


RESEARCH ARTICLE

Accurate, rapid and reliable, fully automated MRI brainstem segmentation for application in multiple sclerosis and neurodegenerative diseases

Laura Sander^{1,2} | Simon Pezold³ | Simon Andermatt³ | Michael Amann^{1,4} |
 Dominik Meier⁴ | Maria J. Wendebourg¹ | Tim Sinnecker^{1,2,4} |
 Ernst-Wilhelm Radue¹ | Yvonne Naegelin¹ | Cristina Granziera^{1,2} | Ludwig Kappos¹ |
 Jens Wuerfel⁴ | Philippe Cattin³ | Regina Schlaeger^{1,2}  | for the Alzheimer's Disease
 Neuroimaging Initiative[†]

¹Neurology Clinic and Policlinic, Departments of Medicine, Clinical Research and Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland

²Translational Imaging in Neurology (ThINK) Basel, Department of Medicine and Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland

³Center for medical Image Analysis & Navigation (CIAN), Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland

⁴Medical Image Analysis Center (MIAC AG), Basel and qbig, Department of Biomedical Engineering, University of Basel, Basel, Switzerland

Correspondence

Regina Schlaeger, Neurology Clinic and
 Policlinic, Departments of Medicine, Clinical
 Research and Biomedical Engineering,
 University Hospital Basel and University of
 Basel, Basel, Switzerland.
 Email: regina.schlaeger@usb.ch

Funding information

Schweizerische Multiple Sklerose Gesellschaft;
 Schweizerischer Nationalfonds zur Förderung
 der Wissenschaftlichen Forschung, Grant/
 Award Number: MHV program; PMPDP3
 171391

Abstract

Neurodegenerative disorders, such as Alzheimer's disease (AD) and progressive forms of multiple sclerosis (MS), can affect the brainstem and are associated with atrophy that can be visualized by MRI. Anatomically accurate, large-scale assessments of brainstem atrophy are challenging due to lack of automated, accurate segmentation methods. We present a novel method for brainstem volumetry using a fully-automated segmentation approach based on multi-dimensional gated recurrent units (MD-GRU), a deep learning based semantic segmentation approach employing a convolutional adaptation of gated recurrent units. The neural network was trained on 67 3D-high resolution T1-weighted MRI scans from MS patients and healthy controls (HC) and refined using segmentations of 20 independent MS patients' scans. Reproducibility was assessed in MR test–retest experiments in 33 HC. Accuracy and robustness were examined by Dice scores comparing MD-GRU to FreeSurfer and manual brainstem segmentations in independent MS and AD datasets. The mean %-change/*SD* between test–retest brainstem volumes were 0.45%/0.005 (MD-GRU), 0.95%/0.009 (FreeSurfer), 0.86%/0.007 (manually edited segmentations). Comparing MD-GRU to manually edited segmentations the mean Dice scores/*SD* were:

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

0.97/0.005 (brainstem), 0.95/0.013 (mesencephalon), 0.98/0.006 (pons), 0.95/0.015 (medulla oblongata). Compared to the manual gold standard, MD-GRU brainstem segmentations were more accurate than FreeSurfer segmentations ($p < .001$). In the multi-centric acquired AD data, the mean Dice score/*SD* for the MD-GRU-manual segmentation comparison was 0.97/0.006. The fully automated brainstem segmentation method MD-GRU provides accurate, highly reproducible, and robust segmentations in HC and patients with MS and AD in 200 s/scan on an Nvidia GeForce GTX 1080 GPU and shows potential for application in large and longitudinal datasets.

KEYWORDS

atrophy, brainstem, deep learning, MD-GRU, multiple sclerosis, segmentation

1 | INTRODUCTION

The brainstem (BS) is situated at the base of the cerebrum and forms the anatomical link between brain and spinal cord. From cranial to caudal, the BS consists of three substructures: midbrain (mesencephalon; M), pons (P), and medulla oblongata (MO).

The BS is a vitally important and complex structure, containing multiple scattered cranial nerve nuclei, white matter tracts for relaying sensory and motor data to and from the brain and spinal cord, and reticular nuclei including monoamine-producing nuclei with widespread connections to all parts of the brain (Naidich et al., 2009; Nieuwenhuys, 1985).

Several primary or secondary neurodegenerative diseases as for example, Alzheimer's disease (AD; Grinberg et al., 2009), Parkinson's disease (Grinberg, Rueb, Alho, & Heinsen, 2010), progressive supranuclear palsy (Williams & Lees, 2009), multisystem atrophy (Ghorayeb et al., 2002), amyotrophic lateral sclerosis (Warabi, Hayashi, Nagao, & Shimizu, 2017) and multiple sclerosis (MS; Noseworthy, Lucchinetti, Rodriguez, & Weinshenker, 2000), can affect the BS or its substructures. During these diseases, progressive neurodegeneration of BS structures can result in dysphagia, dysarthria, autonomic dysfunction and other symptoms that influence not only the patient's quality of life but also survival (Grinberg et al., 2010; Kim et al., 2018).

One of the macroscopic hallmarks of neurodegeneration is tissue loss (atrophy) that can be assessed and quantified by magnetic resonance imaging (MRI) in vivo. Yet, while atrophy of the brain and spinal cord has been extensively studied in neurodegenerative diseases by MRI over the past decades, the BS has been less well investigated. Challenges of BS imaging include the relatively small structure of the BS and frequent artifacts in conventional MR imaging due to cerebrospinal fluid (CSF) or pulsatile blood flow (Herlihy et al., 2001; Tanaka, Abe, Kojima, Nishimura, & Hayabuchi, 2000).

Moreover, the BS is anatomically less well demarcated compared with other brain regions, especially at the level of the BS caudal border toward the spinal cord.

While still regarded as the gold standard, manual segmentations provide high anatomic accuracy, however, are prone to intra- and

inter-rater bias. Several originally for brain segmentation designed methods enable also automated BS segmentation (Akhondi-Asl & Warfield, 2013; Fischl et al., 2002; Iglesias et al., 2015; Patenaude, Smith, Kennedy, & Jenkinson, 2011). However, anatomic accurate segmentations, especially the correct identification of the lower BS border, are still a mayor challenge and limits the applicability of these methods for atrophy assessments in clinical studies. Previous studies suggested that the combination of automated segmentation and corrective learning provides a more accurate segmentation of the BS than automated segmentation alone (Wang, Ngo, Hessl, Hagerman, & Rivera, 2016).

Building on multi-dimensional gated recurrent units (MD-GRU; Andermatt, Pezold, & Cattin, 2016; Andermatt, Pezold, & Cattin, 2018), a recently developed deep learning based semantic segmentation approach, the objective of this study was to develop an accurate, reliable and efficient BS segmentation method from T1-weighted (T1-w) MRI images enabling improved quantification of BS volume loss in the investigation of neurodegenerative diseases.

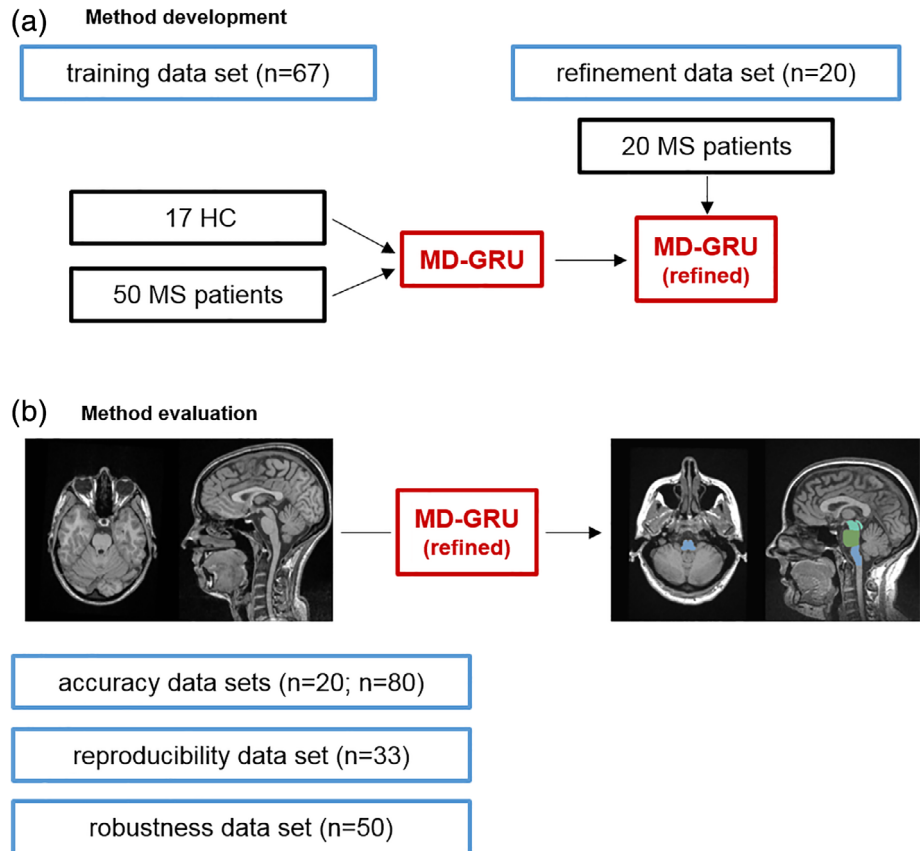
We report the reproducibility, accuracy, and robustness of this novel method in segmentation of the BS and its substructures in healthy controls (HC), MS, and AD patients.

