NeuroImage 264 (2022) 119703

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

CerebNet: A fast and reliable deep-learning pipeline for detailed cerebellum sub-segmentation

Jennifer Faber[a,c,1], David Kügler[a,1], Emad Bahrami[a,b,1], Lea-Sophie Heinz[a], Dagmar Timmann[d], Thomas M. Ernst[d], Katerina Deike-Hofmann[e], Thomas Klockgether[a,c], Bart van de Warrenburg[f], Judith van Gaalen[f], Kathrin Reetz[g,h], Sandro Romanzetti[g], Gulin Oz[i], James M. Joers[i], Jorn Diedrichsen[j], ESMI MRI Study Group[2], Martin Reuter[a,k,l,*]

[a] *German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*
[b] *Computer Science Department, University Bonn, Bonn, Germany*
[c] *Department of Neurology, University Hospital Bonn, Germany*
[d] *Department of Neurology, Center for Translational Neuro, and Behavioral Sciences (C-TNBS), University Hospital Essen, University of Duisburg-Essen, Essen, Germany*
[e] *Department of Neuroradiology, University Hospital Bonn, Germany*
[f] *Department of Neurology, Donders Institute for Brain, Cognition, and Behaviour, Radboud university medical center, Nijmegen, The Netherlands*
[g] *Department of Neurology, RWTH Aachen University, Germany*
[h] *JARA-Brain Institute Molecular Neuroscience and Neuroimaging, Forschungszentrum Jülich, Germany*
[i] *Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN, USA*
[j] *Departments of Computer Science and Statistical and Actuarial Sciences, Western University, London, ON, Canada*
[k] *A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA*
[l] *Department of Radiology, Harvard Medical School, Boston, MA, USA*

ARTICLE INFO

2020 MSC:
00-01
99-00

Keywords:
CerebNet
Cerebellum
Computational neuroimaging
Deep learning

ABSTRACT

Quantifying the volume of the cerebellum and its lobes is of profound interest in various neurodegenerative and acquired diseases. Especially for the most common spinocerebellar ataxias (SCA), for which the first antisense oligonculeotide-base gene silencing trial has recently started, there is an urgent need for quantitative, sensitive imaging markers at pre-symptomatic stages for stratification and treatment assessment. This work introduces *CerebNet*, a fully automated, extensively validated, deep learning method for the lobular segmentation of the cerebellum, including the separation of gray and white matter. For training, validation, and testing, T1-weighted images from 30 participants were manually annotated into cerebellar lobules and vermal sub-segments, as well as cerebellar white matter. *CerebNet* combines *FastSurferCNN*, a UNet-based 2.5D segmentation network, with extensive data augmentation, e.g. realistic non-linear deformations to increase the anatomical variety, eliminating additional preprocessing steps, such as spatial normalization or bias field correction. *CerebNet* demonstrates a high accuracy (on average 0.87 Dice and 1.742mm Robust Hausdorff Distance across all structures) outperforming state-of-the-art approaches. Furthermore, it shows high test-retest reliability (average ICC > 0.97 on OASIS and Kirby) as well as high sensitivity to disease effects, including the pre-ataxic stage of spinocerebellar ataxia type 3 (SCA3). *CerebNet* is compatible with *FreeSurfer* and *FastSurfer* and can analyze a 3D volume within seconds on a consumer GPU in an end-to-end fashion, thus providing an efficient and validated solution for assessing cerebellum sub-structure volumes. We make *CerebNet* available as source-code (https://github.com/Deep-MI/FastSurfer).

* Corresponding author.
E-mail address: martin.reuter@dzne.de (M. Reuter).

[1] Contributed equally.

[2] *ESMI MR Study Group:* Paola Giunti (Ataxia Centre, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology & National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK), Hector Garcia-Moreno (Ataxia Centre, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology & National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK), Heike Jacobi (Department of Neurology, University Hospital of Heidelberg, Heidelberg, Germany), Johann Jende (Department of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany), Jeroen de Vries (Department of Neurology, Expertise Center Movement Disorders Groningen, University Medical Center Groningen, University of Groningen, The Netherlands), Michal Povazan (Johns Hopkins University School of Medicine, Baltimore, MD, U.S.), Peter B. Barker (Johns Hopkins University School of Medicine, Baltimore, MD, U.S.), Katherina Marie Steiner (Department of Neurology, Center for Translational Neuro, and Behavioral Sciences (C-TNBS), University Hospital Essen, University of Duisburg-Essen, Essen, Germany), Janna Krahe (Department of Neurology, RWTH Aachen University, Germany).

https://doi.org/10.1016/j.neuroimage.2022.119703.
Received 12 September 2022; Accepted 18 October 2022
Available online 27 October 2022.

1. Introduction

For decades, the cerebellum was attributed to have an exclusive role in motor control. Recently, growing evidence suggests a more general involvement of the cerebellum in the adaptive control also of cognitive and emotional processing. In fact, morphometric studies demonstrate significant cerebellar atrophy with age and in a number of non-motor brain diseases, e.g. schizophrenia, autism or Alzheimer's disease (Diedrichsen et al., 2015; D'Mello et al., 2015; Han et al., 2020a; Lin et al., 2020; Marek et al., 2018; Okugawa et al., 2003; Toniolo et al., 2018; Webb et al., 2009; Womer et al., 2016). In healthy humans, the representation of cerebral networks and cognitive domains has been investigated using functional connectivity (Buckner et al., 2011) as well as task functional MRI (King et al., 2019). These complementary studies have helped to increase knowledge about the role of the cerebellum in cognitive and emotional processes. This notwithstanding, the cerebellum is crucial for motor control, in particular metric and power of target movements. With regard to movement disorders, cerebellar atrophy is the characterizing feature in ataxias, which manifest as acquired, genetic, or sporadic degenerative diseases. With clinical features including progressive loss of balance, coordination deficits, and slurred speech, ataxia patients suffer substantial restrictions of mobility and communicative skills. In genetic ataxies, such as the worldwide most common autosomal dominantly inherited spinocerebellar ataxia type 3 (SCA3), the manifest or ataxic stage of the disease is preceded by a pre-ataxic stage, in which neurodegeneration is already quantifiable, e.g., as cerebellar atrophy, while manifest ataxia is not yet present (Faber et al., 2021; Kim et al., 2021; Rezende et al., 2018). Preventive interventions that aim to silence the disease gene in pre-ataxic mutation carriers offer a promising treatment option prior to clinical onset (McLoughlin et al., 2018). Now, that the first clinical gene silencing trial has recently started (ClinicalTrials.gov Identifier: NCT05160558), there is an urgent need for non-invasive biomarkers to assess disease manifestation and progression, and to quantify potential treatment effects as clinical scales lack sensitivity during the pre-ataxic stage. Accurate cerebellar volume estimation from structural MRI is a relevant neuroanatomical marker. However, fast automated determination of cerebellar volumes is required, as detailed, manual volumetry, especially of sub-regions, is too time-consuming. Clearly, automated segmentation will benefit various study designs, by reducing workload and by improving reliability.

In the present work, we introduce *CerebNet*, an automated method to sub-segment the cerebellum at the lobular level based on T1-weighted MRI. Our labels focus on a detailed boundary delineation between cerebellar gray matter (CGM) and cerebellar white matter (CWM) capturing the branches of CWM that reach into the cerebellar cortex based on T1-weighted MRI. Our deep learning method leverages the *FastSurfer* approach (Henschel et al., 2020) of multiple 2D networks and minimal pre-processing to obtain detailed boundary segmentations. Since *CerebNet* does not require any preprocessing steps and performs the localization and segmentation of 27 cerebellar regions in only 12 seconds per MRI, it is optimally suited to also efficiently process and screen in large data sets. With very labor-intensive manual reference segmentation, the methodological challenge is to achieve high accuracy and generalizability despite a small reference dataset. To this effect, we perform extensive pre-training on representative cross-study datasets and apply several data augmentation steps including realistic non-linear deformations to ensure wide applicability. Moreover, we validate our method with respect to test-retest reliability and in an association study of neuro-morphometric cerebellum markers across 109 SCA3 mutation carriers, including 42 pre-ataxic participants, as well as 41 healthy controls. Results reveal stronger group differences for *CerebNet* consistent with known patterns of neurodegenerative changes.

1.1. Protocols and anatomical reference

The Schmahmann atlas (Schmahmann et al., 1999) is the standard anatomical reference for cerebellar cortex sub-segmentation protocols (Bogovic et al., 2013b; Park et al., 2014) including the "Spatially Unbiased Infratentorial Template" (SUIT) (Diedrichsen, 2006). It introduces a unified terminology of the nomenclature. Slices of the cerebellum are directly compared with the corresponding slices of MR images, thus facilitating the identification of anatomical landmarks. Briefly summarized, the CGM is macroscopically subdivided into the midline vermis and four hemispheric lobes: the anterior, posterior-superior, posterior-inferior, and the flocculonodular lobe. The anterior and posterior lobes are further subdivided into lobules. The vermis is subdivided analogously to the hemispheres except for the anterior lobe. Like all previous protocols, ours follows the nomenclature introduced by Schmahmann (Schmahmann et al., 1999) and our segmentation is largely comparable to previous protocols (Bogovic et al., 2013b; Diedrichsen, 2006; Park et al., 2014). The protocols for segmenting the cerebellum on MR images differ in the level of detail at which single anterior lobules and vermal subsegments are distinguished or aggregated (Bogovic et al., 2013b; Diedrichsen, 2006; Park et al., 2014). Previous work has largely only differed in finding an aggregation compromise in the level of detail for segments I-V. We detail a comparison between the different segmentation protocols as well as to related automated segmentation procedures in the appendix of our protocol for manual segmentation(Heinz et al., 2022). It should be noted, that all previous protocols ignore the CWM strands projecting into the cerebellar cortex (Bogovic et al., 2013b; Diedrichsen, 2006; Park et al., 2014) simplifying the CGM/CWM boundary to a connection line across the base of CWM strands. In consequence, details at the CGM/CWM boundary of the cerebellum are not captured by any of the previous protocols. To allow deeper analysis of the GM/WM boundary in the cerebellum, we extend our protocol by a fine-grained segmentation of CWM strands projecting into the cerebellar cortex. To foster reproducibility and extensibility, we establish and publish our illustrated segmentation protocol online with this publication (Heinz et al., 2022).

1.2. Automated methods for cerebellar sub-segmentation

Several methods have been presented for segmenting cerebellar substructures including both semi-automated (Pierson et al., 2002) and fully automated (Bogovic et al., 2013c; Carass et al., 2018; Diedrichsen, 2006; Han et al., 2020b) approaches. While previous methods relied on atlas-based registration (Diedrichsen, 2006; Diedrichsen et al., 2009; Park et al., 2014; Plassard et al., 2016; Romero et al., 2017), artificial neural networks (Powell et al., 2008), support vector machines (Powell et al., 2008), level sets (Bogovic et al., 2013a), active appearance models (Price et al., 2014), and patch matching (Romero et al., 2017; Weier et al., 2014), recent work introduced deep learning (Han et al., 2020b; 2019).

