Human Brain Mapping

WILEY

TECHNICAL REPORT OPEN ACCESS

Extended Technical and Clinical Validation of Deep Learning-Based Brainstem Segmentation for Application in Neurodegenerative Diseases

Benno Gesierich[1] | Laura Sander[2,3] | Lukas Pirpamer[1] | Dominik S. Meier[1] | Esther Ruberte[1,3] | Michael Amann[1] | Tim Sinnecker[1,2] | Antal Huck[4] | Frank-Erik de Leeuw[5] | Pauline Maillard[6] | Sue Moy[7] | Karl G. Helmer[7,8] | MarkVCID Consortium | Johannes Levin[9,10,11] | Günter U. Höglinger[9,10,11] | PROMESA Study Group | Michael Kühne[12] | Leo H. Bonati[2,13] | Jens Kuhle[2,14] | Philippe Cattin[4] | Cristina Granziera[2] | Regina Schlaeger[2,3] | Marco Duering[1,3,15]

[1]Medical Image Analysis Center (MIAC), Basel, Switzerland | [2]Neurologic Clinic and Policlinic, Departments of Neurology and Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland | [3]Translational Imaging in Neurology (ThINk) Basel, Department of Biomedical Engineering, University of Basel, Basel, Switzerland | [4]Center for Medical Image Analysis & Navigation (CIAN), Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland | [5]Radboud University Medical Center, Donders Institute for Brain, Cognition and Behaviour, Center for Neuroscience, Department of Neurology and Radboud University, Donders Institute for Brain, Cognition and Behaviour, Center for Cognitive Neuroimaging, Nijmegen, the Netherlands | [6]Department of Neurology, University of California Davis, Davis, California, USA | [7]Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts, USA | [8]Harvard Medical School, Boston, Massachusetts, USA | [9]Department of Neurology, LMU University Hospital, Ludwig-Maximilians-Universität (LMU) München, Munich, Germany | [10]German Center for Neurodegenerative Diseases (DZNE), Munich, Germany | [11]Munich Cluster for Systems Neurology (SyNergy), Munich, Germany | [12]University Heart Center, University Hospital Basel, Basel, Switzerland | [13]Research Department, Reha Rheinfelden, Rheinfelden, Switzerland | [14]Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University of Basel, Basel, Switzerland | [15]Institute for Stroke and Dementia Research (ISD), LMU University Hospital, LMU Munich, Germany

Correspondence: Marco Duering (marco.duering@miac.ch)

Received: 16 July 2024 | **Revised:** 13 December 2024 | **Accepted:** 6 January 2025

Funding: This work was supported by the National Institute of Neurological Disorders and Stroke/National Institute on Aging (U24NS100591).

Keywords: atrophy | brainstem | deep learning | neurodegeneration | segmentation

ABSTRACT

Disorders of the central nervous system, including neurodegenerative diseases, frequently affect the brainstem and can present with focal atrophy. This study aimed to (1) optimize deep learning-based brainstem segmentation for a wide range of pathologies and T1-weighted image acquisition parameters, (2) conduct a systematic technical and clinical validation, (3) improve segmentation quality in the presence of brainstem lesions, and (4) make an optimized brainstem segmentation tool available for public use. An intentionally heterogeneous ground truth dataset ($n=257$) was employed in the training of deep learning models based on multi-dimensional gated recurrent units (MD-GRU) or the nnU-Net method. Segmentation performance was evaluated against ground truth labels. FreeSurfer was used for benchmarking in subsequent validation. Technical validation, including scan-rescan repeatability ($n=46$) and inter-scanner reproducibility ($n=20$, 3 different scanners) in unseen data, was conducted in patients with cerebral small vessel disease. Clinical validation in unseen data was performed in 1-year follow-up data of 16 patients with multiple system atrophy, evaluating the annual percentage volume change. Two lesion filling algorithms were investigated to improve segmentation performance in 23 patients with multiple sclerosis. The MD-GRU and nnU-Net models demonstrated very good segmentation performance (median Dice coefficients ≥ 0.95 each) and outperformed a previously published model trained

Benno Gesierich, Laura Sander, Regina Schlaeger and Marco Duering have contributed equally to this work.

on a narrower dataset. Scan–rescan repeatability and inter-scanner reproducibility yielded similar Bland–Altman derived limits of agreement for longitudinal FreeSurfer (total brainstem volume repeatability/reproducibility 0.68/1.85), MD-GRU (0.72/1.46), and nnU-Net (0.48/1.52). All methods showed comparable performance in the detection of atrophy in the total brainstem (atrophy detected in 100% of patients) and its substructures. In patients with multiple sclerosis, lesion filling further improved the accuracy of brainstem segmentation. We enhanced and systematically validated two fully automated deep learning brainstem segmentation methods and released them publicly. This enables a broader evaluation of brainstem volume as a candidate biomarker for neurodegeneration.

Summary

- We optimized deep learning-based brainstem segmentation approaches based on the MD-GRU and U-Net network types.
- Systematic technical and clinical validation was performed in relevant target populations.
- Our results emphasize the importance of training segmentation models on diverse datasets containing different imaging sequences and pathologies.

1 | Introduction

The brainstem is the anatomical and physiological link between the brain and the spinal cord and regulates vitally important physiological processes. In a cranio-caudal order, it consists of the three substructures, mesencephalon, pons, and medulla oblongata (Figure 1a).

Atrophy is a hallmark of neurodegeneration and can be assessed non-invasively by MRI in vivo (Duering et al. 2023). Several neurodegenerative disorders, such as multiple system atrophy (MSA), primarily affect the brainstem. MSA is an adult-onset, rapidly progressive, fatal-ending neurodegenerative disease presenting clinically with autonomic failure and Parkinsonian or cerebellar features (Wenning et al. 2022) and pathologically with glial cytoplasmic inclusions and neuronal loss predominantly in striatonigral and olivopontocerebellar systems (Fanciulli and Wenning 2015). Previous volumetric MRI studies reported substantial atrophy in the mesencephalon and pons in patients with different MSA subtypes (Krismer et al. 2024) using the segmentation software FreeSurfer. Other structural pathologies of different etiologies, such as demyelinating or ischemic lesions, can result in secondary neurodegeneration (Duering et al. 2015). While MR imaging of the cerebral hemispheres and the spinal cord has been extensively studied (Casserly et al. 2018; Rocca et al. 2017), the brainstem has been less well investigated. However, brainstem imaging can offer a potential diagnostic or therapeutic added value in these diseases, such as the detection of atrophy in early stages of multiple sclerosis (MS; Eshaghi et al. 2018).

