JOURNAL OF
NEUROIMAGING
THE OFFICIAL JOURNAL OF THE AMERICAN SOCIETY OF NEUROIMAGING

WILEY

CLINICAL INVESTIGATIVE STUDY

Automatic deep learning multicontrast corpus callosum segmentation in multiple sclerosis

Irene Brusini[1,2] | Michael Platten[1,3,4] | Russell Ouellette[3,4] | Fredrik Piehl[4,5,6] | Chunliang Wang[1] | Tobias Granberg[3,4]

[1] School of Chemistry, Biotechnology, and Health, Royal Institute of Technology, Stockholm, Sweden

[2] Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden

[3] Department of Neuroradiology, Karolinska University Hospital, Stockholm, Sweden

[4] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

[5] Department of Neurology, Karolinska University Hospital, Stockholm, Sweden

[6] Center for Neurology, Academic Specialist Center, Stockholm Health Services, Stockholm, Sweden

Correspondence
Michael Platten, Kungshamra 56A Lgh 1101, Stockholm, Sweden.
Email: michael.platten@ki.se

Irene Brusini and Michael Platten shared first-authorship.

Abstract

Background and Purpose: Corpus callosum (CC) atrophy is predictive of future disability in multiple sclerosis (MS). However, current segmentation methods are either labor- or computationally intensive. We therefore developed an automated deep learning-based CC segmentation tool and hypothesized that its output would correlate with disability.

Methods: A cohort of 631 MS patients (449 females, baseline age 41 ± 11 years) with both 3-dimensional T1-weighted and T2-weighted fluid-attenuated inversion recovery (FLAIR) MRI was used for the development. Data from 204 patients were manually segmented to train convolutional neural networks in extracting the midsagittal intracranial and CC areas. Remaining data were used to compare segmentations with FreeSurfer and benchmark the outputs with regard to clinical correlations. A 1.5 and 3 Tesla reproducibility cohort of 9 MS patients evaluated the segmentation robustness.

Results: The deep learning-based tool was accurate in selecting the appropriate slice for segmentation (98% accuracy within 3 mm of the manual ground truth) and segmenting the CC (Dice coefficient .88-.91) and intracranial areas (.97-.98). The accuracy was lower with higher atrophy. Reproducibility was excellent (intraclass correlation coefficient > .90) for T1-weighted scans and moderate-good for FLAIR (.74-.75). Segmentations were associated with baseline and future (average follow-up time 6-7 years) Expanded Disability Status Scale ($\rho = -.13$ to $-.24$) and Symbol Digit Modalities Test ($r = .18$-$.29$) scores.

Conclusions: We present a fully automatic deep learning-based CC segmentation tool optimized to modern imaging in MS with clinical correlations on par with computationally expensive alternatives.

KEYWORDS
atrophy, convolutional neural networks, corpus callosum, magnetic resonance imaging, multiple sclerosis, neurodegeneration

INTRODUCTION

Multiple sclerosis (MS) is an immune-mediated disease characterized by neuroinflammation and neurodegeneration.[1] Although focal lesions in the deep white matter are a hallmark of MS, neurodegenerative processes involve cortical and subcortical structures as well as the normal-appearing white matter.[2] Neurodegeneration represents the most important anatomical correlate of permanent disability in MS, as shown by the predictive value of atrophy measures across different anatomical structures.[3] The corpus callosum (CC) consists of heavily myelinated axons that connect the two cerebral hemispheres and their cortical/subcortical networks.[4] It is evident that the CC is highly susceptible to the disease processes occurring in MS with high atrophy rates, which makes it a relevant biomarker and a predictor of neurodegeneration.[5] Not only is the CC an independent biomarker of neurodegeneration, but it also correlates well with other neuroimaging features in MS.[6]

Despite its clinical relevance, CC atrophy is rarely assessed in clinical practice because validated visual rating scales are lacking and although several 2-dimensional manual approaches exist for measuring the CC in MS, annotation of images is labor-intensive and confounded by inter- and intrarater variability.[7] In terms of automatic 3-dimensional segmentations, the freely available and popular research software FreeSurfer segments the CC volume among several other structures,[8] and offers a longitudinal segmentation pipeline,[9] but it is computationally expensive reducing its feasibility in large cohort studies.

In recent years, deep learning has emerged as a machine learning approach that is especially suited for image segmentation tasks with high computational efficiency once trained.[10] Convolutional neural networks (CNNs) and, in particular fully convolutional networks, have become one of the most applied methods.[11] In small sample sizes, a U-net architecture is popular and it is often combined with effective image augmentation for improved classification.[12] Previous studies have applied CNNs to segment the CC in 2 dimensions in healthy volunteers as well as in MS but on older sagittal two-dimensional T2-weighted (T2w) and T1-weighted (T1w) images.[13,14] Meanwhile, the current MAGNIMS-CMSC-NAIMS consensus guidelines recommend that 3-dimensional T2w fluid-attenuated inversion recovery (FLAIR) and T1w images are acquired for diagnosis and monitoring in MS.[15,16]

Therefore, we developed a fully automatic, computationally efficient deep learning-based segmentation tool tailored for modern 3-dimensional FLAIR and T1w scans in MS. The aim was to provide a biomarker of neurodegeneration in MS feasible to be run on large datasets. We hypothesized that the output would correlate with physical and neurological disability in MS to a similar degree as the more computationally expensive FreeSurfer segmentations of the CC volume.

METHODS

Study design and clinical data

The data for this study were acquired within the Stockholm Prospective assessment of Multiple Sclerosis (STOP-MS) study, a prospective population-based cohort study started in January 2001 aiming to identify prognostic factors for long-term outcomes in newly diagnosed MS patients.[17] In total, 631 individual MS patients were available, of which data from 204 patients were randomly selected for training the CNNs. Clinical data were extracted from the Swedish MS registry, which collects prospectively clinical information with high validity.[18] Disease courses, as defined by the treating neurologist, were either relapsing-remitting MS, secondary progressive MS, or primary progressive MS.[19–21] The Expanded Disability Status Scale (EDSS), representing an accumulated score of different neurological subdomain disabilities, was extracted if there were baseline values <6 months of the scan date and/or at follow-up (>6 months from the scan date). Similarly, Symbol Digit Modalities Test (SDMT), representing a measure of cognitive processing speed, was extracted for the baseline and follow-up. The SDMT scores were transformed into sex- and age-adjusted z-scores.[22] Table 1 presents the demographics of the 204 patients in the training data set and the 427 patients who underwent segmentations by both the deep learning algorithms and FreeSurfer. This study was approved by the Regional Ethical Review Board in Stockholm (reg. no. EPN 2009/2017-31/2) (with amendments 2018/2711-32, 2020-01954, 2020-03471, and 2021-02060).

Magnetic resonance imaging

Brain MRIs were acquired using three different Siemens scanners at the Karolinska University Hospital in Huddinge, Stockholm, Sweden. The image acquisition parameters are presented in Table 2.

