Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients
Quentin Vanderbecq, Eric Xu, Sebastian Stroër, Baptiste Couvy-Duchesne, Mauricio Diaz Melo, Didier Dormont, Olivier Colliot

To cite this version:
Quentin Vanderbecq, Eric Xu, Sebastian Stroër, Baptiste Couvy-Duchesne, Mauricio Diaz Melo, et al.. Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. Neuroimage-Clinical, Elsevier, 2020, 27, pp.102357. 10.1016/j.nicl.2020.102357. hal-02926474

HAL Id: hal-02926474
https://hal.archives-ouvertes.fr/hal-02926474
Submitted on 1 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients

Quentin Vanderbecq\textsuperscript{a,b,c,d,e,*}, Eric Xu\textsuperscript{f}, Sebastian Ströer\textsuperscript{g}, Baptiste Couvy-Duchesne\textsuperscript{a,b,c,d,e,g}, Mauricio Diaz Melo\textsuperscript{h,e}, Didier Dormont\textsuperscript{a,b,c,e,h}, Olivier Colliot\textsuperscript{a,b,c,d,e,i}, for the Alzheimer’s Disease Neuroimaging Initiative

\textsuperscript{a} Institut du Cerveau et de la Moelle épinière, ICM, F-75013 Paris, France
\textsuperscript{b} Inserm, U 1127, F-75013 Paris, France
\textsuperscript{c} CNRS, UMR 7225, F-75013 Paris, France
\textsuperscript{d} Sorbonne Université, F-75013 Paris, France
\textsuperscript{e} Inria Paris, Aramis Project-Team, F-75013 Paris, France
\textsuperscript{f} Department of Radiology, University Hospital La Cavale Blanche, F-29200 Brest, France
\textsuperscript{g} Institute for Molecular Bioscience, the University of Queensland, 4072 Brisbane, Australia
\textsuperscript{h} AP-HP, Hôpital de la Pitié-Salpêtrière, Department of Neuroradiology, F-75013 Paris, France
\textsuperscript{i} AP-HP, Hôpital de la Pitié-Salpêtrière, Department of Neurology, Institut de la Mémoire et de la Maladie d’Alzheimer (IM2A), F-75013 Paris, France

ARTICLE INFO

Keywords:
White matter hyperintensity
Dementia
Artificial intelligence
Segmentation
Microvascular

ABSTRACT

Background: Manual segmentation is currently the gold standard to assess white matter hyperintensities (WMH), but it is time consuming and subject to intra and inter-operator variability.

Purpose: To compare automatic methods to segment white matter hyperintensities (WMH) in the elderly in order to assist radiologist and researchers in selecting the most relevant method for application on clinical or research data.

Material and Methods: We studied a research dataset composed of 147 patients, including 97 patients from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) 2 database and 50 patients from ADNI 3 and a clinical routine dataset comprising 60 patients referred for cognitive impairment at the Pitié-Salpêtrière hospital (imaged using four different MRI machines). We used manual segmentation as the gold standard reference. Both manual and automatic segmentations were performed using FLAIR MRI. We compared seven freely available methods that produce segmentation mask and are usable by a radiologist without a strong knowledge of computer programming: LGA (Schmidt et al., 2012), LPA (Schmidt, 2017), BIANCA (Griffanti et al., 2016), UBO detector (Jiang et al., 2018), W2MHS (Ithapu et al., 2014), nicMSlesion (with and without retraining) (Valverde et al., 2017, 2019). The primary outcome for assessing segmentation accuracy was the Dice similarity coefficient (DSC) between the manual and the automatic segmentation software. Secondary outcomes included five other metrics.

Results: A deep learning approach, NicMSlesion, retrained on data from the research dataset ADNI, performed best on this research dataset (DSC: 0.595) and its DSC was significantly higher than that of all others. However, it ranked fifth on the clinical routine dataset and its performance severely dropped on data with artifacts. On the clinical routine dataset, the three top-ranked methods were LPA, SLS and BIANCA. Their performance did not differ significantly but was significantly higher than that of other methods.

Conclusion: This work provides an objective comparison of methods for WMH segmentation. Results can be used by radiologists to select a tool.

Abbreviations: WMH, White Matter Hyperintensities; WM, White Matter; ADNI, Alzheimer’s Disease Neuroimaging Initiative; DSC, Dice similarity coefficient

*Corresponding author: ICM – Brain and Spinal Cord Institute ARAMIS Team, Pitié-Salpêtrière Hospital 47-83, boulevard de l’Hôpital, 75651 Paris Cedex 13, France.

E-mail address: q.vanderbecq@gmail.com (Q. Vanderbecq).

https://doi.org/10.1016/j.nicl.2020.102357
Received 21 May 2020; Received in revised form 16 July 2020; Accepted 20 July 2020
Available online 22 July 2020
2213-1582/ © 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).
1. Introduction

White matter hyperintensities (WMH) are signal abnormalities of white matter (WM) on T2-weighted (T2w) magnetic resonance imaging (MRI) sequences. They are commonly seen in the brain of elderly people. In such populations, the majority of these abnormalities are presumed to be of vascular origin. The STandards for Reporting Vascular changes on nEuroimaging (STRIVE) have provided recommendations to standardize their interpretations (Wardlaw et al., 2013). In clinical practice, visual rating scales are used to evaluate WMH linked to microvascular pathology, the most common being the Fazekas scale (Fazekas et al., 1987). However, it does not give a precise information about the spatial localization and volume of WMH. Manual assessment of cognitive impairment on each of the four MRI machines (GE Signa HDxt 3 T, Siemens Skyra 3 T, GE Optima MR450w 1.5 T, GE Signa PET/MR 3 T) and parameters were heterogeneous (see Supplementary Tables 3 and 4), thereby reflecting the reality of clinical routine.