The brainstem is a small structure, anatomically less well demarcated compared to other brain regions. MRI and automatic segmentation of the brainstem is demanding because of the caudal position and artifacts caused by motion, pulsatile blood, and cerebrospinal fluid flow (Brooks et al. 2013; Herlihy et al. 2001). Segmentation is further impeded in the presence of lesions, which can appear hypointense on T1-weighted imaging. A comparison of the most frequently used atlas-based segmentation approaches, i.e., FreeSurfer (Fischl et al. 2002, 2004; Iglesias et al. 2015), PSTAPLE (Akhondi-Asl and Warfield 2013) and FSL-FIRST (Patenaude et al. 2011), found the highest reproducibility of brainstem segmentations performed by FreeSurfer (Velasco-Annis et al. 2018).

In a previous study, we developed a fully automated, deep learning-based brainstem segmentation method based on multidimensional gated recurrent units (MD-GRU) (Andermatt, Pezold, and Cattin 2016, 2018). The method provided reliable and robust brainstem segmentation in patients with Alzheimer's disease and MS, with more accurate segmentations than FreeSurfer in comparison to manual ground truth labels (Sander et al. 2019). However, the method's broader use has been constrained by a relatively small and homogenous training dataset and limited technical and clinical validation. The U-Net architecture (Ronneberger, Fischer, and Brox 2015) is now widely recognized for its effectiveness in medical image segmentation. Recently, nnU-Net has been introduced (Isensee et al. 2021) as a method to automatically self-configure preprocessing, U-Net architecture setup, training, and post-processing. In multiple segmentation challenges, nnU-Net surpassed specialized segmentation algorithms, establishing itself as an ideal tool for benchmarking.

The goal of this study was to further enhance deep learning-based brainstem segmentation and to facilitate a broader use of the trained models, also by making them publicly available. The first aim was to adapt the methods for diverse image datasets and diseases. The second aim was to conduct a systematic technical and clinical validation, adhering to consensus and state-of-the-art recommendations (FDA-NIH Biomarker Working Group 2016; Smith et al. 2019). The third aim was to improve segmentation performance in the presence of brainstem lesions.

2 | Methods

2.1 | Training of Brainstem Segmentation Models

2.1.1 | Ground Truth Dataset

Details on the ground truth dataset are provided in Table S1. It comprised the “original” dataset used for training of the previously published MD-GRU algorithm (Sander et al. 2019) and an “extension” dataset. The goal of the extension was to cover a wider range of MR acquisitions (1.5 T and 3T, different 3D T1-weighted sequences, including MP2RAGE) and pathologies (MS, age-related brain diseases, including cerebral small vessel disease, and healthy controls). Ground truth brainstem masks were created semi-automatically as previously described (Sander et al. 2019). In short, a first segmentation was created with the brainstem module of the FreeSurfer package (version 7.2.0; Iglesias et al. 2015) and then corrected manually by a neurologist expert rater (LS). The difference between ground truth masks and uncorrected FreeSurfer brainstem segmentation masks was smallest for the pons and largest for the medulla oblongata (Table S2). The ground truth dataset ($n = 257$) was split into subsamples for training ($n = 183$), validation ($n = 35$), and testing ($n = 39$).

2.1.2 | Training of MD-GRU Models

MD-GRU (https://github.com/zubata88/mdgru; commit 791ee4f) was first trained in the training subsample with default parameters (cross-entropy loss, no data augmentation, no filtering), a patch size of 100x100x100 voxels, padding of 20x20x20 voxels, and native T1 images as input. Seven additional training experiments were run by changing parameters from their defaults and/or using preprocessed T1-weighted images. Non-default parameters included the use of Dice loss in combination with cross-entropy loss, the use of data augmentation (deformation, rotation and scaling), and activating high pass filtering of the input images. Added preprocessing steps included resampling T1-weighted images into isotropic 1 mm^3 voxel grids and cropping of isotropic images around the brainstem (as defined by ground truth and leaving a border of at least 20 voxels at each side of the brainstem), resulting in a consistent field of view for all subjects of 94x84x124 voxels, which was also used as patch size for the training. This last training experiment was also done with a combination of Dice loss and cross-entropy loss. In total, 8 training experiments were conducted (Table S3).

During training, the current checkpoint was validated every 5.000 iterations in the "validation" subsample, running inference, and calculating the Dice similarity coefficients (referred to as Dice coefficients hereafter) against ground truth labels. Training experiments were run for at least 100.000 iterations or until the Dice coefficients reached a stable plateau in the "validation" subsample. After completion of the training experiments, the checkpoint with the best performance in the "validation" subsample was selected for each experiment and this performance (in terms of Dice coefficients in the "validation" subsample) was compared across training experiments, to select the best model. The MD-GRU model selected in this way was the one trained with default parameters. In the following, we refer to this newly trained model as "MD-GRU 2024" and the previously published MD-GRU model as "MD-GRU 2019."

2.1.3 | Training of nnU-Net Models

nnU-Net (https://github.com/MIC-DKFZ/nnUNet.git; commit 96d44c2) was trained with the default parameters using the "3d_fullres" configuration. We trained 3 models (Table S3) using different amounts of graphics processing unit memory: 8 (default), 11, and 24 GB. The automated experiment planning carried out by nnU-Net resulted in changes to the U-Net used during training (e.g., larger patch sizes with increasing memory). Patch sizes for the three models were 160x128x112 (with 8 GB), 160x160x128 (with 11 GB) and 192x192x160 voxels (with 24 GB). By default,

