Replicability, repeatability, and long-term reproducibility of cerebellar morphometry

Peter Sörös[1,2,✉], **Louise Wölk**[1], **Carsten Bantel**[2,3], **Anja Bräuer**[2,4], **Frank Klawonn**[5,6 *], and **Karsten Witt**[1,2 *]

[1]Department of Neurology, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
[2]Research Center Neurosensory Science, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
[3]Anesthesiology, Critical Care, Emergency Medicine, and Pain Management, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
[4]Department of Anatomy, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
[5]Biostatistics, Helmholtz Centre for Infection Research, Braunschweig, Germany
[6]Department of Computer Science, Ostfalia University of Applied Sciences, Wolfenbüttel, Germany
[*]These authors share senior authorship.

Abstract

To identify robust and reproducible methods of cerebellar morphometry that can be used in future large-scale structural MRI studies, we investigated the replicability, repeatability, and long-term reproducibility of three fully-automated software tools: FreeSurfer, CERES, and CNN. Replicability was defined as computational replicability, determined by comparing two analyses of the same high-resolution MRI data set performed with identical analysis software and computer hardware. Repeatability was determined by comparing the analyses of two MRI scans of the same participant taken during two independent MRI sessions on the same day for the Kirby-21 study. Long-term reproducibility was assessed by analyzing two MRI scans of the same participant in the longitudinal OASIS-2 study. We determined percent difference, the image intraclass correlation coefficient, the coefficient of variation, and the intraclass correlation coefficient between two analyses. Our results show that CERES and CNN use stochastic algorithms that result in surprisingly high differences between identical analyses for CNN and small differences for CERES. Changes between two consecutive scans from the Kirby-21 study were less than ±5% in most cases for FreeSurfer and CERES (i.e., demonstrating high repeatability). As expected, long-term reproducibility was lower than repeatability for all software tools. In summary, CERES is an accurate, as demonstrated before, and reproducible tool for fully-automated segmentation and parcellation of the cerebellum. We conclude with recommendations for the assessment of replicability, repeatability, and long-term reproducibility in future studies on cerebellar structure.

brain | cerebellum | segmentation | parcellation | FreeSurfer | CERES | MRI

Correspondence: ***peter.soros@gmail.com***

Introduction

Physiology and pathophysiology of the cerebellum have received growing attention in basic and clinical neurosciences (1–3). Early nineteenth century neuroscientists, in particular, Luigi Rolando and Pierre Flourens, have established the crucial role of the cerebellum in motor control (4) and, more specifically, motor coordination (5). More recently, the role of motor learning (6, 7) and the non-motor functions of the cerebellum (8) have been investigated in greater detail. The cerebellar contributions to various cognitive (9) and emotional functions (10) as well as timing (11, 12) have been acknowledged. Moreover, structural changes of the cerebellum in healthy aging (13) and neurodegenerative disease (14, 15) have been studied.

The advent of magnetic resonance imaging (MRI) has opened the door to quantitative, non-invasive investigations of cerebellar morphology. The segmentation of the cerebellum into gray and white matter and the parcellation into lobes and single lobules turned out to be challenging because of its tightly folded structure, consisting of numerous small folia, the equivalent of cerebral gyri. Moreover, the anatomy of the cerebellum is characterized by pronounced inter-individual differences (16, 17). Manual slice-by-slice labeling of MRIs by an expert neuroanatomist is considered the gold standard of cerebellar research (18). Nevertheless, manual segmentation and parcellation have major disadvantages, requiring expert knowledge and being observer-dependent and time-consuming, and are not feasible in large-scale studies.

To overcome the limitations of manual identification of cerebellar structures, several fully automated methods for cerebellar morphometry have been developed and made publicly available (for a review, see Carass et al. (19)). The results of several of these methods have been compared with manually labeled adult and pediatric cerebellar data sets (19). In this comparison, an improved version of the patch-based multi-atlas segmentation tool CERES (CEREbellum Segmentation) (20) exhibited highest accuracy and outperformed established methods, such as the MATLAB toolbox SUIT (Spatially Unbiased Infra-tentorial Template) (16, 21). While the accuracy of CERES and other methods have been established, the reproducibility of fully automated cerebellar morphometry has not been determined so far.

In the present study we investigate the replicability, repeatability, and long-term reproducibility of cerebellar morphometry using three independent MRI data sets and three software packages based on different computational approaches. The definitions of replicability, repeatability, and reproducibility follow the suggestions by Nichols et al. (22). Replicability is defined as computational or analysis replicability, determined by comparing two analyses of the same MRI data set performed with identical analysis software and computer hardware. Repeatability is determined by comparing the analyses of two MRI scans of the same participant taken during two independent MRI sessions on the same day. Long-term reproducibility, finally, is assessed by analyzing two MRI scans of the same participant in a longitudinal study. We decided to test the following three software packages: (1) FreeSurfer, an established and widely used approach of subcortical segmentation, based on a probabilistic atlas, which performs cerebellar segmentation, but not parcellation (23), (2) CERES, a recent segmentation and parcellation method based on a multi-atlas label fusion technique (20), the most accurate software tool in the comparison by Carass et al. (19), and (3) CNN, a very recent and promising parcellation approach based on convolutional neural networks (24), not included in the comparison by Carass et al. (19). The ultimate aim of this study is to identify robust and reproducible methods of fully automated cerebellar morphometry that can be used in MRI studies with large sample sizes.

Methods

MRI data. For this study, three independent data sets of T1-weighted MRIs of the entire brain have been analyzed with three different fully automated software packages: FreeSurfer 7.1.0 (23), CERES (20), and CNN (24).

Replicability: ChroPain2 study. To investigate the analysis replicability of cerebellar morphometry, we performed two separate, but identical analyses of high-resolution structural MRIs of 23 healthy individuals (17 women, 6 men) who served as control participants for the ChroPain2 study. Inclusion and exclusion criteria have been published previously (25). Mean age ± standard deviation was 51 ± 10 years (minimum: 30 years, maximum: 66 years). All participants provided written informed consent for participation in this study. The study was approved by the Medical Research Ethics Board, University of Oldenburg, Germany (2017-059) and was preregistered with the German Clinical Trials Register (DRKS00012791)[1].

MR images of the entire brain were acquired in the Neuroimaging Unit, School of Medicine and Health Sciences, University of Oldenburg[2], on a research-only Siemens MAGNETOM Prisma whole-body scanner (Siemens, Erlangen, Germany) at 3 Tesla with a 64-channel head/neck receive-array coil. A 3-dimensional high-resolution and high-contrast T1-weighted magnetization prepared rapid gradient echo (MPRAGE) sequence was used (26). Imaging parameters were: TR (repetition time; between two successive inversion pulses): 2000 ms, TE (echo time): 2.07 ms, TI (inversion time): 952 ms, flip angle: 9°, isotropic voxel size: $0.75 \times 0.75 \times 0.75$ mm^3, 224 sagittal slices, k-space interpolation-based in-plane acceleration (GRAPPA) with an acceleration factor of 2 (27), time of acquisition: 6:16 min. Siemens' prescan normalization filter was used for online compensation of regional signal inhomogeneities.

Repeatability: Kirby-21 study. To investigate repeatability of cerebellar morphometry, we analyzed data from the Kirby-21 multi-modal MRI reproducibility study (28), performed at the F.M. Kirby Research Center for Functional Brain Imaging, Kennedy Krieger Institute, Baltimore, MD, USA. For this study, each participant received two identical MRI examinations on the same day, each consisting of several sequences, including a T1-weighted MPRAGE sequence. After the first examination, participants left the scanner room for a short break and were then repositioned and scanned with the identical imaging protocol a second time. The time interval between the two T1-weighted images was approximately 1 hour. MRIs were acquired from 21 individuals (10 women, 11 men) with no history of neurological disorders. Mean age ± standard deviation was 32 ± 9 years (minimum: 22 years, maximum: 61 years). For a detailed description of the entire study, see (28). The data set is publicly available for download[3] and has been used in several studies on the reproducibility of MRI analyses (e.g., (29, 30)).

MR images of the entire brain were acquired at 3 Tesla using a Philips Achieva MR scanner (Philips Healthcare, Best, The Netherlands) with an 8-channel receive-array head coil. Imaging parameters for the MPRAGE sequence were: TR (between two successive gradient echoes): 6.7 ms, TE: 3.1 ms, TI: 842 ms, flip angle: 8°, voxel size: $1 \times 1 \times 1.2$ mm^3, image domain-based in-plane acceleration (SENSE) with an acceleration factor of 2, duration: 5:56 min.

Long-term reproducibility: OASIS-2 study. To investigate long-term reproducibility of cerebellar morphometry, we performed analyses of MR images acquired for the Open Access Series of Imaging Studies (OASIS-2) (31), performed at the Washington University School of Medicine, St. Louis, MO, USA. The OASIS-2 study comprises longitudinal MR examinations of patients with Alzheimer's disease and healthy controls. For the present study of cerebellar morphometry, we analyzed the data of 72 individuals (50 women, 22 men) who remained cognitively unimpaired throughout the study, as demonstrated by a Clinical Dementia Rating (CDR) score of 0 (32). Mean age at inclusion ± SD was 75 ± 8 years (minimum: 60 years, maximum: 93 years). For the OASIS-2 study, participants received 2-5 MRI examinations months or years apart; each MRI examination consisted of 3-4 T1-weighted MRI scans. For the present study, we only considered the first two MRI examinations of each participant. If more than one MRI scan was available for one examination, we chose the first one. The mean interval ± SD between the two MRIs was 738 ± 249 days (minimum: 182, maximum: 1510 days). All MRIs were obtained with the same scanner with identical pulse sequences. For a detailed description of the study and the CDR scale, see Marcus et al. (31). OASIS-2 data sets are publicly available for download.[4]

MR images of the entire brain were acquired on a Siemens Vision whole-body scanner (Siemens, Erlangen, Germany) at 1.5 Tesla. Imaging parameters for the MPRAGE sequence were: TR (between two successive gradient echoes): 9.7 ms, TE: 4 ms, TI 20 ms, flip angle: 10°, voxel size: $1 \times 1 \times 1.25$ mm^3, 128 sagittal slices.