2 | MATERIALS AND METHODS

2.1 | Imaging data

For algorithm training and refinement, assessments of accuracy and reproducibility, several independent sets of 3D high-resolution T1w MR imaging data (MPRAGE, all obtained on the same 1.5 T Magnetom Avanto scanner [Siemens Healthineers, Erlangen, Germany] at the Department of Radiology, University Hospital Basel) with identical acquisition parameters were used in this study (see Figure 1 for an overview). Briefly, acquisition parameters were TR = 2080 ms, TI = 1,100 ms, TE = 3.1 ms, $\alpha = 15^\circ$, 160 sagittal slices, spatial resolution of $0.98 \times 0.98 \times 1\text{mm}^3$ (Bendfeldt et al., 2009; Weier et al., 2014). In addition, reproducibility was also assessed in 22 data set pairs that were obtained from one single Siemens 3 T Prisma scanner

FIGURE 1 Schematic figure of (a) training and refinement and (b) accuracy, reproducibility and robustness assessment of the algorithm [Color figure can be viewed at wileyonlinelibrary.com]



(acquisition parameters: TR = 2,700 ms, TI = 950 ms, TE = 5.03 ms, $\alpha = 8^\circ$; spatial resolution of $1.0 \times 1.0 \times 1.0 \text{ mm}^3$).

Written informed consent was obtained from all patients and HC. The study has been approved by the local ethics committee.

2.1.1 | Algorithm training data set

For algorithm training, data sets from 50 patients of an ongoing cohort study, with diagnosis of MS or clinically isolated syndrome (CIS) according to McDonald et al. (2001) (mean age 46.3 years, *SD* 11.18, range 28–67 years, 66% women, median EDSS 3.5, IQR = 3.25) and from 17 HC (mean age 30.1 years, *SD* 10.49, range 18–61 years, 29% women) were used.

2.1.2 | Algorithm refinement data set

For refinement of the algorithm parameters, 20 independent brain MPRAGE data sets with identical acquisition parameters from 20 additional patients with MS or CIS from the same cohort (mean age 41.7 years, *SD* 10.84, range 28–63 years, 70% women, median EDSS 1.75, IQR = 1.5) were assessed.

2.1.3 | Accuracy data sets

For comparison of the automated refined MD-GRU segmentation algorithm and the FreeSurfer segmentation approach with an

independent gold standard (exclusively manual segmentations), 20 independent data sets of patients with MS or CIS from the same cohort (mean age 42.1 years, *SD* 12.42, range 21–65, 70% women, median EDSS 2.0, IQR 1.5) were assessed.

Accuracy of the refined MD-GRU segmentation algorithm was then additionally assessed in a larger dataset of 80 independent brain MPRAGE images with identical acquisition parameters from additional 80 patients with MS or CIS from the same cohort (mean age 42.8 years, *SD* 11.27, range 21–65 years, 68% women, median EDSS 2.5, IQR = 2.5) and compared to manually edited segmentations based on FreeSurfer presegmentations as described in “Manual segmentations.”

2.1.4 | Reproducibility data set

For reproducibility assessments, T1w brain MRI data sets of 33 HC (mean age 30.2 years, *SD* 9.02, range 20–56 years, 61% women) were evaluated. These scans were acquired as a MR test–retest experiment with identical acquisition parameters on identical scanners respectively with repositioning between scans. Eleven data set pairs were acquired on a 1.5 T Siemens Avanto scanner (Keshavan et al., 2016; mean age 38.5 years, *SD* 10.02, range 24–56 years, 64% women), 22 independent data set pairs were obtained from a 3 T Siemens Prisma scanner (mean age 26.1 years, *SD* 4.26, range 21–39 years, 59% women).

2.1.5 | Robustness data set

The robustness of the MD-GRU BS segmentation approach was assessed using multicentric MPRAGE data sets (Siemens, 3 T) of 50 patients with AD (mean age 75.4 years, *SD* 8.55, range 56–90 years, 52% women) that were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

2.2 | Manual segmentations

Manually edited segmentations of the BS and its substructures were used for algorithm training, refinement, and accuracy validation against MD-GRU segmentations in MS and AD patients. Manually edited BS segmentations were performed using the open source software 3D Slicer 4.8.1 (www.slicer.org), based on presegmentations

generated by FreeSurfer (<http://surfer.nmr.mgh.harvard.edu/>; Iglesias et al., 2015). We chose FreeSurfer for presegmentations due to its relatively good reproducibility in BS segmentations compared to other methods (Velasco-Annis, Akhondi-Asl, Stamm, & Warfield, 2018).

The manual editing process followed clear definitions. Particular focus was put on the anatomically correct limitation of the midbrain toward the epiphysis, and the cranio-caudal extension of the MO, defined by the medullo-pontine sulcus and the bilateral exit of the first spinal root (Figure 2, described in detail in Appendix A) and was performed by two independent trained neurologists experienced in neuroimaging. Inter-rater reliability of the manual editing process was assessed on the MR data sets of 10 MS patients from the same cohort by intra-class correlation coefficients (ICC, two-way random, absolute agreement) and coefficients of variation ($COV = \frac{[\max_{Volume} - \min_{Volume}]}{\text{mean}_{Volume}}$). Due to the high ICCs (≥ 0.998 for the BS and all three substructures) and low COVs ($COV_{BS} = 0.077\%$), all further manual edits were performed by one rater only (for detailed information see Appendix C).

Exclusively manual segmentations were performed in a subset of cases to generate a FreeSurfer independent gold standard and enable an independent comparison between the automated segmentation methods MD-GRU and FreeSurfer. Manual segmentation was