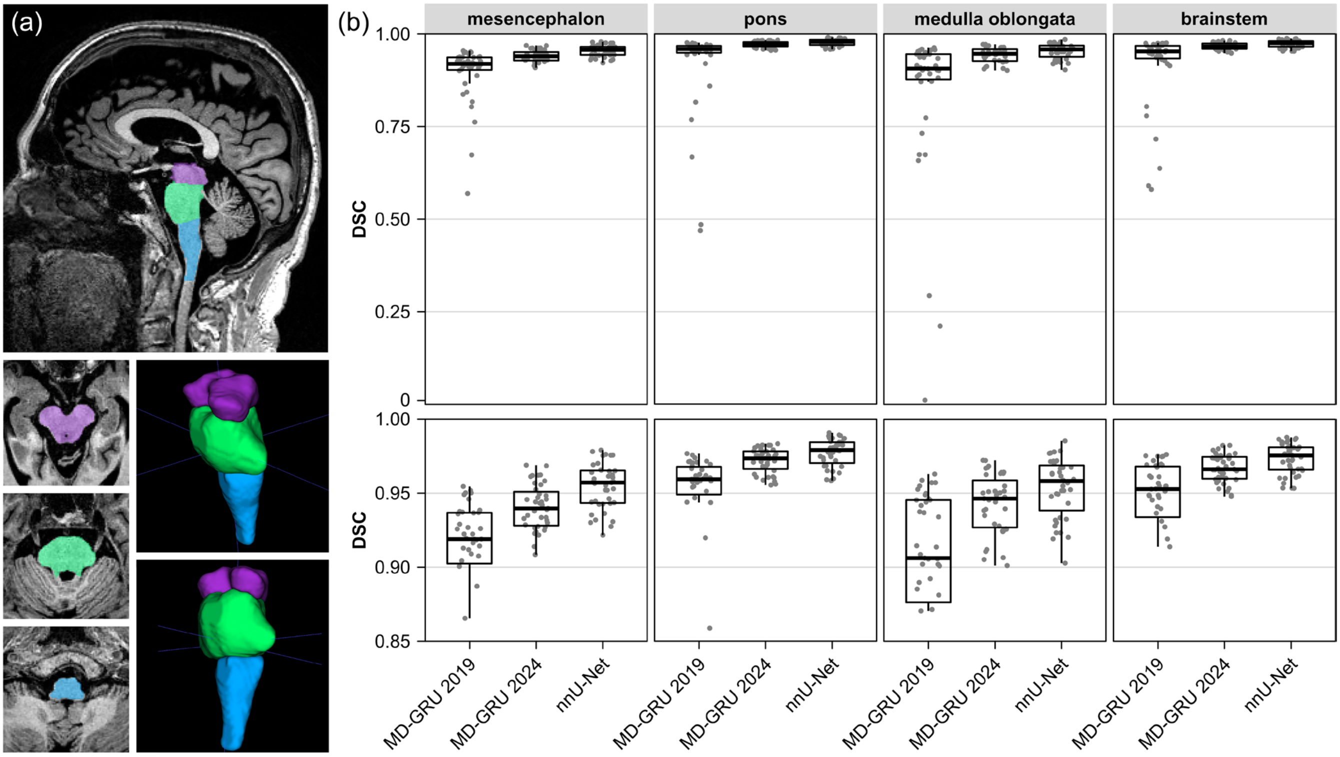

FIGURE 1 | Brainstem segmentation performance. (a) The target structures mesencephalon (purple), pons (green) and medulla oblongata (blue) are shown for a patient with cerebral small vessel disease. (b) Dice similarity coefficient (DSC) for the overlap of inferred with ground truth masks in the three subregions and the total brainstem. Bottom panels zoom into the range above 0.85.

the nnU-Net training terminates automatically after 1.000 epochs and uses 5-fold cross-validation to assess performance of the final checkpoint. Given the use of cross-validation, a validation set was not required, and training was performed in the merged training and validation subsamples. The model trained with 24 GB memory had the best performance in cross-validation and was thus selected for subsequent analysis. Using default settings, nnU-Net inference was performed as ensemble prediction using all 5 models.

2.1.4 | Benchmarking

"MD-GRU 2024" was benchmarked against "MD-GRU 2019" and the newly trained nnU-Net model. We used the test subsample of the ground truth dataset to calculate Dice coefficients between inferred and ground truth labels.

2.2 | Technical Validation

For technical validation of brainstem (subregion) volume as an imaging biomarker, we estimated the bias against a reference method and the precision under repeatability and reproducibility conditions.

2.2.1 | Estimation of Bias Against a Reference Method

Using the test subsample of the ground truth dataset, we estimated the constant and proportional bias of volumes derived from the MD-GRU 2024 and nnU-Net segmentations against reference volumes from the ground truth segmentations.

2.2.2 | Scan-Rescan Repeatability and Inter-Scanner Reproducibility

Scan–rescan repeatability and inter-scanner reproducibility were compared across algorithms in unseen data from a different source than the training data, i.e., two subsamples of the MarkVCID study (Lu et al. 2021; Maillard et al. 2022; Wilcock et al. 2021), calculating Bland–Altman statistics (Deepankar Datta and Love 2018) for the comparison of brainstem volumes between either scan and rescan session or between different scanners. Details on the scanning protocols used in MarkVCID were previously described (Lu et al. 2021). For brainstem segmentation, only the T1-weighted sequence was used. For comparison, we added segmentations from FreeSurfer's (version 7.2.0) cross-sectional (FS cross) and longitudinal pipeline (FS long) using the brainstem module.

The first MarkVCID subsample, used for scan-rescan repeatability, comprised 46 cerebral small vessel disease patients, each with two scans of the same protocol acquired within 14 days on the same scanner. Three patients were excluded because either the FreeSurfer or MD-GRU pipeline failed.

The second MarkVCID subsample, used for inter-scanner reproducibility, comprised 20 patients with scans acquired on 4 different scanners: Philips Achieva, Siemens Trio, Siemens Prisma, GE 750 W. Because the brainstem was substantially cropped caudally in all GE scans, data from this scanner were excluded. Furthermore, the Philips Achieva scan was missing for one patient and the Siemens Prisma scan was excluded for another patient due to insufficient scan quality.

2.3 | Clinical Validation

Clinical validation was conducted using unseen data from a different source than the training data, i.e. from the PROMESA clinical trial (Levin et al. 2019). The trial included patients older than 30 years diagnosed with MSA. The subset from the MRI substudy comprised 21 patients with T1-weighted imaging at baseline and at 1-year follow-up. Details regarding the scanning protocol were published previously (Levin et al. 2019). Due to scan quality issues, 5 patients had to be excluded.

The measure of interest for the clinical validation is the ability to detect pathological atrophy in MSA. To this end, we calculated the annual percentage volume change (PVC) of the brainstem and its subregions for all algorithms. To determine a threshold for pathological atrophy, we referred to a study on healthy aging, which reports two mean PVCs for the brainstem in healthy elderly subjects, −0.31% and −0.43% per annum (Fjell et al. 2009). Thus, we assessed the percentage of patients exceeding the mean of these two PVC values, being −0.37% per annum.

2.4 | Lesion Filling

Although the training dataset included MS patients with brainstem lesions, the trained models might still have problems dealing with the altered tissue signal of larger brainstem lesions. We assessed the severity of this issue and validated lesion filling as a mitigation strategy in scans of 23 MS patients with brainstem lesions from the ongoing Swiss MS cohort (SMSC) (Disanto et al. 2016). The ground truth mask and in addition a T1-hypointense lesion mask was created by expert readers (LS and ER).

Lesion filling was conducted with two different algorithms, from the FMRIB Software Library (FSL, version 6.0.4; function "lesion_filling") and from Advanced Normalization Tools (ANTs, version 2.4.3; function "LesionFilling") (Avants et al. 2011). Both algorithms used the T1-weighted image and lesion mask as input. The FSL algorithm requires additionally a white matter mask, which was created from the FreeSurfer segmentations (aparc+aseg). The manually created ground truth brainstem mask was merged into the white matter mask.

The brainstem was segmented on T1-weighted images with the MD-GRU 2024 and nnU-Net models before and after filling lesions with FSL or ANTs, respectively.

Segmentation quality was then assessed by the Dice coefficients for the overlap of the inferred with the ground truth brainstem masks and by the sensitivity to lesioned areas (i.e., lesion coverage by the inferred mask). Finally, the different segmentation procedures were ranked by significance scores. These scores were calculated by comparing each metric (i.e., Dice coefficients and lesion coverage) pairwise across the six segmentation procedures and counting for each procedure the number of comparisons in which it performed better (i.e., higher Dice coefficients or lesion coverage). Wilcoxon signed rank tests were used for comparisons, with an uncorrected significance level of alpha = 0.05.