Training and validating the CNNs

Preparing the training data

ITK snap v3.6.0 (www.itksnap.org)[23] was used for manually selecting the midsagittal slice and segmenting the intracranial (IC) and CC areas. This was performed by an experienced rater and physician (MP) for each of the training images. Figure 1 provides examples of six manual segmentations.

Automatic midslice selection

Automatic midsagittal slice selection was achieved by training a CNN classifier. It received an MRI sagittal slice as input and was trained

TABLE 1 Cohort demographics of the training and testing dataset

	Training data (n = 204)			Testing data (n = 427)
Scanner	Aera (n = 68)	Avanto (n = 68)	Trio (n = 68)	Aera/Avanto/Trio, % 30/47/23
Age in years, mean ± SD	39 ± 11	39 ± 12	36 ± 12	43 ± 11
Disease duration in years, mean ± SD	6.7 ± 8.0	6.0 ± 5.7	4.8 ± 6.6	4.4 ± 5.4
Sex, % F/M	77/23	74/26	57/43	72/28
Subtype, % RRMS/SPMS/PPMS/NA	69/22/2.9/4.4	75/16/4.4/4.4	79/5.9/2.9/12	70/23/1.8/5.3
Median EDSS within 6 months, IQR	2.0 (1.0-3.0)	1.5 (0.0-2.8)	2.0 (1.5-3.0)	2.5 (1.5-3.5)
	(n = 35)	(n = 37)	(n = 37)	(n = 252)
Median EDSS future score, IQR[a]	2.5 (1.5-4.0)	2.5 (1.5-4.0)	3.0 (1.0-3.5)	2.5 (1.5-4.0)
	(n = 57)	(n = 57)	(n = 59)	(n = 331)
Median SDMT within 6 months, IQR	−0.59 (−1.6 to 0.12)	−0.70 (−1.41 to −0.05)	−0.93 (−2.0 to −0.17)	−0.69 (0.065 to −1.4)
	(n = 31)	(n = 21)	(n = 35)	(n = 172)
Median SDMT future score, IQR[b]	−1.1 (−2.3 to −0.49)	−1.4 (−2.21 to −0.54)	−1.6 (−2.3 to −0.65)	−0.97 (−0.18 to −1.9)
	(n = 55)	(n = 53)	(n = 59)	(n = 304)

Note: n signifies the number of patients.

Abbreviations: EDSS, Expanded Disability Status Scale; F, Female; IQR, interquartile range; M, Male; NA, not available; PPMS, primary-progressing MS; RRMS, relapsing-remitting MS; SD, standard deviation; SDMT, Symbol Digit Modalities Test; SPMS, secondary-progressive MS.

[a]Average number of years between scan and EDSS was 6.7 ± 2.6 years.
[b]Average number of years between scan and SDMT was 5.8 ± 2.8 years.

TABLE 2 MRI scanner settings

	T1-weighted MPRAGE			T2-weighted SPACE FLAIR		
Scanner model	Aera	Avanto	Trio	Aera	Avanto	Trio
Field strength	1.5	1.5	3.0	1.5	1.5	3.0
Voxel size	1.0 × 1.0 × 1.5	1.0 × 1.0 × 1.5	1.0 × 1.0 × 1.5	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0
Echo time	3.02	3.55	3.39	333	333	388
Repetition time	1900	1900	1900	5000	6000	6000
Inversion time	1100	1100	900	1800	2200	2100
Flip angle, °	15	15	9	120	120	120

Note: All times are given as milliseconds. The main magnetic field strength is given as Tesla.

FLAIR, fluid-attenuated inversion recovery; MPRAGE, magnetization-prepared rapid gradient echo; SPACE, sampling perfection with application optimized contrasts using different flip angle evolution.

to predict whether it was a middle or nonmiddle slice. The CNN was modeled using four 2D convolutional layers (each with Rectified Linear Unit, ReLU, activation,[24] 3 × 3 kernels, and 16, 32, 64, and 64 filters, respectively), with max-pooling layers between the layers. This was followed by two dense layers, with 128 and 1 unit(s), respectively. See Figure 2 for an overview of the midslice selection pipeline.

Three separate CNNs were trained to classify the midslices (midCNN) and named based on the MRI sequence(s) provided as input: midCNN$_{T1}$, midCNN$_{FLAIR}$, and midCNN$_{T1/FLAIR}$. All three networks were first trained using 10-fold cross-validation on the available dataset. For this purpose, all available 3-dimensional scans were first randomly split into 10 folds. Then, for each MRI scan, only sagittal slices

