang, H. Zheng, et al. (2018). *Increased level of procalcitonin is associated with total MRI burden of cerebral small vessel disease in patients with ischemic stroke.* Neuroscience letters. **662**: p. 242-246.
7. McKinley, R., R. Wepfer, L. Grunder, F. Aschwanden, T. Fischer, C. Friedli, et al. (2019). *Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence.* NeuroImage. Clinical. **25**: p. 102104-102104.
8. Salem, M., S. Valverde, M. Cabezas, D. Pareto, A. Oliver, J. Salvi, et al. (2019). *A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis.* NeuroImage. Clinical. **25**: p. 102149-102149.
9. Zijdenbos, A.P., R. Forghani, and A.C. Evans. (2002). *Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis.* IEEE transactions on medical imaging. **21**(10): p. 1280-1291.
10. García-Lorenzo, D., S. Francis, S. Narayanan, D.L. Arnold, and D.L. Collins. (2013). *Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging.* Medical image analysis. **17**(1): p. 1-18.
11. Dadar, M., J. Maranzano, K. Misquitta, C.J. Anor, V.S. Fonov, M.C. Tartaglia, et al. *Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging.* (1095-9572 (Electronic)).
12. Diniz, P.H.B., T.L.A. Valente, J.O.B. Diniz, A.C. Silva, M. Gattass, N. Ventura, et al. (2018). *Detection of white matter lesion regions in MRI using SLIC0 and convolutional neural network.* Computer methods and programs in biomedicine. **167**: p. 49-63.
13. Li, H., G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, et al. (2018). *Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images.* NeuroImage. **183**: p. 650-665.
14. Xu, B., Y. Chai, C.M. Galarza, C.Q. Vu, B. Tamrazi, B. Gaonkar, et al. (2018). *ORCHESTRAL FULLY CONVOLUTIONAL NETWORKS FOR SMALL LESION SEGMENTATION IN BRAIN MRI.* Proceedings. IEEE International Symposium on Biomedical Imaging. **2018**: p. 889-892.

15. Duong, M.T., J.D. Rudie, J. Wang, L. Xie, S. Mohan, J.C. Gee, et al. (2019). *Convolutional Neural Network for Automated FLAIR Lesion Segmentation on Clinical Brain MR Imaging.* AJNR. American journal of neuroradiology. **40**(8): p. 1282-1290.
16. Schmidt, P., C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, et al. (2012). *An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis.* NeuroImage. **59**(4): p. 3774-3783.
17. Rachmadi, M.F., M.d.C. Valdés-Hernández, M.L.F. Agan, C. Di Perri, and T. Komura. (2018). *Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology.* Computerized Medical Imaging and Graphics. **66**: p. 28-43.
18. La Rosa, F., A. Abdulkadir, M.J. Fartaria, R. Rahmanzadeh, P.-J. Lu, R. Galbusera, et al. (2020). *Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE.* NeuroImage: Clinical. **27**: p. 102335.
19. Mongan, J.A.-O., L.A.-O. Moy, and C.E.J.A.-O. Kahn. *Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers.* (2638-6100 (Electronic)).
20. Heinen, R., M.D. Steenwijk, F. Barkhof, J.M. Biesbroek, W.M. van der Flier, H.J. Kuijf, et al. (2019). *Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset.* Scientific reports. **9**(1): p. 16742-16742.
21. Le, M., L.Y.W. Tang, E. Hernández-Torres, M. Jarrett, T. Brosch, L. Metz, et al. (2019). *FLAIR(2) improves LesionTOADS automatic segmentation of multiple sclerosis lesions in non-homogenized, multi-center, 2D clinical magnetic resonance images.* NeuroImage. Clinical. **23**: p. 101918-101918.
22. Thompson, A.J., B.L. Banwell, F. Barkhof, W.M. Carroll, T. Coetzee, G. Comi, et al. (2018). *Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria.* The Lancet Neurology. **17**(2): p. 162-173.
23. Wingerchuk, D.M., B. Banwell, J.L. Bennett, P. Cabre, W. Carroll, T. Chitnis, et al. (2015). *International consensus diagnostic criteria for neuromyelitis optica spectrum disorders.* Neurology. **85**(2): p. 177-89.
24. Wardlaw, J.M., E.E. Smith, G.J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, et al. (2013). *Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration.* The Lancet Neurology. **12**(8): p. 822-838.
25. Fedorov, A., R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, et al. (2012). *3D Slicer as an image computing platform for the Quantitative Imaging Network.* Magnetic resonance imaging. **30**(9): p. 1323-1341.
26. Ronneberger, O., P. Fischer, and T. Brox. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing.
27. Sudre, C.H., W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. (2017). *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing.
28. Kingma, D.P. and J. Ba. (2014). *Adam: A Method for Stochastic Optimization.* arXiv e-prints: p. arXiv:1412.6980.
29. Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* arXiv e-prints: p. arXiv:1912.01703.
30. Ribaldi, F., D. Altomare, J. Jovicich, C. Ferrari, A. Picco, F.B. Pizzini, et al. (2021). *Accuracy and reproducibility of automated white matter hyperintensities segmentation with lesion segmentation tool: A European multi-site 3T study.* Magnetic Resonance Imaging. **76**: p. 108-115.