2.5 | Method Availability

The trained models are available for download at Zenodo for MD-GRU https://doi.org/10.5281/zenodo.12578294 and nnU-Net https://zenodo.org/records/13323293. Ready to use container images for both algorithms, the code for building, and instructions for using the images are available at GitHub: https://github.com/miac-research/dl-brainstem.

3 | Results

3.1 | Benchmarking of Brainstem Segmentation Models

While the newly trained MD-GRU 2024 model improved segmentation performance over the previously published MD-GRU 2019 model as measured by Dice coefficients against ground truth in the test subsample, the nnU-Net model showed the best performance (Table 1, Figure 1b). Findings were consistent over the three subregions and the total brainstem. For MD-GRU 2019, there were several outliers with very low Dice coefficients in the extension ground truth dataset (Table 1, Figure S1). Consequently, while Dice coefficients were generally slightly lower in the ground truth extension dataset compared to the original ground truth dataset, this difference was most obvious for the MD-GRU 2019 model, highlighting poorer segmentation performance on image acquisition protocols and diseases not included in the previous ground truth dataset. Direct comparisons between MD-GRU 2019 and MD-GRU 2024 in unseen data from different sources than the training data further highlight the improved generalizability of the newly trained MD-GRU 2024 (Figures S2 and S3). Thus, in the analyses below, we only report results from MD-GRU 2024.

3.2 | Technical Validation: Estimation of Bias Against a Reference Method

Neither MD-GRU nor nnU-Net showed a significant constant bias compared with reference volumes. A proportional bias with slightly negative slope (regression slope −0.12 mL/mL, uncorrected p value = 0.032) was found only for nnU-Net in the medulla oblongata (Table S4, Figure S4).

3.3 | Technical Validation: Scan-Rescan Repeatability

Brainstem volumes obtained by three algorithms (FreeSurfer [cross-sectional "FS cross" and longitudinal "FS long" pipeline], MD-GRU 2024, and nnU-Net) were compared across the repeated scans using Bland–Altman statistics (Table 2 and Figure 2a). No systematic bias was observed. Overlapping confidence intervals for the limits of agreement (LOA) indicate mostly similar repeatability for the total brainstem volume, but LOA were smaller for nnU-Net compared to "FS cross." Looking at the subregions, LOA were smaller for MD-GRU and nnU-Net than both FreeSurfer pipelines in the mesencephalon, smaller for nnU-Net than "FS cross" in the pons, and smaller for nnU-Net than MD-GRU and both FreeSurfer pipelines in the medulla oblongata.

3.4 | Technical Validation: Inter-Scanner Reproducibility

Total brainstem volumes were compared between three scanners using Bland–Altman statistics (Table 3 and Figure 2b).

TABLE 1 | Segmentation performance of the trained models in the test subsample. Dice similarity coefficients (median [IQR]), for the overlap of inferred with ground truth masks in the three subregions and in the total brainstem. Scores were calculated separately for cases from the original ground truth dataset and the newly added extension ground truth dataset. Highest median score in each row in bold.

Region	Ground truth	MD-GRU 2019	MD-GRU 2024	nnU-Net
Mesencephalon	Original	0.93 [0.91,0.94]	0.95 [0.93,0.96]	**0.96** [0.95,0.97]
	Extension	0.91 [0.78,0.92]	0.93 [0.92,0.93]	**0.94** [0.93,0.94]
	Total	0.92 [0.90,0.94]	0.94 [0.93,0.95]	**0.96** [0.94,0.97]
Pons	Original	0.96 [0.96,0.97]	**0.98** [0.97,0.98]	**0.98** [0.98,0.99]
	Extension	0.94 [0.72,0.96]	0.96 [0.96,0.96]	**0.97** [0.96,0.97]
	Total	0.96 [0.95,0.97]	0.97 [0.97,0.98]	**0.98** [0.97,0.98]
Medulla oblongata	Original	0.93 [0.90,0.95]	0.95 [0.95,0.96]	**0.96** [0.95,0.97]
	Extension	0.77 [0.48,0.90]	**0.93** [0.91,0.93]	**0.93** [0.92,0.95]
	Total	0.91 [0.88,0.95]	0.95 [0.93,0.96]	**0.96** [0.94,0.97]
Brainstem	Original	0.96 [0.95,0.97]	0.97 [0.97,0.98]	**0.98** [0.97,0.98]
	Extension	0.93 [0.68,0.95]	**0.96** [0.95,0.96]	**0.96** [0.96,0.96]
	Total	0.95 [0.93,0.97]	0.97 [0.96,0.97]	**0.98** [0.97,0.98]

TABLE 2 | Scan-rescan repeatability. Bias and (one-sided) limit of agreement (LOA) of volumes (in ml) from the Bland–Altman analysis (95% confidence intervals in brackets).

Region	Algorithm	Constant bias	LOA (one-sided)
Mesencephalon	FS cross	0.01 [−0.07,0.09]	0.51 [0.37,0.65]
	FS long	0.03 [−0.02,0.09]	0.34 [0.25,0.43]
	MD-GRU	0.01 [−0.01,0.04]	0.15 [0.11,0.19]
	nnU-Net	0.01 [−0.01,0.03]	0.13 [0.10,0.16]
Pons	FS cross	0.06 [−0.14,0.02]	0.48 [0.35,0.61]
	FS long	0.02 [−0.03,0.06]	0.30 [0.22,0.38]
	MD-GRU	0.02 [−0.04,0.09]	0.41 [0.30,0.52]
	nnU-Net	0.01 [−0.03,0.05]	0.26 [0.19,0.34]
Medulla oblongata	FS cross	−0.07 [−0.14,0.00]	0.47 [0.35,0.60]
	FS long	−0.04 [−0.09,0.02]	0.34 [0.25,0.43]
	MD-GRU	−0.01 [−0.07,0.05]	0.37 [0.27,0.47]
	nnU-Net	−0.01 [−0.03,0.02]	0.17 [0.12,0.21]
Brainstem	FS cross	−0.12 [−0.29,0.05]	1.08 [0.79,1.37]
	FS long	0.01 [−0.09,0.12]	0.68 [0.50,0.86]
	MD-GRU	0.03 [−0.08,0.14]	0.72 [0.53,0.92]
	nnU-Net	0.01 [−0.06,0.09]	0.48 [0.35,0.61]

Volumes segmented from Philips Achieva scans were in most cases larger than from the Siemens (Trio and Prisma) scans. In terms of reproducibility, differences between algorithms were generally small or absent. A significant proportional bias was not found in any comparison.

The Bland–Altman LOA tended to be larger between Siemens Prisma and the two other scanners. There was no clear difference between algorithms. Bland–Altman statistics for brainstem subregions are shown in Table S5.