The reference method "Spatially Unbiased Infra-tentorial Template" (SUIT) (Diedrichsen, 2006) pioneered fully automatic cerebellum sub-segmentation using non-linear registration to an atlas. Powell et al. (2008) compared atlas registration with fully connected neural network and support vector machine segmentation methods, and demonstrated superior performance of learning approaches. ACCLAIM (Bogovic et al., 2013a), which is based on the Multiple object Geometric Deformable Model framework (Bogovic et al., 2013c; Carass and Prince, 2016), adapts a random forest for boundary classification to produce topologically correct results. The Multiple Automatically Generated Templates brain segmentation algorithm (MAGeT) (Chakravarty et al., 2013; Park et al., 2014) creates a template library, then nonlinearly registers the target image to each template. The final segmentation is achieved by fusing multiple segmentations using majority voting. The Cerebellar Analysis Toolkit (CATK) (Price et al., 2014) adapts

Bayesian active appearance modeling (Patenaude et al., 2011) to generate statistical models for shape and texture and their inter-relationship as priors. RASCAL (Weier et al., 2014) utilizes a patch matching-based approach, which improves the multi-atlas segmentation fusion method of Coupe et al. (Coupé et al., 2011) via majority voting for label fusion and nonlinear registration. CERES (Romero et al., 2017), another patch matching-based segmentation tool, employs the Optimized PatchMatch Label fusion (OPAL) method (Giraud et al., 2016; Ta et al., 2014). CERES2 (Carass et al., 2018) improves upon CERES (Romero et al., 2017) by adding a patch-based boosted neural network method for error correction. CGCUTS (Yang et al., 2016) combines multi-atlas labeling and random forest classification in the context of a graph cut framework to produce the segmentation. Van der Lijn et al. (van der Lijn et al., 2009) present a method that combines an appearance model and atlas registration. Carass et al. (Carass et al., 2018) summarize and compare several cerebellum sub-segmentation methods, highlighting CERES2 (Carass et al., 2018; Romero et al., 2017) as the most performant 'traditional' (i.e. non deep-learning) approach. However, while image processing with CERES1 is supported online, CERES2 is unavailable to the scientific community.

The most recent cerebellum sub-segmentation tool, Anatomical Parcellation using a U-Net with Locally Constrained Optimization (ACAPULCO) (Han et al., 2020b), introduces a two-step deep learning method with two 3D convolutional neural networks (CNNs) to first localize and then sub-segment the cerebellum, outperforming the challenge winner CERES2 (Carass et al., 2018) in a head-to-head comparison. Preprocessing steps include bias field inhomogeneity correction and registration to MNI space. However, both the training and the evaluation procedure include some short-comings, e.g. by forcing nearest neighbor label interpolation both during training and evaluation, predominantly reducing detail in fine structures such as thin CWM strands. In fact, reported performance metrics were calculated entirely in MNI space, which required lossy nearest neighbor interpolation of manual reference labels to MNI space potentially mischaracterizing segmentation performance.

In contrast to ACAPULCO, our method does not require any preprocessing steps such as bias field correction or spatial (atlas) normalization/registration during inference. Moreover, to increase the anatomical variety in our training data, we employ various augmentation approaches, e.g. we generate realistic non-linear deformations via cross-subject registration of training images to various images from multiple datasets. To further improve the generalization of our model we pretrain the model on a compiled cross-study dataset. We examine the effect of data augmentations such as non-linear deformation and pre-training in several experiments. In our proposed method, the neural network architecture follows *FastSurferCNN*, a 2.5D approach in which three 2D networks for each axial, coronal, and sagittal view are trained and the final 3D prediction is created by view-aggregation (Henschel et al., 2020).

1.3. Contributions

This work presents five contributions:

- A detailed labeling protocol ensuring replicability and extensibility for the 25 cerebellar cortex labels and 2 cerebellar white matter segmentations including the fine branching, as well as a manual reference dataset of consensus cerebellar subsegmentation labels for training and testing.
- A training methodology with extensive data augmentation including realistic deformations to address the challenge of a small training dataset.
- Detailed method ablation to establish design choices in dedicated experiments on a subset of cases, not overlapping with the training or test sets.
- *CerebNet* consistently and significantly outperforms state-of-the-art cerebellum sub-segmentation methods with respect to accuracy and test-retest reliability.
- Sensitive *CerebNet* segmentations reproduce cerebellar atrophy effects in the pre-ataxic stage of spinocerebellar ataxia type 3 with a superior group separability.

2. Methods

We first describe the datasets for training, validation, and testing of *CerebNet*, then continue with the description of our method, and finally detail the evaluation.

2.1. Datasets

2.1.1. CerebNet dataset

We assemble a diverse cerebellum sub-segmentation dataset for training, validation and testing of models based on acquisitions from ongoing observational studies. This superset includes participants equally distributed between healthy controls as well as pre-ataxic and ataxic SCA3 mutation carriers, thereby covering a broad range of different degrees of cerebellar atrophy.

Participants 32 T1-weighted MRI of SCA3 mutation carriers and healthy controls were acquired at 4 sites: Bonn and Aachen, Germany, Nijmegen, The Netherlands and Minneapolis, MN, US. All participants provided written informed consent according to the guidelines set by the local institutional review boards. Two cases with visible motion artifacts were excluded, resulting in the final CerebNet dataset of 20 SCA3 mutation carriers, covering the whole disease course of SCA3 from early pre-ataxic to late ataxic disease stages, and 10 healthy controls of the same age range. In Table 1, we report demographics (age, sex) and ataxia severity, assessed with the Scale for Assessment and Rating of Ataxia (SARA) (Schmitz-Hübsch et al., 2006) for the three groups. For SCA3 mutation carriers, we also report the CAG repeat length. To divide SCA3 mutation carriers into pre-ataxic (SARA < 3, $N = 11$) and ataxic individuals (SARA ≥ 3, $N = 9$), we follow the established SARA cut-off value of 3, corresponding to the mean plus 2 standard deviations of the healthy control group distribution from the original SARA validation study (Schmitz-Hübsch et al., 2006).

MRI scans All T1-weighted MRI were acquired as MPRAGE on 3T SIEMENS scanners (Siemens Medical Systems, Erlangen, Germany). All scans share an isotropic resolution of 1mm, FOV 256×256 and 192 slices, acquired in sagittal direction with a 32-channel head coil. Bonn ($N = 16$, Skyra), Minnesota ($N = 7$, Prisma Fit), and Aachen ($N = 4$, Prisma) acquired at TR = 2500ms, TE = 4.37ms, TI = 1100ms, FA = 7°, while Nijmegen ($N = 4$, Trio) acquired at TR = 2300ms, TE = 3.03ms, TI = 1100ms, FA = 8°.

Segmentation Protocol Following the Schmahmann atlas (Schmahmann et al., 1999) as anatomical reference, we define 27 disjoint macroscopic subsegments of the cerebellum. In addition to 20 hemispheric lobules (10 for each hemisphere), we include 5 vermis labels and two CWM labels (left and right). The cerebellar segmentation is divided into 6 hierarchical steps, gradually moving from large-scale structures to the subdivision of cerebellar lobules. First, we delineate the CGM cortex with an exact outer boundary separating CGM from cerebrospinal fluid and other subtentorial structures, such as cranial nerves (*Step 1*). In this step, any inwardly projecting CWM branches are ignored. Subsequently, the four lobes (*Step 2*), and the vermis (*Step 3*) are segmented. We conduct the sub-segmentation of the hemispheric lobules (*Step 4*) as well as the subdivision of the vermis (*Step 5*). Finally, the fine delineation of the CWM including its branches into the CGM cortex band and a consistent boundary towards the brainstem is drawn (*Step 6*). The detailed protocol is publicly available for reproducibility (Heinz et al., 2022).

Manual Reference Standard The correct subdivision of the cerebellar cortex into lobules is critical, since the cerebellum shows a high morphological variability of its anatomical structure (Fig. 1). At the isotropic resolution of 1mm, it remains challenging to precisely determine whether a single small folia or branch belongs to one or an adja-

Table 1
Demographic and characterizing data of the *CerebNet* data set cohort consisting of pre-ataxic and ataxic SCA3 mutation carriers as well as healthy controls (HC). [1]Time to onset is given in years. The reported time from onset (defined as the first occurrence of gait disturbances) is given where available and for the remaining seven pre-ataxic mutation carriers, not yet experiencing gait disturbances, we estimated the time to onset following the model introduced by Tezenas et al. [42], which depends on both the number of CAG RL as well as the actual age; SD = standard deviation, CAG RL = CAG repeat length of the longer allele.

Group	N	**age** [years] mean ± SD [range]	sex m/f	SARA mean ± SD [range]	**Time to ataxia onset**[1] mean ± SD [range]	CAG RL mean ± SD
HC	10	43.9 ± 13.22 [22; 63]	4/6	0.3 ± 0.54 [0; 1.5]	n.a.	n.a.
pre-ataxic SCA3	11	31.6 ± 7.1 [20; 43]	4/7	1.4 ± 0.8 [0; 2.5]	-4.5 ± 6.4 [-13.8; 8.0]	72.4 ± 3.1
ataxic SCA3	9	44.6 ± 7.3 [32; 57]	6/3	12.6 ± 4.5 [7; 19]	8.4 ± 5.4 [1.0; 8.0]	70.89 ± 4.11

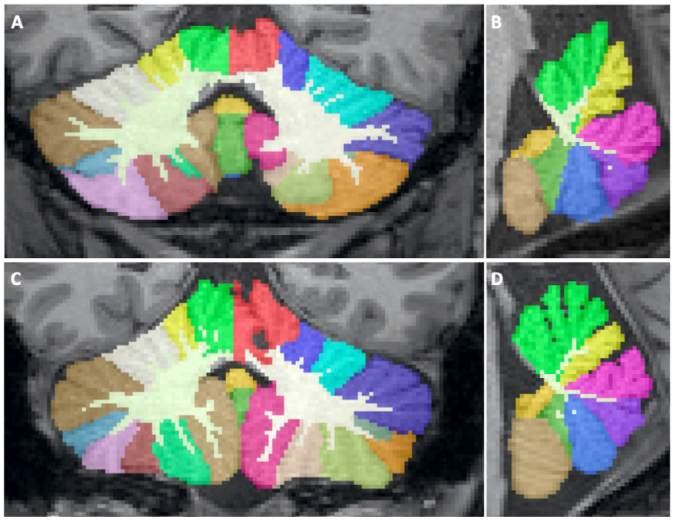

Fig. 1. Segmentation examples of a fully automated segmentation of *CerebNet* in a healthy control (A, B) as well as a symptomatic SCA3 patient (C, D) projected onto a coronal and sagittal slice.

cent lobule. To address this, all lobular boundaries within the cerebellar cortex (*Step 1-5*) are subsegmented by two experienced raters on all MRIs independently. To unify differences between the two raters, cortex segmentations are reviewed by an interdisciplinary team consisting of the experienced raters as well as a neurologist and a neuroradiologist. A consensus was reached for all cases. Furthermore, a team of four trained raters delineated the fine-grained CGM/CWM boundary (*Step 6*). The final consensus segmentation together with the CWM delineations represents the manual reference standard for training, validation, and testing of our method. We split the participants contained in the final reference dataset into 18/4/8 for training, validation, and testing. For individual splits, we preserve the distribution of controls, pre-ataxic and ataxic participants.