were provided as input. Given the strong numerical imbalance between middle (only one per scan) and nonmiddle slices, only half of the nonmiddle slices of every subject were randomly selected to be used during training. All three networks underwent the same image preprocessing: resized to 256 × 256 pixels and normalized by subtracting the mean pixel intensity and dividing them by the standard deviation. All models were trained for 50 epochs using a binary cross-entropy loss function and an Adam optimizer,[25] with a constant learning rate of 0.0001. To address the problem of data imbalance, class weights were used during training. In particular, each of the two classes (middle and nonmiddle) was associated with a weight that was inversely proportional to its frequency in the training dataset, that is, higher weights were given to true
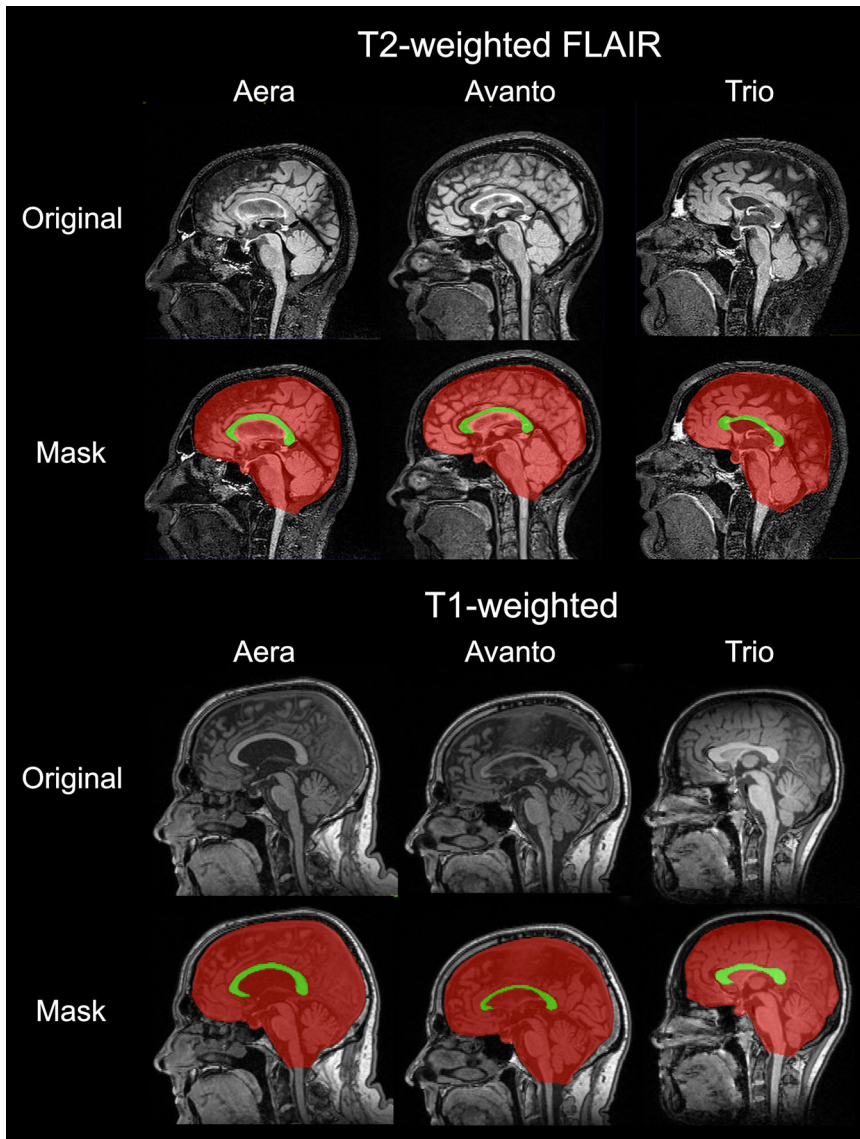
FIGURE 1 Manual corpus callosum and intracranial segmentations of six MS patients. Both T1-weighted and T2-weighted FLAIR scans were segmented using ITK snap (v3.6.0, www.itksnap.org)
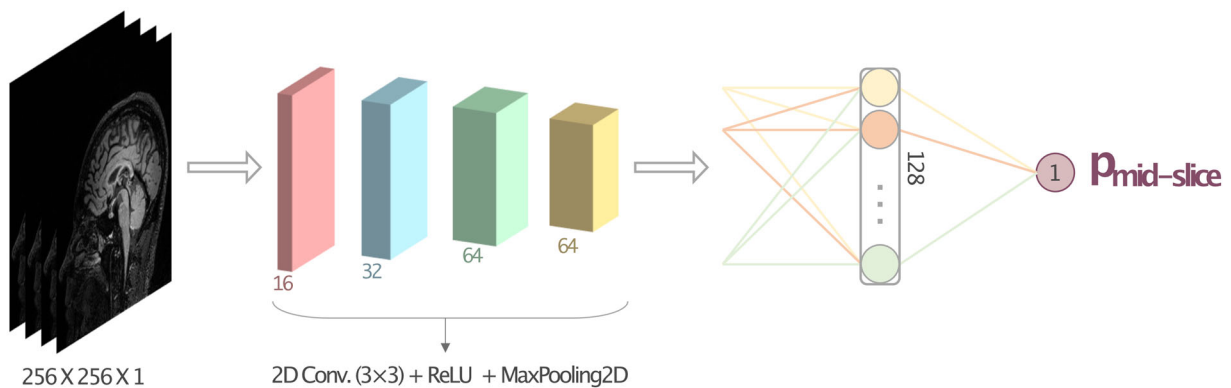

FIGURE 2 Midslice selection pipeline. The input consists of individual slices from the MRI scan, with dimensions $256 \times 256 \times 1$. The input subsequently goes through four convolutional layers, using a 3×3 kernel, ReLU activation, and 2×2 max pooling. At the end, there are two dense layers (128 and 1, respectively), resulting in a high-probability output of the midsagittal slice. The number under each block signifies the number of filters applied in that convolutional layer

middle slices. In the validation phase, all sagittal slices of a test MRI scan were provided as input to the model. The performance of the model was evaluated by computing the absolute error between the index of the predicted midslice slice and the index of the manually annotated midslice. All network implementations, training, and testing were performed using Tensorflow 2.4.1.

Subsequently, the same architectures were re-trained, but now using a scanner-wise (and thus threefold) cross-validation strategy. At each fold, the slices from the scans acquired from one scanner (Aera, Avanto, or Trio) were used as the validation set, whereas the slices from the remaining two scanners were employed as the training set. This was done to investigate the performance of the proposed architecture on data from new unseen scanners. The same data preprocessing and training hyperparameters as those described above were used. However, in this case, data augmentation was also added during training, in order to limit the amount of overfitting toward the two scanners provided as input. This data augmentation consisted of applying random rotation ($-10°$ to $10°$), translation (-10 to 10 pixels along both the x and y axes), and scaling (-5% to 5% of the original size) to all the training slices at each epoch.

U-Net-based CC and IC segmentation

Following the same strategy presented in previous literature,[14] two U-Net architectures were implemented to automatically segment the CC and the IC, respectively, from a single input midsagittal slice. Analogous to the description in the previous section, the algorithms were named according to input data: CC-Net$_{T1}$ and IC-Net$_{T1}$; CC-Net$_{FLAIR}$ and IC-Net$_{FLAIR}$; CC-Net$_{T1/FLAIR}$ and IC-Net$_{T1/FLAIR}$. The CC-Net$_{T1/FLAIR}$ and IC-Net$_{T1/FLAIR}$ received both T1w and FLAIR scans during training, allowing it to segment both types of sequences. The number of filters in each convolutional block (ie, in the encoder, bottleneck, and decoder) was set to 16, 32, 64, 128, 256, 128, 64, 32, and 16, in order from the input to the output. Batch normalization was also applied after each convolutional layer, except for IC-Net$_{T1/FLAIR}$, for which the performance was found to be improved by removing batch normalization. The same image preprocessing performed for the automatic midslice selection was employed for the U-Nets. Two-dimensional data augmentation with random rotation ($-10°$ to $10°$), translation (-10 to 10 pixels along both the x and y axes), and scaling (-5% to 5% of the original size) was introduced to reduce overfitting. When combining T1w and FLAIR data (ie, for CC-Net$_{T1/FLAIR}$ and IC-Net$_{T1/FLAIR}$), different sample weights were assigned to the different MRI sequences. This was done to improve the segmentation performance on FLAIR images, which turned out to be a more challenging task. All U-Nets were trained using a Dice loss function and a stochastic gradient descent optimizer with a decaying learning rate. Similar to the previous section, the networks were trained and evaluated by performing 10-fold cross-validation. For the CC-Net$_{T1}$ and CC-Net$_{FLAIR}$, the cross-validations were also stratified based on atrophy levels. This stratification entailed assigning an atrophy level of high, medium, or low based on whether the normalized CC area within the cohort was in the top, middle, or bottom third.

On the other hand, for training IC-Net$_{T1/FLAIR}$ and CC-Net$_{T1/FLAIR}$, an equal distribution of T1w and FLAIR samples was maintained across folds. Figure 3 provides a schematic overview of the entire algorithm pipeline. All network implementations, training, and testing were performed using Tensorflow 2.4.1.