3.5 | Clinical Validation

Our requirement was to detect pathological brainstem atrophy in 16 MSA patients from the PROMESA MRI substudy over a 1-year follow-up, defined as an annual PVC below −0.37%. All three algorithms (FreeSurfer longitudinal pipeline, MD-GRU 2024, and nnU-Net) fulfilled the requirement in all patients, when considering the entire brainstem (Table 4 and Figure 3). nnU-Net also detected pathological atrophy in the pons for all patients, while one patient was missed by FreeSurfer and MD-GRU. In the mesencephalon and medulla oblongata, all algorithms performed worse, with only small differences between them.

3.6 | Lesion Filling

Figure 4a depicts the number of MS patients with lesions in the brainstem and its subregions, as well as the percentage of the lesioned volume per (sub-)region. In the subsequent analysis, only the total brainstem was considered as lesions often crossed subregion borders. Dice coefficients indicated better segmentation performance with lesion filling (Figure 4b). ANTs lesion filling performed slightly better than FSL and the nnU-Net model performed generally better than the MD-GRU 2024 model. Also, the percentage of lesion volume covered by the inferred brainstem mask improved with lesion filling, especially when using ANTs.

4 | Discussion

This study optimized and systematically validated deep learning-based brainstem segmentation approaches based on the MD-GRU and U-Net network types. For the U-Net, data preprocessing and network configuration were optimized using the nnU-Net framework. We provide ready-to-use containers for straightforward deployment of these anatomically accurate, highly reliable, and robust segmentation methods for the brainstem and its substructures.

Segmentation quality for the brainstem and all three of its substructures was accurate, yielding high Dice scores compared with ground truth for MD-GRU and nnU-Net. The main improvement over the previously published method is the re-training on a ground truth dataset with increased heterogeneity in terms of MRI acquisition and disease. The vastly improved segmentation performance of MD-GRU 2024 over MD-GRU 2019 on data added during the current training (extension ground truth dataset) highlights the importance of addressing input data shift in this segmentation task, i.e., accurate model performance can be expected only for input data within the distribution of the

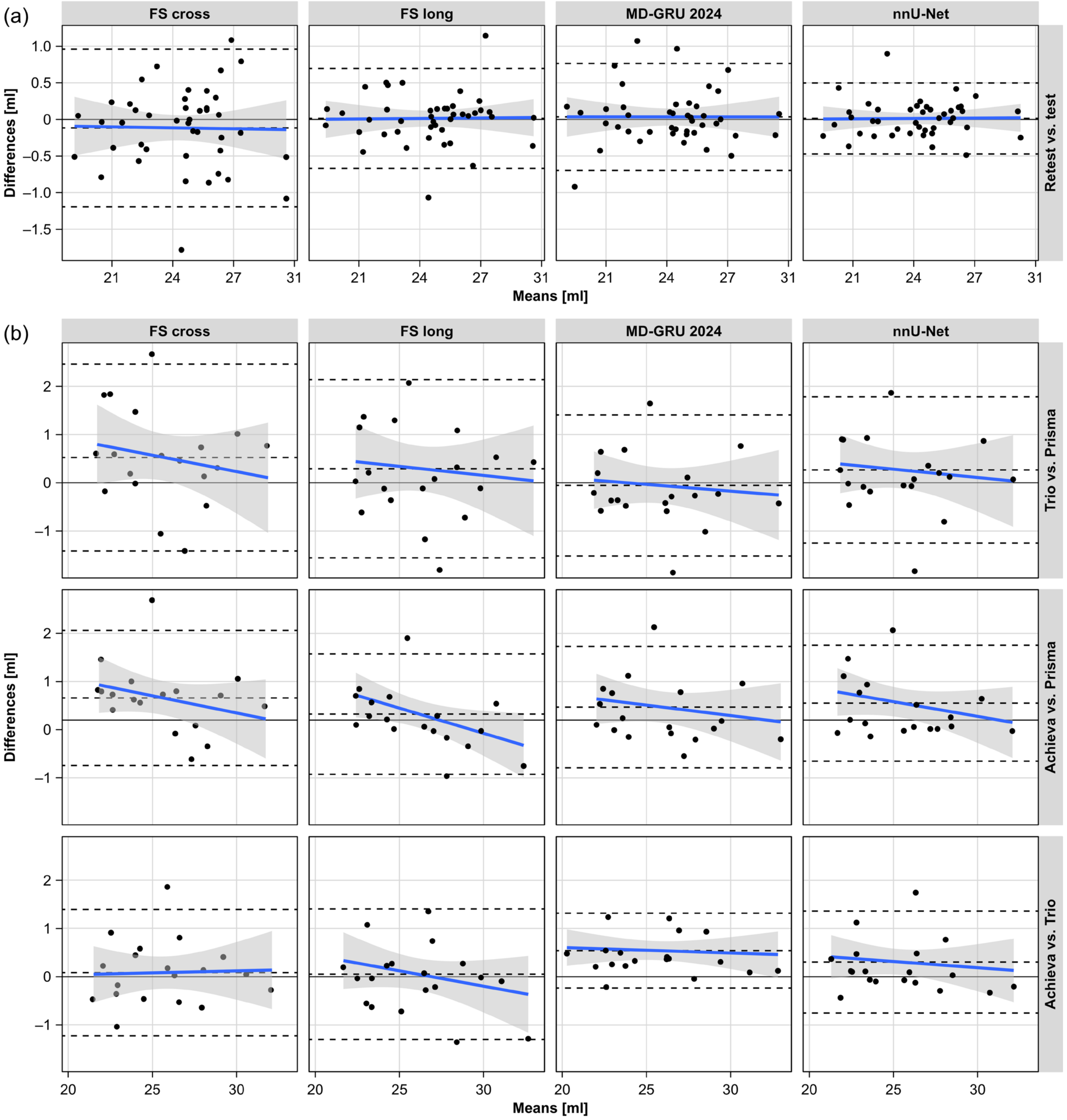

FIGURE 2 | Technical validation. Bland–Altman plots for scan-rescan repeatability (a) and inter-scanner reproducibility (b) of total brainstem volume. Dashed lines indicate (from top to bottom) upper limit of agreement (LOA), constant bias, and lower LOA. The blue line indicates the proportional bias with confidence interval (grey).

training dataset. By using a purposely heterogeneous dataset in the re-training, we are confident that the method can be applied to a wide range of acquisitions and diseases. The improved generalizability was corroborated in unseen data from different sources than the training data: MD-GRU 2024 demonstrated better scan–rescan repeatability and enhanced detection of pathological atrophy, while MD-GRU 2019 had a much higher failure rate in the unseen data sources.

In technical validation, we found a weak negative proportional bias of the nnU-Net model in the medulla oblongata, indicating that small volumes in this subregion are overestimated and large volumes underestimated. While not significant, the same tendency was also present for the MD-GRU model. This finding might be related to difficulties of the algorithms in exactly defining the border of the brainstem against the spinal cord.

TABLE 3 | Inter-scanner reproducibility. Bias and (one-sided) limit of agreement (LOA) of volumes (in mL) from the Bland–Altman analysis (95% confidence intervals in brackets). Only results for total brainstem volume are shown.