We assessed visually the presence of artifacts on T1 and FLAIR images. Specifically, a participant was assigned to the “artifact group” if either the T1 or the FLAIR image had artifacts that could limit the interpretation. This assessment was made independently by two radiologists (QV, EX) and both readers agreed on all cases. Ten participants were assigned to the “artifact group”. Three participants had artifacts on the T1w image, two had artifacts on the FLAIR image while for the remaining five participants, both images were artifacted. One example of an artifacted image is shown in Supplementary Fig. 1 and the characteristics of those participants are reported in Supplementary Table 2.

2. Material and methods

2.1. Participants

We used two different datasets: a research dataset from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) (Mueller et al., 2005) and a clinical routine dataset. For the research dataset, we randomly selected 97 participants from ADNI 2 and 50 participants from ADNI 3. We extracted a training subset by randomly selecting 20 participants from ADNI 2 and 20 from ADNI 3. We randomly selected 97 participants from ADNI 2

patients with stroke, tumor, hematoma, inflammatory and infectious pathology. For the clinical routine dataset, all clinical and biological data were generated during a routine clinical workup and were retrospectively extracted for the purpose of this study. Therefore, according to French legislation, explicit consent was waived.

The main characteristics of these three populations are summarized in Table 1 and detailed information is provided in Supplementary Tables 1 and 2.

2.2. MRI acquisition

In the research dataset (ADNI), all patients had a 3D T1-weighted (T1w) and a FLAIR sequence acquired at 3 T. FLAIR sequences were 2D in ADNI2 and 3D in ADNI 3. Acquisition protocols have been previously described (Jack et al., 2015, 2008) (http://adni.loni.usc.edu/methods/documents/mri-protocols/).

In the clinical routine dataset, all patients had a 3D T1-weighted sequence and a 3D FLAIR sequence (except for two patients who had a 2D FLAIR). The acquisitions were performed on different MRI machines (GE Signa HDxt 3 T, Siemens Skyra 3 T, GE Optima MR450w 1.5 T, GE Signa PET/MR 3 T) and parameters were heterogeneous (see Supplementary Tables 3 and 4), thereby reflecting the reality of clinical routine.

We selected segmentation methods by reviewing the literature from 2012 to November 2018, from which we identified 33 different methods. A summary of our literature review can be found in Supplementary Table 5. For inclusion in the comparison, methods needed to be freely available and to produce the segmentation mask as output (and not only the volume), which reduced the list to fourteen methods. Moreover, we included only user-friendly methods, which had to be usable by a radiologist without a strong knowledge of computer programming. Thus, we removed methods for which the user needed to perform specific image preprocessing. This left seven methods that we considered in our study.

We included: 1) the lesion growth algorithm (LGA) (Schmidt et al., 2012) from the lesion segmentation toolbox (LST) (www.statisticalmodelling.de/lst.html), included in SPM12 and based on probabilistic modeling and a region growing algorithm; 2) the lesion prediction

Table 1

Demographic information for the research dataset (from ADNI) and the clinical routine dataset. Continuous values are displayed as average with the min–max range within parentheses. For ADNI, we also display the characteristics of the training and testing datasets separately.

|                | ADNI   | ROUTINE |
|----------------|--------|---------|
|                | All    | Training| Testing| All   |
| N              | 147    | 40      | 107     | 60    |
| Age            | 74     | 74.7    | 73.7    | 78.2  |
| (range)        | (58–90)| (58–90) | (59–90) | (52–101) |
| Sex            | M/F    | F/M     | F/M     | F/M   |
| Sex            | 85     | 9      | 66      | 30    |
| (range)        | (58–90)| (58–90)| (59–90) | (52–101) |
| Age            | 74     | 74.7    | 73.7    | 78.2  |
| (range)        | (58–90)| (58–90)| (59–90) | (52–101) |
| Sex            | F/M    | F/M     | F/M     | F/M   |
| Sex            | 85     | 9      | 66      | 30    |
| (range)        | (58–90)| (58–90)| (59–90) | (52–101) |

The main characteristics of these three populations are summarized in Table 1 and detailed information is provided in Supplementary Tables 1 and 2.
algorithm (LPA) (Schmidt, 2017) (Schmidt, 2017, Chapter 6.1), also from the SPM LST toolbox and based on logistic regression; 3) the Brain Intensity AbNormality Classification Algorithm (BIANCA) (Griffanti et al., 2016) included in FSL, based on the K nearest neighbors (K-nn) algorithm; 4) the UBO detector (Jiang et al., 2018), also based on K-nn; 5) the Wisconsin White Matter Hyperintensities Segmentation Toolbox (W2MHS) (Itthapu et al., 2014) a method based on the random forest algorithm; 6) the multiple Sclerosis Lesion Segmentation toolbox (Roura et al., 2015) (SLS) based on thresholding of a WM segmentation map; 7) the nicMSlesion toolbox based on a cascade of two 3D patchwise convolutional neural networks (Valverde et al., 2019, 2017). For nicMSlesion, we used two different models: the original model (trained by the authors of the original publication on a multiple sclerosis dataset and directly available), a retrained model for which we retrained the last three fully connected layers using our training dataset. To note, BIANCA specifically requires a training set, to create a set of feature vectors for lesion and non-lesion classes, for which we also used the ADNI training subset. Computation times for the different methods are reported in Supplementary Text 2.