31. Steenwijk, M.D., P.J.W. Pouwels, M. Daams, J.W. van Dalen, M.W.A. Caan, E. Richard, et al. (2013). *Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs).* NeuroImage. Clinical. **3**: p. 462-469.
32. Damangir, S., E. Westman, A. Simmons, H. Vrenken, L.-O. Wahlund, and G. Spulber. (2017). *Reproducible segmentation of white matter hyperintensities using a new statistical definition.* Magma (New York, N.Y.). **30**(3): p. 227-237.

Tables and figure legend:

Table 1 Data distribution from multiple clinical centers. 2D FLAIR images of MS, NMOSD and CSVD are collected on scanners from multiple vendors (Philips, GE, Siemens), with slice thickness ranging from 3 to 8 mm, in-plane resolution ranging from 0.4102 × 0.4102 mm^2/pixel to 0.9375 × 0.9375 mm^2/pixel.

Lesion type	Center code	# patients	# image slices	Slice thickness (mm)	In-plane resolution (mm*mm/pixel)	Scanner type
MS	Age (mean (SD)) 37.2 (12.7) ys	Gender - female ratio 82/135 (61%)				
	Ctr_MS#1	30	588	5.5, 6, 6.5	(0.4688*0.4688), (0.9375*0.9375)	GE Discovery MR750 3.0T
	Ctr_MS#2	37	592	8	(0.4688*0.4688)	GE Discovery MR750 3.0T
	Ctr_MS#3	27	541	6.5	(0.4492*0.4492),	Siemens

					(0.7188*0.7188)	Skyra 3.0T
	Ctr_MS#4	27	603	4, 5.2, 5.36, 5.5, 6.5	(0.4102*0.4102), (0.4297*0.4297), (0.4395*0.4395), (0.6875*0.6875)	Siemens Skyra 3.0T
	Ctr_MS#5	14	295	3, 4, 5, 6, 6.5, 7.5	(0.4297*0.4297), (0.4688*0.4688)	Siemens TrioTIm 3.0T
NMOSD	Age (mean (SD)) 39.5 (13.5) ys	Gender - female ratio 62/74 (84%)				
	Ctr_NMOSD#1	12	246	5, 6.5	(0.4688*0.4688), (0.5078*0.5078)	GE Discovery MR750 3.0T
	Ctr_NMOSD#2	12	192	8	(0.4688*0.4688)	GE Discovery MR750 3.0T
	Ctr_NMOSD#3	6	120	6.5	(0.7188*0.7188)	Siemens Skyra 3.0T
	Ctr_NMOSD#4	43	1373	4, 5, 6, 6.5, 7.5	(0.4297*0.4297), (0.4688*0.4688), (0.5*0.5), (0.8*0.8), (0.9375*0.9375)	Siemens TrioTIm 3.0T