[1] www.drks.de
[2] uol.de/en/medicine/biomedicum/neuroimaging-unit
[3] www.nitrc.org/projects/multimodal
[4] www.oasis-brains.org

Table 1. Cerebellar regions parcellated in CERES and CNN. CERES determines the entire volume (cm^3), the mean cortical thickness (mm), and the gray matter volume (cm^3) of each region. CNN determines the volume (mm^3) of each region. CERES and CNN make use of the cerebellar nomenclature proposed by Schmahmann et al. (34). In addition, the traditional names of vermical regions according to the Terminologia Anatomica (35) are listed. The less common names of hemispheric regions were omitted.

CERES	CNN		Terminologia Anatomica
	Vermis	Hemisphere	Vermis
Anterior Lobe			
Lobules I-II		Lobule I-III	Lobulus I: Lingula Lobulus II: Centralis
Lobule III			Lobulus III: Centralis
Lobule IV		Lobule IV	Lobulus IV: Culmen
Lobule V		Lobule V	Lobulus V: Culmen
Posterior Lobe			
Lobule VI	Vermis VI	Lobule VI	Lobulus VI: Declive
Crus I		Lobule VIIAf	Lobulus VIIA: Folium vermis
Crus II	Vermis VII	Lobule VIIAt	
Lobule VIIB		Lobule VIIB	Lobulus VIIB: Tuber
Lobule VIIIA	Vermis VIII	Lobule VIIIA	Lobulus VIIIA: Pyramis
Lobule VIIIB		Lobule VIIIB	Lobulus VIIIB: Pyramis
Lobule IX	Vermis IX	Lobule IX	Lobulus IX: Uvula
Flocculonodular Lobe			
Lobule X	Vermis X	Lobule X	Lobulus X: Nodulus

Data analysis. FreeSurfer and CNN analyses were performed on the high-performance computer cluster CARL[5] at the University of Oldenburg, Germany, running Red Hat Enterprise Linux. CERES was run through the online MRI Brain Volumetry System volBrain (33). CERES can only be used through the volBrain website and was not available for installation on our computer cluster. All analyses were done fully automated. Manual editing of output images was not performed, because the aim of this study was to assess reproducibility of cerebellar morphometry for future use in large-scale data sets.

FreeSurfer. For automated analysis of subcortical structures, including the cerebellum, the FreeSurfer 7.1.0 image analysis suite was used, which is freely available for download online[6] (36). Processing was done with the `recon-all -all` command. For the ChroPain2 and the Kirby-21 data sets, the `-3T` and `-mprage` flags were used. For the OASIS-2 data sets, the `-mprage` flag was used. Processing started with automated transformation to Talairach space, followed by intensity normalization of the output images and removal of non-brain tissue using a hybrid approach that combines watershed algorithms and deformable surface models (37). During segmentation, a neuroanatomical label is assigned to all voxels of the T1-weighted MRI based on a probabilistic atlas, derived from a manually labeled training set (23), using a Bayesian approach. Details of atlas construction, registration of the probabilistic atlas to the individual MRI, and segmentation based on the assumption that spatial distribution of labels can be approximated by an anisotropic non-stationary Markov random field are given by Fischl et al. (23). FreeSurfer reports the volumes of the left and right cerebellar cortex and the left and right cerebellar white matter (Figure 1).

Parallelization was not used, all processes were run on a single computer core of a high-performance computer cluster. Processing of the first MRI of the first MR examination of participant OAS2_0095 failed due to an error during topology correction (with and without the `-mprage` flag). We analyzed the second MRI of the first examination instead; processing finished without error.

CERES (CEREbellum Segmentation). CERES is an automated pipeline for cerebellar segmentation and parcellation (20) and is part of the volBrain Automated MRI Brain Volumetry System (33). In brief, CERES receives an anonymized T1-weighted MRI brain volume in NIfTI format through the volBrain website[7], performs image preprocessing, and labels cerebellar voxels based on Optimized PatchMatch Label fusion (38).

Preprocessing includes (1) denoising (39), (2) bias field correction using the N4 algorithm (40), (3) linear registration to the MNI152 standard space template using Advanced Normalization Tools (ANTs) (41, 42), (4) cropping of the cerebellum area, (5) non-linear registration to the cropped MNI152 template using ANTs (41, 42), and (6) local intensity normalization. Labelling of cerebellar voxels was performed with non-local patch-based label fusion, a multi-atlas segmentation technique combining segmentations from multiple reference atlases, initially developed for hippocampal segmentation (43, 44). The atlases were created based on manually segmented high-resolution MR images from 5 healthy volunteers (3 women, 2 men, aged 29–57 years) (18), available for download[8]. CERES determines the entire volume, cortical thickness, and gray matter volume of all regions listed in Table 1, separately for the left and right side of the cerebellum (Figure 2). Of note, we have used the publicly available first version of CERES. All analysis steps have been determined by the developers; changes of analysis methods or parameters are not possible. In the study on accuracy of cerebellar morphometry performed by Carass et al. (19), an improved version (CERES2) was tested, which employs an improved intensity normalization method and a systematic error correction step; CERES2 has not been released for public use so far.

CNN (cerebellum parcellation with Convolutional Neural Networks). Han et al. (24) developed a method using convolutional neural networks for cerebellar parcellation (CNN). CNN processes T1-weighted images of the brain in NIfTI format, preferentially acquired with an MPRAGE sequence. A Singularity image of this software is publicly available[9]. We ran this image on University of Oldenburg's HPC cluster using Singularity 2.6.

As suggested by the developers, all images were first cropped with the `robustfov` command provided by FSL[10] to remove the lower head and neck in MRIs with large field-of-view. Processing within CNN included (1) estimation of a

[5] uol.de/en/school5/sc/high-perfomance-computing/hpc-facilities/carl

[6] https://surfer.nmr.mgh.harvard.edu/fswiki/rel7downloads

[7] volbrain.upv.es

[8] cobralab.ca/atlases/Cerebellum

[9] hwww.iacl.ece.jhu.edu/index.php/Cerebellum_CNN

[10] fsl.fmrib.ox.ac.uk/fsl/fslwiki/InitialProcessing

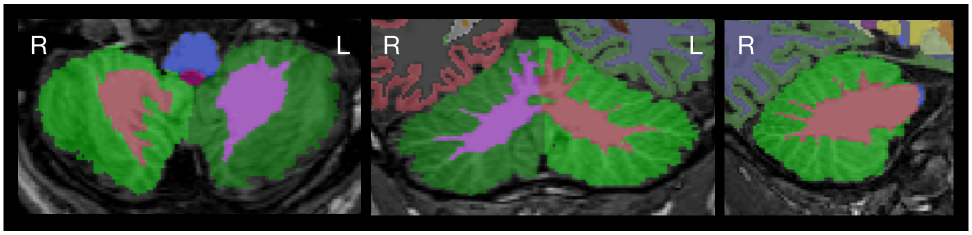

Fig. 1. Cerebellar segmentation as determined by FreeSurfer. Images were created with FSLeyes. The left image shows a horizontal, the middle image a coronal, and the right image a sagittal section of the cerebellum. Images are in radiological convention (the left side of the cerebellum is on the right side of the image). Cerebellar cortex is displayed in green color.

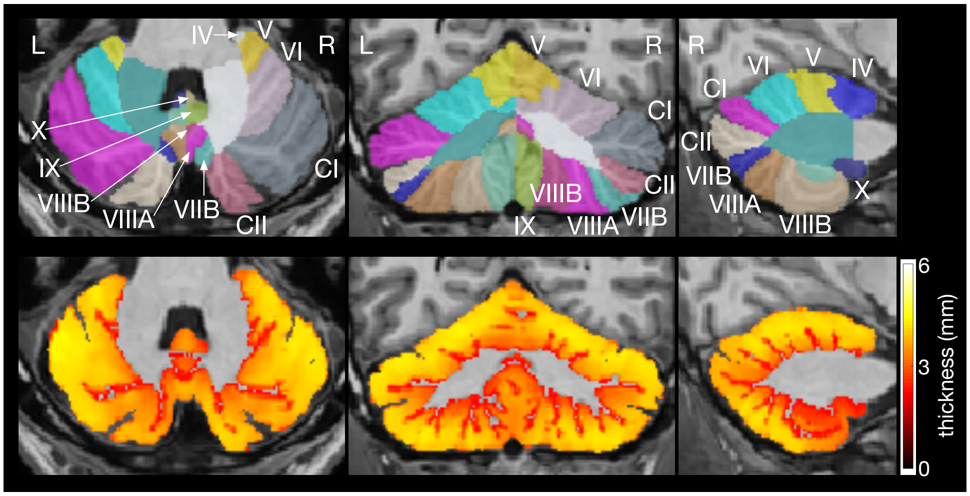

Fig. 2. Cerebellar parcellation (upper row) and thickness of cerebellar cortex (lower row) as determined by CERES. The left images show horizontal, the middle images coronal, and the right images sagittal sections of the cerebellum. Images were produced by CERES and are in neurological convention (the left side of the cerebellum is on the left side of the image). The roman numerals of the cerebellar lobules were added. CI denotes Crus I; CII, Crus II.