2.1.2. Cross-study pre-training dataset

For pre-training purposes, we compile a dataset of 160 T1-weighted images gathered from the Autism Brain Imaging Data Exchange II (ABIDE II) (Di Martino et al., 2017), the Alzheimers Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005), the UCLA Consortium for Neuropsychiatric Phenomics LA5c Study (LA5c) (Poldrack et al., 2016), the Open Access Series of Imaging Studies 1 and 2 (OASIS-1[3] (Marcus et al., 2007) and OASIS-2 (Marcus et al., 2010)), and the Minimal Interval Resonance Imaging in Alzheimers Disease (MIRIAD) (Malone et al., 2013). For these 160 cases, we automatically generate cerebellar sub-segmentation labels using SUIT v3.3 (Diedrichsen, 2006; Diedrichsen et al., 2009). SUIT is an atlas-based segmentation tool which provides segmentations of 28 sub-regions of the cerebellar cortex at the level of cerebellar lobules and a sub-segmentation of the vermis according to the Schmahmann atlas (Schmahmann et al., 1999). Since SUIT does not provide segmentations for CWM, we additionally process all 160 images with *FreeSurfer* (FS) (Fischl et al., 2002) and merge FS-generated CWM with the cerebellar sub-regions labels from SUIT. Gaps between cerebellar CWM and CGM are resolved by mapping them to the nearest CGM structure. The compiled external dataset is split into 140 training and 20 validation cases and is exclusively used for pre-training of our model.

2.1.3. Deformation dataset for augmentation

To increase variability of our training data, we generate realistic non-linear deformations for data augmentation. For this we generated an auxiliary dataset of 100 cases selected from ABIDE II (Di Martino et al., 2017), ADNI (Mueller et al., 2005), LA5c (Poldrack et al., 2016), OASIS-1[3] (Marcus et al., 2007) and OASIS-2 (Marcus et al., 2010), MIRIAD (Malone et al., 2013), and the Human Connectome Project (HCP) (Van Essen et al., 2012).

2.1.4. Test-retest dataset

We use the OASIS-1 (reliability subset) (Marcus et al., 2007) and Kirby (Landman et al., 2011) datasets for test-retest analysis. OASIS-1 contains 20 participants that were scanned no more than 90 days apart (all except 5 less than 30 days). The Kirby dataset consists of scan-rescan MPRAGE images of 21 healthy participants with one hour break between scanning sessions.

2.2. Cerebellar sub-segmentation method

This section introduces our cerebellar sub-segmentation pipeline, consisting of an initial localization step to extract the relevant cerebellum region, a subsequent multi-view ensemble for CNN-based segmentation, and a final view-aggregation step to merge the predictions. The pipeline accepts unprocessed 1.0mm T1-weighted images and outputs segmentation maps and tabulated volume reports. To achieve high accuracy, the relatively small size of the manual reference standard requires special consideration. We address it by pre-training with a representative cross-study dataset as well as applying data augmentation. Specifically, intensity and spatial data augmentation techniques such as realistic deformations increase the diversity presented to the network during training and, thus, its performance.

Localization To constrain the sub-segmentation network to the cerebellum and reduce memory and computational requirements, we crop a bounding box of $128 \times 128 \times 128$ isotropic 1mm voxels containing both cerebelli. The bounding box is placed symmetrically around the full cerebellar region obtained from a quick single-view (coronal) *FastSurfer* segmentation (Henschel et al., 2020). A visual inspection of this localization approach confirms the cerebellum is always correctly localized and fully contained within the bounding box in all cases.

Cerebellar Sub-segmentation Network The method for cerebellum sub-segmentation follows *FastSurfer* (Henschel et al., 2020). Briefly, in its

[3] Excluding cases from the OASIS-1 reliability section.

2.5D approach, *FastSurfer* utilizes an ensemble of three two-dimensional CNNs (*FastSurferCNN*) – each of these processing the MRI images sliced in a different direction (axial, coronal, and sagittal views). A final view-aggregation step combines the resulting label probability maps in probability space. *FastSurferCNN* is a U-Net-based fully CNN architecture with a dense encoder and decoder block per depth-level. In contrast to its predecessor (Guha Roy et al., 2019), the architecture extends the dense blocks and unpooling operations with a local competition approach (Estrada et al., 2018; 2020) and gathers information in the third dimension via spatial information aggregation (SPI). The SPI approach provides the network with a wider volumetric context by stacking additional three preceding and three succeeding neighboring slices for a total of 7 input channels. Both the view aggregation and the SPI approach together allow the method to process 3D information, while at the same time retaining the computational advantages of 2D networks, primarily lower memory requirements and sample efficiency, i.e. 1. a lower number of parameters compared to 3D networks and 2. 3D MRI are split into slices increasing the number of samples presented to the network.

Spatial Augmentations Spatial augmentations such as flipping, translation, rotation, and scaling were used during training to improve the robustness of our model. We encode these transformation as a 3×3 in-slice transformation matrix in homogeneous coordinates with coefficients uniformly sampled from predefined ranges. Random offsets along the in-slice-axes for translation are selected from -12 to 12mm to simulate cerebellum centroid variation. We sample the scaling factor from 0.95 to 1.2. The image is rotated in-slice with respect to its center with angles uniformly sampled from -20° to 20°. We also apply a random left-right flip, i.e., both the image and its labels are mirrored, but label IDs are swapped with respect to the mid-plane separating the two hemispheres keeping left-labels on the left.

Augmentation with non-linear Deformation To increase variability of our training data, we perform static augmentation with 500 non-linearly deformed training images. For this, we first non-linearly register each image of the Deformation (see Section 2.1.3) to 5 randomly selected images from the training split of the CerebNet dataset (see Section 2.1.1) using ANTs v2.3.1 (Avants et al., 2008). We ensure each manually labeled *CerebNet* case is at least paired once. For each of the resulting 500 anatomically realistic deformation fields, we then map both image and manual label from the training split of the CerebNet dataset using the obtained deformation field. In effect, this procedure drastically increases the anatomical variance presented to the network during training.

Intensity Augmentations Random MRI magnetic field inhomogeneities are synthesized and linearly superimposed to the images to increase the robustness of the model to bias field artifacts. We generate the augmented inhomogeneity field by linear-combination of randomly weighted cubic polynomial basis functions (Van Leemput et al., 1999). The coefficients of the basis functions are uniformly sampled from a -0.5 to 0.5 range.

2.3. Metrics for evaluation

To establish the quality and accuracy of *CerebNet* with respect to volumetric and geometric features, we evaluate the resulting segmentations with three common segmentation metrics: The *Dice Score*, calculated as the general label overlap, is well established as a good compromise between volumetric and geometric segmentation properties; the *Hausdorff Distance* serves as a metric for geometric and spatial similarity, and finally the *Volume Similarity* completely ignores overlap and spatial distance, but most directly evaluates the reliability for volumetric measures commonly used in statistical modeling.

Dice Score The Dice score (Dice) (Dice, 1945; Sørensen and Julius, 1948), is one of the most frequently used metrics in validating semantic segmentations. If P and G are the segmentation maps of the network prediction and ground-truth respectively, then Dice is defined as:

$$\text{Dice} = 2 \times \frac{|G \cap P|}{|G| + |P|}, \tag{1}$$

where $|.|$ represents cardinality. It measures overlap of 3D volumes on a scale between 0 and 1, where a value of 1 indicates exact agreement and 0 disjoint segmentations.

Hausdorff Distance To evaluate the quality of segmentation boundaries we calculate the distance between the manual and the automatic segmentation boundaries. In particular, this distance metric allows to test the overall accuracy of the boundary delineation emphasizing the correct contour. As this distance-metric decreases, segmentation boundaries more closely correspond to each other locally, i.e. more agreement of geometric details. Boundary distances can be quantified by the standard Hausdorff Distance (HD) or the Robust Hausdorff Distance (HD95). The standard HD measures the maximum distance and therefore is strongly affected by local outliers. HD95 – the 95% percentile of distances between surfaces (Huttenlocher et al., 1993) – is less sensitive to outliers and consequently more informative when analyzing the general trend. Formally, for boundaries B_G (of the ground-truth label map G) and B_P (of its predicted correspondent), we use their distances $D_{G \leftrightarrow P} = \{\min_{g \in B_G} \; d(p,g) \mid \forall p \in B_P\} \cap \{\min_{p \in B_P} \; d(p,g) \mid \forall g \in B_G\}$ to compute HD and HD95 as $d_{\text{HD}} = \max D_{G \leftrightarrow P}$ and $\text{P}(d < d_{\text{HD95}}) < 0.95$, $d \in D_{G \leftrightarrow P}$.

Volume Similarity Volume similarity (vol_{sim}) compares the absolute volume difference with the sum of volumes. Given V_G and V_P, the volumes of the ground-truth and predicted segmentations (G and P), vol_{sim} is calculated as

$$vol_{\text{sim}} = 1 - \frac{|V_G - V_P|}{V_G + V_P}. \tag{2}$$

Since this metric ignores overlap and geometric information, the optimal similarity (a value of 1) can be achieved for two segmentations of the same size, even if their spatial overlap is zero. However, its independence from spatial correspondence enables cross-acquisition comparison, e.g. for test-retest analysis, without requiring image alignment.

Intraclass Correlation Coefficient The Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss, 1979) evaluates the reliability and agreement between measurements. Its values range from 0 to 1 with larger values representing higher reliability. We also compute the 95% confidence interval around the ICC. For test-retest scenarios, we calculate the ICC as a measure of agreement between two repeated scans of the same participant (relative agreement, single fixed rater). Since scans are acquired in close temporal proximity, we assume only little volumetric changes and thus a high ICC.

2.4. Implementation details

Here, we detail the training of *CerebNet* and our adaptations of state-of-the-art methods to establish compatibility with our labeling protocol.

CerebNet Training We train each network for axial, coronal, sagittal views independently for 70 epochs with a batch size of 128 using one NVIDIA Tesla V100 GPU with 32 GB RAM. We use the AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2019) optimizer with a weight decay of 10^{-4} and an initial learning rate (LR) of 0.01. The *reduce on plateau* strategy for scheduling updates the LR based on the Dice score on the validation set. This strategy reduces the LR by a factor of 0.01, if there is no improvement in Dice score for 4 epochs.