Finally, similar to what was performed for the automatic midslice selection, a scanner-wise cross-validation strategy was also investigated for the IC and CC segmentations.

Reproducibility: Scan-rescan precision

To discern the reproducibility of the segmentation tool in estimating the normalized CC area, a separate dataset of 9 MS patients (6 females, age 38 ± 13 years, disease duration 7.3 ± 5.2 years) scanned with both T1w and FLAIR in all three MRI scanners on the same day was used. Thereafter, the intraclass correlation coefficient (ICC) was calculated as a metric of precision, using the ICC(A,1) model.[26] We also aimed to investigate if the automatic midslice selection (midCNN) introduced bias into the CC-Net and IC-Net segmentations. Thus, the ICC was also calculated on the results obtained from the algorithm segmentations performed on manually selected midslices on the same 9 patients. We will refer to this approach as the semiautomatic pipeline.

Applying the full pipeline on real-life data

The full pipeline was re-trained on all available training data and was thereafter applied on a dataset of 427 additional unique patients who all had both T1w and FLAIR scans available. These patients' T1w scan also underwent FreeSurfer 3-dimensional segmentation of the CC and an estimation of the total IC volume, for computation of a normalized CC volume (nCCV). The performance of each segmentation technique was evaluated by correlating it to neurologic disability (SDMT and EDSS). The processing time was less than 1 minute per slice extraction and segmentation compared to just over 10 hours per CPU core for FreeSurfer on a Mac-Book Pro with 3.3 GHz Dual-Core Intel Core i7 and 8 GB 2133 MHz LPDDR3 RAM (Apple Inc., Cupertino, CA, USA).

Statistical analysis

Normality was assessed through histograms and Shapiro-Wilk's test. Pearson's and Spearman's rho correlation coefficients were applied for parametric and nonparametric data, respectively. A Dice coefficient was applied to evaluate the accuracy of the segmentations. An ICC was used to compare intrarater variability, as well as to evaluate the midslice selection algorithm. A paired t-test was applied to evaluate the performance between algorithms, and an analysis of variance (ANOVA test) was applied to evaluate the algorithms' performance across scanners. A P-value of $<.05$ was considered statistically significant.
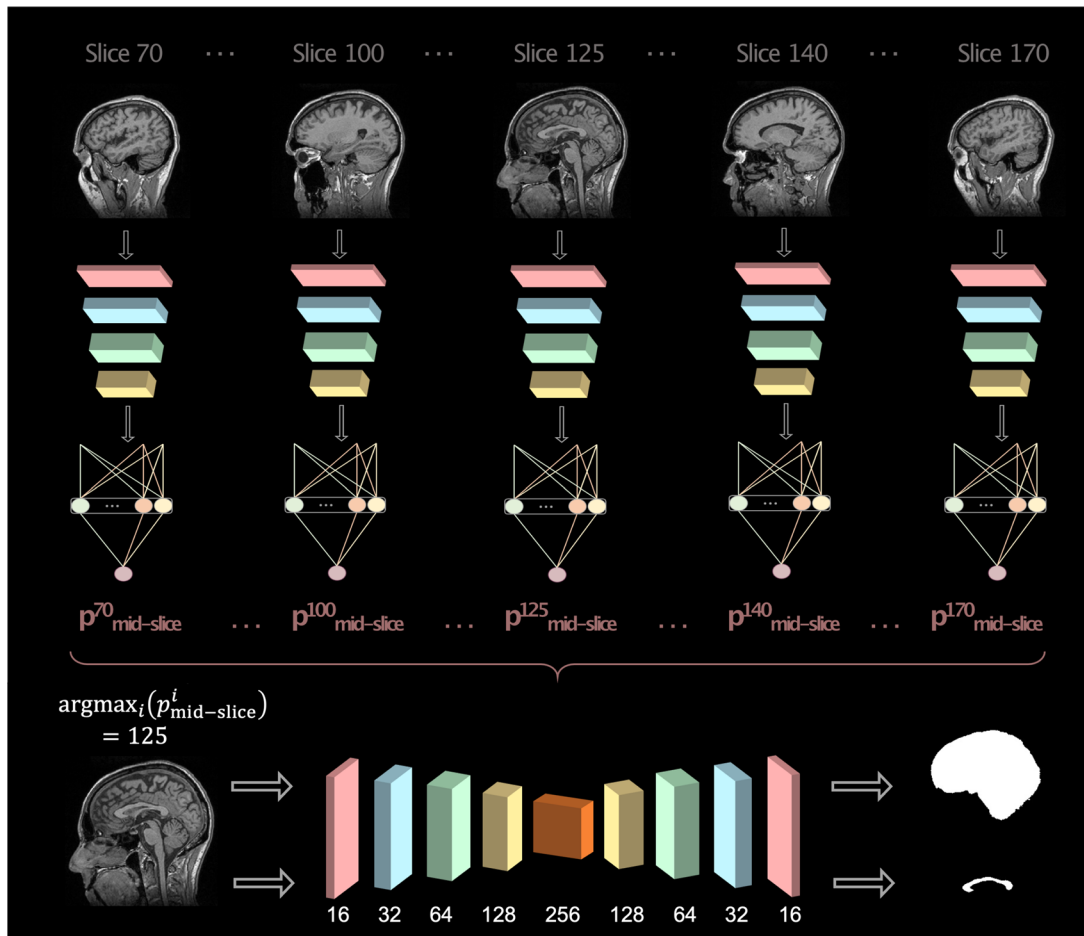
FIGURE 3 Full midslice and segmentation pipeline. Initially the midslice selection algorithm picks the slice with the highest probability of being middle. This is then fed into the segmentation pipelines that segment both the intracranial and corpus callosum area separately. This U-net is based on the model by Ronneberger et al.,[12] where each convolutional layer in the downsampling path applies twice as many filters as the previous layer. The numbers underneath each box represent the number of filters present. Concatenation is applied in order to retain spatial information in the upsampling path. A kernel size of 3×3, along with a max pooling of 2×2, was applied. CC, corpus callosum: IC, intracranial

RESULTS

Midslice selection algorithm

The performance of the three implemented midslice selection networks was first obtained through 10-fold cross-validation. All three architectures showed a mean absolute error (MAE; referring to the absolute difference between ground-truth and predicted midslice index) ranging between 0.60 and 1.07 mm. Only a very small portion of the data showed an absolute error above 3 mm (<1%, <1%, and <2% for $midCNN_{T1}$, $midCNN_{FLAIR}$, and $midCNN_{T1/FLAIR}$, respectively), indicating a limited presence of outliers. The maximum absolute error ranged between 2 and 6 mm. The mean normalized absolute error ranged between 0.20% and 0.59%.