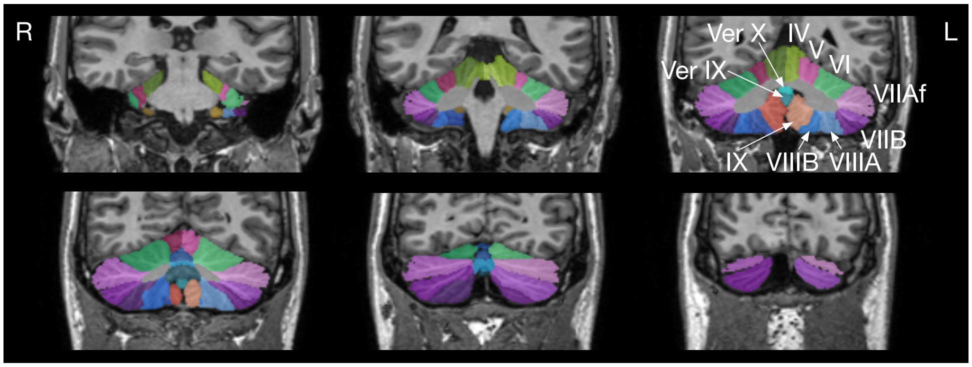

Fig. 3. Cerebellar parcellation as determined by CNN. The upper left image is the most anterior coronal section. Images were produced by CNN. The upper part of each image was cropped and the roman numerals of the cerebellar lobules were added. Ver IX denotes Vermis IX I; Ver X, Vermis X. Images are in radiological convention (left side of the cerebellum is on the right side of the image).

brain mask using Robust Brain Extraction (ROBEX)[11] (45) for subsequent bias field correction, (2) bias field correction using the N4 algorithm (40), (3) linear registration to MNI space using the 1 mm isotropic ICBM 2009c nonlinear symmetric template[12] using ANTs (41, 42), (4) parcellation of the cerebellum as described by Han et al. (24), and (5) transformation of the parcellation into original space using ANTs with the `MultiLabel` interpolation. For cerebellar parcellation, CNN employs two three-dimensional convolutional neural networks. First, a locating network is used to predict a bounding box around the cerebellum. Second, a parcellating network is used to parcellate the cerebellum using the entire region within the bounding box (24). CNN employs the TensorFlow software library for Python[13] and the GNU Parallel tool (46). The cerebellar regions identified by CNN are summarized in Table 1; CNN reports the entire volume of the left and right lobules and the vermis regions in the midline (Figure 3). Of note, all analysis steps have been determined by the developers; changes of analysis methods or parameters are not possible.

Unexpectedly, we found large differences between the first and second analysis of the high-resolution ChroPain2 data set using CNN. To investigate CNN's analysis replicability with a different data set using larger voxel sizes, we performed another two separate analyses of the T1-weighted MRIs available in the Kirby-21 study using Singularity 3.4 (in the meantime, Singularity 2.6 had been deleted from the cluster). For scan KKI2009-33, the locating network of CNN predicted an incorrect bounding box in one of these analyses, placing it well above the cerebellum, leading to erroneous results of the parcellating network. This scan was excluded from the assessment of analysis replicability in the Kirby-21 study (Figure 4C).

Statistical analysis. For further data analyses, we calculated the percent difference between the first and second analysis of one MRI (ChroPain2 study) or the first and second

[11] www.nitrc.org/projects/robex/
[12] www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009
[13] www.tensorflow.org

MRI (Kirby-21 and OASIS-2 studies), and determined the coefficient of variation (CV) and the intraclass correlation coefficient (ICC) for each cerebellar region. We also computed the image intraclass correlation coefficient (I2C2) (47) for cerebellar parcellations obtained by CERES.

Coefficient of variation. The coefficient of variation (CV) describes the level of variability within a sample independently of the absolute values of the observations. To calculate the CV, the standard deviation across all measurements of one parameter (including the results of the first and second analysis of one MRI or the analyses of the first and second MRI) was divided by the (absolute) mean across all measurements and expressed as %. In addition, the lower and upper 95% confidence intervals were estimated using R for Windows (48).

Intraclass correlation coefficient. The intraclass correlation coefficient (ICC) is a measure of within-subject variability relative to between-subject variability. ICC estimates and their lower and upper 95% confidence intervals were calculated using the R package *psych*[14] and the function `ICC` (49). Following the suggestions of Liljequist et al. (50) we first calculated all three single-measurement ICCs (51, 52). The results of all three formulas were very similar, indicating the absence of bias (systematic error). Hence, we report the one-way random effects, absolute agreement, single measurement ICC according to McGraw and Wong (52) or the ICC(1,1) according to Shrout and Fleiss (51). ICC confidence intervals indicate poor reliability (<0.5), moderate reliability (0.5 - 0.75), good reliability (0.75 - 0.9), or excellent reliability (>0.9) (53).

Image intraclass correlation coefficient. The I2C2 has been developed as a global measure of reliability for imaging data (47). The (I2C2)[15] was calculated for all cerebellar parcellations obtained by CERES for the Kirby-21 and OASIS-2 data sets using the *I2C2* package version 0.2.4 (47) for Neuroconductor (54).

First, all parcellated images created by CERES were split into 24 image files containing one parcellation only (labels 1-12 for the left cerebellum, labels 101-112 for the right cerebellum). Then, .nii files were imported into R using the `readnii` function of the *neurobase* package for Neuroconductor. Finally, the I2C2 and the nonparametrically bootstrapped 95% confidence interval of the I2C2 (with 1000 repetitions) between the first and second image of each participant were estimated.

Results

In this section, we will visualize results of cerebellar morphometry for the right lobules V, VI, VIIIA obtained by CERES and CNN (Figures 4, 5, and 6). These lobules were chosen because of their critical role in motor and non-motor functions of the cerebellum. According to an activation likelihood estimate meta-analysis of neuroimaging studies (55), (1) right lobule V is associated with motor and somatosensory processing, (2) right lobule VI is associated with motor, spatial, language, working memory, and emotional processing, and (3) right lobule VIIIA is associated with motor and working memory processing.

We will also present the coefficient of variation (CV) and the intraclass correlation coefficient (ICC) for the results of all analyses with FreeSurfer, CERES, and CNN (Tables 3-4). For CERES parcellations, we will also provide the image intraclass correlation coefficients (I2C2) (Table 2).

Supplementary data are available at the Open Science Framework (OSF)[16].

Visual inspection. All fully automated analyses resulted in anatomically broadly correct segmentations and parcellations except one FreeSurfer analysis (analysis failure) and one CNN analysis (incorrect placement of the bounding box localizing the cerebellum). In several FreeSurfer analyses, voxels containing dura and surrounding non-brain tissue were mislabeled as cerebellum, in particular, in the midline. In single CNN analyses, the parcellation algorithm mislabeled voxels located in the neck as cerebellum, even after postprocessing (e.g., the second examination of participant OAS2_0013 in the OASIS-2 study). Visual inspection of all CERES analyses did not reveal remarkable inaccuracies.

Replicability. Using FreeSurfer with data from the ChroPain2 study, two identical analyses of the same T1-weighted image provided identical results for all participants regarding gray and white matter volumes.

Using CERES with data from the ChroPain2 study, two identical analyses provided identical results for lobular volumes, cortical thickness, and gray matter volumes in most participants (Figure 4D, E, F for right lobules V, VI, VIIIA). For lobular volumes, differences for all regions were smaller than ± 0.1%. For cortical thickness, maximum differences were found in the left lobules I-II (-4.8 - 3.7%). Maximum differences in gray matter volume were also found in the left lobules I-II (-5.8 - 11.8%, data not shown).

Using CNN with data from the ChroPain2 study, two identical analyses provided different results for all regions. Differences were larger than those found with CERES (Figure 4B). Differences were between -36.6% (right lobule VIIIB) and 20.6% (right lobule IX). To confirm these results, analysis replicability of CNN was also assessed with all T1-weighted images of the Kirby-21 study (Figure 4C). For this data set, differences were between -11.5% (vermis X) and 19.4% (left lobule VIIIB, data not shown).

Repeatability. Comparing the FreeSurfer results of the first and second T1-weighted MRI in the Kirby-21 study, differences in gray matter volumes were below ±5% (Figure 5A). Differences in white matter volumes were higher, between

[14] https://cran.r-project.org/web/packages/psych/index.html

[15] https://neuroconductor.org/package/I2C2

[16] https://osf.io/n8y5h/

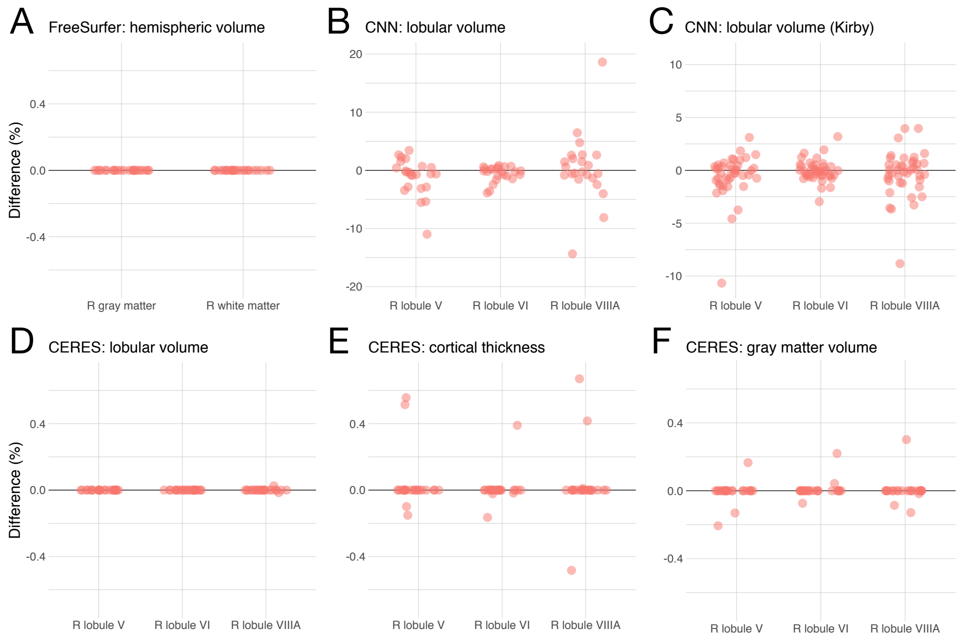

Fig. 4. Analysis replicability of cerebellar morphometry in the ChroPain2 study using FreeSurfer (A), CNN (B), and CERES (D, E, F). Analysis replicability was also assessed with all T1-weighted MRIs of the Kirby-21 study using CNN (C). The graphs show percent difference between the first and second analysis of the same data set for right gray and white matter (FreeSurfer) and the right lobules V, VI, and VIIIA (CNN, CERES). **Note:** the scales of the y-axes differ across graphs.