2.4. Determination of hyper-parameters

Some methods have hyper-parameters that can be adjusted. We determined the optimal value of the parameters which maximized the DSC on the ADNI training subset of 40 patients. We performed a “grid-search”, i.e. testing several possible combinations of parameters using fixed intervals within the range of possible hyperparameters. We determined the optimal parameters for all methods, except for SLS for which there is no adjustable parameter. For all other methods, which return a continuous prediction, we estimated the optimal probability which there is no adjustable parameter. For all other methods, which determined the optimal parameters for all methods, except for SLS for fixed intervals within the range of possible hyperparameters. We determined the optimal parameters for all methods, except for SLS for which there is no adjustable parameter. For all other methods, which return a continuous prediction, we estimated the optimal probability threshold used to define WMH. Other hyper parameters included the number of K-nn neighbours (UBO), a threshold on the result of the registration of the segmentation mask to FLAIR space (nicMSlesion, LGA), a threshold of the WM mask registered to FLAIR space (BIANCA), a “cleaning threshold” which removes hyperintensities that are closer than a given distance from the grey matter (W2MHS). More details regarding the parameters and their optimization are provided in Supplementary Text 3.

2.5. Manual segmentation

The reference standard was built using manual segmentation of WMH by a radiology resident trained rater (QV). WMH masks were manually segmented from the FLAIR sequence using the ITK SNAP editing tool (Yushkevich et al., 2006). A segmentation protocol was designed from the advice of two experienced neuroradiologists (SS and DD). It included the following rules: 1) exclusion of hyperintense lines adjacent to the ventricles that are one voxel thick; 2) exclusion of WMH in the septum pellucidum, at the junction of the genu of the corpus callosum and the septum pellucidum and at the junction of the splenium of the corpus callosum and the ventricles.

We evaluated inter and intra-rater reproducibility on the ADNI training subset of 40 patients. For inter-rater agreement, images were segmented by two raters (QV and EX). For intra-rater, QV segmented the images twice, with a minimum interval of 4 weeks between the two evaluations.

2.6. Statistical analysis

The primary outcome for assessing segmentation accuracy was the Dice similarity coefficient (DSC) calculated as:

\[
\text{DSC} = \frac{2 \times |A \cap B|}{|A| + |B|}
\]

where \(A\) and \(B\) are the manual and automatic segmentations, respectively. We used paired t-tests to compare DSC between methods, or between clinical images with and without artifact. In post-hoc analyses, we further adjusted for WMH volume and site/scanner, we also stratified the analysis by high/low WMH volumes (cut off at 10,000 mm\(^3\), which correspond to Fazekas score < 3 (Hernández et al., 2013) commonly used in clinical practice). On the clinical routine dataset, we estimated the proportion of DICE variance attributable to scanner variability (partial eta-square effect sizes) and tested its significance using ANOVAs. The significance level was corrected for multiple comparisons using Bonferroni correction.

Secondary outcomes included the following metrics:

- **Volume Similarity** (Taha and Hanbury, 2015):
  \[
  \text{Volume Similarity} = 1 - \frac{|\text{Manual Volume} - \text{Automatic Volume}|}{\text{Manual Volume} + \text{Automatic Volume}}
  \]

- **Absolute volume error rate**:
  \[
  \text{Absolute volume error rate} = \frac{\text{number of False positive voxels}}{\text{number of automatic WMH voxels}} + \frac{\text{number of False negative voxels}}{\text{number of manual WMH voxels}}
  \]

- **Voxel-level false positive ratio** :
  \[
  \text{Voxel-level false positive ratio} = \frac{\text{number of False positive voxels}}{\text{number of automatic WMH voxels}}
  \]

- **Voxel-level false negative ratio** :
  \[
  \text{Voxel-level false negative ratio} = \frac{\text{number of False negative voxels}}{\text{number of manual WMH voxels}}
  \]

- **Intra-class correlation coefficient between volumes using a two-way model with absolute agreement definition and single rater** (Koo and Li, 2016)(Shrout and Fleiss, 1979)

3. Results

3.1. Reproducibility of manual segmentation

Table 2 displays the intra and inter-observer agreement on 40 patients from the training ADNI dataset. In all cases, the DSC indicated a substantial, nonetheless imperfect agreement of WMH maps, Though the intraclass correlation between volumes indicated an excellent agreement of WMH volumes (Koo and Li, 2016).

3.2. WMH volume distribution

The average WMH volume was 9.6 ml (SD 14.3, median 3.8) on the ADNI training dataset and 8.0 ml (SD 11.9, median 4.2) on ADNI testing dataset. WMH distributions were highly skewed to the right and with many outliers (Supplementary Fig. 2) which led us test to use the Mann-Whitney-Wilcoxon test for difference between groups. WMH volumes were not significantly different between the ADNI training and testing dataset (p = 0.5, Mann-Whitney-Wilcoxon test). The average WMH volume was 16.2 ml (SD 24.8, median 8.1) on the clinical routine dataset. WMH volumes were 16.2 ml (SD 24.8, median 8.1) on the clinical routine dataset. WMH distributions were highly skewed to the right and with many outliers (Supplementary Fig. 2) which led us test to use the Mann-Whitney-Wilcoxon test to test for differences between groups. WMH volumes were not significantly different between the ADNI training and testing dataset (p = 0.001, Mann-Whitney-Wilcoxon test).