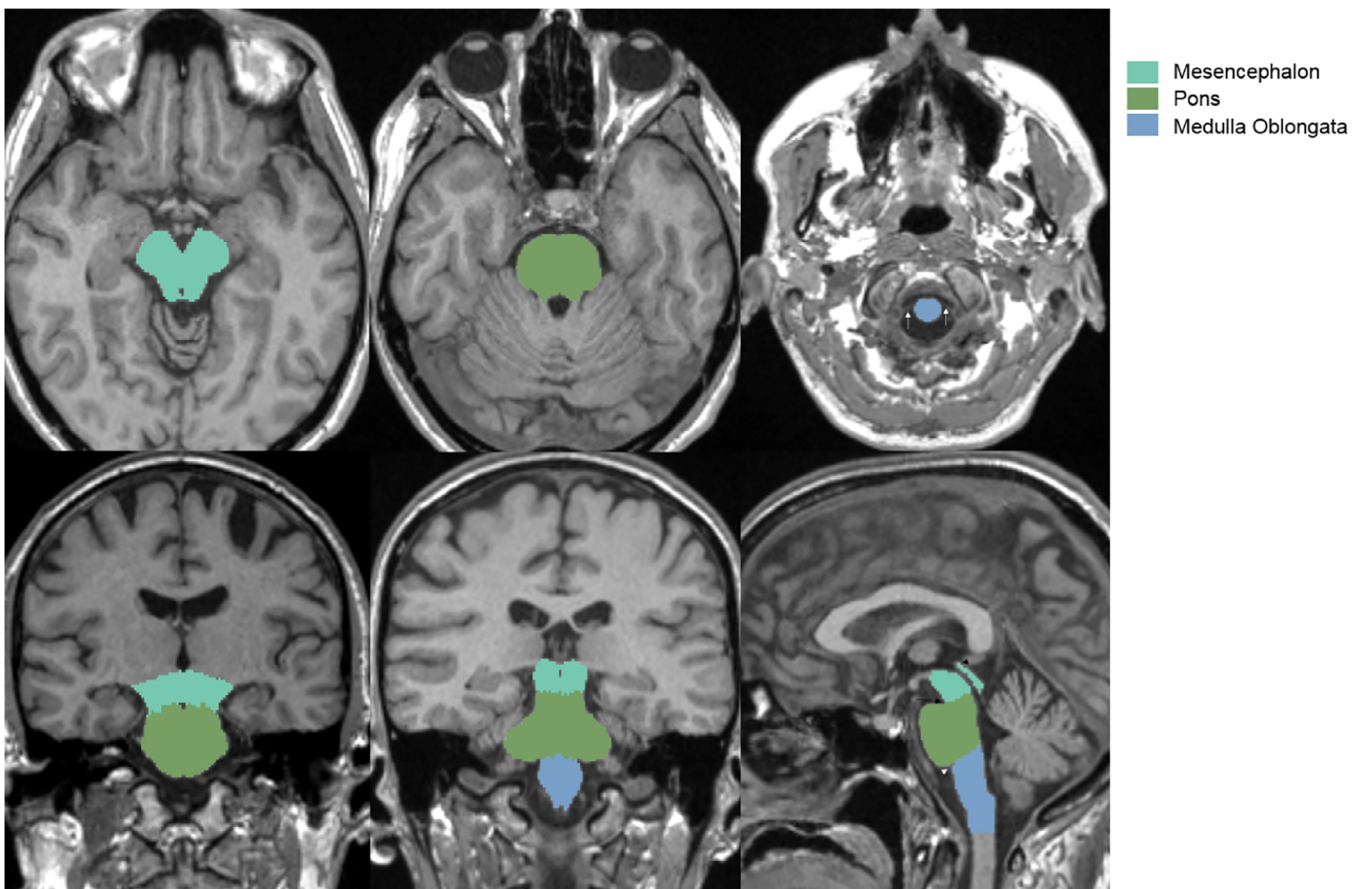


FIGURE 2 Illustrations of anatomical landmarks used in the manual segmentation of the mesencephalon, pons and medulla oblongata. Caudal delimitation of the MO is defined as the bilateral exit of the first spinal root (white arrows) in axial slices. The pontomedullary sulcus marks the cranial delimitation (white arrowheads). The pontomesencephalic junction is marked by the black arrow, the cranial delimitation of the mesencephalon toward the pineal gland is shown by the black arrowhead

performed using 3D Slicer followed clear anatomical definitions as described in the Appendix B and were performed by a trained neurologist with experience in neuroimaging, not aware of the automated segmentation results.

2.3 | Segmentation algorithm

Segmentation was done using multi-dimensional gated recurrent units (MD-GRU; Andermatt et al., 2016), adopting the network architecture from Andermatt et al. (2018). In brief, MD-GRU is a deep learning-based, fully-automated semantic segmentation approach employing a convolutional adaptation of gated recurrent units (GRU; Cho et al., 2014). Each MD-GRU layer traverses an image forward and backward along each of its spatial dimensions to infer the current segmentation class label from the local appearance and its surrounding context.

As a preprocessing step, images were filtered with a high-pass filter as described in Andermatt et al. (2016), the result of which was passed to the network as an additional input channel. Blocks of $80 \times 80 \times 80$ voxels were sampled from training images allowing a sampling of 10 voxels beyond the image boundaries with zero-padding. Data augmentation with random deformations (grid spacing 64, deformation vector components sampled from $N(0,3)$ [Andermatt et al., 2016], small rotations ($\pm 10^\circ$) and scaling ($\pm 10\%$) was used to increase the size of the training data set. To prevent overfitting, DropConnect was performed on input and previous state with a keep rate of 0.5, multiplying the weights with a Gaussian random variable (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Selective sampling and residual learning on the MD-GRU level were applied as described in Andermatt et al. (2018).

The respective neural network was trained for 100,000 iterations, using AdaDelta (Zeiler, 2012), on the training data set described above. For the final 40,000 iterations, the training state of the network was evaluated every 3,000 iterations against the refinement data set described above, and the state scoring the highest mean Dice score across all labeled subregions of the BS in all refinement images was chosen for further use in our experiments. To segment unseen data with the trained network in our experiments, each image was regularly subdivided into blocks of $80 \times 80 \times 80$ voxels with an overlap of 10 voxels in each dimension. The blockwise segmentation results were then stitched together into the original image size.

2.4 | Accuracy analysis

The accuracy of the automated segmentation was assessed by two different approaches. First, a comparison of exclusively manual BS segmentations with MD-GRU and FreeSurfer segmentations was done in a data set of 20 MS patients that were not part of the algorithm training and refinement (accuracy data set, *manual segmentations*) to enable a comparison between the two automated segmentation methods with an independently generated gold standard. Second, we compared the MD-GRU segmentations with expert-labeled manually edited and FreeSurfer BS segmentations in another larger dataset from 80 MS patients, also independent from the algorithm training and refinement (accuracy data set, *manually edited segmentations*). All MD-GRU segmentations were visually inspected and considered successful. Then Dice coefficients and mean surface distances (MSD) were each calculated for total BS volumes, M, P, and MO volumes.

Pearson correlation was used to assess the association between MD-GRU calculated BS volumes and manually edited segmented volumes. In addition, Bland-Altman plots were created to graphically compare the two BS segmentation approaches by plotting the difference between corresponding measurements obtained by the two segmentation methods against their averages.

2.5 | Reproducibility analysis

The reproducibility of the MD-GRU segmentations was assessed in 33 healthy subjects that underwent a MR test-retest experiment being scanned twice in the same session after repositioning, 11 subjects at 1.5 T, and 22 subjects at 3 T as described above. All segmentations were visually inspected and considered successful and anatomically adequate. Reproducibility of the BS volumes was assessed as percent change between test and retest scans. The mean difference and mean percentage change between test and retest and the respective ICCs (two-way random, absolute agreement) were calculated.

2.6 | Robustness analysis

By applying the algorithm on multicentric acquired MPRAGE images of 50 AD patients, we further tested the robustness of the novel method in a more diverse dataset including different acquisition platforms. Again, all MD-GRU segmentations were visually controlled and

TABLE 1A Mean Dice scores comparing MD-GRU and FreeSurfer segmentations to the manual gold standard in 20 MS patients

| | Mean Dice score/ <i>SD</i> /95%CI comparing MD-GRU vs. manual segmentations | Mean Dice score/ <i>SD</i> /95%CI comparing FreeSurfer vs. manual segmentations |
|-------------------|---|---|
| Brainstem | 0.94/0.010/0.937–0.945 | 0.93/0.010/0.926–0.935 |
| Mesencephalon | 0.87/0.024/0.864–0.884 | 0.87/0.029/0.852–0.876 |
| Pons | 0.93/0.011/0.927–0.936 | 0.93/0.011/0.920–0.931 |
| Medulla oblongata | 0.92/0.015/0.912–0.924 | 0.89/0.022/0.876–0.894 |

Note: Mean Dice scores and *SD* for the total brainstem, mesencephalon, pons, and medulla oblongata comparing exclusively manual segmentations to MD-GRU and FreeSurfer (Iglesias et al., 2015) segmentations.

considered anatomically adequate. Dice coefficients and MSD comparing automated to manually edited and FreeSurfer segmentations were calculated.

Statistical analysis was performed using SPSS 22 and JMP Pro 14.1.

3 | RESULTS

3.1 | Accuracy

For validation, we compared exclusively manual segmentations, as independent gold standard, to MD-GRU and FreeSurfer segmentations in a separate accuracy dataset characterized above that was not used during the development of the segmentation approach.

Dice scores for the MD-GRU-manual segmentation comparison (mean Dice score/*SD*/95%CI 0.94/0.01/0.937–0.945) were slightly but significantly higher than for the FreeSurfer-manual segmentation comparison (0.93/0.01/0.926–0.935), $p < .001$. The corresponding Dice scores for all substructures can be found in Table 1a.