Subsequently, intrarater variability was analyzed by comparing the midslice selections performed by the same expert rater on two occasions separated by 2 months. These two different sets of annotations showed an ICC of .999 and .993 (P-value < .01 for both) for T1w and FLAIR data, respectively. For comparison, the ICCs between the predictions from $midCNN_{T1}$, $midCNN_{FLAIR}$, and $midCNN_{T1/FLAIR}$ to their

ground-truth annotations were .991, .968, and .998 (P-value < .01 for all), respectively. Finally, between the two sets of manual annotations, we observed a mean absolute difference of 0.12 mm (maximum difference of 3 mm) between the selected midslice indices on T1w data, whereas on FLAIR data, the mean absolute difference was 0.15 mm (maximum difference of 3 mm).

Furthermore, Table 3 shows the performance of the three networks when following a scanner-wise cross-validation strategy. In this case, the proposed algorithms show a similar MAE range of 0.64-1.16 mm, and again a small minority of prediction errors (<1%, <2%, and <2% for $midCNN_{T1}$, $midCNN_{FLAIR}$, and $midCNN_{T1/FLAIR}$, respectively) were higher than 3 mm. None of the scanners showed a significant difference when tested by ANOVA.

Segmentation algorithm

Similar to the midslice selection algorithm, the performance of the three scanners' IC and CC segmentation networks was evaluated using 10-fold cross-validation. For the IC segmentation, all three networks

TABLE 3 Performance of the midslice selection algorithms using scanner-wise cross-validation

			Overall	Aera	Avanto	Trio
			(n = 102)	(n = 34)	(n = 34)	(n = 34)
midCNN$_{T1}$		MAE	1.16 mm	1.11 mm	1.11 mm	1.28
		Mean NAE	0.32%	0.31%	0.29%	0.35%
		Max AE	6 mm	3 mm	3 mm	6 mm
		N. AE ≤ 3 mm	101	34	34	33
midCNN$_{FLAIR}$		MAE	0.68 mm	0.62 mm	0.59 mm	0.85 mm
		Mean NAE	0.40%	0.36%	0.34%	0.50%
		Max AE	4 mm	2 mm	3 mm	4 mm
		N. AE ≤ 3 mm	101	34	34	33
midCNN$_{T1/FLAIR}$	On T1w scans	MAE	1.14 mm	1.07 mm	0.98 mm	1.41 mm
		Mean NAE	0.31%	0.29%	0.26%	0.39%
		Max AE	6 mm	3 mm	3 mm	6 mm
		N. AE ≤ 3 mm	99	34	34	31
	On FLAIR scans	MAE	0.64 mm	0.76 mm	0.44 mm	0.71 mm
		Mean NAE	0.37%	0.44%	0.25%	0.42%
		Max AE	2 mm	2 mm	1 mm	2 mm
		N. AE ≤ 3 mm	102	34	34	34

Note: For this scanner-wise cross-validation, at each fold the data from one scanner (Aera, Avanto or Trio) were used as validation set, whereas those from the remaining scanners were used as training set. *n* signifies the number of patients. For midCNN$_{T1/+FLAIR}$, the AEs are analyzed separately for each MRI sequence (T1-weighted [T1w] and FLAIR) as this is an important metric for a multicontrast algorithm. No significant difference was found across scanners as tested by analysis of variance (ANOVA).

Abbreviations: CNN, convolutional neural network; MAE, mean absolute error (between ground-truth and predicted midslice); Mean NAE, average normalized absolute error (obtained by dividing each AE by the total image size along the sagittal view); Max AE, maximum absolute error; N. AE ≤ 3 mm, number of cases that reported an error that was less or equal to 3 mm.

showed similar performance overall (ie, mean Dice coefficient in the range between .970 and .978). When applying IC-Net$_{T1/FLAIR}$ to segment both T1w and FLAIR data, the obtained average Dice was significantly, but not substantially, better than the two other algorithms (.98 vs. .97, $P < .01$, by independent *t*-test).

The CC-Net$_{T1}$ exhibited a significantly higher performance on average than FLAIR segmentations, with a mean Dice coefficient of .91 vs. .88 for both CC-Net$_{FLAIR}$ and CC-Net$_{T1/FLAIR}$ (*P*-value: <.01, by independent samples *t*-test). Moreover, when applying CC-Net$_{T1/FLAIR}$, the differences between T1w and FLAIR persisted, with T1w images showing a significantly higher segmentation accuracy on average (Dice coefficient of .895 ± .070) compared to FLAIR scans (.865 ± .074) (*P*-value: <.05, by independent samples *t*-test). The higher the level of CC atrophy, the lower the performance, across all three of the implemented architectures (Figure 4). Representative examples from the CC and IC algorithm are presented in Figure 5.

Finally, the results from the scanner-wise analysis are presented in Table 4. The performance obtained on the IC segmentation is consistent with that observed using 10-fold cross-validation across all three networks and scan contrasts. Similarly, the CC segmentations performed on T1w data, using both CC-Net$_{T1}$ and CC-Net$_{T1/FLAIR}$, were largely unchanged compared to the 10-fold cross-validation. In all these cases, no significant differences in performance were found across scanners, as tested by ANOVA. In contrast, the accuracy of the CC segmentations on FLAIR scans using the CC-Net$_{T1/FLAIR}$ resulted in significant differences between scanners, as tested by ANOVA. For this dataset, the highest accuracy was observed when testing the network on Trio data (mean Dice of .852) and the lowest accuracy on Aera (mean Dice of .742).

Reproducibility

Table 5 presents the ICCs that were computed between normalized CC area measures (automatic and semiautomatic) from 9 patients who were scanned using all three scanners on the same day. Despite the high accuracy of the automatic midslice selection algorithm, the use of the semiautomatic pipeline led to numerically higher ICC in the FLAIR-specific pipeline (.828 vs. .739). In the segmentation of T1w data using IC-Net$_{T1/FLAIR}$ + CC-Net$_{T1/FLAIR}$, the ICCs were similar to the semiautomatic approach (.910 with 95% confidence interval of [.734, .977] vs. .908 with 95% confidence interval of [.617, .979]).