ACAPULCO Re-Training In order to detach protocol and training data differences from method features, we retrain the deep learning method ACAPULCO (Han et al., 2020b) with our data following identical splits, hereafter referred to as ACAPULCO$^{\text{rt}}$. For this, we start with the published method code of ACAPULCO and configure the dataset loader to accept our dataset. Since we do not have access to the ACAPULCO training data, training our method on their data for comparison is not possible. We trained ACAPULCO with the publicly available source code and therein defined hyperparameters.

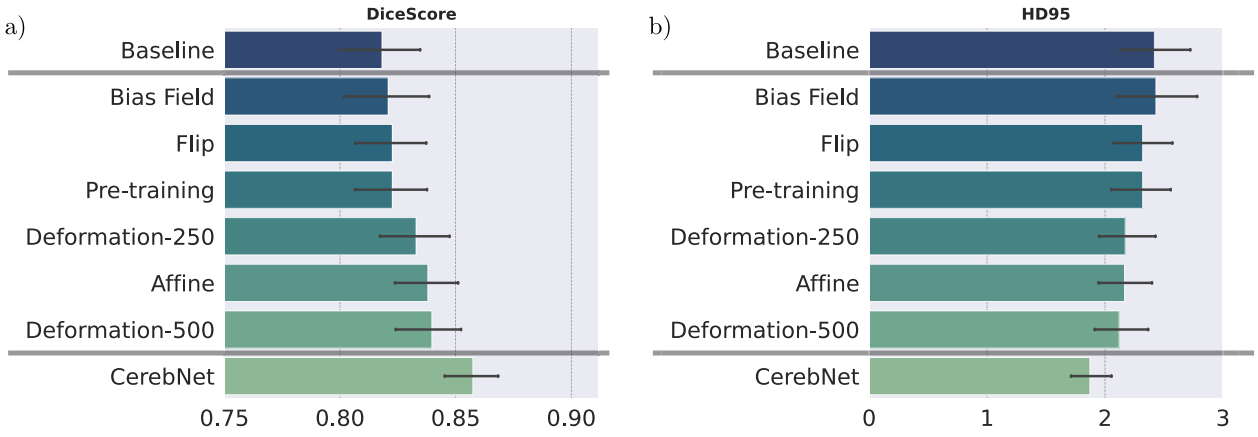

Fig. 2. Dice score (larger values are better) and Robust Hausdorff Distances (HD95, smaller values are better) on validation cases for comparison of baseline, individual method contributions (not cumulative) and *CerebNet*. *CerebNet* combines multiple data augmentations with pre-training on a representative cross-study dataset. Deformation-250/500 indicates the number of realistic deformation fields used for static augmentation. The baseline model is our network without augmentation or pre-training. Error bars indicate 95% confidence intervals.

SUIT + FS The leading traditional method (Carass et al., 2018), SUIT (Diedrichsen et al., 2009) does not rely on a deep learning approach and is only compatible with the CerebNet labels after combination with *FreeSurfer* (Fischl et al., 2002). In analogy to Section 2.1.2, we therefore merge SUIT cerebellum sub-segmentation labels with *FreeSurfer*'s CWM segmentation (SUIT + FS) to obtain the full set of labels.

3. Results

We report detailed results for multiple experiments to assess the performance of *CerebNet*. First, we ablatively establish the configuration and parameters of *CerebNet* on a validation hold-out set. Second, keeping method parameters fixed from here on, we compare the average performance of *CerebNet* with the state-of-the-art using four volumetric and geometric metrics: the Dice Score, two Hausdorff distances, and volume similarity. We investigate regional performance differences of these methods for all cerebellar sub-structures. Third, we contextualize the accuracy of *CerebNet* with differences between raters. Fourth, we compare the test-retest reliability of *CerebNet* with the state-of-the-art method ACAPULCO (Han et al., 2020b). Finally, we validate whether *CerebNet* reproduces known group differences between pre-ataxic and ataxic patients and healthy controls.

3.1. Ablation experiments

We perform several experiments to determine, how different changes to the data augmentation impact the performance of our method. In specific, we isolate the individual effects of different data augmentation methods and pre-training. We assess random flipping (Flip), bias field, affine deformation, and realistic non-linear deformation (Deformation-N, we test $N = 250$ and $N = 500$ deformation fields). While the baseline foregoes all data augmentation and pre-training, *CerebNet* combines all data augmentations with pre-training. All individual contributions improve results over the baseline (Fig. 2) in both Dice and Robust Hausdorff Distance (HD95) evaluations. Finally, the combination of all contributions clearly improves the results over any individual approach. We exclusively evaluate on validation cases for this analysis to avoid data-leakage.

3.2. Comparison with the state-of-the-art

In a summary evaluation, we compare the overall performance of *CerebNet*, ACAPULCO^{rt} (Han et al., 2020b) and SUIT + FS (Diedrichsen, 2006; Fischl et al., 2002) on the test subset of the CerebNet dataset (Section 2.1.1). On average across all segmented structures, *CerebNet* achieves a 0.870 per-structure Dice score and a 1.742mm Robust Hausdorff distance, which is the 95% percentile of surface-to-surface distances. In comparison with both state-of-the-art approaches, *CerebNet* outperforms either approach significantly in all four metrics ($p < .01$, see Fig. 3): the Dice score, Hausdorff distance (HD), Robust Hausdorff Distance (HD95) and volume similarity.

Results for individual structures are very consistent across all four metrics. Therefore, we focus further analysis and discussion on the Dice score and the Robust Hausdorff Distance. Specifically, we favor the robust implementation, since its robustness to outliers better reflects the accuracy across the surface, yet the high margin of 95% ensures larger structures (like CWM strands) are captured. Additionally, conclusions and reported significance values (derived by a Wilcoxon signed-ranked test) are completely independent of the choice of Hausdorff metric.

Dice Score CerebNet surpasses a 0.75 Dice score for all 27 individual structures and exceeds 0.95 Dice for the joint CWM. In fact, the least performing structures (specifically lobes VIIb and VIIIa/b) are "thin structures" sharing predominantly hard to define boundaries with other gray matter regions. *CerebNet* outperforms ACAPULCO^{rt} in 22 of 27 individual structures. In 14 out of 27 structures the improvement is significant (10 times $p < .01$ and 4 times $p < .05$, Fig. 4). For regions, with better performance of ACAPULCO^{rt}, the difference is usually small and never significant. *CerebNet* also significantly improves over ACAPULCO^{rt} segmentations for both merged gray matter and merged vermis regions ($p < .01$, Fig. 4). In comparison to the traditional SUIT + FS method, *CerebNet* always achieves better Dice scores, which are also statistically significant for all but three sub-structures (Fig. 4).

Robust Hausdorff Distance (HD95) On average, *CerebNet* achieves a HD95 distance of 1.742mm improving substantially over ACAPULCO^{rt} by 0.779mm. Across different vermis regions, the Crus I and lobe X regions and the merged gray matter region, *CerebNet* even exceeds a 1.25mm threshold. Larger distances remain in the lobes, where hard to reproduce lobe-to-lobe boundaries dominate the evaluation. Across all 27 structures, *CerebNet* outperforms ACAPULCO^{rt} in 26 of 27 structures (significantly for 14 structures, 5 times $p < .01$, 9 times $p < .05$, Fig. 4). In fact, the large performance differences for the merged CWM and CGM (both $\gtrapprox 2mm$) clearly indicate the differences between methods (Fig. 4). Compared with SUIT + FS, *CerebNet* demonstrates a superior performance consistently.

To quantify the robustness of *CerebNet*, we perform an outlier analysis for the Dice and Robust Hausdorff results. All data points are within 2.5 standard deviations of the per class mean.

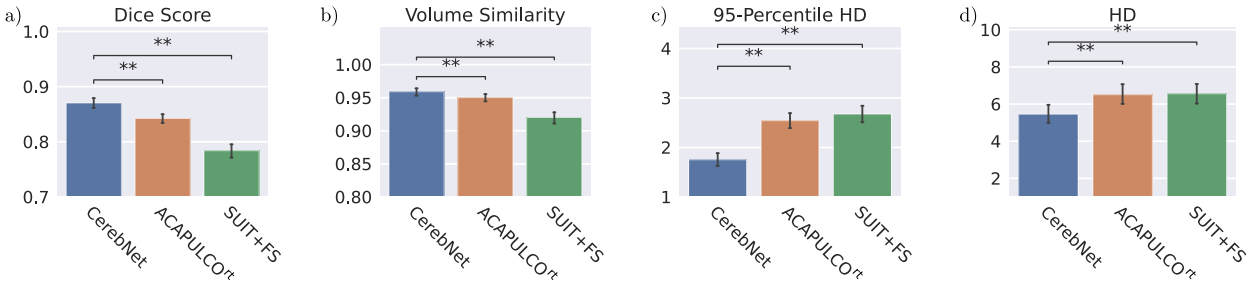

Fig. 3. Comparison of mean a) Dice score (larger values are better), b) Volume similarity (larger values are better), c) Robust Hausdorff Distance (HD95, smaller values are better), and d) Hausdorff Distance (HD, smaller values are better) over all structures and participants. *CerebNet* outperforms both ACAPULCO^{rt} (which is retrained on our dataset for direct comparison) and SUIT + FS. Error bars indicate 95% confidence intervals. Statistical significance for all results is confirmed by two-sided non-parametric Wilcoxon signed-rank tests (**: $p < .01$).

3.3. Inter-rater reproducibility

In our experience, delineation of cerebellar sub-structures, manual or automatic, is a challenging task due to the inherent uncertainty and lack of information to determine boundaries between cerebellar lobules even at 1mm isotropic resolution. To evaluate the *CerebNet* performance in the context of the reliability of manual segmentation, we analyze *CerebNet* segmentation errors together with the inter-rater variability. Figs. 4 and 5 share the same evaluation results for *CerebNet* in both cases comparing *CerebNet* predictions with the final "consensus segmentation" (after Step 6, see Section 2.1.1). However, for best annotation quality, labels from multiple raters are merged and harmonized in Step 6 of the protocol (see Section 2.1.1). To consistently and comparably represent the inter-rater reliability, we compare labels from one rater prior to this harmonization (after Step 5) to the "consensus". Since Step 5 data also does not include CWM labels, we exclude segmentation errors along the CGM/CWM boundary in the inter-rater evaluation (i.e. we mask out the CWM as defined in the "consensus") and only consider CGM regions for evaluation.

To evaluate the inter-rater variability, we utilize the Dice score and the Robust Hausdorff Distance (HD95). Fig. 5 illustrates per-region *CerebNet* and inter-rater Dice scores and HD95 on the *CerebNet* test set. Both volumetric and geometric segmentation scores are strongly correlated and – in most cases – at similar levels. Specifically, lower *CerebNet* performance values also map to lower inter-rater reliability in lobes V, VIIb and VIIIa/b. Results to lobe X as well as vermis VII/IX/X are outliers to this observation, where *CerebNet* provides good segmentations despite – in comparison – low inter-rater reliability.