In the second, independent larger data set of MS patients, MD-GRU total BS volumes were highly correlated with independently manually edited volumes ($R^2 = 0.99$) (Figure 3a and Figure 3b). The mean Dice scores and MSD comparing the MD-GRU to the manually edited and FreeSurfer segmentations are shown in Tables 1b and 1c. The mean Dice scores comparing the MD-GRU to the manually edited segmentations were 0.97 for the total BS, 0.95 for the mesencephalon, 0.98 for the

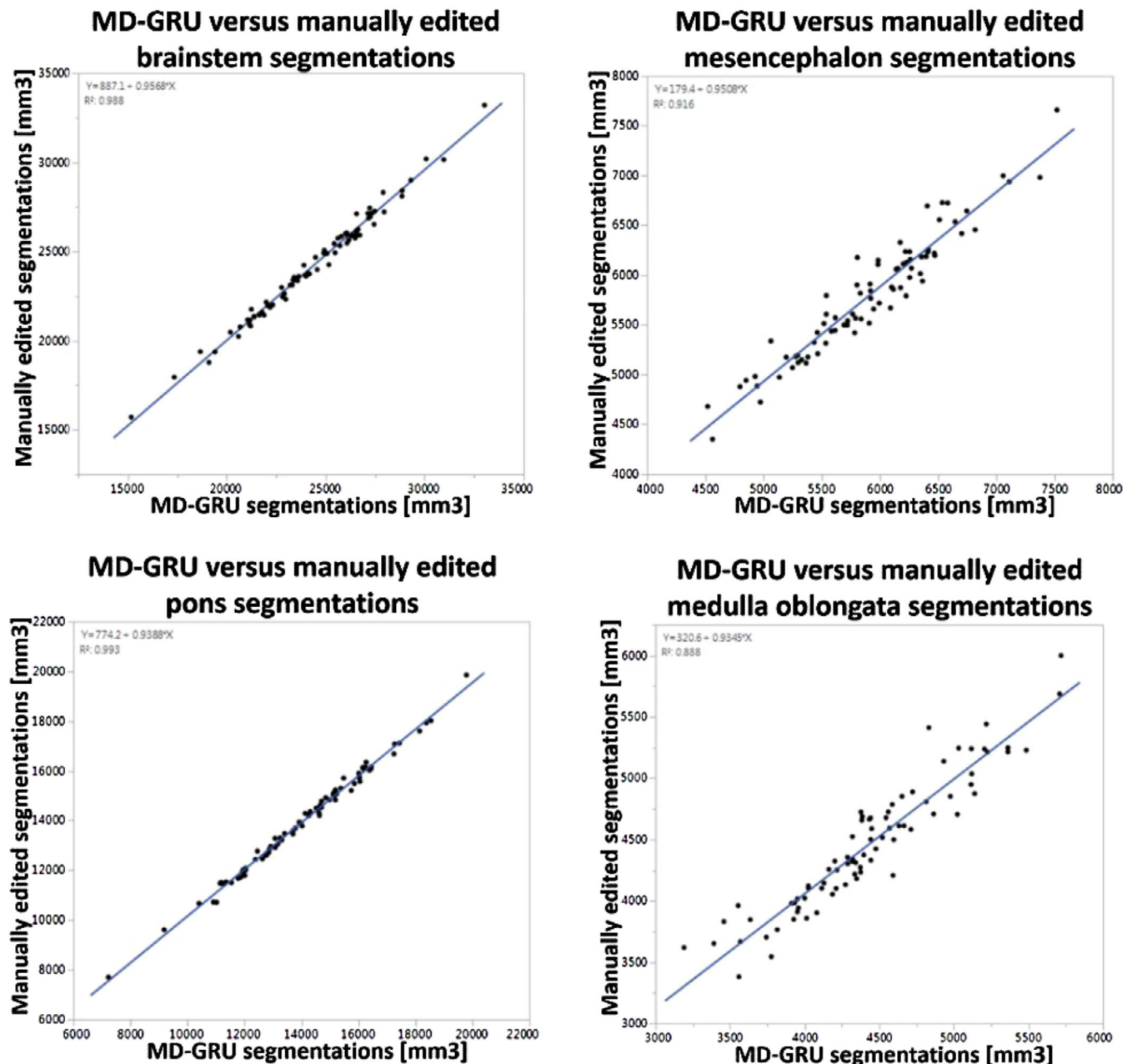


FIGURE 3A Association between manually edited segmentations and MD-GRU segmentations for the total brainstem and its substructures mesencephalon, pons and medulla oblongata ($n = 80$) [Color figure can be viewed at wileyonlinelibrary.com]

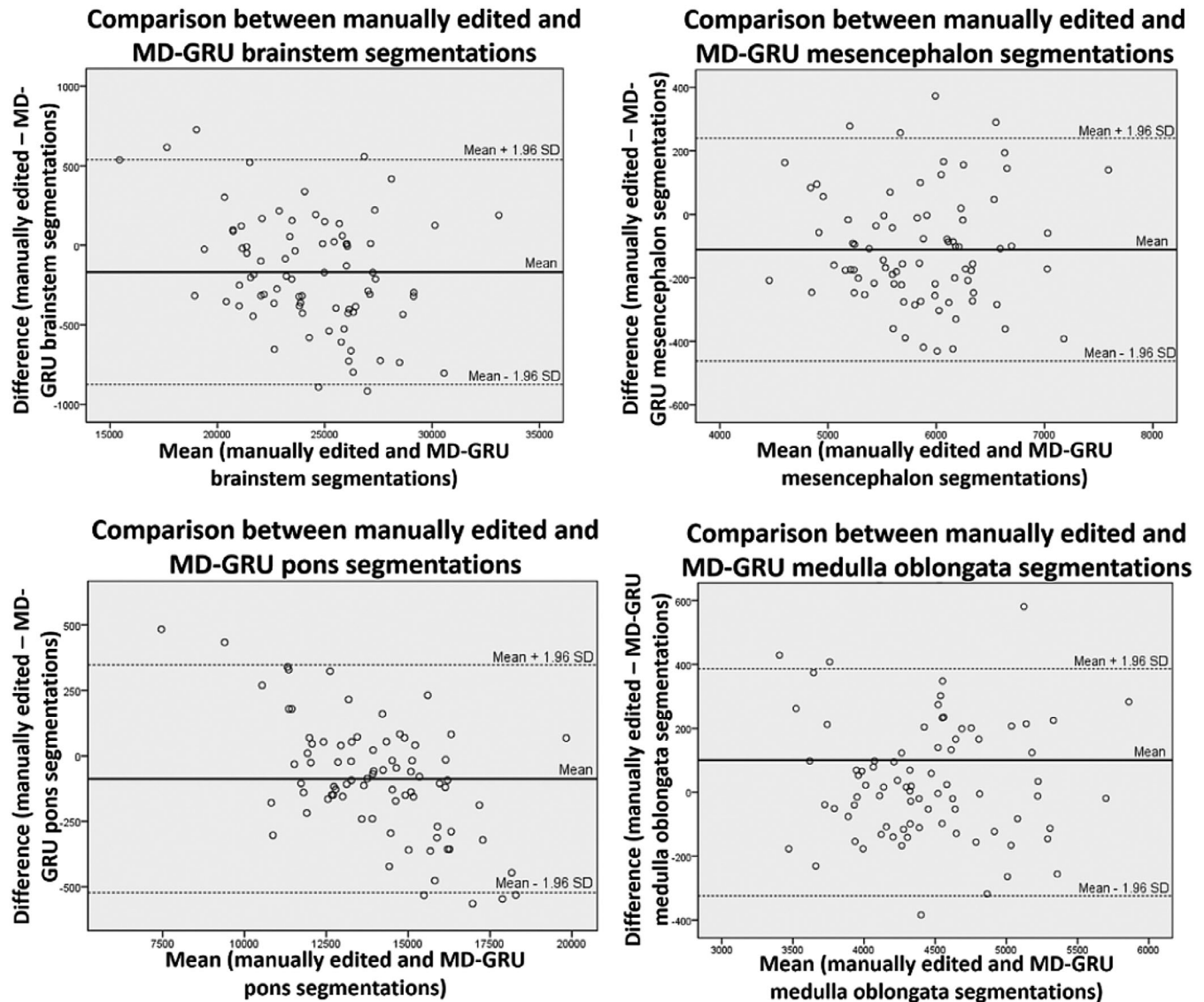


FIGURE 3B Bland-Altman plots showing the absolute differences between manually edited and MD-GRU based segmentations plotted against their averages. The dashed lines indicate the limits of agreement ($\text{mean} \pm 1.96 \text{ SD}$), ($n = 80$)

TABLE 1B Mean Dice scores comparing MD-GRU segmentations to manually edited and FreeSurfer segmentations in 80 MS patients

| | Mean Dice score/ <i>SD</i> comparing MD-GRU vs. manually edited segmentations | Mean Dice score/ <i>SD</i> comparing MD-GRU vs. FreeSurfer segmentations |
|-------------------|--|--|
| Brainstem | 0.97/0.005 | 0.97/0.005 |
| Mesencephalon | 0.95/0.013 | 0.95/0.013 |
| Pons | 0.98/0.006 | 0.97/0.006 |
| Medulla oblongata | 0.95/0.015 | 0.94/0.015 |

Note: Mean Dice scores and *SD* for the total brainstem, mesencephalon, pons, and medulla oblongata comparing segmentations to MD-GRU versus manually edited and FreeSurfer (Iglesias et al., 2015) segmentations.

TABLE 1C Mean surface distances (MSD) comparing segmentation methods in 80 MS patients

| | MSD [mm]/ <i>SD</i> comparing MD-GRU vs. manually edited segmentations | MSD/ <i>SD</i> comparing MD-GRU vs. FreeSurfer segmentations |
|-------------------|---|---|
| Brainstem | 0.24/0.043 | 0.26/0.04 |
| Mesencephalon | 0.30/0.066 | 0.31/0.065 |
| Pons | 0.24/0.051 | 0.24/0.047 |
| Medulla oblongata | 0.27/0.085 | 0.37/0.083 |

Note: Mean surface distance (MSD) [mm] and *SD* for the total brainstem, mesencephalon, pons and medulla oblongata comparing segmentations with MD-GRU versus manually edited and FreeSurfer (Iglesias et al., 2015) segmentations.