Comparison	Algorithm	Constant bias	LOA (one-sided)
Trio versus Prisma	FS cross	0.52 [0.05,1.00]	1.94 [1.11,2.77]
	FS long	0.29 [−0.16,0.75]	1.85 [1.06,2.64]
	MD-GRU	−0.05 [−0.41,0.31]	1.46 [0.84,2.09]
	nnU-Net	0.27 [−0.11,0.64]	1.52 [0.87,2.16]
Achieva versus Prisma	FS cross	0.66 [0.30,1.01]	1.40 [0.78,2.02]
	FS long	0.32 [0.01,0.64]	1.25 [0.70,1.80]
	MD-GRU	0.47 [0.15,0.79]	1.26 [0.70,1.82]
	nnU-Net	0.55 [0.25,0.86]	1.20 [0.67,1.74]
Achieva versus Trio	FS cross	0.08 [−0.24,0.40]	1.31 [0.75,1.87]
	FS long	0.05 [−0.28,0.39]	1.35 [0.77,1.93]
	MD-GRU	0.54 [0.35,0.73]	0.78 [0.44,1.11]
	nnU-Net	0.31 [0.05,0.57]	1.06 [0.60,1.51]

TABLE 4 | Clinical validation. Number of PROMESA study participants ($n = 16$) in which pathological atrophy could be detected.

Label	FreeSurfer	MD-GRU	nnU-Net
Mesencephalon	14 (87.5%)	14 (87.5%)	13 (81.3%)
Pons	15 (93.8%)	15 (93.8%)	16 (100%)
Medulla oblongata	9 (56.3%)	11 (68.8%)	12 (75.0%)
Brainstem	16 (100%)	16 (100%)	16 (100%)

We further compared scan–rescan repeatability and inter-scanner reproducibility between FreeSurfer, MD-GRU, and nnU-Net. A previous study showed highest repeatability of brainstem segmentations in healthy subjects obtained by FreeSurfer compared to two other automated brainstem segmentation methods (Velasco-Annis et al. 2018). Noteworthy, using the MarkVCID dataset, we were able to perform technical validation in a target population of patients, rather than young, healthy controls. This ensures that the technical validation results are applicable in the clinical setting. Overall, all three segmentation approaches showed similar scan–rescan variability in the total brainstem and pons, with nnU-Net and MD-GRU showing better scan–rescan repeatability in the mesencephalon than FreeSurfer, and nnU-Net segmentations showing the best scan–rescan variability in the anatomically most challenging region, the medulla oblongata. Inter-scanner reproducibility was similar across all three methods.

For clinical validation, we chose to evaluate MRI scans from patients with MSA, a rapidly progressive, neurodegenerative disease affecting the brainstem and presenting with cerebellar, mesencephalic, and pontine atrophy (Krismer et al. 2024; Mascalchi, Vella, and Ceravolo 2012). Percent volume changes for the brainstem and its substructures were assessed by the three segmentation methods between baseline and 1-year follow-up. All three segmentation methods consistently detected atrophy in the total brainstem over 1 year, exceeding the atrophy expected in healthy subjects. The same performance level was reached in the pons, while performance slightly dropped for the mesencephalon. In the medulla oblongata, the detected volume loss was less pronounced with all three methods, consistent with the clinical profile of the disorder. Brainstem volumetry is crucial in parkinsonian disorders, and automated mesencephalon segmentation using FreeSurfer was found to perform better than planimetric measurements in separating progressive supranuclear palsy from Parkinson's disease (Sjöström et al. 2020). Future studies are needed to investigate if brainstem volumetry assessed by the presented automated brainstem segmentation methods can be used in diagnosing parkinsonian disorders and can potentially be applied as outcome measures in clinical trials.

Although brainstem lesions were present in the training data by including patients with MS, severe hypointense brainstem lesions often caused incomplete segmentations, particularly when in an MS-typical location at the edge of the brainstem. Our results show that a lesion-filling algorithm can be applied in MS patients to improve the subsequent brainstem segmentation performance. Lesion filling was shown previously to improve segmentation accuracy in patients with MS using different segmentation approaches (Battaglini, Jenkinson, and De Stefano 2012). In demyelinating disorders, brainstem segmentation is a promising biomarker for differential diagnosis, showing different atrophy patterns in patients with MS and neuromyelitis optica spectrum disorders (Lee et al. 2018).

As a key element of our work, we make both brainstem segmentation algorithms, based on MD-GRU and nnU-Net, publicly available for non-commercial research use. While the nnU-Net-based algorithm performed overall slightly better than the MD-GRU-based, the differences were mostly small and often had overlapping confidence intervals. To facilitate the setup and reproducible science, we provide ready-to-use and versioned software containers for download at a public container registry. Installing deep learning algorithms can be challenging due

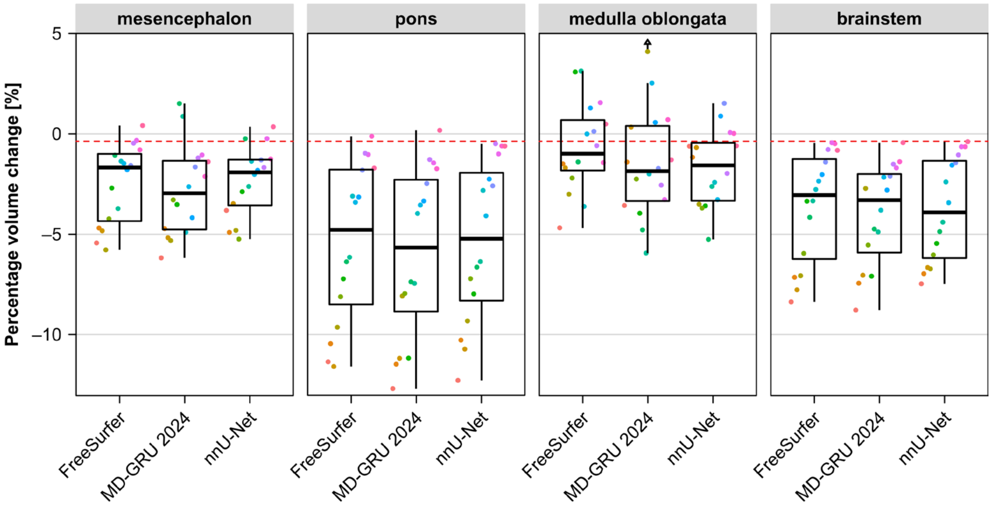

FIGURE 3 | Clinical Validation in MSA patients ($n = 16$). Percent volume change (PVC) over 1 year in the subregions and the total brainstem. Individual patients are shown as colored dots with the same color code across regions. The dashed red line indicates threshold for pathological brainstem atrophy at −0.37% per annum. One extreme outlier (PVC = +14.1%) observed in the medulla oblongata analysis using MD-GRU is depicted with an arrow and was the result of a larger segmentation error on the baseline image.