Table 2

| Intraclass correlation | Volume similar | Volume error rate | False positive rate | False negative rate |
|------------------------|----------------|------------------|---------------------|---------------------|
| Intra-operator reproducibility | 0.744 | 0.899 | 0.987 | 0.185 | 0.196 | 0.292 |
| First segmentation first operator vs second operator | 0.723 | 0.884 | 0.984 | 0.277 | 0.324 | 0.199 |
| Second segmentation first operator vs second operator | 0.701 | 0.844 | 0.974 | 0.310 | 0.262 | 0.290 |

DSC: Dice similarity coefficient. For each metric, the table displays the average and the 95% confidence interval within parentheses.
3.3. Determination of hyper-parameters

Initially, we aimed to determine optimal parameters separately for 2D (from ADNI2) and 3D data (from ADNI3). However, we found that the optimal parameter were very similar for 2D and 3D data (Supplementary Fig. 3) and yielded comparable DSC. Thus, we used a single optimal value based on the merged dataset of 2D and 3D training dataset and did not separate 2D and 3D scans in the rest of the analysis. We report the distribution of absolute volume error rate and intraclass correlation according to the different parameter values on Supplementary Fig. 4 (LPA), 5 (DSC), 6 (Volume error rate) and 7 (Intraclass correlation).

We report the best DSC results and the optimal parameters for each method in Supplementary Table 6. We used these parameters for our evaluations in both research and clinical routine datasets.

3.4. Performance on the research dataset (ADNI)

The performance of retrained nicMSlesion was significantly better than that of algorithm LPA that came second (DSC of 0.595 vs. 0.533 Fig. 1, Table 3; p < 0.001 Supplementary Table 7). nicMSlesion also achieved the best performance according to secondary outcomes except for the false negative ratio, where it was superseded by LPA, UBO and SLS (Supplementary Figure 8, Table 3). The DSC difference between LPA and SLS (that came third) did not reach significance, however they performed better than all other algorithms (p < 0.001). Adjusting for site and WMH volume load did change the ranking, though the superiority of retrained nicMSlesion over LPA (1st vs. 2nd) and SLS (3rd) over LGA, nicMSlesion (original) and UBO could not be deemed significant anymore (Supplementary Table 7).

To complement our analysis controlling for WMH volume load, we studied the performance separately for patients with low (< 10,000 mm³, which correspond to Fazekas score < 3 (Hernández et al., 2013)) and high WMH volume load. Again, the retrained nicMSlesion performed best for both low and high volume load groups followed by SLS and LPA. (Supplementary Table 8).

In order to appreciate the spatial distribution of errors, we constructed maps of false positive and false negative rate for each algorithm (Fig. 2). A comparison of manual and automatic segmentation for a single individual is shown in Supplementary figure 9. Overall, we note that the errors remain localized around the true WMH location. We observe that retrained nicMSlesion had a low level of false negative and that all errors remained around true WMH location. We note that retraining nicMSlesion reduced massively the frequency of errors and avoided large errors in unusual locations (e.g. cerebellum). LPA and SLS good DSC performance came from a relatively low false negative rate (Fig. 2, Table 3). However, LPA and SLS resulted in false positives located mainly in posterior regions.

Table 3

| Method            | DSC       | Volume similarity | Volume error rate | Intraclass correlation | False positive rate | False negative rate |
|-------------------|-----------|-------------------|-------------------|------------------------|---------------------|---------------------|
| LPA               | 0.539     | 0.734             | 0.850             | 0.812                  | 0.438               | 0.366               |
|                   | (0.505–0.573) | (0.691–0.775)   | (0.570–1.131)     | (0.709–0.876)          | (0.399–0.477)       | (0.321–0.410)       |
| LGA               | 0.474     | 0.759             | 0.426             | 0.680                  | 0.444               | 0.535               |
|                   | (0.441–0.509) | (0.719–0.798)   | (0.361–0.490)     | (0.561–0.770)          | (0.406–0.480)       | (0.494–0.574)       |
| BIANCA            | 0.469     | 0.638             | 0.760             | 0.417                  | 0.393               | 0.481               |
|                   | (0.430–0.506) | (0.588–0.686)   | (0.609–0.912)     | (0.249–0.560)          | (0.349–0.436)       | (0.428–0.533)       |
| SLS               | 0.527     | 0.732             | 0.903             | 0.890                  | 0.564               | 0.277               |
|                   | (0.495–0.559) | (0.696–0.766)   | (0.729–1.078)     | (0.507–0.957)          | (0.531–0.596)       | (0.239–0.314)       |
| W2MHS             | 0.351     | 0.663             | 2.219             | 0.292                  | 0.539               | 0.569               |
|                   | (0.318–0.385) | (0.551–0.654)   | (1.139–3.299)     | (0.108–0.456)          | (0.482–0.594)       | (0.529–0.608)       |
| nicMSlesion(original) | 0.454     | 0.787             | 0.694             | 0.948                  | 0.517               | 0.503               |
|                   | (0.419–0.490) | (0.746–0.826)   | (0.382–1.007)     | (0.924–0.964)          | (0.476–0.557)       | (0.463–0.543)       |
| nicMSlesion(retrained) | 0.595     | 0.889             | 0.270             | 0.979                  | 0.384               | 0.402               |
|                   | (0.537–0.921) | (0.867–0.910)   | (0.199–0.381)     | (0.968–0.986)          | (0.351–0.416)       | (0.376–0.427)       |
| UBO               | 0.486     | 0.762             | 0.907             | 0.881                  | 0.587               | 0.360               |
|                   | (0.459–0.514) | (0.730–0.793)   | (0.575–1.239)     | (0.652–0.945)          | (0.559–0.615)       | (0.328–0.392)       |