Clinical correlates

A total of 427 unique patients, having both T1w and FLAIR scans available and not belonging to the training data, were available for segmentation by our algorithm as well as FreeSurfer. Both the FreeSurfer and
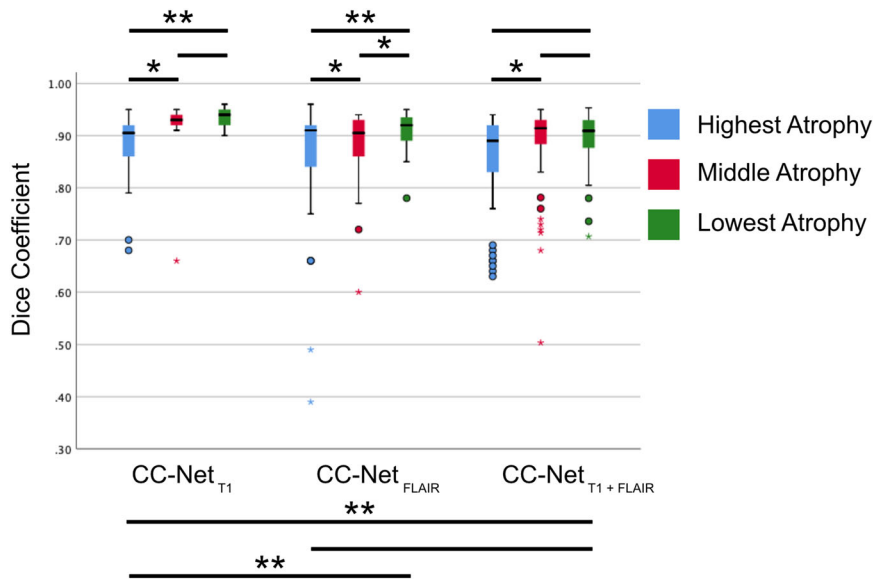
FIGURE 4 A clustered boxplot showing how the segmentation accuracy significantly decreases with higher atrophy. All patients in the training cohort were split into one of three atrophy levels, based on whether their normalized CC area was in the top, middle, or bottom third of the cohort. *P-value < .05; **P-value < .01; CC, corpus callosum

TABLE 4 Performance of the IC and CC segmentation networks using scanner-wise cross-validation

		Overall	Aera	Avanto	Trio
		(n = 102)	(n = 34)	(n = 34)	(n = 34)
T1 only	IC-Net$_{T1}$	.974 ± .019	.973 ± .011	.976 ± .012	.971 ± .029
	CC-Net$_{T1}$	.902 ± .065	.908 ± .049	.902 ± .050	.896 ± .088
FLAIR only	IC-Net$_{FLAIR}$	.965 ± .028	.970 ± .014	.957 ± .043	.968 ± .013
	CC-Net$_{FLAIR}$	.828 ± .110	.812 ± .144	.846 ± .098	.827 ± .094
T1 using T1/ FLAIR networks	IC-Net$_{T1/+FLAIR}$	.974 ± .013	.975 ± .006	.973 ± .011	.973 ± .019
	CC-Net$_{T1/+FLAIR}$	.894 ± .062	.911 ± .028	.888 ± .067	.884 ± .078
FLAIR using T1/FLAIR networks	IC-Net$_{T1/FLAIR}$	.967 ± .019	.961 ± .029	.970 ± .010	.971 ± .010
	CC-Net$_{T1/FLAIR}$	.808 ± .149	.742 ± .196**	.830 ± .128**	.852 ± .079**

Note: For this scanner-wise cross-validation, at each fold the data from one scanner (Aera, Avanto or Trio) were used as validation set, whereas those from the remaining scanners were used as training set.

Abbreviations: CC, corpus callosum; FLAIR, fluid-attenuated inversion recovery; IC, intracranial; n, number of patients; T1, T1-weighted scan.

**P < .01, only the CC segmentation of FLAIR scans using the CC-Net$_{T1/+FLAIR}$ algorithm showed a significant difference between scanners, as tested by analysis of variance (ANOVA). All other segmentations did not vary significantly across scanners.
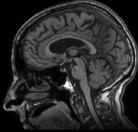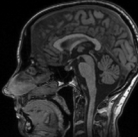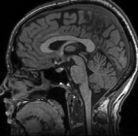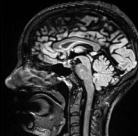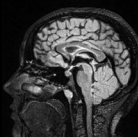
TABLE 5 Intraclass correlation coefficients of automatic and semi-automatic pipelines

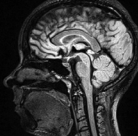|		Automatic midslice selection	Manual midslice selection
IC-Net$_{T1}$ + CC-Net$_{T1}$		ICC = .942	ICC = .883
		95% CI: .602-.988	95% CI: .679-.97
IC-Net$_{FLAIR}$ + CC-Net$_{FLAIR}$		ICC = .739	ICC = .828
		95% CI: .421-.925	95% CI: .501-.956
IC-Net$_{T1/FLAIR}$ + CC-Net$_{T1/FLAIR}$	T1 data	ICC = .908	ICC = .910
		95% CI: .617-.979	95% CI: .734-.977
	FLAIR data	ICC = .753	ICC = .633
		95% CI: .421-.931	95% CI: .235-.889

Note: Results presented do not significantly differ between algorithms, as tested by independent samples t-tests.

Abbreviations: CC, corpus callosum; CI, confidence interval; FLAIR, fluid-attenuated inversion recovery; IC, intracranial; ICC, intraclass correlation coefficient; T1, T1-weighted scan.

FIGURE 5 Segmentation output for each sequence and atrophy level. nCCA, normalized corpus callosum area

the FLAIR segmentation algorithms displayed significant correlations with both current and future physical and cognitive disabilities. The combined T1w and FLAIR segmentation pipeline for T1w scans was not significantly correlated with EDSS and SDMT at baseline, whereas the T1w-specific algorithm correlated with baseline EDSS. However, both the combined and T1w-specific pipelines were significantly associated with future physical and cognitive scores (Table 6).

DISCUSSION

We report a fully automated deep learning-based segmentation tool tailored for 3-dimensional T1w and T2w-FLAIR scans, including an automatic midslice selection, that outputs a normalized CC area, an established metric of neurodegeneration in MS. The automatic slice selection produced an output with over 98% extraction accuracy within 3 mm of the true midslice. This level of accuracy was also observed

when using a scanner-wise cross-validation approach, for which the networks' performance was evaluated on data from unseen scanners. The segmentation pipeline further segments the IC area with a mean Dice coefficient of .97-.98. The CC, a smaller and more difficult biomarker to segment, presented significantly lower mean Dice coefficients with FLAIR (.87-.90) as compared to T1w scans (.91). As hypothesized, the CC segmentation accuracy dropped in patients with greater atrophy. A higher level of accuracy and reliability for CC segmentations performed on unseen scanners was observed on T1w data, as opposed to FLAIR, as tested by scanner-wise cross-validation. The same approach showed high and consistent performance across scanners and contrasts for the IC segmentation. Furthermore, in our subcohort of scan-rescan T1w data, the fully automatic algorithm's precision performed well with a high ICC (above .90), indicating excellent reliability. On the other hand, FLAIR provided lower reproducibility with an ICC ranging between .74 and .75. FreeSurfer and the FLAIR