3.4. Test-retest reliability

With the substantial time and labor requirements of manual segmentation, crucial external validation of methods is not easily possible. However, test-retest datasets with multiple scans of the same underlying anatomy and acquisition/machine properties offer the opportunity to test the reliability of methods. The OASIS-1 reliability dataset (Marcus et al., 2007) and Kirby dataset (Landman et al., 2011) are not only acquired at sites and in studies independent of the CerebNet dataset, but also feature 1.5T Siemens and 3T Philips scanners, respectively. To avoid influences of potentially error-prone image registration and interpolation, only per-structure volumes will be compared, as all geometric analysis would require alignment of baseline and follow-up scans.

Here, we compare the reliability of regional volumes with the intra-class correlation coefficient (ICC) and the volume differences derived from the two test-retest images. In Fig. 6, we plot the ICC values (and its 95% confidence interval) of *CerebNet* and ACAPULCO^{rt} for the two datasets. Statistical significance tests, however, are directly performed on volume differences using a Wilcoxon signed-rank test to compare the methods. The ICCs of *CerebNet* and ACAPULCO^{rt} range between 0.635 and 0.997 across both datasets with – in most cases – more consistent results (higher ICC) for *CerebNet*. In fact, the ICC is superior for *CerebNet* over ACAPULCO^{rt} in 24 of 27 sub-structures for the Kirby data and in 23 out of 27 sub-structures for the OASIS1 data set set as well as for the combined regions of the vermis and the left and right hemispheric CGM. This difference was significant in 17 (9) out of all 30 structures for the OASIS1 (Kirby) data set (only once in favor of ACAPULCO^{rt}, Fig. 6). In particular, *CerebNet* was more consistent as evidenced by much lower standard deviations and smaller 95% confidence intervals of the ICC for each sub-structure in comparison to ACAPULCO^{rt} (Fig. 6).

3.5. Volumetric changes in pre-ataxic and ataxic spinocerebellar ataxia type 3 (SCA3)

We analyzed the cerebellar volumes of 109 SCA3 mutation carriers and 41 healthy controls (HC), who are participants of ongoing observational studies and gave their written informed consent. MRI were acquired at 7 EU and 2 US sites. All T1-weighted images were acquired on 3T SIEMENS scanners (Siemens Medical Systems, Erlangen, Germany) with an isotropic resolution of 1mm. To establish generalizability of *CerebNet* to this dataset, we visually inspect a random subset of 5 cases per group (total $N = 15$) finding good segmentation quality with no outliers.

To investigate group differences between pre-ataxic and ataxic SCA3 as well as healthy controls, we used a linear mixed-effects model with the co-variables age and estimated total intracranial volume (eTIV) as well as group (pre-ataxic SCA3, ataxic SCA3 and HC) and sex as fixed and scanner as random factors, respectively. Ataxia severity was assessed with the Scale for Assessment and Rating of Ataxia (SARA) (Schmitz-Hübsch et al., 2006). We applied the common SARA cut-off value of 3 to divide the group of SCA3 mutation carriers into pre-ataxic (SARA < 3) and ataxic (SARA ≥ 3) individuals (Jacobi et al., 2020). The eTIV was assessed using *FreeSurfer* 6.0 (Buckner et al., 2004). Cerebellar volumes were compared between pre-ataxic SCA3 (N= 42, mean age 38.02 years, 62.91% female, mean SARA 1.25) and ataxic SCA3 (N= 67, mean age 49.94 years, 35.82% female, mean SARA 12.05) as well as healthy controls (N= 41, mean age 43.95, female 43.90%, mean SARA 0.27). In the post-hoc analyses of pairwise comparisons, we applied Bonferroni correction for multiple comparisons. P-values smaller than $p < .05$ after Bonferroni correction were considered significant.

For *CerebNet*-derived per-region volumes, pre-ataxic SCA3 mutation carriers already showed significant volume reduction in comparison to HC in the right lobules I-IV, left and right lobule X, vermis IX as well as the left and right CWM. We detected significant volume reduction of ataxic patients in comparison to pre-ataxic SCA3 mutation carriers in left and right lobule VI, Crus II, VIIb, VIIIa and left VIIIb, left and right X and the left and right CWM. These results reaffirm that cerebellar neurodegeneration already starts before the clinical onset of the

Fig. 4. Dice score (larger values are better) and Robust Hausdorff Distance (HD95) (smaller values are better) per sub-structure for *CerebNet*, ACAPULCO^{rt} and SUIT+FS. Illustrations show the cross-subject average of the metric (bar) and corresponding, bootstrapped 95% confidence intervals (error bars), data points (eight per bar, may overlap) as well as the significance level calculated by a Wilcoxon signed-rank test ($*$: $p < .05$ and $**$: $p < .01$). CGM: Cerebellar Gray Matter; CWM: Cerebellar White Matter.

disease and is ongoing throughout the disease course with a very early and continuous involvement of cerebellar white matter.

In Fig. 7, we evaluate the power of our neuro-morphometric measures to separate between different groups: HC and pre-ataxic SCA3 mutation carriers (top) as well as pre-ataxic SCA3 and ataxic SCA3 (bottom). While in a direct competition of methods, only *CerebNet* and the original ACAPULCO (Han et al., 2020b) are available publicly, we also include ACAPULCO^{rt} (which is retrained with our labels, see Section 2.4) to illustrate the impact of both our high-quality training data and its interaction with our segmentation pipeline. A clear difference between the methods is already apparent in the varying details of the highly significant CWM segmentations and its branches. Even ACAPULCO^{rt} does not achieve the degree of detail available in *CerebNet*, in spite of it using the same training data. While a direct comparison of p-values is usually not possible, here it is meaningful as the methods operate on exactly the same input images. This is because the p-values of the group effect are monotonically connected to the absolute value of the t-statistic (effect size divided by standard error). More significant effects, i.e. smaller p-values illustrated in Fig. 7 by more saturated colors, indicate better group separation. In the group comparison of pre-ataxic mutation carriers to HC, ACAPULCO showed unexpected, non-significant volume increases in two structures and ACAPULCO^{rt} in one structure (blue regions in Fig. 7). Comparing the p-values for group separation between pre-ataxic SCA3 mutation carriers and healthy controls, *Cereb-*

Fig. 5. Comparison of Inter-rater reliability and *CerebNet* by Dice score and Robust Hausdorff Distance (HD95) per sub-structure. Error bars indicate 95% confidence intervals. CGM is Cerebellar Gray Matter and CWM is Cerebellar White Matter (∗: $p < .05$ and ∗∗: $p < .01$).

Net showed smaller p-values in more structures than ACAPULCO (10 versus 4) as well as ACAPULCO^{rt} (9 versus 6). For the group separation between pre-ataxic and ataxic SCA3 mutation carriers, *CerebNet* showed smaller p-values in 15 structures compared to 6 for ACAPULCO and 13 compared to 11 for ACAPULCO^{rt}. Given that the true group differences for each sub-structure are unknown, these results cannot establish a final superiority, but they can assure that known and expected effects can be reliably detected and that this signal is recovered most strongly with *CerebNet*.

4. Discussion

Neuroanatomical volumetry is a promising imaging biomarker candidate to assess progressive neurodegeneration in clinical trials. The advantages are, first, that non-invasive T1-weighted MRI is widely available, second, that precise quantitative estimates can aid studies into disease progression even at early stages, and third, that these volume estimates permit assessing subtle changes to quantify atrophy rates in various disease stages and effects of potential interventions and disease modifying therapies. Especially quantitative estimates of cerebellar structures are highly relevant for studying ataxia, in particular for those ataxia disorders where clinical trials have already been initiated, such as SCA3. Therefore, with this work, we introduce a multi-stage protocol for reliable and repeatable cerebellum segmentation with carefully drawn and quality-assured boundaries, establish a manually segmented reference dataset, and develop and validate *CerebNet*, a fast and accurate method to automatically sub-segment the cerebellum into its lobules and the cerebellar WM from a T1-weighted MRI.

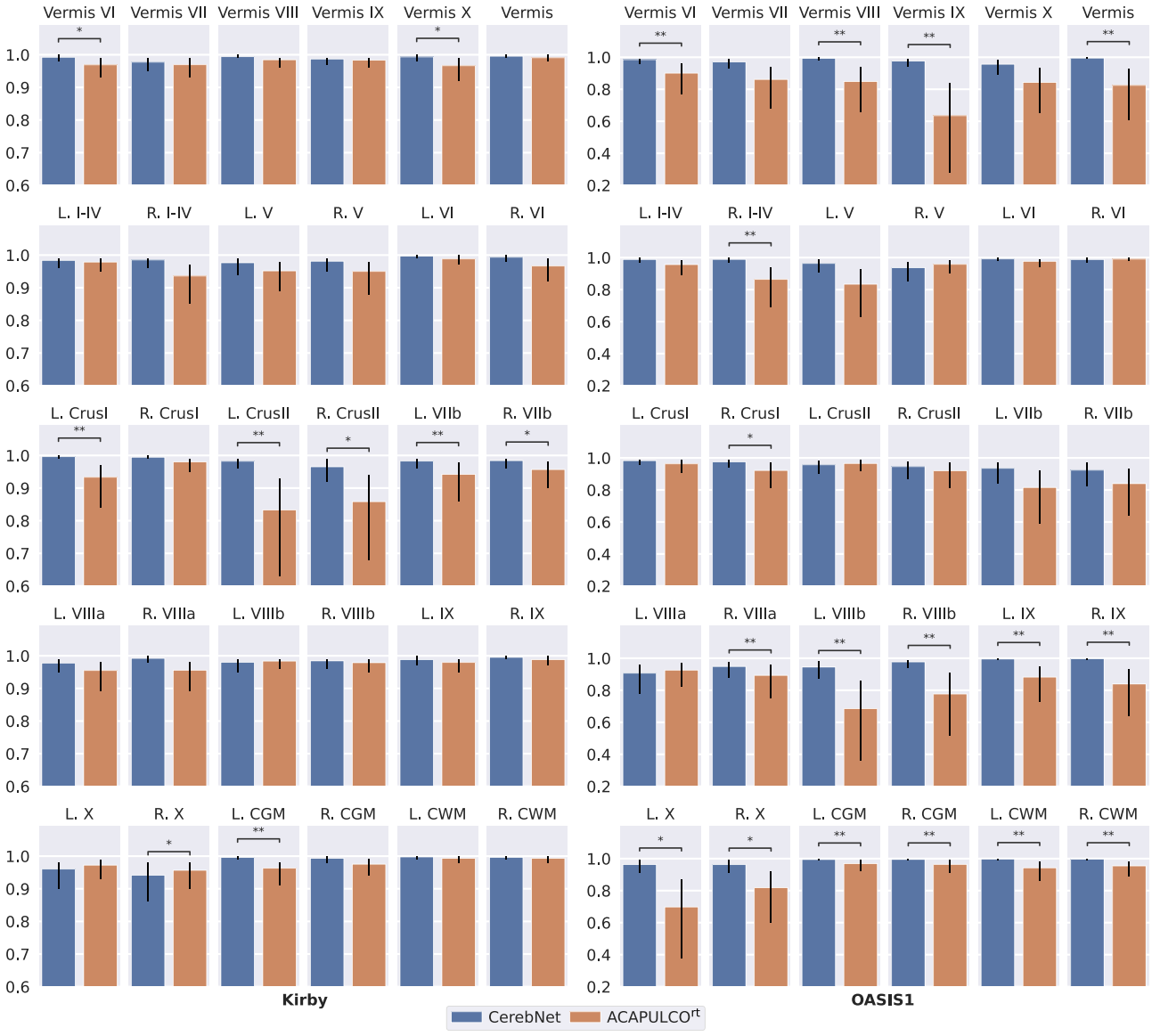

Fig. 6. Intraclass correlation coefficient (ICC) on volume of Kirby and OASIS1 datasets for test-retest analysis. Error bars indicate the 95% confidence interval. Statistical significance is calculated with a two-sided non-parametric Wilcoxon signed-rank test over the absolute volume difference, since ICC values cannot provide significance information. $*$ and $**$ annotations represent statistical significance for better volume consistency with $p < .05$ and $p < .01$, respectively.