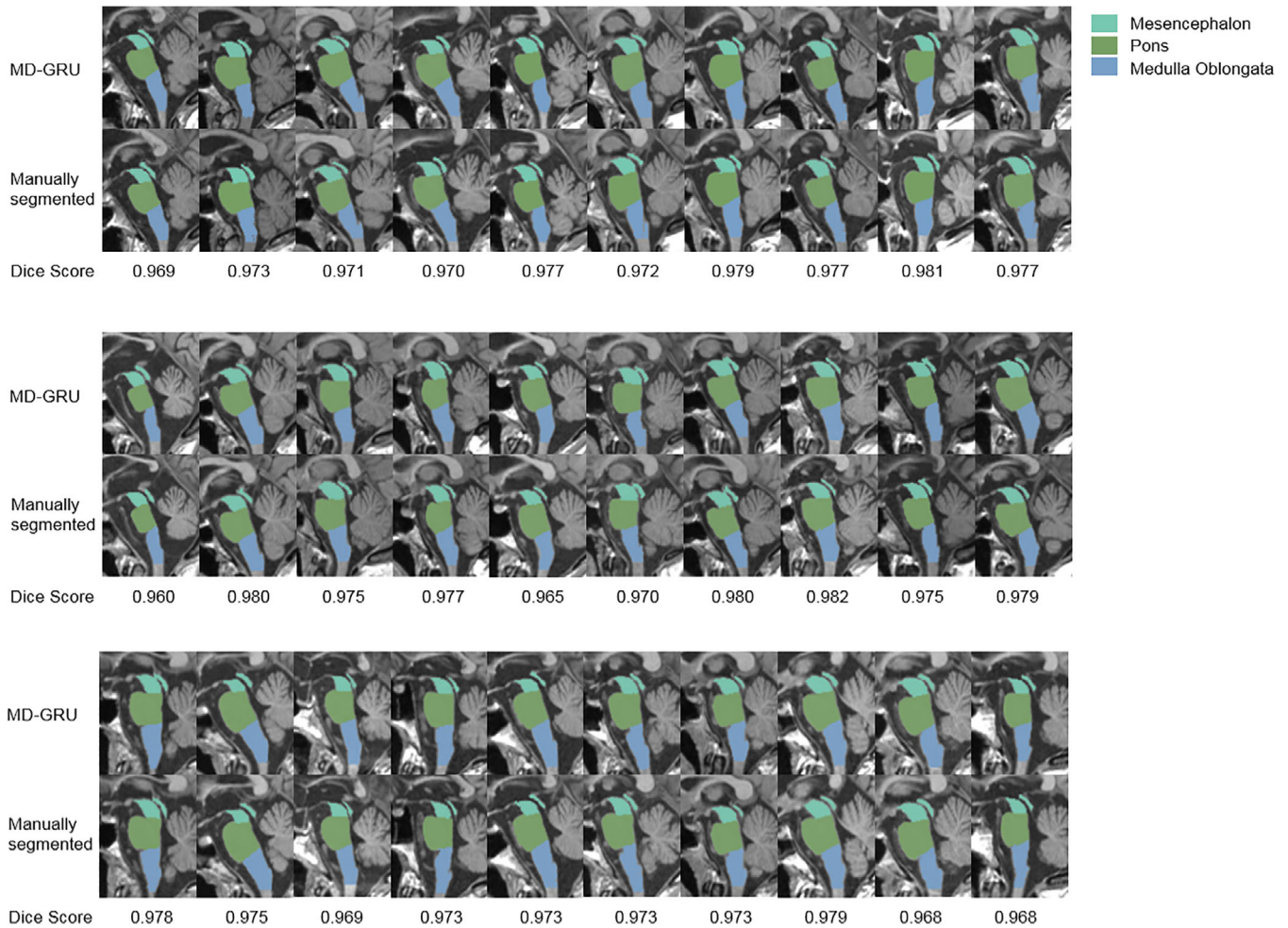


FIGURE 4A Comparison of MD-GRU and manually edited segmentations of 30 exemplary subjects in the accuracy data set ($n = 80$) and corresponding Dice scores for total brainstem volumes [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Mean volumes of the brainstem and its substructures in 80 MS patients

| | FreeSurfer | | Manually edited segmentations | | MD-GRU | |
|-------------------|--------------------------------|----------|--------------------------------|----------|--------------------------------|----------|
| | Mean volume [mm ³] | SD | Mean volume [mm ³] | SD | Mean volume [mm ³] | SD |
| Brainstem | 24,246.86 | 3,041.96 | 24,250.25 | 3,046.03 | 24,418.10 | 3,164.43 |
| Mesencephalon | 5,803.48 | 603.40 | 5,799.21 | 609.89 | 5,910.43 | 613.95 |
| Pons | 14,005.89 | 2,084.54 | 13,994.06 | 2,085.17 | 14,081.53 | 2,213.21 |
| Medulla oblongata | 4,437.50 | 541.35 | 4,456.98 | 531.86 | 4,426.15 | 536.38 |

Note: Mean brainstem and brainstem substructure volumes, obtained by the different segmentation methods with FreeSurfer (Iglesias et al., 2015), manually edited segmentations and MD-GRU.

pons and 0.95 for the medulla oblongata. The corresponding MSD were 0.24 mm for the total BS, 0.30 mm for the mesencephalon, 0.24 mm for the pons and 0.27 mm for the medulla oblongata.

Each MD-GRU-generated segmentation mask was visually inspected for segmentation errors. Figures 4a and 4b show exemplary BS segmentations obtained by MD-GRU.

In general, the segmentation quality was high (see Figures 4a and 4b). Please note slightly imprecise segmentations in a minority of cases without major impact on the respective volumes (see Figure 5). The MD-GRU generated caudal delimitation of the medulla oblongata was

usually within 1–2 slices above or below the corresponding manually defined slice of delimitation containing the exit of the first spinal roots.

The mean volumes/SD of the BS and its substructures obtained by the different segmentation methods are summarized in Table 2.

3.2 | Reproducibility

The results of the test–retest experiment are summarized in Tables 3a and 3b. In brief, the MD-GRU based mean percent BS volume change between the two scans was 0.45% (IQR = 0.42, SD 0.005). For the

FIGURE 4B Exemplary axial and coronal views of the brainstem MD-GRU segmentations [Color figure can be viewed at wileyonlinelibrary.com]