For each metric, we present the average and the 95% confidence interval within parentheses. DSC Dice similarity coefficient. Results in bold indicates the best score for each metric.
Fig. 2. Maps of False negative and False positive rate from each method on the ADNI research dataset. We represent masks of segmentation on MNI template. The first row of the plot represents an overlay of manual segmentation in the ADNI testing set. The greyscale ranges from 0% (white) to 33% (black) of WMH at any particular voxel. The left column of the plot represents the false negative rate map for each method in ADNI testing set. The right column shows the false positive rate map for each method on ADNI dataset. Scale ranges from 0 to 33% of errors at each voxel, which corresponds to the maximal error rates observed.
not possible to test whether this superior performance was statistically significant. On images with artifacts, SLS performance of BIANCA and nicMSlesion significantly dropped on artifacted images and in particular on data with strong artifacts (Table 3, Figs. 1 and 2). It was the opposite for SLS (limited false positives, Fig. 3), which contrasted with results of the research dataset. Valverde and colleagues (Valverde et al., 2019), already demonstrated this for NicMSlesion, on two different multiple sclerosis datasets, that one obtains lower performances when using a model trained on a dataset that is too different from the test set.

On the clinical routine dataset, LPA, SLS and BIANCA exhibited the highest DSC and their performances were not significantly different (respectively 0.65, 0.61, 0.61). To note, LPA and SLS ranked second in term of DSC performance on the ADNI sample (0.54, 0.53), which may be the most accurate when trained on data similar to the data used for testing but they do not generalize well. Valverde and colleagues (Valverde et al., 2019), already demonstrated this for NicMSlesion, on two different multiple sclerosis datasets, that one obtains lower performances when using a model trained on a dataset that is too different from the test set.

We compared seven tools for automatic WMH segmentation to determine which is the most efficient. All tools are freely available and usable by a radiologist without advanced knowledge in computer programming. Our evaluation used both a research dataset (ADNI) and a routine practice dataset of patients with cognitive impairment.

On the research dataset, nicMSlesion, a cascade of convolutional networks (with a specific re-training on a subset of the sample) achieved the highest performance on the primary criterion (DSC). However, its performance did not generalize well on clinical routine images and in particular on data with strong artifacts (Table 3, Figs. 1 and 3a). One important lesson from our study is that complex models (such as neural network) may be the most accurate when trained on data similar to the data used for testing but they do not generalize well. Valverde and colleagues (Valverde et al., 2019), already demonstrated this for NicMSlesion, on two different multiple sclerosis datasets, that one obtains lower performances when using a model trained on a dataset that is too different from the test set.

4. Discussion

We compared seven tools for automatic WMH segmentation to determine which is the most efficient. All tools are freely available and usable by a radiologist without advanced knowledge in computer programming. Our evaluation used both a research dataset (ADNI) and a routine practice dataset of patients with cognitive impairment.

On the research dataset, nicMSlesion, a cascade of convolutional networks (with a specific re-training on a subset of the sample) achieved the highest performance on the primary criterion (DSC). However, its performance did not generalize well on clinical routine images and in particular on data with strong artifacts (Table 3, Figs. 1 and 3a). One important lesson from our study is that complex models (such as neural network) may be the most accurate when trained on data similar to the data used for testing but they do not generalize well. Valverde and colleagues (Valverde et al., 2019), already demonstrated this for NicMSlesion, on two different multiple sclerosis datasets, that one obtains lower performances when using a model trained on a dataset that is too different from the test set.

On the clinical routine dataset, LPA, SLS and BIANCA exhibited the highest DSC and their performances were not significantly different (respectively 0.65, 0.61, 0.61). To note, LPA and SLS ranked second in term of DSC performance on the ADNI sample (0.54, 0.53), which may be the most accurate when trained on data similar to the data used for testing but they do not generalize well. Valverde and colleagues (Valverde et al., 2019), already demonstrated this for NicMSlesion, on two different multiple sclerosis datasets, that one obtains lower performances when using a model trained on a dataset that is too different from the test set.

On the clinical routine dataset, LPA, SLS and BIANCA exhibited the highest DSC and their performances were not significantly different (respectively 0.65, 0.61, 0.61). To note, LPA and SLS ranked second in term of DSC performance on the ADNI sample (0.54, 0.53), which may be the most accurate when trained on data similar to the data used for testing but they do not generalize well. Valverde and colleagues (Valverde et al., 2019), already demonstrated this for NicMSlesion, on two different multiple sclerosis datasets, that one obtains lower performances when using a model trained on a dataset that is too different from the test set.