Our method *CerebNet* employs a *FastSurferCNN* deep-learning model customized to our cerebellum training dataset. In contrast to state-of-the-art methods (Carass et al., 2018; Diedrichsen, 2006; Han et al., 2020b), *CerebNet* does not require any preprocessing, such as spatial normalization or bias field correction, thus preserving sufficient detail to segment even the fine branches of the white matter and simultaneously allowing rapid processing at only 12 seconds per MRI with one GPU (Nvidia Titan Xp). Fast MRI segmentation in general opens up multiple avenues of potential applications, ranging from direct feedback or field-of-view localization during image acquisition or fast clinical decision support by quantitative personalized measurements. In addition to speed, we demonstrate in an extensive validation that the *CerebNet* pipeline outperforms state-of-the-art approaches and provides detailed segmentation masks especially for white matter strands.

Our quantitative analysis illustrates *CerebNet*'s superior segmentation quality in both volumetric and geometric metrics. Furthermore, we demonstrate *CerebNet*'s superior test-retest reliability and show-case its utility to down-stream group analysis: While clinical scales lack sensitivity in pre-ataxic cases, simply due to the absence of symptoms, *CerebNet* reliably identifies patterns of cerebellar degeneration consistent with previous studies (Faber et al., 2020; Kim et al., 2021; Rezende et al., 2018). Consequently volumetric estimates of the cerebellum, especially subtle longitudinal changes, are promising imaging biomarker candidates to assess the effect of preventive genetic therapies during the pre-ataxic stage and might play a central role as stratification markers or even as secondary outcome parameters in clinical trials.

A qualitative inspection of the predicted segmentation maps illustrates the different character of the presented pipelines. *CerebNet*-derived segmentation maps feature the highest level of detail, especially visible at the intricate boundary between CWM and CGM (see Fig. 7). In fact, comparing predictions of *CerebNet*, ACAPULCO and ACAPULCO^{rt}[4], we find that the level of detail of ACAPULCO^{rt} lies between *CerebNet* and ACAPULCO (see Fig. 7), highlighting both the added value of our dataset with its manual segmentations (ACAPULCO^{rt} vs. ACAPULCO) and of our method (*CerebNet* vs. ACAPULCO^{rt}). In contrast to volu-

[4] Since the ACAPULCO training dataset is not available publicly, we cannot retrain *CerebNet* with this data.

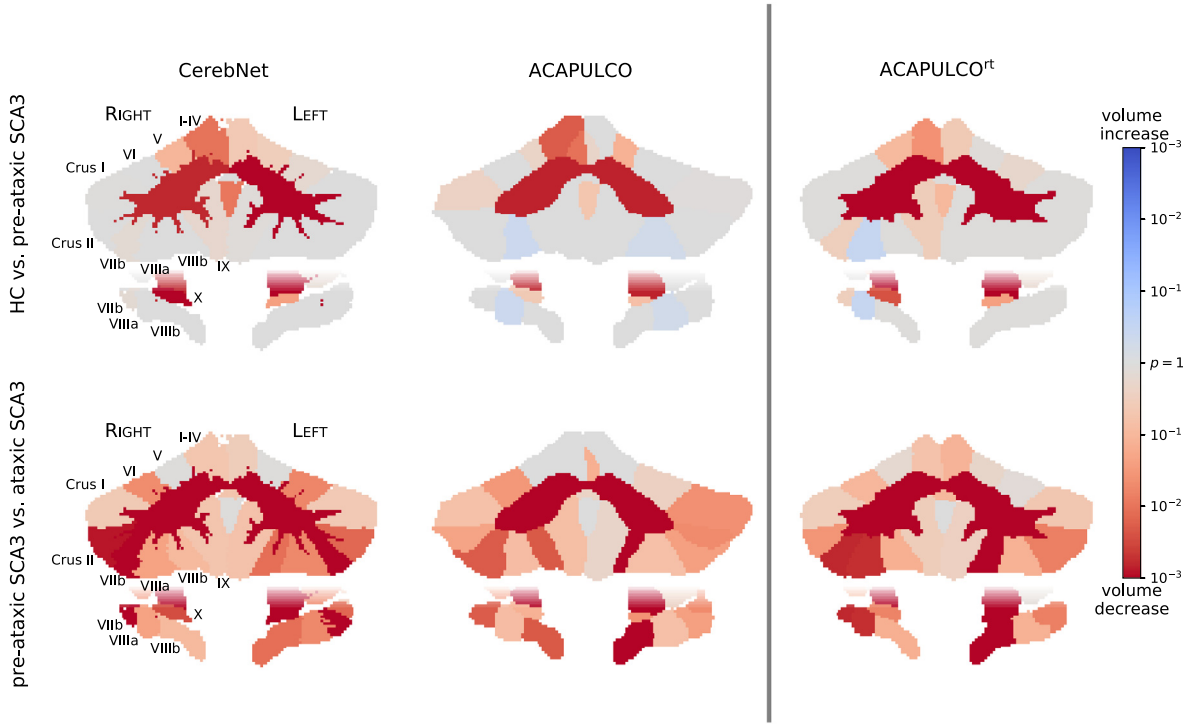

Fig. 7. Map of volume change in HC vs. pre-ataxic SCA3 (top) and pre-ataxic vs. ataxic SCA3 (bottom). Per-region p-values of the respective group comparisons are shown for 3 different methods: *CerebNet*, ACAPULCO (as distributed by Han et al., 2020b) and ACAPULCO^{rt} (ACAPULCO retrained on our dataset). Red colors indicate atrophy, blue colors indicate volume increase (color saturation corresponds to significance). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

metric analyses, which are relatively robust to limited detail in segmentation maps, structural and geometric analyses, including thickness analysis, rely on accurate and detailed boundaries (Sörös et al., 2021). Because of the fine-grained furcations of CWM, the delineation of the CGM/CWM boundary is particularly challenging both for manual raters and automated methods. *CerebNet* especially improves these boundaries as proven by significantly improved Dice and Robust Hausdorff metrics over both ACAPULCO^{rt} and FreeSurfer (see L./R. CGM/CWM in Fig. 4).

In general, a critical limitation of learning-based approaches remains in the uncertain generalizability beyond images similar to those encountered during training. This limitation also applies to *CerebNet*. While our reference dataset features diversity in terms of severity of ataxia and cerebellar atrophy, the generalizability to other datasets is not guaranteed, since we only included T1w MRI of healthy controls and SCA3 mutation carriers acquired on SIEMENS scanners. Therefore, as for any method, dedicated experimental validation is required to confirm the validity under differing conditions, i.e. at least rigorous, manual quality checks of generated segmentations. Given the convincing test-retest performance (c.f. Fig. 6 Kirby dataset, which was acquired with Philips scanners), we are optimistic *CerebNet*'s extensive augmentation may already enable basic generalizability to other scan-settings. Furthermore, we visually inspected automatically generated segmentations of several clinically diagnosed sporadic and hereditary ataxias to verify whether *CerebNet* generalizes to other pathologies ($N = 14$: two randomly selected cases of MSA-C, RFC1, SCA1, SCA2 and SCA6, AOA2 as well as one case of SYNE1 and CTX each, including cases with severe atrophy, see also Fig. 9 in the Appendix). While not a formal validation for these pathologies, we found the segmentation quality among these cases comparable to our SCA3 cases without fails or unacceptable quality, further supporting the generalizability of *CerebNet*.

Obviously, volumetric analyses of other sporadic and hereditary neurodegenerative ataxias are canonical further research questions. Moreover, *CerebNet* may enable cerebellar analyses of aging, non-motor diseases (e.g. Alzheimer's disease or attention deficit hyperactivity disorder) and combined analyses of imaging and neuropsychological data. For these applications the focus may shift to parts of the cerebellum primarily involved in the adaptive control of non-motor processes. Since the functional representation of cognitive tasks is oriented across lobules along a parasagittal axis (King et al., 2019), the utility of segmentation along the anatomical boundaries of the hemispheric lobules is unclear for studies of the cerebellar involvement in cognitive and emotional processes.

In summary, *CerebNet* offers significant improvements and advantages for users in terms of runtime, accuracy, reliability and sensitivity to subtle cerebellar atrophy. Thus, we are confident, that *CerebNet* will enable and simplify the detailed morphometric analysis of the cerebellum.