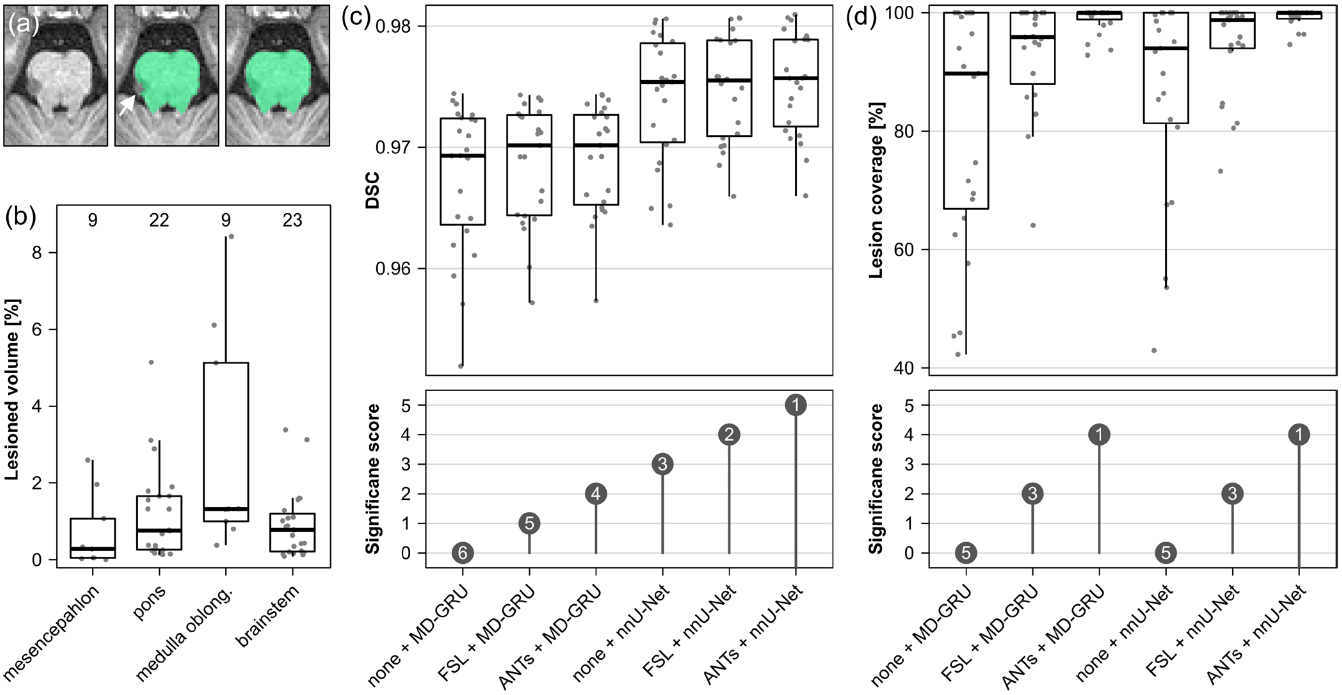

FIGURE 4 | Lesion filling. (a) Segmentation of the pons in the presence of an MS-typical brainstem lesion without (middle) and with (right) lesion filling. (b) Lesioned volume as percentage of the total brainstem and its subregions. The number of patients with lesions in each region is shown above the boxplots. (c) Dice similarity coefficient (DSC) for the fit of the inferred with the ground truth brainstem segmentations before (none) and after filling lesions with FSL or ANTs. Lower panels show significance scores with ranks (numbers in the dots). (d) Coverage of lesions by the inferred brainstem mask as percentage of the lesion volume.

to specific and sometimes conflicting software dependencies. Prebuilt containers simplify this process and ensure consistent segmentation results regardless of the local software environment (Moreau, Wiebels, and Boettiger 2023).

Main strengths of our work include the systematic validation approach following the HARNESS initiative guidelines (Smith et al. 2019). Importantly, we perform both clinical and technical validation in relevant target populations to ensure that results apply to the clinical setting. Limitations include the ground truth creation by just one reader. The intra-rater reliability of the rater was previously shown to be excellent (Sander et al. 2019) and the impact of the reader was reduced by using pre-segmentations. The lack of data from a GE scanner in the technical validation is another limitation. Although T1-weighted scans from a GE scanner are available in the MarkVCID dataset, the brainstem was unfortunately cropped due to a relatively small field-of-view. Finally, the small number of patients with longitudinal

MRI data in the PROMESA study precluded analyzing correlations with clinical outcomes. This currently confines clinical validation to atrophy assessments and a monitoring biomarker use case. Future studies with larger sample sizes are necessary to assess clinical correlations and the utility of brainstem volumetry as a surrogate endpoint in clinical trials.

5 | Conclusion

The common and clinically relevant involvement of the brainstem in neurodegenerative diseases makes brainstem volumetry an interesting biomarker candidate for neurodegeneration. Our findings emphasize the need to train segmentation models on diverse datasets that include various imaging sequences and the target pathologies. Through our study, we make two validated, fully automated, and fast brainstem segmentation algorithms publicly available, packaged as containers for convenient application. This will facilitate future studies that assess the extent of brainstem atrophy and its association with clinical outcome and prognosis in different neurodegenerative diseases.

Acknowledgments

Data used in preparation of this article were obtained from the MarkVCID consortium. A complete listing of MarkVCID investigators can be found at www.markvcid.org. Data collection and sharing for MarkVCID was funded by NINDS/NIA as part of the MarkVCID Consortium (U24NS100591).

Disclosure

The authors have nothing to report.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated during the current study.

References

Akhondi-Asl, A., and S. K. Warfield. 2013. "Simultaneous Truth and Performance Level Estimation Through Fusion of Probabilistic Segmentations." *IEEE Transactions on Medical Imaging* 32: 1840–1852.

Andermatt, S., S. Pezold, and P. Cattin. 2016. "Multi-Dimensional Gated Recurrent Units for the Segmentation of Biomedical 3D-Data." In *Deep Learning and Data Labeling for Medical Applications*, edited by G. Carneiro, D. Mateus, L. Peter, et al., vol. 10008, 142–151. Cham: Springer International Publishing. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-46976-8_15.

Andermatt, S., S. Pezold, and P. C. Cattin. 2018. "Automated Segmentation of Multiple Sclerosis Lesions Using Multi-Dimensional Gated Recurrent Units." In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, edited by A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, vol. 10670, 31–42. Cham: Springer International Publishing. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-75238-9_3.

Avants, B. B., N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. 2011. "A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration." *NeuroImage* 54: 2033–2044.

Battaglini, M., M. Jenkinson, and N. De Stefano. 2012. "Evaluating and Reducing the Impact of White Matter Lesions on Brain Volume Measurements." *Human Brain Mapping* 33: 2062–2071.