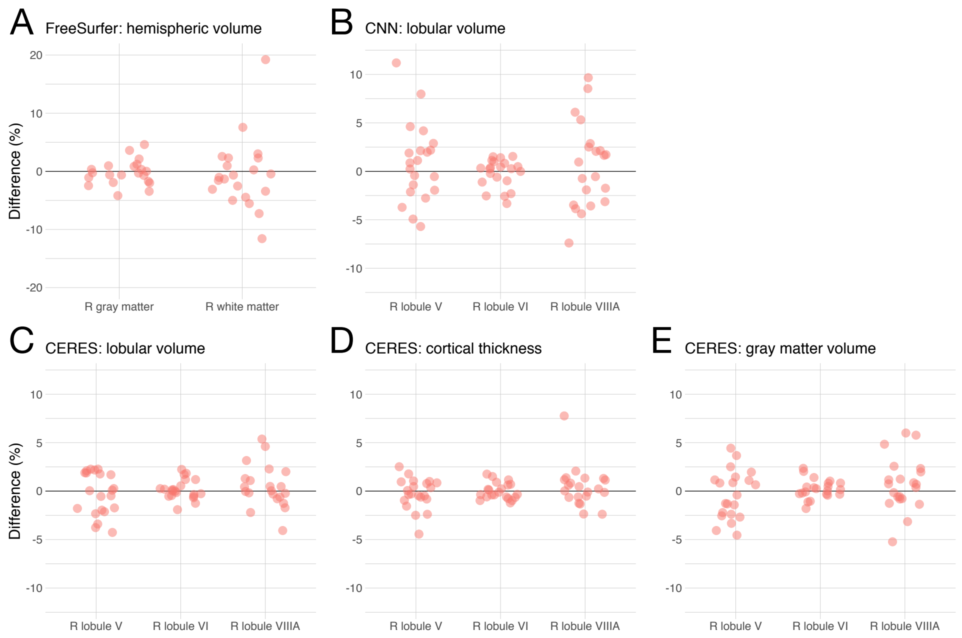

Fig. 5. Repeatability of cerebellar morphometry in the Kirby-21 study. The graphs show percent difference between the first and second MRI acquired on the same day for right gray and white matter using FreeSurfer (A) and the right lobules V, VI, and VIIIA using CNN (B) and CERES (C, D, E). **Note:** the scales of the y-axes differ across graphs.

-12.1% and 19.2%. With CERES, differences in lobular volumes, cortical thickness, and gray matter volume were below ±5% in most cases (Figure 5C, D, E). In some cases, differences were considerably higher, in particular for the small lobules I-II. With CNN, differences in lobular volumes were also below ±5% in most cases (Figure 5B). Maximum differences were between -20% (left lobule VIIB) and 35.1% (left lobule VIIIA, data not shown).

The image intraclass correlation coefficients (I2C2) for repeatability using the CERES parcellations are presented in Table 2. The coefficients of variation and the intraclass correlation coefficients for repeatability are presented in Table 3. Most lower 95% confidence intervals suggest good or even excellent repeatability.

Long-term reproducibility. Comparing the FreeSurfer results of the first and second T1-weighted MRI in the OASIS-2 study, most differences in gray matter volume were below ±5% (Figure 6A). Maximum differences in gray matter volume were between -12.3% and 9.7%, in white matter volume between -15.6% and 25.5%. With CERES, most differences for lobular volumes, cortical thickness, and gray matter volumes were below ±10%, many even below ±5% (Figure 6C, D, E). Maximum differences for lobular volumes (-60.1%, 167.5%), cortical thickness (-54.4%, 190.8%), and for gray matter volumes (-55.3%, 139.5%) were considerably higher.

Table 2. Image intraclass correlation coefficients (I2C2) and 95% confidence intervals for cerebellar regions obtained by CERES with data from the Kirby-21 study.

Region	Left	Right
Lobules I-II	0.83 (0.79 - 0.86)	0.83 (0.80 - 0.86)
Lobule III	0.87 (0.85 - 0.90)	0.86 (0.84 - 0.88)
Lobule IV	0.84 (0.82 - 0.86)	0.85 (0.83 - 0.87)
Lobule V	0.83 (0.79 - 0.85)	0.86 (0.84 - 0.88)
Lobule VI	0.88 (0.85 - 0.90)	0.89 (0.88 - 0.91)
Crus I	0.88 (0.87 - 0.90)	0.89 (0.88 - 0.91)
Crus II	0.89 (0.87 - 0.91)	0.89 (0.87 - 0.90)
Lobule VIIB	0.87 (0.85 - 0.89)	0.87 (0.85 - 0.89)
Lobule VIIIA	0.89 (0.87 - 0.90)	0.89 (0.87 - 0.91)
Lobule VIIIB	0.89 (0.87 - 0.91)	0.89 (0.87 - 0.91)
Lobule IX	0.89 (0.87 - 0.90)	0.88 (0.86 - 0.90)
Lobule X	0.89 (0.86 - 0.90)	0.88 (0.85 - 0.90)

With CNN, differences were also below ±10% in the majority of cases, many even below ±5% (Figure 6B). Maximum differences were between -96.4% and 180.9%.

The image intraclass correlation coefficients using the CERES parcellations suggest moderate reproducibility (data not shown). The coefficients of variation and the intraclass correlation coefficients for reproducibility are presented in Table 4.

Discussion

We present a detailed analysis of the reproducibility of fully automated cerebellar morphometry using three different software packages regarding (1) replicability (two analyses of one data set with identical hardware and software), (2) repeatability (analyses of two data sets taken on the same day), and (3) long-term reproducibility (analyses of two data sets taken months or years apart).

Regarding analysis replicability, we found that the results of FreeSurfer segmentations were identical in all analyses. Replicability was high for CERES parcellations and segmentations in most regions (Figure 4D-F), although the PatchMatch algorithm employed by CERES is non-deterministic and involves a random search step that is performed iteratively (20). By contrast, we found substantial differences when performing two identical CNN analyses of the high-resolution ChroPain2 data sets (Figure 4B). We hypothesized that the submillimeter resolution (0.75 mm isotropic voxel size) of this data set might have caused problems for CNN's parcellating network which has been trained with MPRAGE images resampled to 1 mm isotropic resolution (24). Therefore, we assessed CNN's analysis replicability with data from the Kirby-21 study (1 × 1 × 1.2 mm^3 voxel size). Differences between two identical CNN analyses were lower in the Kirby-21 study compared to the ChroPain 2 study (Figure 4B-C) but still relatively high, with most differences <±5%.

Assessment of repeatability revealed a remarkably similar picture for all software packages (Figure 5). Most differences between the first and the second MRI taken on the same day were <±5%. This result presents an estimation of the reproducibility with which cerebellar subdivisions can be determined with a recent MRI system at 3 Tesla, a widely-used MPRAGE sequence, and a fully automated segmentation and/or parcellation software for individual participants today. For comparison, estimation of cerebral cortical thicknesses using FreeSurfer demonstrated an overall higher reproducibility with differences between scans taken within minutes of $\leq$±1.9% and between scans taken within weeks of $\leq$±2.3% (56). Of course, the reported differences between two scans of one person is a complex mixture of several factors, including not only imperfections of the image analysis software used, but also of scanner hardware and MRI sequences, and differences in the positioning of the head. Using a high-resolution sequence (e.g. with a 0.75 mm isotropic voxel size) and/or a higher magnetic field strength (i.e., 7 Tesla) is expected to improve not only assessment of cerebral cortical thicknesses (57) but also of cerebellar volumes and cortical thicknesses due to reduced partial volume effects or increased signal-to-noise-ratios. As the developers of CERES acknowledge, the main limitation of their analysis software is the small library of only five manually labeled cerebellar templates on which CERES relies at present (20). Hopefully, the developers will include additional templates in future versions of their software, likely improving segmentation and parcellation results.

As expected, long-term reproducibility of cerebellar morphometry was lower than repeatability on the same day. Brain volumes and cortical thicknesses change over time, not only due to aging, but also due to factors unrelated to aging, such as diurnal factors (58), hydration (59), or alcohol intake (60). In single cases, both CERES and CNN analyses resulted in dramatic differences, suggesting mislabeling of large parts of cerebellar regions.