Acknowledgments

We would like to thank Beate Brol, Tim Elter, Isabelle Finkel, and Sophia Wismeth for their contribution to the manual segmentation. This work was supported by the National Ataxia Foundations SCA Young Investigator Award as well as by DZNE institutional funds, by the Federal Ministry of Education and Research of Germany (031L0206, 01GQ1801), and by NIH (R01 LM012719, R01 AG064027, R56 MH121426, and P41 EB030006). JF is fellow of the Hertie Network of excellence in clinical Neuroscience. This publication is an outcome of ESMI, an EU Joint Programme - Neurodegenerative Disease Research (JPND) project (see www.jpnd.eu). The project is supported through the following funding organisations under the aegis of JPND: Germany, Federal Ministry of Education and Research (BMBF; funding codes 01ED1602A/B); Netherlands, The Netherlands Organisation for Health Research and Development; Portugal, Foundation for Science and Technology and Regional Fund for Science and Technology of the Azores; United Kingdom, Medical Research Council. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 643417. For the

contribution of the Minnesota site, this work was in part supported by the National Ataxia Foundation and the National Institute of Neurological Disorders and Stroke (NINDS) grant R01NS080816. The Center for Magnetic Resonance Research is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) grant P41 EB027061, and the Institutional Center Cores for Advanced Neuroimaging award P30 NS076408 and S10OD017974 grant. Data used in the preparation of this article for pre-training and augmentation were obtained in part by the OASIS Cross-Sectional with principal investigators D. Marcus, R, Buckner, J, Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382, and OASIS: Longitudinal: Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. Further, data used in the preparation of this article were obtained from the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC). Data used in preparation of this article were obtained from the Alzheimers 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. Data collection and sharing for this project 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, Alzheimers Association; Alzheimers 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 Alzheimers 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. Data were also provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

5. Appendix

5.1. Dice score and Hausdorff metric compared between CerebNet, ACAPULCO^{rt}, original ACAPULCO and SUIT + FS

To motivate our choice of pre-training with SUIT + FS and illustrate the impact of the dataset, we compare four methods on our test set: *CerebNet*, the ACAPULCO^{rt} (Han et al., 2020b) (both trained on our training set), as well as the original ACAPULCO (Han et al., 2020b) and SUIT + FS (Diedrichsen et al., 2009) (both trained on their individual datasets). We compare each prediction with our manually labeled reference segmentation to obtain average Dice and Robust Hausdorff metrics for each method in Fig. 8. Two observations are notable: 1. SUIT + FS outperforms the original ACAPULCO; 2. ACAPULCO^{rt} outperforms ACAPULCO (and SUIT + FS). Both results are expected and illustrate how much different labeling protocols (e.g. along the CGM/CWM border) can impact the performance and ranking of methods. These results confirm that inconsistent labeling protocols between training and test significantly impact the measured performance even to the level of contradicting previous rankings (Carass et al., 2018; Han et al., 2020b). Therefore, we 1. choose SUIT + FS for pre-training and 2. retrain ACAPULCO (yielding ACAPULCO^{rt}) so that our methodological comparison are fair and not impacted by the choice of protocols.

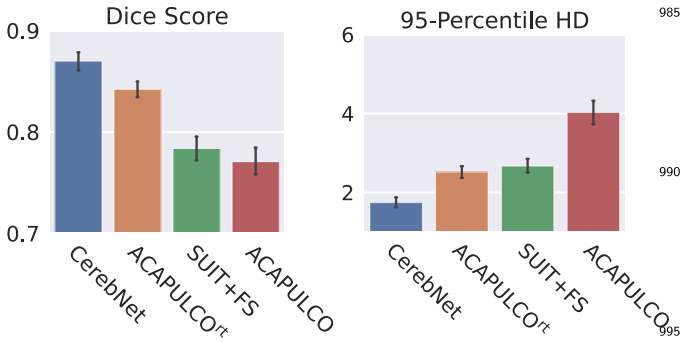

Fig. 8. Comparison of Dice and Robust Hausdorff Distance (HD95) metrics for CerebNet, the retrained ACAPULCO^{rt}, SUIT + FS as well as the original (published) ACAPULCO on our test set. Note, the direct comparison of *CerebNet* and the original ACAPULCO does not correct for the differences in the training datasets; ACAPULCO^{rt} corrects for this difference (see text).

This analysis also raises the question of how the performance of two pipelines may be compared fairly (a pipeline evaluation includes the impact of both the training dataset and the method). This is specifically difficult if protocols differ and higher quality reference standards are not available. While retraining on the same dataset yields a fair, direct comparison of methods (see for example section 3.2), pipeline comparisons, in situations where retraining is not feasible, require indirect evaluations based on segmentation-derived metrics instead of segmentation maps, e.g. whether and how well volume estimates can be used to differentiate between patient groups as done for SCA3 in Section 3.5.

5.2. Correlation of SARA sum scores with volumetric estimates

We perform a correlation analysis between the SARA sum score and regional volumes. Table 2 shows individual Kendell Tau coefficients for three methods: *CerebNet*, ACAPULCO (original) and ACAPULCO^{rt} (retrained). We also report whether the analysis achieved statistical significance. However, we would like to note that SCA3 is not a pure cerebllar disease like for example SCA6. The patterns of neurodegeneration in SCA3 include non-cerebellar structures, e.g. the basal ganglia or the peripheral nerve system. Progressive neurodegeneration of the cerebellum might be the main driver of ataxia severity in SCA3, but symptoms resulting from non-cerebellar manifestations, like e.g. spasticity or polyneuropathy have a direct impact on SARA items such as gait and stance. This should be taken into account in the interpretation of the correlation.

5.3. CerebNet segmentation in sporadic and hereditary ataxias

To illustrate the robustness of *CerebNet* to "out-of-distribution" samples, we show some qualitative examples of segmentations in Fig. 9.

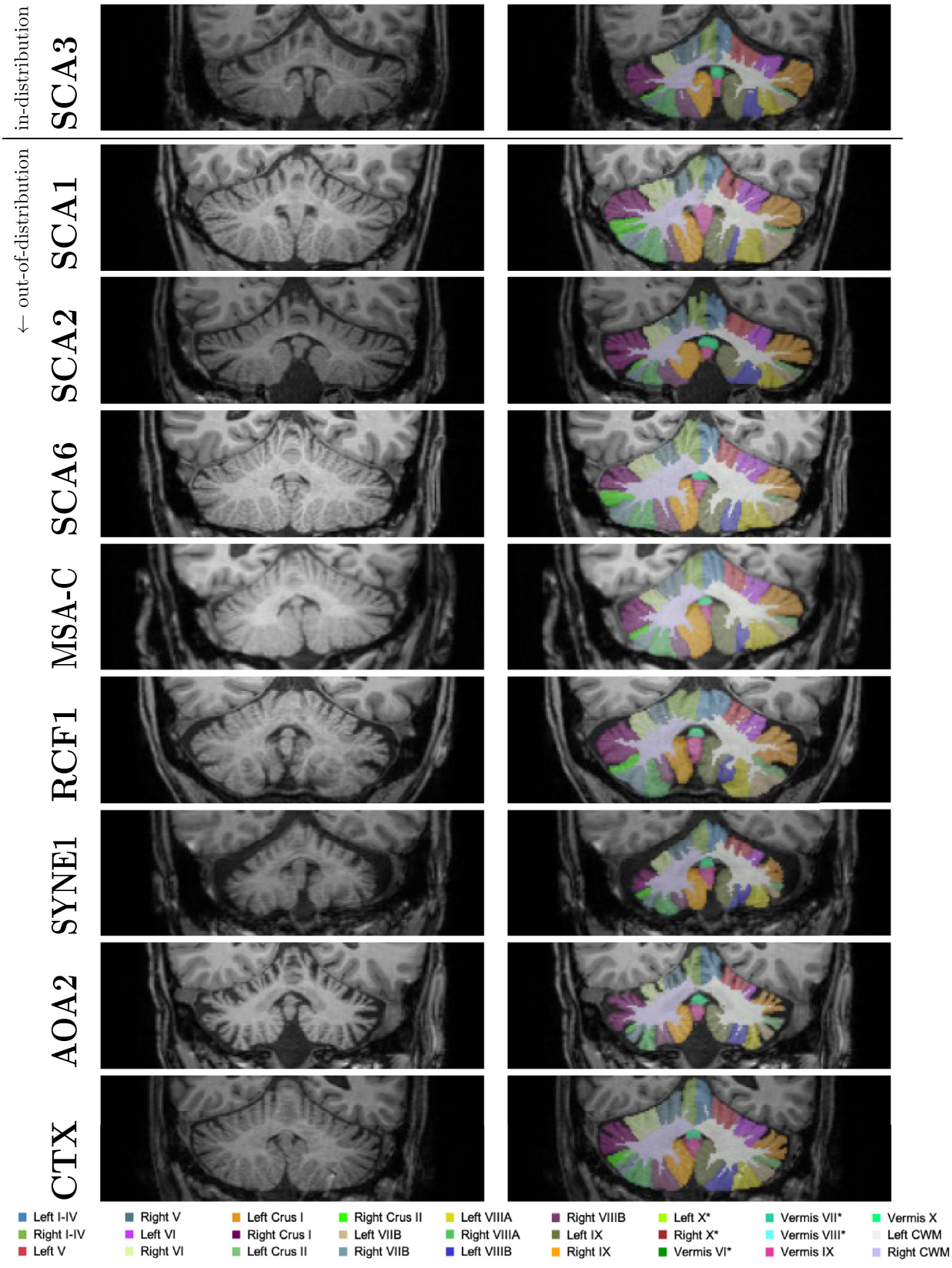

Fig. 9. Qualitative "out-of-distribution" evaluation of *CerebNet*: Segmentation maps for pathologies, which are not part of the training (SCA1, SCA2, SCA6, MSA-C, RFC1, SYNE1, AOA2 and CTX) together with an in-distribution example (SCA3). Images illustrated here are randomly picked from a larger repository of images and represent average performance. *: Label is not visible in the shown slice.

Table 2
Correlation of SARA sum score with each cerebellar volume for CerebNet, ACAPULCO and ACAPULCO^{rt}. Kendell Tau correlation coefficients are given, and statistical significance of the correlation is indicated by * ($p < .05$) and ** ($p < .01$). For each volume, the most negative, statistically significant correlation coefficient is printed in boldface.

	CerebNet	ACAPULO (original)	ACAPULCO^{rt} (retrained)
Left I-IV	-0.204**	-0.134*	**-0.211****
Right I-IV	**-0.235****	-0.167*	-0.208**
Left V	-0.0602	**-0.213****	-0.0593
Right V	-0.00294	**-0.194****	-0.0401
Left VI	**-0.306****	-0.227**	-0.284**
Vermis VI	-0.0954	-0.0197	-0.0874
Right VI	**-0.322****	-0.225**	-0.264**
L. Crus I	-0.252**	**-0.287****	-0.214**
R. Crus I	**-0.200****	-0.199**	-0.151*
L. Crus II	-0.219**	**-0.267****	-0.145*
R. Crus II	**-0.214****	-0.179**	-0.154*
Left VIIb	**-0.385****	-0.241**	-0.341**
Right VIIb	**-0.410****	-0.364**	-0.314**
Vermis VII	**-0.144***	-0.139*	-0.0765
Left VIIIa	**-0.299****	-0.113	-0.264**
Right VIIIa	-0.315**	-0.100	**-0.408****
Left VIIIb	-0.317**	-0.311**	**-0.374****
Right VIIIb	-0.254**	**-0.344****	-0.323**
Vermis VIII	**-0.182****	-0.0941	-0.170**
Left IX	**-0.258****	-0.159*	-0.232**
Vermis IX	-0.0650	-0.0327	-0.0760
Right IX	**-0.292****	-0.248**	-0.240**
Left X	**-0.265****	0.0101	-0.207**
Vermis X	-0.115	-0.0667	-0.0270
Right X	**-0.298****	-0.0598	-0.253**
CWM	**-0.583****	-0.440**	-0.549**

While these results indicate good performance across many pathologies, studies utilizing *CerebNet* to segment patients with these or other diseases should still ensure the performance also translates to their datasets by performing a formal validation, or, at least, rigorous quality assurance as laid out in the Discussion.