Overall, our results demonstrate that several tools achieved acceptable performances on both research and clinical datasets. A reasonable first choice of WMH segmentation tool can be either LPA or SLS, even though one drawback is that they require a Matlab license (Table 5). Based on our results when the image dataset could contain patients with artifacts, SLS may be the method of choice. (Fig. 3a, Table 5). On the other hand, in a dataset with many different scanners, BIANCA may be preferred because of its robustness (Fig. 3b, Table 5). As for the neural network nicMSlesion, it seems to be performant only when retrained on data that is similar to the data to be segmented (Fig. 1, Table 4).
Fig. 3. Maps of False negative and False positive rate from each method on the clinical routine dataset. We represent masks of segmentation on MNI template. The first row of the plot represents an overlay of manual segmentation in the ADNI testing set. The greyscale ranges from 0% (white) to 33% (black) of WMH at any particular voxel. The left column of the plot represents the false negative rate map for each method in ADNI testing set. The right column shows the false positive rate map for each method on ADNI dataset. Scale ranges from 0 to 33% of errors at each voxel, which corresponds to the maximal error rates observed.
Overall, we reported lower performances than in the majority of previously published papers (Griffanti et al., 2016; Jiang et al., 2018; Kuijf et al., 2019). However, Rachmadi and colleagues obtained results similar to ours when evaluating LST LGA on the ADNI data (Rachmadi et al., 2018). In the same way, Heinen and coworkers had also similar result on their evaluation of LST LGA and LPA (Heinen et al., 2019). One of the reasons could be the lower volume of WMH in our two datasets (and especially in ADNI). However, Jiang and colleagues (Jiang et al., 2018), in the original publication describing the UBO software, also reported higher performance for low WMH volume. Another possible explanation is that the methods and performance reported may not generalize that well and may be somewhat optimistic or simply not comparable between publications or with our results that used different training and test samples. We also benchmarked prediction accuracy on the same research and clinical samples. Finally, we cannot rule out that the lower performances we report are attributable to differences in manual segmentation protocols between our study and previous ones.

One should note that the performances we report of most tools is moderate (DSC between 0.4 and 0.6 in most cases, Table 3, 4, Figure 11) and always below intra and inter-rater reproducibility (Table 2). This suggests more work is needed to improve performance of automated algorithms for WMH parcellation, which may include considering other models (Supplementary Table 5) or larger training samples. We provided maps of false positive and false negative rates for each of the methods (Figs. 2 and 3), which may represent a useful feedback for method developers. In short, we note that on the ADNI dataset, errors were really close to the WMH identified by the radiographers which suggests some of the errors are on the boundaries of WMH regions. To improve performance of their methods, apart from the fact that the medical definition of WMH could be better homogenized, they could use a mask to eliminate some regions where WMH is impossible or presumed to be of vascular origin (e.g. septum pellucidum, cortex or cerebellum white matter). Secondly, they should find ways to better standardize white matter intensity, either using extensive training datasets or using the intensity of the cortex for instance.

In clinical use, it might be important to progress preprocessing to reduce artifact rate. Importantly, we should not necessarily discard those methods because of low DSC, as DSC may be overly sensitive to limited WMH parcellation errors. The boundaries of WMH regions are always being debated. Thus, we reported several metrics throughout the manuscript that may advise on which method is best for different applications. For

---

**Table 5.**

Overall, we reported lower performances than in the majority of previously published papers (Griffanti et al., 2016; Jiang et al., 2018; Kuijf et al., 2019). However, Rachmadi and colleagues obtained results similar to ours when evaluating LST LGA on the ADNI data (Rachmadi et al., 2018). In the same way, Heinen and coworkers had also similar result on their evaluation of LST LGA and LPA (Heinen et al., 2019). One of the reasons could be the lower volume of WMH in our two datasets (and especially in ADNI). However, Jiang and colleagues (Jiang et al., 2018), in the original publication describing the UBO software, also reported higher performance for low WMH volume. Another possible explanation is that the methods and performance reported may not generalize that well and may be somewhat optimistic or simply not comparable between publications or with our results that used different training and test samples. We also benchmarked prediction accuracy on the same research and clinical samples. Finally, we cannot rule out that the lower performances we report are attributable to differences in manual segmentation protocols between our study and previous ones.

One should note that the performances we report of most tools is moderate (DSC between 0.4 and 0.6 in most cases, Table 3, 4, Figure 11) and always below intra and inter-rater reproducibility (Table 2). This suggests more work is needed to improve performance of automated algorithms for WMH parcellation, which may include considering other models (Supplementary Table 5) or larger training samples. We provided maps of false positive and false negative rates for each of the methods (Figs. 2 and 3), which may represent a useful feedback for method developers. In short, we note that on the ADNI dataset, errors were really close to the WMH identified by the radiographers which suggests some of the errors are on the boundaries of WMH regions. To improve performance of their methods, apart from the fact that the medical definition of WMH could be better homogenized, they could use a mask to eliminate some regions where WMH is impossible or presumed to be of vascular origin (e.g. septum pellucidum, cortex or cerebellum white matter). Secondly, they should find ways to better standardize white matter intensity, either using extensive training datasets or using the intensity of the cortex for instance.

In clinical use, it might be important to progress preprocessing to reduce artifact rate. Importantly, we should not necessarily discard those methods because of low DSC, as DSC may be overly sensitive to limited WMH parcellation errors. The boundaries of WMH regions are always being debated. Thus, we reported several metrics throughout the manuscript that may advise on which method is best for different applications. For
example, several algorithms achieved good performance on WMH volume evaluation [Table 3, 4], which is the main criterion used in clinical assessment. For example, one could choose to use UBO, given that it obtained good volumetric results and that it directly provides additional features, such as segmenting WMH by vascular territories or anatomical regions. In addition, one could select the method that minimizes the false negative rate (SLS or LPA) as a way of reducing the search space for manual segmentation.