Recommendations for use of automated cerebellar morphometry. Based on the presented analyses, we recommend the following steps to improve the design, data analysis, and interpretation of future neuroimaging studies:

1. Quality control through visual inspection of all labeled regions. Corroborating the results of Kavaklioglu et al. (61), we recognized that FreeSurfer frequently mislabeled voxels representing the dura mater or the dural sinuses as cerebellar gray matter. The number of these voxels is usually small compared to the entire gray matter of the left or right cerebellum. Manual correction of labels and recomputing of cerebellar volumes is possible, but would require substantial expertise and time (62), and is therefore not feasible in large-scale studies. Of note, the locating network used in CNN failed in one analysis. In this case, the parcellating network mislabeled all voxels and finished without error message. Thus, we strongly recommend the visual inspection of all results of neuroimaging pipelines, including automated cerebellar morphometry. Visual inspection of subcortical FreeSurfer results requires manual loading of .mgz files in FreeSurfer's Freeview file viewer or in another viewer capable of displaying .mgz files (e.g., FSLeyes). Visual inspection of CERES and CNN results is less time-consuming because both analysis packages create report pages in pdf or html format for convenient inspection.

2. Assessment of analysis replicability. Many MRI analysis

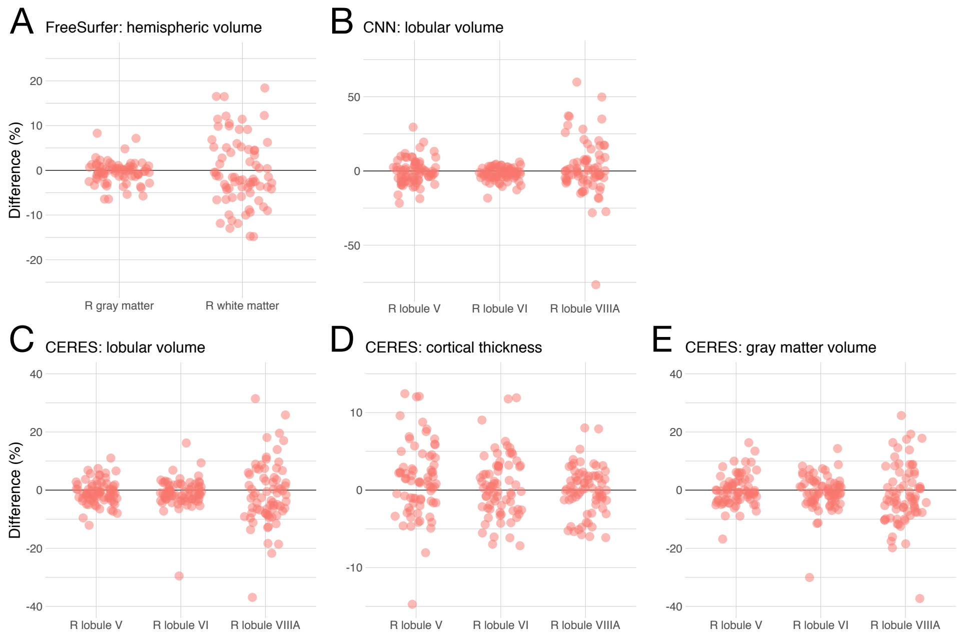

Fig. 6. Long-term reproducibility of cerebellar morphometry in the OASIS-2 study. The graphs show percent difference between the first and second MRI for right gray and white matter using FreeSurfer (A) and the right lobules V, VI, and VIIIA using CNN (B) and CERES (C, D, E). The mean interval between the two MRIs was 738 ± 249 days (minimum: 182, maximum: 1510 days). **Note:** the scales of the y-axes differ across graphs.

packages include stochastic algorithms, such as random seed generation for the initialization of analyses (63). Given the remarkable differences found in identical analyses by CNN, we recommend reporting the analysis replicability for every neuroimaging pipeline, including cerebellar morphometry.

3. Assessment of repeatability. For cross-sectional studies, we recommend reporting the repeatability of the selected neuroimaging pipeline in addition to its analysis replicability. The data set for assessment of repeatability should include two identical MRI scans taken on the same day, ideally directly one after another, but with repositioning in between, to minimize true changes in brain volumes or cortical thicknesses.

4. Assessment of long-term reproducibility. For the design of a longitudinal study, we recommend investigating the long-term reproducibility of the selected neuroimaging pipeline in addition to its analysis replicability. The data set for estimation of long-term reproducibility should include two or more scans, taken in time intervals comparable to the planned longitudinal study. The obtained results should guide the decision if the expected changes may be observed with the sample size and the study design under consideration (64).

Conclusions. Based on its high accuracy (19), its overall high reproducibility shown here, and its ability to differentiate between entire lobular volumes, gray matter lobular volumes, and lobular cortical thicknesses, CERES is a powerful tool to investigate cerebellar morphometry. Cerebellar morphometry is expected to provide important biomarkers for cerebellar aging and disease. Reliable neuroimaging biomarkers depend on reproducible analyses. For every neuroimaging pipeline, not only for cerebellar morphometry, reproducibility should be investigated, reported, and utilized for the interpretation of its results.

ACKNOWLEDGEMENTS

The authors wish to thank Stefan Harfst and Fynn Schwietzer, Scientific Computing, University of Oldenburg, Germany, for continuous support with the HPC cluster. The ChroPain2 study was supported by the Neuroimaging Unit, University of Oldenburg, funded by grants from the German Research Foundation (DFG; 3T MRI INST 184/152-1 FUGG and MEG INST 184/148-1 FUGG). The high-performance computer cluster CARL, University of Oldenburg, is funded by a grant from the German Research Foundation (DFG; INST 184/157-1 FUGG).

References

1. N Ahmadian, K van Baarsen, M van Zandvoort, and PA Robe. The cerebellar cognitive affective syndrome - A meta-analysis. *Cerebellum*, 18(5):941–950, 2019.
2. X Guell and J Schmahmann. Cerebellar functional anatomy: A didactic summary based on human fMRI evidence. *Cerebellum*, 19(1):1–5, 2020.
3. GPD Argyropoulos, K van Dun, M Adamaszek, M Leggio, M Manto, M Masciullo, M Molinari, CJ Stoodley, F Van Overwalle, RB Ivry, and JD Schmahmann. The cerebellar cognitive affective/Schmahmann syndrome: A task force paper. *Cerebellum*, 19(1):102–125, 2020.
4. L Rolando. *Saggio sopra la vera struttura del cervello dell'uomo e degl' animali e sopra le funzioni del sistema nervoso.* Stamperìa da S.S.R.M. Privilegiata, Sassari, 1809.
5. P Flourens. *Recherches expérimentales sur les propriétés et les fonctions du système nerveux, dans les animaux vertébrés.* Crevot, Paris, 1824.
6. WT Thach. A role for the cerebellum in learning movement coordination. *Neurobiol Learn Mem*, 70(1-2):177–188, 1998.
7. H Imamizu, S Miyauchi, T Tamada, Y Sasaki, R Takino, B Pütz, T Yoshioka, and M Kawato. Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature*, 403 (6766):192–195, 2000.
8. JD Schmahmann and JC Sherman. The cerebellar cognitive affective syndrome. *Brain*, 121 (Pt 4):561–579, 1998.
9. JD Schmahmann. The cerebellum and cognition. *Neurosci Lett*, 688:62–75, 2019.
10. M Adamaszek, F D'Agata, R Ferrucci, C Habas, S Keulen, KC Kirkby, M Leggio, P Mariën, M Molinari, E Moulton, L Orsi, F Van Overwalle, C Papadelis, A Priori, B Sacchetti, DJ Schutter, C Styliadis, and J Verhoeven. Consensus paper: Cerebellum and emotion. *Cerebellum*, 16(2):552–576, 2017.
11. E D'Angelo and CI De Zeeuw. Timing and plasticity in the cerebellum: Focus on the granular layer. *Trends Neurosci*, 32(1):30–40, 2009.
12. M Bareš, R Apps, L Avanzino, A Breska, E D'Angelo, P Filip, M Gerwig, RB Ivry, CL Lawrenson, ED Louis, NA Lusk, M Manto, WH Meck, H Mitoma, and EA Petter. Consensus paper: Decoding the contributions of the cerebellum as a time machine. From neurons to clinical applications. *Cerebellum*, 18(2):266–286, 2019.
13. BB Andersen, HJ Gundersen, and B Pakkenberg. Aging of the human cerebellum: A stereological study. *J Comp Neurol*, 466(3):356–365, 2003.
14. HM Gellersen, CC Guo, C O'Callaghan, RH Tan, S Sami, and M Hornberger. Cerebellar atrophy in neurodegeneration – A meta-analysis. *J Neurol Neurosurg Psychiatry*, 88(9): 780–788, 2017.
15. K Kansal, Z Yang, AM Fishman, HI Sair, SH Ying, BM Jedynak, JL Prince, and CU Onyike.

Structural cerebellar correlates of cognitive and motor dysfunctions in cerebellar degeneration. *Brain*, 140(3):707–720, 2017.