Brooks, J. C. W., O. K. Faull, K. T. S. Pattinson, and M. Jenkinson. 2013. "Physiological Noise in Brainstem FMRI." *Frontiers in Human Neuroscience* 7: 623.

Casserly, C., E. E. Seyman, P. Alcaide-Leon, et al. 2018. "Spinal Cord Atrophy in Multiple Sclerosis: A Systematic Review and Meta-Analysis." *Journal of Neuroimaging* 28: 556–586.

Datta, D., and J. Love. 2018. "deepankardatta/blandr: Version 0.5.1." [object Object]. https://zenodo.org/record/824514.

Disanto, G., P. Benkert, J. Lorscheider, et al. 2016. "The Swiss Multiple Sclerosis Cohort-Study (SMSC): A Prospective Swiss Wide Investigation of Key Phases in Disease Evolution and New Treatment Options." *PLoS One* 11: e0152347.

Duering, M., G. J. Biessels, A. Brodtmann, et al. 2023. "Neuroimaging Standards for Research Into Small Vessel Disease-Advances Since 2013." *Lancet Neurology* 22: 602–618.

Duering, M., R. Righart, F. A. Wollenweber, V. Zietemann, B. Gesierich, and M. Dichgans. 2015. "Acute Infarcts Cause Focal Thinning in Remote Cortex via Degeneration of Connecting Fiber Tracts." *Neurology* 84: 1685–1692.

Eshaghi, A., R. V. Marinescu, A. L. Young, et al. 2018. "Progression of Regional Grey Matter Atrophy in Multiple Sclerosis." *Brain* 141: 1665–1677.

Fanciulli, A., and G. K. Wenning. 2015. "Multiple-System Atrophy." *New England Journal of Medicine* 372: 249–263.

FDA-NIH Biomarker Working Group. 2016. *BEST (Biomarkers, EndpointS, and other Tools) Resource*. Silver Spring (MD): Food and Drug Administration (US). http://www.ncbi.nlm.nih.gov/books/NBK326791/.

Fischl, B., D. H. Salat, E. Busa, et al. 2002. "Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain." *Neuron* 33: 341–355.

Fischl, B., D. H. Salat, A. J. W. van der Kouwe, et al. 2004. "Sequence-Independent Segmentation of Magnetic Resonance Images." *NeuroImage* 23, no. Suppl 1: S69–S84.

Fjell, A. M., K. B. Walhovd, C. Fennema-Notestine, et al. 2009. "One-Year Brain Atrophy Evident in Healthy Aging." *Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29: 15223–15231.

Herlihy, A. H., J. V. Hajnal, W. L. Curati, et al. 2001. "Reduction of CSF and Blood Flow Artifacts on FLAIR Images of the Brain With k-Space Reordered by Inversion Time at Each Slice Position (KRISP)." *AJNR. American Journal of Neuroradiology* 22: 896–904.

Iglesias, J. E., K. Van Leemput, P. Bhatt, et al. 2015. "Bayesian Segmentation of Brainstem Structures in MRI." *NeuroImage* 113: 184–195.

Isensee, F., P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. 2021. "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation." *Nature Methods* 18: 203–211.

Krismer, F., P. Péran, V. Beliveau, et al. 2024. "Progressive Brain Atrophy in Multiple System Atrophy: A Longitudinal, Multicenter, Magnetic Resonance Imaging Study." *Movement Disorders: Official Journal of the Movement Disorder Society* 39: 119–129.

Lee, C.-Y., H. K.-F. Mak, P.-W. Chiu, H.-C. Chang, F. Barkhof, and K.-H. Chan. 2018. "Differential Brainstem Atrophy Patterns in Multiple Sclerosis and Neuromyelitis Optica Spectrum Disorders." *Journal of Magnetic Resonance Imaging* 47: 1601–1609.

Levin, J., S. Maaß, M. Schuberth, et al. 2019. "Safety and Efficacy of Epigallocatechin Gallate in Multiple System Atrophy (PROMESA):

A Randomised, Double-Blind, Placebo-Controlled Trial." *Lancet Neurology* 18: 724–735.

Lu, H., A. H. Kashani, K. Arfanakis, et al. 2021. "MarkVCID Cerebral Small Vessel Consortium: II. Neuroimaging Protocols." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 17: 716–725.

Maillard, P., H. Lu, K. Arfanakis, et al. 2022. "Instrumental Validation of Free Water, Peak-Width of Skeletonized Mean Diffusivity, and White Matter Hyperintensities: MarkVCID Neuroimaging Kits." *Alzheimer's & Dementia (Amsterdam, Netherlands)* 14: e12261.

Mascalchi, M., A. Vella, and R. Ceravolo. 2012. "Movement Disorders: Role of Imaging in Diagnosis." *Journal of Magnetic Resonance Imaging: JMRI* 35: 239–256.

Moreau, D., K. Wiebels, and C. Boettiger. 2023. "Containers for Computational Reproducibility." *Nature Reviews Methods Primers* 3: 50.

Patenaude, B., S. M. Smith, D. N. Kennedy, and M. Jenkinson. 2011. "A Bayesian Model of Shape and Appearance for Subcortical Brain Segmentation." *NeuroImage* 56: 907–922.

Rocca, M. A., M. Battaglini, R. H. B. Benedict, et al. 2017. "Brain MRI Atrophy Quantification in MS: From Methods to Clinical Application." *Neurology* 88: 403–413.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." https://arxiv.org/abs/1505.04597.

Sander, L., S. Pezold, S. Andermatt, et al. 2019. "Accurate, Rapid and Reliable, Fully Automated MRI Brainstem Segmentation for Application in Multiple Sclerosis and Neurodegenerative Diseases." *Human Brain Mapping* 40: 4091–4104.

Sjöström, H., T. Granberg, F. Hashim, E. Westman, and P. Svenningsson. 2020. "Automated Brainstem Volumetry Can Aid in the Diagnostics of Parkinsonian Disorders." *Parkinsonism & Related Disorders* 79: 18–25.

Smith, E. E., G. J. Biessels, F. De Guio, et al. 2019. "Harmonizing Brain Magnetic Resonance Imaging Methods for Vascular Contributions to Neurodegeneration." *Alzheimer's & Dementia (Amsterdam, Netherlands)* 11: 191–204.

Velasco-Annis, C., A. Akhondi-Asl, A. Stamm, and S. K. Warfield. 2018. "Reproducibility of Brain MRI Segmentation Algorithms: Empirical Comparison of Local MAP PSTAPLE, FreeSurfer, and FSL-FIRST." *Journal of Neuroimaging* 28: 162–172.

Wenning, G. K., I. Stankovic, L. Vignatelli, et al. 2022. "The Movement Disorder Society Criteria for the Diagnosis of Multiple System Atrophy." *Movement Disorders: Official Journal of the Movement Disorder Society* 37: 1131–1148.

Wilcock, D., G. Jicha, D. Blacker, et al. 2021. "MarkVCID Cerebral small vessel consortium: I. Enrollment, clinical, fluid protocols." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 17: 704–715.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.