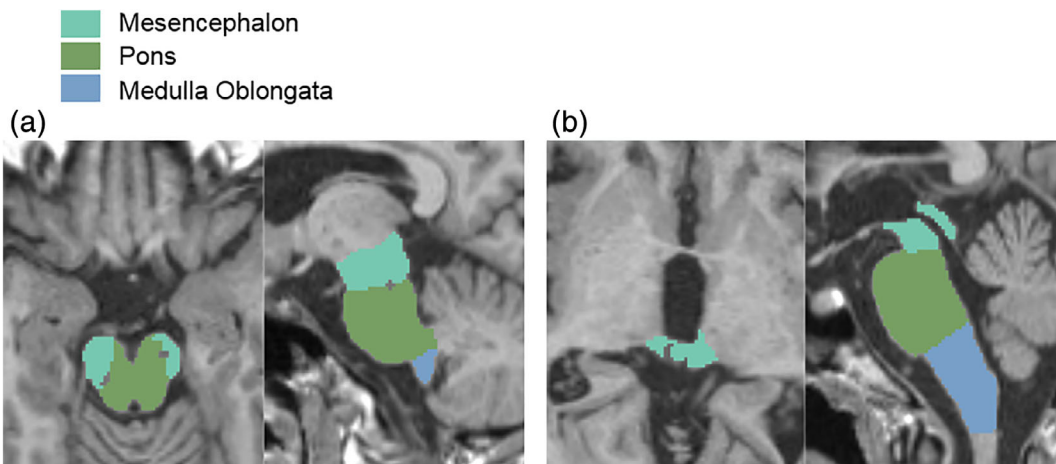
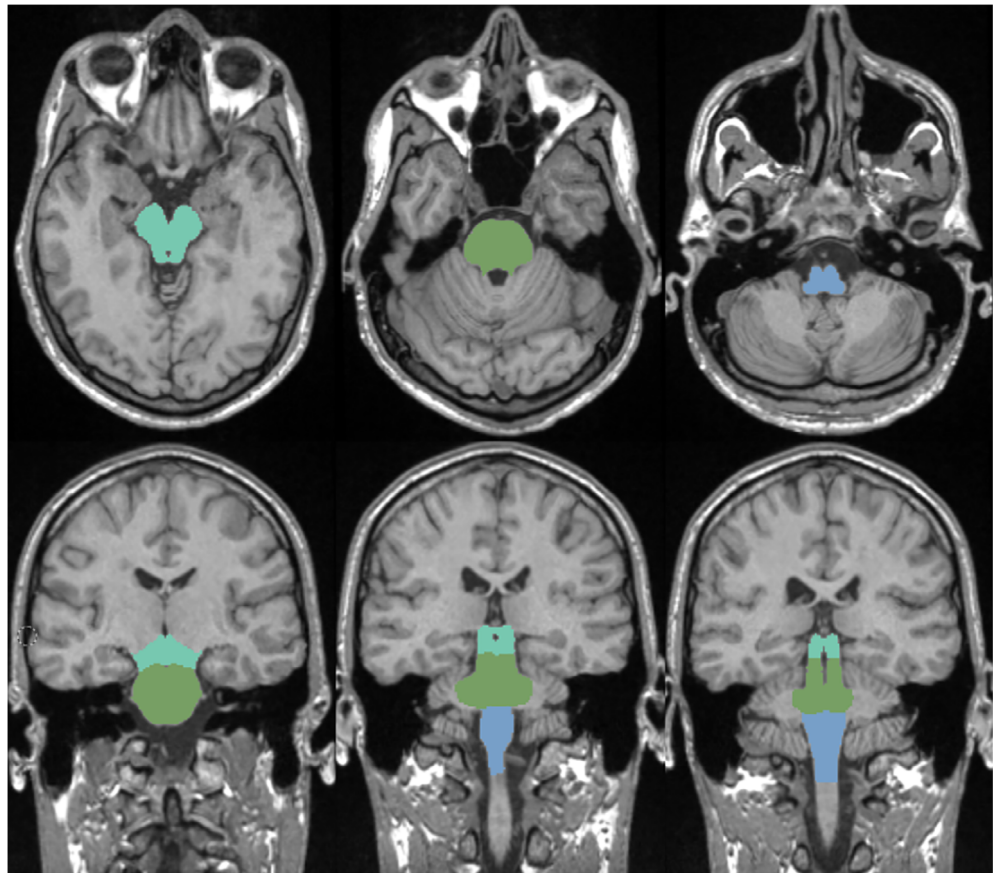


FIGURE 5 Example of random imprecise segmentations of the segmentation algorithm MD-GRU. (a) Missing voxels in the ponto-mesencephalic junction. (b) Segmented voxels in the pineal gland

1.5 T data ($n = 11$), the BS volume change/SD was 0.48%/0.004, for the 3 T data ($n = 22$) 0.43%/0.006.

The corresponding ICC for the MD-GRU based test-retest BS, M, and P volumes were all >0.99 , of MO volumes 0.972.

Figure 6 shows exemplary segmentations of the 11 scan-rescans obtained from the 1.5 T Avanto scanner and corresponding mean percentage changes for total BS volumes.

3.3 | Robustness

For the multicentric data of 50 AD patients, the mean Dice scores/SD and MSD comparing the MD-GRU to the manually edited and FreeSurfer segmentations are shown in Tables 4a and 4b. The mean Dice scores comparing the MD-GRU to the manually edited segmentations were 0.97 for the total BS, 0.94 for the mesencephalon, 0.97

TABLE 3A Mean percentage volume change between test- and retest scans

| | FreeSurfer | | Manually edited segmentations | | MD-GRU | |
|-------------------|----------------------|-------|-------------------------------|-------|----------------------|-------|
| | Mean % volume change | SD | Mean % volume change | SD | Mean % volume change | SD |
| Brainstem | 0.95 | 0.009 | 0.86 | 0.007 | 0.45 | 0.005 |
| Mesencephalon | 1.57 | 0.017 | 1.65 | 0.016 | 0.63 | 0.006 |
| Pons | 0.60 | 0.005 | 0.59 | 0.005 | 0.35 | 0.003 |
| Medulla oblongata | 3.42 | 0.028 | 2.70 | 0.020 | 1.47 | 0.025 |

Note: Mean percentage brainstem volume change between test- and retest scans and SD for segmentations performed with FreeSurfer (Iglesias et al., 2015), manually edited and with the novel segmentation approach MD-GRU, 33 healthy subjects.

TABLE 3B Intraclass correlation coefficient for test–retest segmentations

| | FreeSurfer | | Manually edited segmentations | | MD-GRU | |
|-------------------|------------|-------------|-------------------------------|-------------|--------|-------------|
| | ICC | 95%CI | ICC | 95%CI | ICC | 95%CI |
| Brainstem | >0.99 | 0.987–0.998 | >0.99 | 0.991–0.998 | >0.99 | 0.996–0.999 |
| Mesencephalon | 0.975 | 0.951–0.988 | 0.976 | 0.953–0.988 | >0.99 | 0.993–0.998 |
| Pons | >0.99 | 0.996–0.999 | >0.99 | 0.996–0.999 | >0.99 | 0.999–1 |
| Medulla oblongata | 0.923 | 0.850–0.961 | 0.946 | 0.895–0.973 | 0.972 | 0.941–0.987 |

Note: Intraclass correlation coefficients (ICC; two-way random, absolute agreement) and 95% confidence intervals (CI) for test–retest segmentations performed with FreeSurfer (Iglesias et al., 2015), manually edited and with the novel segmentation approach MD-GRU, 33 healthy subjects.

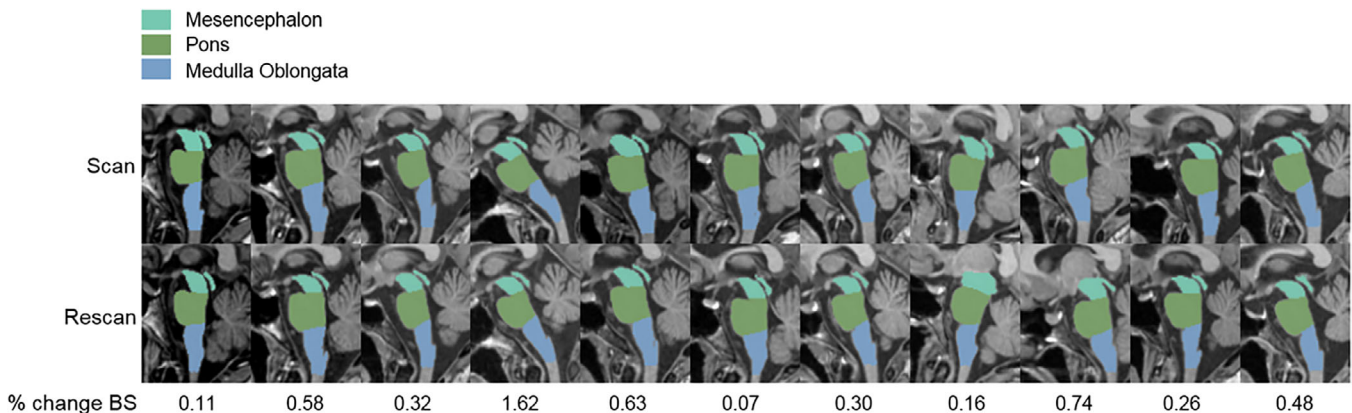


FIGURE 6 MD-GRU segmentations of the eleven 1.5 T scan-rescan experiments and corresponding mean percentage changes for total BS volumes. Please note one outlier with 1.62% BS volume change between the two scans, associated with a pronounced anteversion of the brainstem axis of this subject compared to the other subjects

for the pons and 0.94 for the medulla oblongata. The corresponding MSD were 0.28 mm for the total BS and pons, and 0.33 mm for the mesencephalon and medulla oblongata.

4 | DISCUSSION

The segmentation approach MD-GRU presented in this paper provides accurate, highly reproducible, and robust fully-automated deep learning-based segmentations of the BS and its substructures in HC and patients with MS and AD.

Compared to manually edited segmentations, MD-GRU reduces sources of error by operator-dependent factors like intra- or

interexpert variability. The use of an automatic segmentation method ensures the use of a consistent anatomical border definition as determined by the manually segmented BS templates during training. MD-GRU also dramatically reduces analysis time. With an application time of about 200 s/scan on Nvidia GeForce GTX 1080 GPU, it is not only faster than manual segmentations but also faster than the automated BS segmentation with FreeSurfer requiring 15 min per data set (Iglesias et al., 2015), allowing effective application in large datasets.