16. J Diedrichsen. A spatially unbiased atlas template of the human cerebellum. *Neuroimage*, 33(1):127–138, 2006.
17. CJ Steele and MM Chakravarty. Gray-matter structural variability in the human cerebellum: Lobule-specific differences across sex and hemisphere. *Neuroimage*, 170:164–173, 2018.
18. MT Park, J Pipitone, LH Baer, JL Winterburn, Y Shah, S Chavez, MM Schira, NJ Lobaugh, JP Lerch, AN Voineskos, and MM Chakravarty. Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates. *Neuroimage*, 95:217–231, 2014.
19. A Carass, JL Cuzzocreo, S Han, CR Hernandez-Castillo, PE Rasser, M Ganz, V Beliveau, J Dolz, I Ben Ayed, C Desrosiers, B Thyreau, JE Romero, P Coupé, JV Manjón, VS Fonov, DL Collins, SH Ying, CU Onyike, D Crocetti, BA Landman, SH Mostofsky, PM Thompson, and JL Prince. Comparing fully automated state-of-the-art cerebellum parcellation from magnetic resonance images. *Neuroimage*, 183:150–172, 2018.
20. JE Romero, P Coupé, R Giraud, VT Ta, V Fonov, MTM Park, MM Chakravarty, AN Voineskos, and JV Manjón. Ceres: A new cerebellum lobule segmentation method. *Neuroimage*, 147:916–924, 2017.
21. J Diedrichsen, JH Balsters, J Flavell, E Cussans, and N Ramnani. A probabilistic MR atlas of the human cerebellum. *Neuroimage*, 46(1):39–46, 2009.
22. TE Nichols, S Das, SB Eickhoff, AC Evans, T Glatard, M Hanke, N Kriegeskorte, MP Milham, RA Poldrack, JB Poline, E Proal, B Thirion, DC Van Essen, T White, and BT Yeo. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci*, 20(3): 299–303, 2017.
23. B Fischl, DH Salat, E Busa, M Albert, M Dieterich, C Haselgrove, A van der Kouwe, R Killiany, D Kennedy, S Klaveness, A Montillo, N Makris, B Rosen, and AM Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.
24. S Han, Y He, A Carass, SH Ying, and JL Prince. Cerebellum parcellation with convolutional neural networks. *Proc SPIE Int Soc Opt Eng*, 10949, 2019.
25. P Sörös and C Bantel. Chronic noncancer pain is not associated with accelerated brain aging as assessed by structural magnetic resonance imaging in patients treated in specialized outpatient clinics. *Pain*, 161(3):641–650, 2020.
26. JP Mugler and JR Brookeman. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magn Reson Med*, 15(1):152–157, 1990.
27. J Hamilton, D Franson, and N Seiberlich. Recent advances in parallel imaging for MRI. *Prog Nucl Magn Reson Spectrosc*, 101:71–95, 2017.
28. BA Landman, AJ Huang, A Gifford, DS Vikram, IA Lim, JA Farrell, JA Bogovic, J Hua, M Chen, S Jarso, SA Smith, S Joel, S Mori, JJ Pekar, PB Barker, JL Prince, and PC van Zijl. Multi-parametric neuroimaging reproducibility: A 3-T resource study. *Neuroimage*, 54 (4):2854–2866, 2011.
29. L Palumbo, P Bosco, ME Fantacci, E Ferrari, P Oliva, G Spera, and A Retico. Evaluation of the intra- and inter-method agreement of brain MRI segmentation software packages: A comparison between SPM12 and FreeSurfer v6.0. *Phys Med*, 64:261–272, 2019.
30. TJR Rezende, BM Campos, J Hsu, Y Li, C Ceritoglu, K Kutten, MC França Junior, S Mori, MI Miller, and AV Faria. Test-retest reproducibility of a multi-atlas automated segmentation tool on multimodality brain MRI. *Brain Behav*, 9(10):e01363, 2019.
31. DS Marcus, AF Fotenos, JG Csernansky, JC Morris, and RL Buckner. Open Access Series of Imaging Studies: Longitudinal MRI data in nondemented and demented older adults. *J Cogn Neurosci*, 22(12):2677–2684, 2010.
32. CP Hughes, L Berg, WL Danziger, LA Coben, and RL Martin. A new clinical scale for the staging of dementia. *Br J Psychiatry*, 140:566–572, 1982.
33. JV Manjón and P Coupé. volBrain: An online MRI brain volumetry system. *Front Neuroinform*, 10:30, 2016.
34. JD Schmahmann, J Doyon, D McDonald, C Holmes, K Lavoie, AS Hurwitz, N Kabani, A Toga, A Evans, and M Petrides. Three-dimensional MRI atlas of the human cerebellum in proportional stereotaxic space. *Neuroimage*, 10(3 Pt 1):233–260, 1999.
35. Federative International Programme on Anatomical Terminologies. *Terminologia Anatomica*. Thieme, Stuttgart, Germany, 2011.
36. B Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.
37. F Ségonne, AM Dale, E Busa, M Glessner, D Salat, HK Hahn, and B Fischl. A hybrid approach to the skull stripping problem in MRI. *Neuroimage*, 22(3):1060–1075, 2004.
38. R Giraud, VT Ta, N Papadakis, JV Manjón, DL Collins, P Coupé, and Alzheimer's Disease Neuroimaging Initiative. An Optimized PatchMatch for multi-scale and multi-feature label fusion. *Neuroimage*, 124(Pt A):770–782, 2016.
39. JV Manjón, P Coupé, L Martí-Bonmatí, DL Collins, and M Robles. Adaptive non-local means denoising of MR images with spatially varying noise levels. *J Magn Reson Imaging*, 31(1): 192–203, 2010.
40. NJ Tustison, BB Avants, PA Cook, Y Zheng, A Egan, PA Yushkevich, and JC Gee. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*, 29(6):1310, 2010.
41. BB Avants, NJ Tustison, G Song, PA Cook, A Klein, and JC Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.
42. NJ Tustison, PA Cook, A Klein, G Song, SR Das, JT Duda, BM Kandel, N van Strien, JR Stone, JC Gee, and BB Avants. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage*, 99:166–179, 2014.
43. P Coupé, JV Manjón, V Fonov, J Pruessner, M Robles, and DL Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *Neuroimage*, 54(2):940–954, 2011.
44. VT Ta, R Giraud, DL Collins, and P Coupé. Optimized PatchMatch for near real time and accurate label fusion. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 105–112, 2014.
45. JE Iglesias, CY Liu, PM Thompson, and Z Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging*, 30(9):1617–1634, 2011.
46. O Tange. *GNU Parallel*. DOI: https://doi.org/10.5281/zenodo.1146014, 2018.
47. H Shou, A Eloyan, S Lee, V Zipunnikov, AN Crainiceanu, NB Nebel, B Caffo, MA Lindquist, and CM Crainiceanu. Quantifying the reliability of image replication studies: The image intraclass correlation coefficient (I2C2). *Cogn Affect Behav Neurosci*, 13(4):714–724, 2013.
48. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.
49. W Revelle. *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois, USA, 2019.
50. D Liljequist, B Elfving, and K Skavberg Roaldsen. Intraclass correlation - A discussion and demonstration of basic features. *PLoS One*, 14(7):e0219854, 2019.
51. PE Shrout and JL Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*, 86(2):420–428, 1979.
52. KO McGraw and SP Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1:30–46, 1996.
53. TK Koo and MY Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*, 15(2):155–163, 2016.
54. J Muschelli, A Gherman, J-P Fortin, B Avants, B Whitcher, J D Clayden, B S Caffo, and C M Crainiceanu. Neuroconductor: An R platform for medical imaging analysis. *Biostatistics*, 20: 218–239, 2019.
55. CJ Stoodley and JD Schmahmann. Functional topography in the human cerebellum: A meta-analysis of neuroimaging studies. *Neuroimage*, 44(2):489–501, 2009.
56. X Wang, W Bauer, N Chiaia, M Dennis, M Gerken, J Hummel, J Kane, C Kenmuir, S Khuder, R Lane, R Mooney, P Bazeley, V Apkarian, and J Wall. Longitudinal MRI evaluations of human global cortical thickness over minutes to weeks. *Neurosci Lett*, 441(2):145–148, 2008.
57. N Zaretskaya, B Fischl, M Reuter, V Renvall, and JR Polimeni. Advantages of cortical surface reconstruction using submillimeter 7 T MEMPRAGE. *Neuroimage*, 165:11–26, 2018.
58. K Nakamura, RA Brown, S Narayanan, DL Collins, DL Arnold, and Alzheimer's Disease Neuroimaging Initiative. Diurnal fluctuations in brain volume: Statistical analyses of MRI from large populations. *Neuroimage*, 118:126–132, 2015.
59. K Nakamura, RA Brown, D Araujo, S Narayanan, and DL Arnold. Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: Implications for monitoring atrophy in clinical studies. *Neuroimage Clin*, 6:166–170, 2014.
60. S Geibprasert, M Gallucci, and T Krings. Alcohol-induced changes in the brain as assessed by MRI and CT. *Eur Radiol*, 20(6):1492–1501, 2010.
61. T Kavaklioglu, T Guadalupe, M Zwiers, AF Marquand, M Onnink, E Shumskaya, H Brunner, G Fernandez, SE Fisher, and C Francks. Structural asymmetries of the human cerebellum in relation to cerebral cortical asymmetries and handedness. *Brain Struct Funct*, 222(4): 1611–1623, 2017.
62. JY Wang, MM Ngo, D Hessl, RJ Hagerman, and SM Rivera. Robust machine learning-based correction on automatic segmentation of the cerebellum and brainstem. *PLoS One*, 11(5):e0156123, 2016.
63. T Glatard, LB Lewis, R Ferreira da Silva, R Adalat, N Beck, C Lepage, P Rioux, ME Rousseau, T Sherif, E Deelman, N Khalili-Mahani, and AC Evans. Reproducibility of neuroimaging analyses across operating systems. *Front Neuroinform*, 9:12, 2015.
64. C Velasco-Annis, A Akhondi-Asl, A Stamm, and SK Warfield. Reproducibility of brain MRI segmentation algorithms: Empirical comparison of local MAP PSTAPLE, FreeSurfer, and FSL-FIRST. *J Neuroimaging*, 28(2):162–172, 2018.