CSVD	Age (mean (SD)) 43.0 (16.1) ys	Gender - female ratio 177/298 (59%)				
	Ctr_CSVD#1	52	1067	6, 6.5	(0.4688*0.4688), (0.8976*0.8976)	GE Discovery MR750 3.0T
	Ctr_CSVD#2	34	620	6.5, 7	(0.4492*0.4492), (0.4688*0.4688)	Philips Achieva 1.5T
	Ctr_CSVD#3	70	1562	6	(0.4492*0.4492)	Philips Achieva 1.5T
	Ctr_CSVD#4	142	2954	6.5, 7, 7.8, 8	(0.4688*0.4688)	GE Optima MR360 /Signa HDxt 1.5T

Table 2 Median DSC, Sensitivity and Precision of the WM lesions from three disease types of multi-center data. For DSC and Sensitivity, there is no significant difference across the three disease lesion types (P > 0.05).

	MS						NMOSD					CSVD					Overall
	MS Overall	Ctr_MS#1	Ctr_MS#2	Ctr_MS#3	Ctr_MS#4	Ctr_MS#5	**NMOSD Overall**	Ctr_NMOSD#1	Ctr_NMOSD#2	Ctr_NMOSD#3	Ctr_NMOSD#4	**CSVD Overall**	Ctr_CSVD#1	Ctr_CSVD#2	Ctr_CSVD#3	Ctr_CSVD#4	**Overall**
Dice	**0.80**	0.79	0.79	0.75	0.88	0.86	**0.77**	0.80	0.80	0.83	0.75	**0.78**	0.78	0.79	0.74	0.78	**0.78**
Sensitivity	**0.83**	0.73	0.82	0.80	0.87	0.92	**0.84**	0.79	0.78	0.86	0.86	**0.85**	0.84	0.86	0.81	0.86	**0.84**
Precision	**0.86**	0.89	0.85	0.75	0.89	0.83	**0.79**	0.89	0.88	0.77	0.75	**0.79**	0.77	0.84	0.79	0.80	**0.81**

Fig.1 The fully convoluted network architecture with U-net strategy. Each gray box represents a multi-channel feature map. Numbers on top of the boxes denote number of channels; width and height dimensions are denoted at bottom. Light gray boxes represent the copied features through skip connections. The operations are denoted using different types of arrows

Fig.2 Representative cases of WM lesion automated segmentation result (in green) vs. manual labeling (in red). Four typical cases in each disease type are randomly selected and covered both lower and upper position of the brain structures. The U-net based automated segmentation accords with the manual labels for the various lesion patterns and lesion sizes. The DSC for these cases are above 0.8. The segmentation time for each case is within 0.3 s

Fig.3 Distribution of the DSC (top row), Sensitivity (mid row) and Precision (bottom row) across three disease types on multi-center data. Each column corresponds to center-wise result for one type of disease. The right-most column shows result for three diseases. In Fig.3D, 3E, no significant difference in segmentation performance found in DSC or Sensitivity ($p > 0.05$); In Fig.3F, significantly better Precision in MS lesion type is found (*$p < 0.05$ after Bonferroni correction)

Fig.4 DSC, Sensitivity and Precision with regards to lesion volume (LV) partitions. The testing data were partitioned according to the manually labeled WM lesion volumes with 5 groups: G1) LV < 0.2ml; G2) 0.2ml ≤ LV < 0.7ml; G3) 0.7ml ≤ LV < 2ml; G4) 2ml ≤ LV < 5ml; G5) LV ≥ 5ml. The segmentation performance improves along with the increasing WM lesion volume groups, with detailed statistical results marked in the figures. *$p < 0.05$ after Bonferroni correction