TABLE 6 Segmentation output correlation with FreeSurfer and clinical disability

Pipeline		FreeSurfer nCCV (n = 427)	EDSS ± 6 months (n = 252)	EDSS future[a] (n = 331)	SDMT ± 6 months (n = 172)	SDMT future[b] (n = 304)
FreeSurfer nCCV		N/A	$\rho = -.19^{**}$	$\rho = -.18^{**}$	$r = .18^{*}$	$r = .24^{**}$
FLAIR-specific		$r = .69^{**}$	$\rho = -.13^{*}$	$\rho = -.19^{**}$	$r = .18^{**}$	$r = .28^{**}$
T1w-specific		$r = .80^{**}$	$\rho = -.15^{*}$	$\rho = -.18^{**}$	$r = .12$	$r = .18^{**}$
Combined T1w and FLAIR	FLAIR	$r = .68^{**}$	$\rho = -.21^{**}$	$\rho = -.24^{**}$	$r = .25^{**}$	$r = .29^{**}$
	T1w	$r = .81^{**}$	$\rho = -.12$	$\rho = -.18^{**}$	$r = .12$	$r = .21^{**}$

Abbreviations: nCCV, normalized corpus callosum volume; EDSS, Expanded Disability Status Scale; n, number of patients; r, Pearson's correlation coefficient; SDMT, Symbol Digit Modalities Test; T1w, T1-weighted scan; ρ, Spearman's rank correlation coefficient.
*P < .05; **P < .01.
[a]EDSS follow-up time was 6.7 ± 2.6 years.
[b]SDMT follow-up time was 5.8 ± 2.8 years.

algorithm produced similar clinical correlations, tending to be better than the T1w algorithm that did not significantly correlate with baseline SDMT scores.

Compared to a previous study by our group,[14] the main technical novelties of the present work are the introduction of an automatic midslice selection, the possibility of running multiple MRI sequences, and an adaptation for modern 3-dimensional volumetric sequences. The implemented CNNs can receive, as input, raw 3-dimensional T1w and/or FLAIR sequences, and subsequently select the midsagittal slice of interest for CC segmentation, with high accuracy. In a clinical or research setting, the use of this algorithm would, therefore, greatly limit the manual interrater disagreement that otherwise may exist when selecting a slice. In all three midslice algorithms, a very high accuracy was achieved, with an overall mean absolute prediction error ranging between 0.64 and 1.16 mm. The FLAIR midslice selection tended to be better than the T1w, most likely secondary to having a higher resolution (1 mm isotropic vs. $1 \times 1 \times 1.5$ mm). An excellent performance was observed when testing the algorithm on data from scanners that were not used for training. In some cases, this method even outperformed the original 10-fold cross-validation approach, suggesting that the addition of appropriate data augmentation may help generalize the network performance to unseen data. This is a promising finding that suggests that our slice selection method could potentially be applied in other cohorts. Moreover, given the potential suggested by the present results, in the future we aim to further diversify the type of input that can be fed into this type of CNN by training it using different variations of T1w, T2w, and proton density-weighted sequences.

FreeSurfer is one of the most frequently applied brain segmentation software and it provides a cross-sectional stream as well as a longitudinal stream to improve segmentation results.[9] Although FreeSurfer CC segmentation is precise, with an average coefficient of variation ranging from 1% to 4%, the Dice accuracy range has been observed between .79 and .84 for the cross-sectional and longitudinal streams, respectively.[27] Our Dice coefficient varied depending on the patient's atrophy level (.85-.94) but was overall at least on par, or better, with similar studies applying CNNs to segment the CC, such as Platten et al. with a Dice of .89,[14] and

Maruyama et al. with a Dice of .79 (Jaccard index of .652).[13] Overall, the segmentations performed on T1w data were shown to be significantly more accurate than those performed on FLAIR data (independently from the type of segmentation network used). This difference is likely secondary to unsuppressed cerebrospinal fluid directly inferior to the CC, making its delineation inherently more difficult. This finding was also reflected in the results of the scanner-wise cross-validation: FLAIR data showed numerically lower CC segmentation performance, not only compared to T1w scans, but also compared to the results obtained on the same FLAIR data using a 10-fold cross-validation strategy (overall mean Dice of .828 vs. .881 for CC-Net$_{FLAIR}$, and .808 vs. .865 for CC-Net$_{T1/FLAIR}$). This suggests that although our proposed network can generalize very well on new unseen T1w data, it still remains challenging to obtain an equally high level of accuracy on FLAIR scans from unseen cohorts. This result was expected, considering the following two aspects jointly: (1) CC segmentation on FLAIR scans was already found to be the most challenging task of the present work; (2) the networks were trained not only on cohorts that differ from those used for validation but also using a lower amount of data (2/3 of the dataset instead of 9/10). In the future, this issue may be addressed by applying further changes to the architecture; for example, adding CC shape prior information as an additional input to the network, an approach that was found to improve the segmentation accuracy in previous MRI segmentation studies.[28,29]

Moreover, discrepancies were found when analyzing the performance across different atrophy levels, as patients with more atrophy featured lower segmentation accuracy. This result is intuitive because a high heterogeneity in the CC atrophy can make it particularly challenging for the networks to segment properly. It should, however, also be noted that there was not a dramatic loss in performance in patients with high atrophy. In future studies, we plan to expand our dataset to include more patients, while maintaining a homogenous distribution of the atrophy levels in order to improve segmentation performance in cases with high atrophy.

In our reproducibility analysis, we compared the normalized CC area that was obtained from segmentations performed on manually

selected slices against automatically selected slices. Our results indicate that automatic midslice selection algorithm may constitute a valid and quick alternative to reduce the work performed by radiologists, who often must manually extract and annotate datasets of hundreds (or thousands) of MRI scans in large studies. An excellent level of reproducibility (ICC > .90) was observed on T1w data by both pipelines. Moderate to good agreement was observed on the measures obtained from FLAIR data (ICC between .739 and .753).

The CC is a promising marker for several neurodegenerative diseases such as Alzheimer's disease,[30] Parkinson's disease,[31] amyotrophic lateral sclerosis,[32] and particularly MS.[33] Its significance in MS is intimately tied to the CC being a large white matter structure in a disease that predominantly affects myelin. Several studies have shown that the extent of damage to the CC significantly correlates to both physical and cognitive disabilities,[7,27,34] and that the CC is a predictor of cognitive disability 8.5 years later.[5] Our observations corroborate this notion, showing significant correlations with both EDSS and SDMT at both baseline and follow-up on average 6-7 years later. Of note, our data consist of real-world clinical data, which inherently introduces variability in MRI and clinical parameter acquisition, such as different MRI technicians and raters. This may partly explain the low, albeit significant, correlation coefficients. Of interest is that the correlations tended to be numerically higher for the FLAIR pipelines. This could be a result of clinically more disabled patients also having more atrophied CC, which the FLAIR algorithm under-segments further (relative to the T1w algorithm), leading to an artificially inflated association between their clinical score and the normalized CC area.