Comparing the novel segmentation approach MD-GRU to the manual gold standard segmentations yielded very high Dice scores indicating a high anatomic accuracy of the method. Visual inspection

TABLE 4A Mean Dice scores comparing segmentation methods in 50 AD patients

| | Mean Dice score/ <i>SD</i> comparing MD-GRU vs. manually edited segmentations | Mean Dice score/ <i>SD</i> comparing MD-GRU vs. FreeSurfer segmentations |
|-------------------|---|--|
| Brainstem | 0.97/0.006 | 0.97/0.007 |
| Mesencephalon | 0.94/0.017 | 0.94/0.017 |
| Pons | 0.97/0.004 | 0.97/0.004 |
| Medulla oblongata | 0.94/0.021 | 0.93/0.045 |

Note: Mean Dice scores and *SD* for the total brainstem, mesencephalon, pons, and medulla oblongata comparing segmentations with MD-GRU versus manually edited and FreeSurfer (Iglesias et al., 2015) segmentations.

TABLE 4B Mean surface distances (MSD) comparing segmentation methods in 50 AD patients

| | MSD [mm]/ <i>SD</i> comparing MD-GRU vs. manually edited segmentations | MSD/ <i>SD</i> comparing MD-GRU vs. FreeSurfer segmentations |
|-------------------|--|--|
| Brainstem | 0.28/0.065 | 0.30/0.090 |
| Mesencephalon | 0.33/0.083 | 0.34/0.085 |
| Pons | 0.28/0.042 | 0.28/0.038 |
| Medulla oblongata | 0.33/0.152 | 0.43/0.090 |

Note: Mean surface distance (MSD) (mm) and *SD* for the total brainstem, mesencephalon, pons, and medulla oblongata comparing segmentations with MD-GRU versus manually edited and FreeSurfer (Iglesias et al., 2015) segmentations.

of each MD-GRU generated segmentation mask confirmed precise segmentation even of the most error-prone regions that is, the caudal border of MO and the occipital ponto-medullar transition with MD-GRU, both in healthy controls and in patients with MS and AD. Compared to the manual segmentations, however, MD-GRU tended to slightly over-segment the BS (0.69% for total BS volume compared to manually edited segmentations). The reason for this might lie in our use of selective sampling during training: with selective sampling, we ensured that each second training sample block contained at least one BS voxel, to handle class imbalance between BS and non-BS voxels. This, in turn, might have caused a slight bias in the network to assume voxels as belonging to the BS. Other methods for handling class imbalance, such as integrating Dice loss in the MD-GRU training (Horvath et al., 2018), might be used to mitigate this effect in a future version of our network. Comparing MD-GRU and FreeSurfer based BS segmentations against the manual gold standard indicated a superior accuracy of the segmentations based on MD-GRU compared to FreeSurfer.

A recent study by Velasco-Annis et al. (2018) compared three originally for brain segmentation designed, automated BS segmentation methods (FSL-FIRST (Patenaude et al., 2011), PSTAPLE (Akhond-Asl & Warfield, 2013), FreeSurfer (Fischl et al., 2002; Fischl et al.,

2004; Iglesias et al., 2015) and reported highest reproducibility of BS segmentations performed by FreeSurfer.

Compared to FreeSurfer and to manually edited segmentations, MD-GRU segmentations of all three BS substructures and of the total BS showed consistently lower variability between test- and retest scans. Three scan pairs showed a relatively high percent volume change of 1.62, 2.44, and 1.26% between scans, most likely due to either anatomically pronounced anteversion of the BS compared to all other subjects in the first and poor MR image quality due to motion artifacts in the two others. In the remaining, variability ranged between 0.008 and 0.85% BS volume change between test-retest scans.

In particular, the MD-GRU segmentation approach showed improved reproducibility in the anatomically most challenging region of the BS—the medulla oblongata. Despite relatively high reproducibility compared to other available techniques in earlier studies (Velasco-Annis et al., 2018), FreeSurfer segmentations can present inconsistencies especially in the caudal part of the MO. In this part, as well as in the region at or below the level of the central canal entry into the medulla, the differences between medulla and CSF tissue characteristics may lead to anatomically incorrect segmentations. In the manually edited segmentations that were used for training MD-GRU, a continuous, anatomically correct segmentation around and caudal of the central canal was taken account of with the first spinal root pair as anatomic definition of the caudal delimitation of the medulla. As the first nerve root often exits the medulla in several branches we defined the most cranial axial slice in which both (left and right) first nerve roots were both visible as caudal delimitation of the medulla oblongata.

The MD-GRU generated caudal delimitation of the medulla oblongata was usually within 1–2 slices above or below the corresponding manually defined slice of delimitation containing the exit of the first spinal roots.

In this study, the algorithm was trained on MRI data obtained from a single 1.5 T scanner. However, MD-GRU segmentation reproducibility was shown to be high also for 3 T data.

The robustness of our method was further assessed in a multi-centric, multi-scanner AD dataset. Taken together with our own data, the high Dice scores comparing MD-GRU to manually edited segmentations highlight the applicability and anatomic accuracy of our method in both 1.5 and 3 T settings with different acquisition platforms.

The functional importance for vital functions, the frequent and clinically relevant involvement of the BS or its substructures in neurodegenerative diseases, and the sensitivity to volume changes (Liptak et al., 2008) render the BS and its substructures' volumes a potentially attractive biomarker candidate for neurodegeneration and a potential endpoint candidate for clinical trials. As a further development of our work, we are currently planning to evaluate the clinical importance of BS atrophy in MS and other neurodegenerative diseases with BS involvement.

5 | CONCLUSIONS

This fully automated BS segmentation method provides anatomically accurate, highly reproducible BS segmentations in HC and patients with

MS in 200 s/scan on an Nvidia GeForce GTX 1080 GPU and shows high potential for application in large datasets and longitudinal studies.

ACKNOWLEDGMENTS

Data collection and sharing for the robustness experiment was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

DECLARATION OF INTEREST

Sander L: nothing to disclose. Pezold S holds a grant by the Novartis Research Foundation. Andermatt S, Amann M, Meier D, Wendebourg MJ: nothing to disclose. Sinnecker T has received travel support from Actelion and Roche, and speaker fees from Biogen. Naegelin Y, Granziera C: nothing to disclose. Kappos L: Ludwig Kappos' Institution (University Hospital Basel) received in the last 3 years and used exclusively for research support at the Department: steering committee, advisory board, and consultancy fees from Actelion, Almirall, Bayer, Biogen, Celgene/Receptos, df-mp, Excemed, Genzyme, Japan Tobacco, Merck, Minoryx, Mitsubishi Pharma, Novartis, Roche, sanofi-aventis, Santhera, Teva, Vianex and royalties for Neurostatus-UHB products. For educational activities the institution received payments and honoraria from Allergan, Almirall, Baxalta, Bayer, Biogen, CSL-Behring, Desitin, Excemed, Genzyme, Merck, Novartis, Pfizer, Roche, Sanofi-Aventis, Teva. Wuerfel J: CEO of MIAC AG Basel, Switzerland. He served on scientific advisory boards of Actelion, Biogen, Genzyme-Sanofi, Novartis, and Roche. He is or was supported by grants of the EU (Horizon2020), German Federal Ministries of Education and Research (BMBF) and of Economic Affairs and Energy (BMWi). Cattin P: nothing to disclose. Schlaeger R is supported by

the Swiss National Science Foundation (MHV program, PMPDP3 171391), the University of Basel, and the Swiss MS Society.