Table 3. Coefficient of variation (CV) and intraclass correlation coefficient (ICC) for cerebellar regions obtained by FreeSurfer, CERES, and CNN with data from the Kirby-21 study.

Region	Left		Right	
	CV (%)	ICC	CV %	ICC
FreeSurfer				
Gray matter	1.23 (0.95 - 1.76)	>0.99	1.43 (1.10 - 2.04)	0.99 (0.99 - 1)
White matter	3.16 (2.43 - 4.52)	0.96 (0.91 - 0.98)	3.99 (3.07 - 5.70)	0.95 (0.90 - 0.98)
CERES				
Lobular volume				
Lobules I-II	12.18 (9.37 - 17.41)	0.93 (0.85 - 0.96)	11.87 (9.14 - 16.97)	0.93 (0.86 - 0.97)
Lobule III	2.44 (1.87 - 3.48)	>0.99	3.29 (2.53 - 4.70)	0.97 (0.93 - 0.98)
Lobule IV	2.00 (1.54 - 2.86)	0.99 (0.97 - 0.99)	2.11 (1.62 - 3.01)	0.98 (0.97 - 0.99)
Lobule V	2.95 (2.27 - 4.21)	0.95 (0.89 - 0.97)	1.51 (1.17 - 2.16)	0.99 (0.98 - 1)
Lobule VI	2.01 (1.55 - 2.88)	0.98 (0.97 - 0.99)	0.7 (0.54 - 1.00)	>0.99
Crus I	1.94 (1.49 - 2.77)	0.99 (0.97 - 0.99)	1.45 (1.12 - 2.07)	0.99 (0.99 - 1)
Crus II	1.61 (1.24 - 2.3)	0.99 (0.98 - 1)	1.68 (1.29 - 2.40)	0.99 (0.98 - 1)
Lobule VIIB	2.92 (2.25 - 4.17)	0.98 (0.96 - 0.99)	1.62 (1.25 - 2.32)	0.99 (0.99 - 1)
Lobule VIIIA	1.76 (1.36 - 2.52)	0.99 (0.98 - 1)	1.50 (1.16 - 2.15)	0.99 (0.99 - 1)
Lobule VIIIB	1.42 (1.09 - 2.02)	0.99 (0.99 - 1)	1.43 (1.10 - 2.04)	>0.99
Lobule IX	1.15 (0.89 - 1.65)	>0.99	1.74 (1.34 - 2.49)	0.99 (0.99 - 1)
Lobule X	1.94 (1.49 - 2.77)	0.99 (0.97 - 0.99)	1.84 (1.41 - 2.63)	0.98 (0.97 - 0.99)
Cortical thickness				
Lobules I-II	11.95 (9.19 - 17.07)	0.80 (0.62 - 0.90)	10.30 (7.92 - 14.71)	0.83 (0.68 - 0.92)
Lobule III	4.23 (3.26 - 6.05)	0.91 (0.82 - 0.96)	4.62 (3.55 - 6.60)	0.90 (0.81 - 0.95)
Lobule IV	1.6 (1.23 - 2.29)	0.90 (0.79 - 0.95)	1.43 (1.10 - 2.05)	0.94 (0.89 - 0.97)
Lobule V	1.02 (0.78 - 1.46)	0.87 (0.75 - 0.94)	1.11 (0.85 - 1.58)	0.94 (0.89 - 0.97)
Lobule VI	0.57 (0.44 - 0.81)	0.88 (0.76 - 0.94)	0.58 (0.45 - 0.83)	0.93 (0.85 - 0.96)
Crus I	1.22 (0.94 - 1.74)	0.82 (0.66 - 0.91)	1.36 (1.04 - 1.94)	0.86 (0.72 - 0.93)
Crus II	1.02 (0.78 - 1.46)	0.82 (0.66 - 0.91)	1.02 (0.79 - 1.46)	0.93 (0.85 - 0.96)
Lobule VIIB	0.77 (0.59 - 1.10)	0.89 (0.79 - 0.95)	0.86 (0.66 - 1.23)	0.79 (0.60 - 0.89)
Lobule VIIIA	0.66 (0.51 - 0.95)	0.96 (0.92 - 0.98)	1.39 (1.07 - 1.98)	0.77 (0.57 - 0.88)
Lobule VIIIB	3.33 (2.56 - 4.76)	0.81 (0.64 - 0.91)	3.39 (2.60 - 4.84)	0.75 (0.54 - 0.87)
Lobule IX	5.01 (3.85 - 7.15)	0.84 (0.69 - 0.92)	4.56 (3.51 - 6.51)	0.75 (0.54 - 0.87)
Lobule X	6.02 (4.63 - 8.60)	0.80 (0.63 - 0.90)	7.25 (5.58 - 10.37)	0.84 (0.70 - 0.92)
Gray matter volume				
Lobules I-II	20.92 (16.09 - 29.89)	0.88 (0.77 - 0.94)	12.98 (9.99 - 18.55)	0.94 (0.88 - 0.97)
Lobule III	2.83 (2.18 - 4.05)	0.98 (0.95 - 0.99)	3.72 (2.86 - 5.32)	0.96 (0.92 - 0.98)
Lobule IV	1.57 (1.21 - 2.25)	0.99 (0.98 - 1)	1.80 (1.39 - 2.58)	0.99 (0.98 - 0.99)
Lobule V	3.52 (2.70 - 5.02)	0.92 (0.83 - 0.96)	1.72 (1.32 - 2.46)	0.99 (0.98 - 0.99)
Lobule VI	2.29 (1.76 - 3.27)	0.98 (0.96 - 0.99)	0.72 (0.55 - 1.02)	>0.99
Crus I	2.01 (1.54 - 2.87)	0.99 (0.97 - 0.99)	1.62 (1.25 - 2.32)	0.99 (0.98 - 1)
Crus II	1.62 (1.25 - 2.32)	0.99 (0.98 - 1)	1.50 (1.15 - 2.14)	0.99 (0.98 - 1)
Lobule VIIB	3.04 (2.34 - 4.34)	0.98 (0.95 - 0.99)	1.68 (1.29 - 2.40)	0.99 (0.99 - 1)
Lobule VIIIA	1.76 (1.36 - 2.52)	0.99 (0.98 - 1)	1.91 (1.47 - 2.73)	0.99 (0.98 - 1)
Lobule VIIIB	2.44 (1.88 - 3.49)	0.98 (0.96 - 0.99)	1.95 (1.50 - 2.78)	0.99 (0.98 - 1)
Lobule IX	2.76 (2.13 - 3.95)	0.98 (0.95 - 0.99)	2.56 (1.97 - 3.67)	0.98 (0.96 - 0.99)
Lobule X	3.16 (2.43 - 4.52)	0.97 (0.93 - 0.98)	3.10 (2.39 - 4.43)	0.97 (0.93 - 0.98)
CNN				
Lobules I-III	2.90 (2.23 - 4.15)	0.98 (0.97 - 0.99)	3.76 (2.89 - 5.37)	0.95 (0.90 - 0.98)
Lobule IV	2.60 (2 - 3.72)	0.95 (0.90 - 0.98)	2.40 (1.85 - 3.43)	0.98 (0.96 - 0.99)
Lobule V	3.08 (2.37 - 4.40)	0.95 (0.91 - 0.98)	2.71 (2.08 - 3.87)	0.97 (0.94 - 0.99)
Lobule VI	1.26 (0.97 - 1.81)	0.99 (0.99 - 1)	1.08 (0.83 - 1.54)	0.99 (0.99 - 1)
Lobule VIIAf	1.16 (0.89 - 1.65)	>0.99	1.51 (1.16 - 2.16)	0.99 (0.98 - 1)
Lobule VIIAt	2.11 (1.62 - 3.01)	0.99 (0.98 - 0.99)	2.79 (2.14 - 3.98)	0.98 (0.96 - 0.99)
Lobule VIIB	4.76 (3.66 - 6.80)	0.93 (0.85 - 0.96)	3.12 (2.40 - 4.47)	0.96 (0.91 - 0.98)
Lobule VIIIA	4.64 (3.57 - 6.63)	0.97 (0.93 - 0.98)	3.13 (2.40 - 4.47)	0.99 (0.98 - 0.99)
Lobule VIIIB	3.67 (2.83 - 5.25)	0.98 (0.96 - 0.99)	3.88 (2.98 - 5.54)	0.97 (0.95 - 0.99)
Lobule IX	2.13 (1.64 - 3.04)	0.99 (0.97 - 0.99)	2.28 (1.76 - 3.26)	0.99 (0.97 - 0.99)
Lobule X	2.04 (1.57 - 2.92)	0.98 (0.97 - 0.99)	3.07 (2.36 - 4.39)	0.96 (0.92 - 0.98)
	Midline			
Vermis VI	1.32 (1.02 - 1.89)	0.99 (0.98 - 1)		
Vermis VII	2.39 (1.84 - 3.42)	0.98 (0.96 - 0.99)		
Vermis VIII	1.62 (1.25 - 2.32)	0.99 (0.99 - 1)		
Vermis IX	1.53 (1.17 - 2.18)	0.99 (0.98 - 1)		
Vermis X	2.81 (2.16 - 4.01)	0.97 (0.94 - 0.99)		

The table presents the CV and ICC with lower and upper 95% confidence intervals.

Table 4. Coefficient of variation (CV) and intraclass correlation coefficient (ICC) for cerebellar regions obtained by FreeSurfer, CERES, and CNN with data from the OASIS-2 study.