Many teams develop automatic WMH segmentation methods, mainly for multiple sclerosis or microvascular pathology associated with aging and cognitive impairment. However, few have compared the performance of these different methods. Many challenges are comparing automatic WMH segmentation methods, both for multiple sclerosis (MS segmentation challenge MICCAI 2008 (http://www.ia.unc.edu/MSseg/), ISBI 2015 longitudinal multiple sclerosis lesion segmentation challenge (Carass et al., 2017) (http://iacl.ece.jhu.edu/index.php/MSChallenge), MSSEG MICCAI Challenge 2016 (Commowick et al., 2018) (https://portal.fli-iam.iris.fr/msseg-challenge) and age-related WMH (MICCAI 2017 WMH Challenge (Kuijf et al., 2019)). Such challenges are very useful to assess which are the most efficient methodological approaches. But most of the participants to these challenges do not provide easily usable codes implementing their tools. Thereby, while these challenges are very useful to the methodological community of researchers developing new algorithms, they are of less use to a radiologist who would like to choose an easy-to-use tool.

To our knowledge, this is the first study to compare software while including data with artifacts, which reflects the reality of clinical routine. Indeed, artifacts are common, in particular during MRI acquisitions of patients with cognitive impairment. For all methods, there was a performance drop on data with artifacts (Fig. 3a). Such reduction in performance was significant for the deep neural network and BIANCA, losing over 0.2 point of DSC. On the other hand, SLS performs best on these data and the performance drop between data without and with artifacts was only 0.05 points of DSC.

Beyond sheer performances of the algorithms, we also evaluated how robust their performances were when using several MRI scanners, which is one of the principal factor of heterogeneity in MRI intensity. We found that the performance of most algorithms was sensitive to scanner types, though, LGA and BIANCA appeared the most robust. Our results align and extend those of a recent publication which suggested a possible scanner effect on algorithms performance on a sample of 42 participants from 7 different scanners (Heinen et al., 2019). Robustness may be an important criterion for algorithm selection in multi-centric studies, and in particular when the proportion of cases and controls varies between sites/scanners, thus when site/scanner may confound WMH association analyses because. More work is needed to further study if training on each/several scanner type could improve the performance of algorithms sensitive to scanner type. We evaluated heterogeneity related to the type of scanner, although it is a major factor of heterogeneity, it is not the only one. Thus, many factors, such as TE, TR, matrix size, etc., can influence image quality and contrast, and warrant further investigation.

Our study has the following limitations. First, the imperfect reproducibility of manual parcellation (at a vertex level, Table 2) calls for a more precise definition of the WMH and their boundaries, even though it was similar to that reported in the literature (Commowick et al., 2018; Coupé et al., 2018; Kuijf et al., 2019). When we visually inspected our different manual segmentation, we found that most of the differences to be at the limit of the WMH regions with an unclear intensity gradient. There is no precise standardized recommendation about the periventricular hyperintensities, whether they should be considered microvascular pathology or not. To overcome this limitation, we designed a protocol with experienced neuroradiologists. We discussed the pathogenesis of some hyperintensities, such as one-pixel hyperintensities close to the ventricles or in touch with the genu of the corpus callosum. Whether these very thin WMH represent microvascular pathology remains debated in the field. Kim et al. demonstrated that non-ischemic WMH are often located in juxtaventricular areas because they likely result from cerebrospinal fluid leakage (Kim et al., 2008) while Hernandez et al. came to more mixed conclusions (Hernández et al., 2014). Overall, the segmentation accuracy remains substantially lower than what is reported for other medical image segmentation tasks, such as subcortical grey matter structures (Pagnozzi et al., 2019) or brain tumors (Wadhw et al., 2019). This illustrates that WMH segmentation remains a challenging task. Second, having only one reader for the test set may also be seen as a limitation. Nevertheless, considering the limited inter- and intra-observer reproducibility that we report it may have led to an overly conservative definition of WMH. Lastly, another limitation is to have limited ourselves to a set of methods, which are directly usable and allowing access to the segmentation mask. For example, this led us to

### Table 5

Summary of evaluation, and some selected information to choose a method.

| Method            | Ranking on research data | Ranking on routine data | Robustness artifacts | Robustness different scanner | Sequences needed | Need training data | Limitations/Requirements | Proc. time |
|-------------------|--------------------------|-------------------------|----------------------|-----------------------------|------------------|-------------------|--------------------------|------------|
| LPA               | 2                        | 1                       | –                    | –                           | FLAIR            | No                | Matlab                  | 1 min      |
| LGA               | 4                        | 5                       | –                    | –                           | FLAIR/T1w        | No                | Matlab                  | 6 min      |
| BIANCA            | 4                        | 1                       | –                    | –                           | FLAIR/T1w        | Yes               | Need mask of WM         | 17 min     |
| SLS               | 2                        | 1                       | –                    | –                           | FLAIR/T1w        | No                | Matlab                  | 8 min      |
| W2MBS             | 8                        | 8                       | –                    | –                           | FLAIR/T1w        | No                | Matlab                  | 5 min      |
| nicMSlesion (original) | 4                        | 7                       | –                    | –                           | FLAIR/T1w        | No                | GPU                     | 10 min*2   |
| nicMSlesion (retrained) | 4                        | 7                       | –                    | –                           | FLAIR/T1w        | Yes               | GPU                     | 10 min*2   |
| UBO               | 4                        | 4                       | –                    | –                           | FLAIR/T1w        | No                | Matlab                  | 9 min      |

Ranking performed using t-test comparison on the primary criterion (DSC) (see Supplementary Tables 7 and 9 for details). We started by looking at the method with the best DSC. Then all methods not significantly different from it were given the same rank classified, and so on.

Processing time were evaluated on MacBook Pro laptop with a 2.2 GHz Intel Core i7 2018 CPU, without a graphic processing unit (GPU), with 16 Go RAM except for the nicMSlesion for which we used a GPU-equipped computer, namely a Linux workstation with an Intel Xeon E5-2699 @ 2.30 GHz CPU, with NVIDIA Quadro M4000 GPU, 256 Go RAM.