A main and important limitation of the algorithms is the fluctuation that may be introduced through alignment and subject placement. Our MRI scans are aligned along the anterior-posterior commissures, providing a relatively consistent CC angle. Although not tested, introducing an angle to this alignment would likely affect the output. Future improvements to this pipeline would include a registration step to minimize the effect of acquisition or reconstruction angles. Similarly, an axially or coronally acquired sequence would have to be reconstructed to sagittal before introduction into the algorithm. Although we systematically introduce three different scanner types, they are all from the same manufacturer. It is, thus, unknown how the algorithm will fare with scans from different manufacturers. However, our results promisingly indicate that there may be minimal effect between scanners and MRI strengths. Likewise, it is clear from our study that the degree of atrophy affects the segmentation accuracy, which in turn may affect the association between the normalized CC area and the corresponding clinical scores. To our knowledge, this is the first study to examine this relationship between atrophy and segmentation accuracy in MS, but a similar bias may be present in other segmentation algorithms. It is reasonable to think that our algorithms are in fact relatively robust to the atrophy levels compared to other algorithms, as they were specifically trained to handle a heterogeneous MS population. Another limitation is that our algorithms are only trained to handle T1w and FLAIR scans in MS. There is also an inherent discrepancy in comparing a volume (FreeSurfer) to an area (our algorithms); however, both of these methods are clinically relevant.

In conclusion, we present a quick, fully automatic, and accurate deep learning-based segmentation tool that may be used in monitoring neurodegeneration in MS. Due to its computational efficiency, it may be feasible to implement in large MS studies. It also performs similarly on both 1.5 and 3.0 T. Future directions may include applying the algorithm prospectively to evaluate the neuroprotective effect of MS therapies.

ORCID

Michael Platten https://orcid.org/0000-0001-6297-487X
Russell Ouellette https://orcid.org/0000-0001-9217-1445
Tobias Granberg https://orcid.org/0000-0001-6700-1022

REFERENCES

1. Filippi M, Bar-Or A, Piehl F, et al. Multiple sclerosis. Nat Rev Dis Primers 2018;4:43.
2. Friese MA, Schattling B, Fugger L. Mechanisms of neurodegeneration and axonal dysfunction in multiple sclerosis. Nat Rev Neurol 2014;10:225-38.
3. Rocca MA, Comi G, Filippi M. The role of T1-weighted derived measures of neurodegeneration for assessing disability progression in multiple sclerosis. Front Neurol 2017;8:433.
4. Georgy BA, Hesselink JR, Jernigan TL. MR imaging of the corpus callosum. AJR Am J Roentgenol 1993;160:949-55.
5. Ouellette R, Bergendal Å, Shams S, et al. Lesion accumulation is predictive of long-term cognitive decline in multiple sclerosis. Mult Scler Relat Disord 2018;21:110-6.
6. Klawiter EC, Ceccarelli A, Arora A, et al. Corpus callosum atrophy correlates with gray matter atrophy in patients with multiple sclerosis. J Neuroimaging 2015;25:62-7.
7. Granberg T, Bergendal G, Shams S, et al. MRI-defined corpus callosal atrophy in multiple sclerosis: a comparison of volumetric measurements, corpus callosum area and index. J Neuroimaging 2015;25:996-1001.
8. Reuter M, Rosas HD, Fischl B. Highly accurate inverse consistent registration: a robust approach. Neuroimage 2010;53:1181-96.
9. Reuter M, Schmansky NJ, Rosas HD, et al. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 2012;61:1402-18.
10. Shi S, Wang Q, Xu P, et al. Benchmarking state-of-the-art deep learning software tools. arXiv. http://arxiv.org/abs/1608.07249.Accessed August 11, 2021.
11. Anwar SM, Majid M, Qayyum A, et al. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018;42:226.
12. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv. https://arxiv.org/abs/1505.04597.Accessed December 20, 2021.
13. Maruyama T, Hayashi N, Sato Y, et al. Simultaneous brain structure segmentation in magnetic resonance images using deep convolutional neural networks. Radiol Phys Technol. 14:358-365.

14. Platten M, Brusini I, Andersson O, et al .Deep learning corpus callosum segmentation as a neurodegenerative marker in multiple sclerosis. J Neuroimaging 2021;31:493-500.

15. Filippi M, Rocca MA, Ciccarelli O, et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. Lancet Neurol 2016;15:292-303.

16. Wattjes MP, Ciccarelli O, Reich DS, et al. 2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. Lancet Neurol 2021;20:653-70.

17. Kavaliunas A, Manouchehrinia A, Stawiarz L, et al. Importance of early treatment initiation in the clinical course of multiple sclerosis. Mult Scler 2017;23:1233-40.

18. Alping P, Piehl F, Langer-Gould A,et al. Validation of the Swedish Multiple Sclerosis Register: further improving a resource for pharmacoepidemiologic evaluations. Epidemiology 2019;30:230-3.

19. Mcdonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 2001;50: 121-7.

20. Polman CH, Reingold SC, Edan G, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the 'McDonald Criteria'. Ann Neurol 2005;58:840-6.

21. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald Criteria. Ann Neurol 2011;69:292-302.

22. Lezak MD, Howieson DB, Bigler ED, et al. Neuropsychological assessment. Oxford: Oxford University Press; 2012.

23. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006;31:1116-28.

24. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. Presented at the 27th International Conference on Machine Learning; June 21–24, 2010; Haifa.

25. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. http://arxiv.org/abs/1412.6980.Accessed December 7, 2020

26. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60.

27. Platten M, Martola J, Fink K, et al. MRI-based manual versus automated corpus callosum volumetric measurements in multiple sclerosis. J Neuroimaging. 30:198-204.

28. Wang C, Smedby Ö. Automatic whole heart segmentation using deep learning and shape context. In: Pop M, Sermesant M, Jodoin P-M, et al., editors. Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Cham: Springer International Publishing; 2018. p. 242-9.

29. Brusini I, Lindberg O, Muehlboeck J-S, et al. Shape information improves the cross-cohort performance of deep learning-based segmentation of the hippocampus. Front Neurosci 2020;14:15.

30. Hampel H, Teipel SJ, Alexander GE, et al. Corpus callosum atrophy is a possible indicator of region- and cell type-specific neuronal degeneration in Alzheimer disease: a magnetic resonance imaging analysis. Arch Neurol 1998;55:193-8.

31. Goldman JG, Bledsoe IO, Merkitch D, et al. Corpus callosal atrophy and associations with cognitive impairment in Parkinson disease. Neurology 2017;88:1265-72.

32. Spinelli EG, Riva N, Rancoita PMV, et al. Structural MRI outcomes and predictors of disease progression in amyotrophic lateral sclerosis. Neuroimage Clin 2020;27:102315.

33. Gonçalves LI, Dos Passos GR, Conzatti LP, et al. Correlation between the corpus callosum index and brain atrophy, lesion load, and cognitive dysfunction in multiple sclerosis. Mult Scler Relat Disord 2018;20:154-8.

34. Sugijono SE, Mulyadi R, Firdausia S, et al. Corpus callosum index correlates with brain volumetry and disability in multiple sclerosis patients. Neurosciences 2020;25:193-9.