Region	Left		Right	
	CV (%)	ICC	CV %	ICC
FreeSurfer				
Gray matter	2.34 (2.01 - 2.80)	0.98 (0.96 - 0.98)	1.84 (1.58 - 2.20)	0.97 (0.96 - 0.98)
White matter	5.21 (4.48 - 6.23)	0.38 (0.20 - 0.53)	5.42 (4.66 - 6.48)	0.42 (0.25 - 0.57)
CERES				
Lobular volume				
Lobules I-II	27.34 (23.51 - 32.66)	0.58 (0.44 - 0.70)	15.89 (13.66 - 18.99)	0.73 (0.63 - 0.81)
Lobule III	8.15 (7.01 - 9.74)	0.84 (0.77 - 0.89)	8.55 (7.36 - 10.22)	0.77 (0.68 - 0.84)
Lobule IV	3.87 (3.33 - 4.62)	0.94 (0.91 - 0.96)	4.54 (3.91 - 5.43)	0.94 (0.92 - 0.96)
Lobule V	3.17 (2.72 - 3.78)	0.96 (0.95 - 0.97)	2.95 (2.53 - 3.52)	0.95 (0.93 - 0.97)
Lobule VI	2.55 (2.20 - 3.05)	0.97 (0.96 - 0.98)	3.37 (2.90 - 4.03)	0.96 (0.94 - 0.97)
Crus I	3.04 (2.61 - 3.63)	0.96 (0.94 - 0.97)	3.37 (2.90 - 4.03)	0.96 (0.94 - 0.97)
Crus II	4.20 (3.61 - 5.02)	0.95 (0.92 - 0.96)	5.13 (4.41 - 6.13)	0.90 (0.85 - 0.93)
Lobule VIIB	6.95 (5.98 - 8.31)	0.89 (0.84 - 0.92)	7.84 (6.74 - 9.36)	0.82 (0.74 - 0.87)
Lobule VIIIA	7.15 (6.15 - 8.54)	0.87 (0.81 - 0.91)	7.97 (6.85 - 9.52)	0.84 (0.78 - 0.89)
Lobule VIIIB	5.68 (4.89 - 6.79)	0.89 (0.84 - 0.92)	7.68 (6.61 - 9.18)	0.82 (0.75 - 0.88)
Lobule IX	4.19 (3.60 - 5.00)	0.96 (0.94 - 0.97)	4.80 (4.13 - 5.73)	0.94 (0.92 - 0.96)
Lobule X	5.32 (4.57 - 6.36)	0.88 (0.82 - 0.92)	5.08 (4.37 - 6.07)	0.88 (0.83 - 0.92)
Cortical thickness				
Lobules I-II	22.13 (19.03 - 26.45)	0.35 (0.16 - 0.51)	24.73 (21.26 - 29.55)	0.22 (0.03 - 0.40)
Lobule III	8.50 (7.31 - 10.15)	0.48 (0.32 - 0.62)	8.85 (7.61 - 10.57)	0.36 (0.18 - 0.52)
Lobule IV	3.19 (2.75 - 3.82)	0.52 (0.37 - 0.65)	3.36 (2.89 - 4.02)	0.61 (0.47 - 0.72)
Lobule V	2.53 (2.18 - 3.03)	0.67 (0.55 - 0.77)	3.21 (2.76 - 3.84)	0.65 (0.52 - 0.75)
Lobule VI	2.75 (2.36 - 3.28)	0.38 (0.20 - 0.53)	2.70 (2.33 - 3.23)	0.48 (0.31 - 0.61)
Crus I	3.87 (3.33 - 4.62)	0.71 (0.59 - 0.79)	3.45 (2.97 - 4.12)	0.71 (0.60 - 0.79)
Crus II	5.49 (4.72 - 6.56)	0.56 (0.41 - 0.68)	4.39 (3.77 - 5.24)	0.55 (0.40 - 0.67)
Lobule VIIB	3.26 (2.81 - 3.90)	0.45 (0.28 - 0.59)	2.13 (1.84 - 2.55)	0.45 (0.28 - 0.59)
Lobule VIIIA	1.86 (1.60 - 2.22)	0.58 (0.44 - 0.70)	2.14 (1.84 - 2.56)	0.37 (0.19 - 0.52)
Lobule VIIIB	2.12 (1.83 - 2.54)	0.53 (0.37 - 0.66)	2.10 (1.81 - 2.51)	0.48 (0.32 - 0.62)
Lobule IX	4.81 (4.14 - 5.75)	0.37 (0.19 - 0.53)	4.66 (4.01 - 5.57)	0.42 (0.24 - 0.56)
Lobule X	10.08 (8.67 - 12.05)	0.72 (0.61 - 0.80)	10.15 (8.73 - 12.12)	0.71 (0.60 - 0.80)
Gray matter volume				
Lobules I-II	32.49 (27.94 - 38.82)	0.60 (0.46 - 0.71)	22.63 (19.46 - 27.04)	0.54 (0.39 - 0.67)
Lobule III	9.39 (8.07 - 11.22)	0.81 (0.73 - 0.87)	8.53 (7.33 - 10.19)	0.79 (0.70 - 0.85)
Lobule IV	4.39 (3.78 - 5.25)	0.92 (0.89 - 0.95)	4.83 (4.16 - 5.78)	0.94 (0.91 - 0.96)
Lobule V	4.13 (3.55 - 4.93)	0.94 (0.91 - 0.96)	3.64 (3.13 - 4.35)	0.92 (0.89 - 0.95)
Lobule VI	2.81 (2.41 - 3.36)	0.97 (0.95 - 0.98)	3.97 (3.41 - 4.74)	0.94 (0.91 - 0.96)
Crus I	3.00 (2.58 - 3.58)	0.96 (0.95 - 0.98)	3.00 (2.58 - 3.59)	0.97 (0.95 - 0.98)
Crus II	5.17 (4.45 - 6.18)	0.92 (0.88 - 0.94)	5.70 (4.90 - 6.82)	0.87 (0.82 - 0.91)
Lobule VIIB	7.48 (6.44 - 8.94)	0.87 (0.81 - 0.91)	8.03 (6.91 - 9.60)	0.80 (0.71 - 0.96)
Lobule VIIIA	7.11 (6.12 - 8.50)	0.87 (0.81 - 0.91)	7.66 (6.59 - 9.15)	0.85 (0.78 - 0.90)
Lobule VIIIB	6.12 (5.26 - 7.31)	0.87 (0.81 - 0.91)	7.75 (6.66 - 9.26)	0.82 (0.75 - 0.87)
Lobule IX	4.04 (3.48 - 4.83)	0.96 (0.94 - 0.97)	4.46 (3.83 - 5.33)	0.95 (0.92 - 0.96)
Lobule X	5.03 (4.33 - 6.01)	0.89 (0.85 - 0.93)	5.55 (4.78 - 6.64)	0.87 (0.81 - 0.91)
CNN				
Lobules I-III	12.54 (10.79 - 14.99)	0.74 (0.64 - 0.82)	10.48 (9.01 - 12.52)	0.81 (0.74 - 0.87)
Lobule IV	8.17 (7.02 - 9.76)	0.81 (0.73 - 0.87)	6.23 (5.35 - 7.44)	0.86 (0.80 - 0.91)
Lobule V	7.90 (6.80 - 9.44)	0.80 (0.71 - 0.86)	5.83 (5.01 - 6.96)	0.86 (0.80 - 0.91)
Lobule VI	4.19 (3.60 - 5.00)	0.93 (0.90 - 0.95)	3.48 (2.99 - 4.16)	0.95 (0.93 - 0.97)
Lobule VIIAf	2.62 (2.25 - 3.13)	0.97 (0.96 - 0.98)	2.84 (2.44 - 3.39)	0.96 (0.95 - 0.98)
Lobule VIIAt	7.38 (6.34 - 8.82)	0.81 (0.74 - 0.87)	7.84 (6.75 - 9.37)	0.82 (0.75 - 0.88)
Lobule VIIB	8.41 (7.23 - 10.05)	0.80 (0.72 - 0.86)	11.70 (10.06 - 13.98)	0.69 (0.58 - 0.78)
Lobule VIIIA	8.85 (7.61 - 10.57)	0.81 (0.73 - 0.87)	11.28 (9.70 - 13.48)	0.78 (0.69 - 0.84)
Lobule VIIIB	11.32 (9.73 - 13.52)	0.81 (0.73 - 0.87)	10.56 (9.08 - 12.62)	0.74 (0.63 - 0.81)
Lobule IX	5.02 (4.31 - 5.99)	0.93 (0.91 - 0.96)	5.86 (5.04 - 7.00)	0.91 (0.87 - 0.94)
Lobule X	9.26 (7.97 - 11.07)	0.74 (0.63 - 0.81)	8.60 (7.40 - 10.28)	0.70 (0.59 - 0.79)
	Midline			
Vermis VI	6.83 (5.87 - 8.16)	0.81 (0.73 - 0.87)		
Vermis VII	39.77 (34.20 - 47.52)	0.11 (-0.09 - 0.29)*		
Vermis VIII	3.39 (2.92 - 4.05)	0.95 (0.92 - 0.96)		
Vermis IX	4.93 (4.24 - 5.89)	0.90 (0.85 - 0.93)		
Vermis X	8.11 (6.97 - 9.69)	0.76 (0.66 - 0.83)		

The table presents the CV and ICC with lower and upper 95% confidence intervals. * Calculation of the ICC for Vermis VII (CNN) includes one outlier (median: 1.35 cm^3, outlier: 7.94 cm^3). Without the outlier, the ICC is 0.83 (0.76 - 0.88).