ORCID

Regina Schlaeger  <https://orcid.org/0000-0003-2056-5765>

REFERENCES

- Akhondi-Asl, A., & Warfield, S. K. (2013). Simultaneous truth and performance level estimation through fusion of probabilistic segmentations. *IEEE Transactions on Medical Imaging*, 32(10), 1840–1852. <https://doi.org/10.1109/TMI.2013.2266258>
- Andermatt, S., Pezold, S., & Cattin, P. C. (2016). Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-data. In G. Carneiro et al. (Eds.), *Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science* (Vol 10008). Cham: Springer. https://doi.org/10.1007/978-3-319-46976-8_15
- Andermatt, S., Pezold, S., & Cattin, P. C. (2018). Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. In A. Crimi, S. Bakas, H. Kuijf, B. Menze, & M. Reyes (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. BrainLes 2017. Lecture notes in computer science* (p. 10670). Cham: Springer. https://doi.org/10.1007/978-3-319-75238-9_3
- Bendfeldt, K., Kuster, P., Traud, S., Egger, H., Winkhofer, S., Mueller-Lenke, N., ... Borgwardt, S. J. (2009). Association of regional gray matter volume loss and progression of white matter lesions in multiple sclerosis - a longitudinal voxel-based morphometry study. *NeuroImage*, 45(1), 60–67. <https://doi.org/10.1016/j.neuroimage.2008.10.006>
- Cho K., van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. , & Bengio Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv: 1406.1078
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., ... Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3), 341–355.
- Fischl, B., Salat, D. H., van der Kouwe, A. J., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23(Suppl 1), S69–S84.
- Ghorayeb, I., Yekhelef, F., Chrysostome, V., Balestre, E., Bioulac, B., & Tison, F. (2002). Sleep disorders and their determinants in multiple system atrophy. *Journal of Neurology, Neurosurgery, and Psychiatry*, 72(6), 798–800.
- Grinberg, L. T., Rueb, U., Alho, A. T., & Heinsen, H. (2010). Brainstem pathology and non-motor symptoms in PD. *Journal of the Neurological Sciences*, 289(1–2), 81–88. <https://doi.org/10.1016/j.jns.2009.08.021>
- Grinberg, L. T., Rueb, U., Ferretti, R. E., Nitrini, R., Farfel, J. M., Polichiso, L., ... Heinsen, H. (2009). The dorsal raphe nucleus shows phospho-tau neurofibrillary changes before the transentorhinal region in Alzheimer's disease. A precocious onset? *Neuropathology and Applied Neurobiology*, 35(4), 406–416.
- Herlihy, A. H., Hajnal, J. V., Curati, W. L., Virji, N., Oatridge, A., Puri, B. K., & Bydder, G. M. (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). *American Journal of Neuroradiology*, 22(5), 896–904.
- Horvath A. , Tsagkas C. , Andermatt S. , Pezold S. , Parmar K. , & Cattin P. (2018). Spinal cord gray matter-white matter segmentation on magnetic resonance AMIRA images with MD-GRU. arXiv:1808.02408.

- Iglesias, J. E., van Leemput, K., Bhatt, P., Casillas, C., Dutt, S., Schuff, N., ... Fischl, B. (2015). Bayesian segmentation of brainstem structures in MRI. *NeuroImage*, 113, 184–195. <https://doi.org/10.1016/j.neuroimage.2015.02.065>
- Keshavan, A., Paul, F., Beyer, M. K., Zhu, A. H., Papinutto, N., Shinohara, R. T., ... Henry, R. G. (2016). Power estimation for non-standardized multisite studies. *NeuroImage*, 134, 281–294. <https://doi.org/10.1016/j.neuroimage.2016.03.051>
- Kim, Y., Kim, Y. E., Park, E. O., Shin, C. W., Kim, H. J., & Jeon, B. (2018). REM sleep behavior disorder portends poor prognosis in Parkinson's disease: A systematic review. *Journal of Clinical Neuroscience*, 47, 6–13. <https://doi.org/10.1016/j.jocn.2017.09.019>
- Liptak, Z., Berger, A. M., Sampat, M. P., Charil, A., Felsovalyi, O., Healy, B. C., ... Guttmann, C. R. (2008). Medulla oblongata volume: A biomarker of spinal cord damage and disability in multiple sclerosis. *American Journal of Neuroradiology*, 29(8), 1465–1470. <https://doi.org/10.3174/ajnr.A1162>
- McDonald, W. I., Compston, A., Edan, G., Goodkin, D., Hartung, H. P., Lublin, F. D., ... Wolinsky, J. S. (2001). Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology*, 50(1), 121–127.
- Naidich, T. P., Duvernoy, H. M., Delman, B. N., Sorensen, A. G., Kollias, S. S., & Haacke, E. M. (2009). *Duvernoy's atlas of the human brain stem and cerebellum* (p. 54). Wien New York: Springer.
- Nieuwenhuys, R. (1985). *Chemoarchitecture of the brain*. Berlin Heidelberg New York Tokyo: Springer.
- Noseworthy, J. H., Lucchinetti, C., Rodriguez, M., & Weinshenker, B. G. (2000). Multiple sclerosis. *New England Journal of Medicine*, 343(13), 938–952.
- Patenaude, B., Smith, S. M., Kennedy, D. N., & Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56(3), 907–922. <https://doi.org/10.1016/j.neuroimage.2011.02.046>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929–1958.
- Tanaka, N., Abe, T., Kojima, K., Nishimura, H., & Hayabuchi, N. (2000). Applicability and advantages of flow artifact-insensitive fluid-attenuated inversion-recovery MR sequences for imaging the posterior fossa. *American Journal of Neuroradiology*, 21(6), 1095–1098.
- Velasco-Annis, C., Akhondi-Asl, A., Stamm, A., & Warfield, S. K. (2018). Reproducibility of brain MRI segmentation algorithms: Empirical comparison of local MAP PSTAPLE, FreeSurfer, and FSL-FIRST. *Journal of Neuroimaging*, 28(2), 162–172. <https://doi.org/10.1111/jon.12483>
- Wang, J. Y., Ngo, M. M., Hessel, D., Hagerman, R. J., & Rivera, S. M. (2016). Robust machine learning-based correction on automatic segmentation of the cerebellum and brainstem. *PLoS ONE*, 11(5), e0156123. <https://doi.org/10.1371/journal.pone.0156123>
- Warabi, Y., Hayashi, K., Nagao, M., & Shimizu, T. (2017). Marked widespread atrophy of the cerebral cortex and brainstem in sporadic amyotrophic lateral sclerosis in a totally locked-in state. *British Medical Journal Case Reports*. pii: bcr2016218952. <https://doi.org/10.1136/bcr-2016-218952>
- Weier, K., Penner, I. K., Magon, S., Amann, M., Naegelin, Y., Andelova, M., ... Sprenger, T. (2014). Cerebellar abnormalities contribute to disability including cognitive impairment in multiple sclerosis. *PLoS ONE*, 9(1), e86916. <https://doi.org/10.1371/journal.pone.0086916>
- Williams, D. R., & Lees, A. J. (2009). Progressive supranuclear palsy: Clinicopathological concepts and diagnostic challenges. *Lancet Neurology*, 8(3), 270–279. [https://doi.org/10.1016/S1474-4422\(09\)70042-0](https://doi.org/10.1016/S1474-4422(09)70042-0)
- Zeiler M.D. (2012). ADADELTA: An adaptive learning rate method. arXiv: 1212.5701.

How to cite this article: Sander L, Pezold S, Andermatt S, et al. Accurate, rapid and reliable, fully automated MRI brainstem segmentation for application in multiple sclerosis and neurodegenerative diseases. *Hum Brain Mapp*. 2019;40: 4091–4104. <https://doi.org/10.1002/hbm.24687>

APPENDIX

A. GUIDELINES USED IN THE MANUALLY EDITED SEGMENTATIONS OF THE MESENCEPHALON, PONS, AND MEDULLA OBLONGATA BASED ON FREESURFER PRESEGMENTATIONS

Segmentations were performed mainly in the axial and sagittal plane and were visually controlled in the coronal plane.

1. In axial and sagittal view, the anatomically correct cranial limitations of the midbrain toward the epiphysis was traced.
2. For the medullo-pontine transition, the presegmented transition zone was followed occipitally to the most superior axial slice; presegmented voxels of the MO cranial of this slice were deleted.
3. For caudal delineation of the MO, the most cranial axial slice was identified in which both (left and right) first nerve roots were both visible (see Figure 2).
4. Segmentation irregularities were visually controlled in all three planes. Missing voxels inside the anatomically defined substructures were added. Presegmented voxels outside the BS, in particular in the cerebellum, pendunculus cerebelli superior or pineal gland, were deleted.

B. GUIDELINES USED IN THE MANUAL SEGMENTATIONS OF THE MESENCEPHALON, PONS, AND MEDULLA OBLONGATA

1. For caudal delineation of the MO, the most cranial axial slice was identified in which the left and right first nerve roots were both visible (see Figure 2).
2. For cranial delimitation of the MO, in the midsagittal view, the pontomedullary sulcus (see Figure 2) was identified and followed in both directions.
3. For cranial delimitation of the pons, in the midsagittal plane, the ponto-mesencephalic junction was identified with posterior delimitation below the quadriminal plate (see Figure 2). Segmenting was continued toward more lateral planes. In sagittal and axial slices, the upper anterior delimitation was set below the exit of the third cranial nerve. The separation between mesencephalon and pons was continued from lateral toward median slices.

4. In axial and sagittal view (see Figure 2), the anatomically correct cranial limitations of the midbrain toward the pineal gland was segmented with no segmented voxels above the superior colliculus. In sagittal slices, the anterior delineation was marked posterior of the mammillary bodies on both sides.
5. Segmentation irregularities were visually controlled in all three planes with respect to the anatomical contrasts.

C. MANUALLY EDITED SEGMENTATION

The intra-class correlation coefficients (two-way random, absolute agreement) ICC/95%CI of inter-rater reliability of the manual segmentation of the BS and its substructures M and p was $>0.99/1.0$, for MO $0.998/0.992-0.999$.

The COV of the BS substructures were: $COV_M = 0.035\%$, $COV_P = 0.068\%$, $COV_{MO} = 0.495\%$.