– indicates that the DSC is sensitive to artifacts or scanner type at p < 0.05 uncorrected for multiple comparisons, on routine dataset.

– indicates that the DSC is sensitive to artifacts or scanner type at after correction for multiple testing, on routine dataset.

* Best DSC in our evaluation (though not necessarily significantly better which explains equal first).

1 2 min for segmentation and 15 min for generation of the exclusion mask.

2 With graphic processing unit (GPU, NVIDIA Quadro M4000).

3 3.5 min for segmentation and 6.5 min for preprocessing.

4 Retraining time.
exclude the tool which ranked first in the MICCAI 2017 WMH Segmentation Challenge (Li et al., 2018) because it was not directly usable. It uses a U-net convolutional neural network (Ronneberger et al., 2015) and four of the top ten challengers of this challenge used this architecture, but they have not released a user-friendly version. We also did not include the promising LesionBRAIN (Coupé et al., 2018) (Supplementary Table 4) because it does not provide the required segmentation mask. More work is needed to make new software available and user-friendly to the community. In addition, future work could focus on the integration in the clinical workflow, allowing for example, to use these tools directly with DICOM. We have released all hyper-parameters used in our analysis, which should facilitate replication of our results and reuse of the algorithms we considered here.

In conclusion, we compared seven automatic WMH segmentation algorithms on both research and clinical routine data. These results can provide useful information to researchers and radiologists looking to choose an automatic WMH segmentation method.

Source of funding

The research leading to these results has received funding from the program "Investments of d'avenir” reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10 IA Institut Hospitalo-Universitaire-6), from the European Union H2020 program (project EuroPOND, grant number 666992, from the joint NSF/NIH/ANR program “Collaborative Research in Computational Neuroscience” (project HIPLA17, grant number ANR-16-NEUC-0001-01), from Agence Nationale de la Recherche (project PREVDEMALS, grant number ANR-14-CE15-0016-07), from Fonction pour la Recherche sur Alzheimer (project HistoMRI), from the Abeona Foundation (project Brain@Scale), from the Fondation Vaincre Alzheimer (grant number FR-18006CB), and from the "Contrat d’Interface Local” program (to Dr Colliot) from Assistance Publique-Hôpitaux de Paris (AP-HP). Q. Vanderbecq is supported by a "Bourse de Recherche Alain Rahmouni" from the Société Française de Radiologie-College des Enseignants de Radiologie de France (SPR-CERF). BCD is supported by the NHRMC (CJ Martin Fellowship, APP1161356).

CRediT authorship contribution statement

Quentin Vanderbecq: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition.

Eric Xu: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition.

Sebastian Ströer: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision.

Baptiste Couvy-Duchesne: Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Supervision.

Olivier Colliot: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgements

Ours thoughts go to Anne Bertrand, MD, PhD, a brilliant radiologist, professor and researcher, without whom this project would never have taken shape.

We thank Simona Bottani, Johann Fauzou, Alexis Guyot, Arnaud Marcoux, Benoit Martin, Alexandre Routier and Adam Wild for the IT support and assistance in programming during this study. We also thank Ninn Burgos for her helpful advice.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2020.102357.

References

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. Neuroimaging 13, 261–276. https://doi.org/10.1016/j.neurimag.2016.12.064.

Coupé, P., Toudias, T., Linck, P., Romero, J.E., Manjon, J.V., 2018. LesionBrain: An Online Tool for White Matter Lesion Segmentation, in: Bai, W., Sanroma, G., Wu, G., Mummel, B.C., Zhang, Z., Coupé, P. (Eds.), Patch-Based Techniques in Medical Image Analysis, Lecture Notes in Computer Science. Springer International Publishing, pp. 95–103.

Duong, M.T., Rudie, J.D., Wang, J., Xie, L., Mohan, S., Gee, J.C., Rauschecker, A.M., 2019. Convolutional Neural Network for Automated FLAIR Lesion Segmentation on Clinical Brain MR Imaging. Am. J. Neuroradiol. https://doi.org/10.3174/ajnr.A6138.

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 ageing. J. Neurol. Neurosurg. Psychiat. 50, 267–271. https://doi.org/10.1136/jnnp.50.3.267.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Weller, R., Bittner, P., Allali, A., Aliev, A., Humms, M., 2018. Preclinical and clinical stroke research using neurological imaging: a review. Neuroimage 194, 351–356. https://doi.org/10.1016/j.neuroimage.2018.09.060.

Hernández, M. del C.V., Morris, Z., Dickie, D.A., Royle, N.A., Manjica, S.M., Aribisala, B.S., Bastin, M.E., Deary, I.J., Wardlaw, J.M., 2013. Close Correlation between Quantitative and Qualitative Assessments of White Matter Lesions. Neuroinformatics 11, 381–396. https://doi.org/10.1007/s12021-013-9186-0.

Huang, Y., Shyr, Y., Jiang, H., Lin, W., Yang, Z., Liang, H., 2014. Longitudinal white matter hyperintensity progression in MS patients. Mult. Scler. J. 20, 671–678. https://doi.org/10.1177/1352458513515762.

Huang, Y., Shyr, Y., Jiang, H., Lin, W., Yang, Z., Liang, H., 2014. Longitudinal white matter hyperintensity progression in MS patients. Mult. Scler. J. 20, 671–678. https://doi.org/10.1177/1352458513515762.