Data and Code Availability Statement

The MRI data is not publicly available because of data protection regulations. Access can be provided upon reasonable request to scientists in accordance with our Data Use and Access Policy. Requests to access the data should be directed to Jennifer Faber at Jennifer.Faber@dzne.de.

The source code of CerebNet will be made publicly available on Github (https://github.com/Deep-MI/FastSurfer) upon acceptance.

References

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41.

Bogovic, J.A., Bazin, P.L., Ying, S.H., Prince, J.L., 2013. Automated segmentation of the cerebellar lobules using boundary specific classification and evolution. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 7917 LNCS, pp. 62–73.

Bogovic, J.A., Jedynak, B., Rigg, R., Du, A., Landman, B.A., Prince, J.L., Ying, S.H., 2013. Approaching expert results using a hierarchical cerebellum parcellation protocol for multiple inexpert human raters. NeuroImage 64, 616–629. doi:10.1016/j.neuroimage.2012.08.075.

Bogovic, J.A., Prince, J.L., Bazin, P.L., 2013. A multiple object geometric deformable model for image segmentation. Comput. Vision Image Understanding 117 (2), 145–157.

Buckner, R., Head, D., Parker, J., Fotenos, A., Marcus, D., Morris, J., Snyder, A., 2004. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. NeuroImage 23, 724–738. doi:10.1016/j.neuroimage.2004.06.018.

Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Yeo, B.T., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. J. Neurophysiol. 106 (5), 2322–2345. doi:10.1152/jn.00339.2011.

Carass, A., Cuzzocreo, J.L., Han, S., Hernandez-Castillo, C.R., Rasser, P.E., Ganz, M., Beliveau, V., Dolz, J., Ben Ayed, I., Desrosiers, C., Thyreau, B., Romero, J.E., Coupé, P., Manjón, J.V., Fonov, V.S., Collins, D.L., Ying, S.H., Onyike, C.U., Crocetti, D., Landman, B.A., Mostofsky, S.H., Thompson, P.M., Prince, J.L., 2018. Comparing fully automated state-of-the-art cerebellum parcellation from magnetic resonance images. NeuroImage 183, 150–172.

Carass, A., Prince, J.L., 2016. An overview of the multi-object geometric deformable model approach in biomedical imaging. In: Medical Image Recognition, Segmentation and Parsing, pp. 259–279.

Chakravarty, M.M., Steadman, P., van Eede, M.C., Calcott, R.D., Gu, V., Shaw, P., Raznahan, A., Collins, D.L., Lerch, J.P., 2013. Performing label-fusion-based segmentation using multiple automatically generated templates. Hum. Brain Mapp. 34 (10), 2635–2654.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954.

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L.M.E., Bookheimer, S.Y., Braden, B.B., Byrge, L., Castellanos, F.X., Dapretto, M., Delorme, R., Fair, D.A., Fishman, I., Fitzgerald, J., Gallagher, L., Keehn, R.J.J., Kennedy, D.P., Lainhart, J.E., Luna, B., Mostofsky, S.H., Müller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Solomon, M., Toro, R., Vaidya, C.J., Wenderoth, N., White, T., Craddock, R.C., Lord, C., Leventhal, B., Milham, M.P., 2017. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci. Data 4 (1), 1–15.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302. doi:10.2307/1932409.

Diedrichsen, Zotow, J., Ewa, 2015. Surface-based display of volume-averaged cerebellar imaging data. PLoS ONE 10, 1–18.

Diedrichsen, J., 2006. A spatially unbiased atlas template of the human cerebellum. NeuroImage 33 (1), 127138. doi:10.1016/j.neuroimage.2006.05.056.

Diedrichsen, J., Balsters, J.H., Flavell, J., Cussans, E., Ramnani, N., 2009. A probabilistic MR atlas of the human cerebellum. NeuroImage 46 (1), 39–46. doi:10.1016/j.neuroimage.2009.01.045.

D'Mello, A.M., Crocetti, D., Mostofsky, S.H., Stoodley, C.J., 2015. Cerebellar gray matter and lobular volumes correlate with core autism symptoms. NeuroImage 7, 631–639. doi:10.1016/j.nicl.2015.02.007.

Estrada, S., Conjeti, S., Ahmad, M., Navab, N., Reuter, M., 2018. Competition vs. concatenation in skip connections of fully convolutional networks. In: Shi, Y., Suk, H.-I., Liu, M. (Eds.), Machine Learning in Medical Imaging. Springer International Publishing, Cham, pp. 214–222.

Estrada, S., Lu, R., Conjeti, S., Orozco-Ruiz, X., Panos-Willuhn, J., Breteler, M.M.B., Reuter, M., 2020. Fatsegnet: a fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. Magn. Reson. Med. 83 (4), 1471–1483.

Faber, J., Giordano, I., Jiang, X., Kindler, C., Spottke, A., Acosta-Cabronero, J., Nestor, P.J., Machts, J., Düzel, E., Vielhaber, S., Speck, O., Dudesek, A., Kamm, C., Scheef, L., Klockgether, T., 2020. Prominent white matter involvement in multiple system atrophy of cerebellar type. Movement Disorders 35 (5), 816–824. doi:10.1002/mds.27987.

Faber, J., Schaprian, T., Berkan, K., Reetz, K., França Jr, M.C., de Rezende, T.J.R., Hong, J., Liao, W., van de Warrenburg, B., van Gaalen, J., Durr, A., Mochel, F., Giunti, P., Garcia-Moreno, H., Schoels, L., Hengel, H., Synofzik, M., Bender, B., Oz, G., Joers, J., de Vries, J.J., Kang, J.-S., Timmann-Braun, D., Jacobi, H., Infante, J., Joules, R., Romanzetti, S., Diedrichsen, J., Schmid, M., Wolz, R., Klockgether, T., 2021. Regional brain and spinal cord volume loss in spinocerebellar ataxia type 3. Movement Disorders doi:10.1002/mds.28610.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355. doi:10.1016/S0896-6273(02)00569-X.

Giraud, R., Ta, V.T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P., 2016. An optimized patchmatch for multi-scale and multi-feature label fusion. NeuroImage 124, 770–782.

Guha Roy, A., Conjeti, S., Navab, N., Wachinger, C., 2019. QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage 186, 713–727.

Han, S., An, Y., Carass, A., Prince, J.L., Resnick, S.M., 2020. Longitudinal analysis of regional cerebellum volumes during normal aging. NeuroImage 220, 117062. doi:10.1016/j.neuroimage.2020.117062.

Han, S., Carass, A., He, Y., Prince, J.L., 2020. Automatic cerebellum anatomical parcellation using U-Net with locally constrained optimization. NeuroImage 218, 116819.

Han, S., He, Y., Carass, A., Ying, S.H., Prince, J.L., 2019. Cerebellum parcellation with convolutional neural networks. In: Proceedings of SPIE–the International Society for Optical Engineering, Vol. 10949, p. 19.

Heinz, L., Faber, J., Timmann, D., Ernst, T., Deike-Hofmann, K., Klockgether, T., 2022. Manual sub-segmentation of the cererbellum. medRxiv doi:10.1101/2022.05.09.22274814.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219, 117012.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15 (9), 850–863.

Jacobi, H., du Montcel, S.T., Romanzetti, S., Harmuth, F., Mariotti, C., Nanetti, L., Rakow-

icz, M., Makowicz, G., Durr, A., Monin, M.L., Filla, A., Roca, A., Schols, L., Hengel, H., Infante, J., Kang, J.S., Timmann, D., Casali, C., Masciullo, M., Baliko, L., Melegh, B., Nachbauer, W., Burk-Gergs, K., Schulz, J.B., Riess, O., Reetz, K., Klockgether, T., 2020. Conversion of individuals at risk for spinocerebellar ataxia types 1, 2, 3, and 6 to manifest ataxia (RISCA): a longitudinal cohort study. Lancet Neurol. 19 (9), 738–747. doi:10.1016/S1474-4422(20)30235-0.

Kim, D.-H., Kim, R., Lee, J.-Y., Lee, K.-M., 2021. Clinical, imaging, and laboratory markers of premanifest spinocerebellar ataxia 1, 2, 3, and 6: a systematic review. J. Clin. Neurol. 17 (2), 187–199. doi:10.3988/jcn.2021.17.2.187.

King, M., Hernandez-Castillo, C.R., Poldrack, R.A., Ivry, R.B., Diedrichsen, J., 2019. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. Nat. Neurosci. 22 (8), 1371–1378. doi:10.1038/s41593-019-0436-x.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), International Conference on Learning Representations, ICLR.

Landman, B.A., Huang, A.J., Gifford, A., Vikram, D.S., Lim, I.A.L., Farrell, J.A.D., Bogovic, J.A., Hua, J., Chen, M., Jarso, S., Smith, S.A., Joel, S., Mori, S., Pekar, J.J., Barker, P.B., Prince, J.L., van Zijl, P.C.M., 2011. Multi-parametric neuroimaging reproducibility: a 3-t resource study. NeuroImage 54 (4).

van der Lijn, F., de Bruijne, M., Hoogendam, Y.Y., Klein, S., Hameeteman, R., Breteler, M.M.B., Niessen, W.J., 2009. Cerebellum segmentation in MRI using atlas registration and local multi-scale image descriptors. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 221–224.

Lin, C.-Y., Chen, C.-H., Tom, S.E., Kuo, S.-H., for the Alzheimer's Disease Neuroimaging Initiative, 2020. Cerebellar volume is associated with cognitive decline in mild cognitive impairment: results from ADNI. Cerebellum 19 (2), 217–225. doi:10.1007/s12311-019-01099-1.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: International Conference on Learning Representations.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M., et al., 2013. MIRIAD-Public Release of a multiple time point Alzheimer's MR imaging dataset. NeuroImage 70, 33–36. doi:10.1016/j.neuroimage.2012.12.044.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J. Cogn. Neurosci. 22 (12), 2677–2684.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507.

Marek, S., Siegel, J.S., Gordon, E.M., Raut, R.V., Gratton, C., Newbold, D.J., Ortega, M., Laumann, T.O., Adeyemo, B., Miller, D.B., Zheng, A., Lopez, K.C., Berg, J.J., Coalson, R.S., Nguyen, A.L., Dierker, D., Van, A.N., Hoyt, C.R., McDermott, K.B., Norris, S.A., Shimony, J.S., Snyder, A.Z., Nelson, S.M., Barch, D.M., Schlaggar, B.L., Raichle, M.E., Petersen, S.E., Greene, D.J., Dosenbach, N.U.F., 2018. Spatial and temporal organization of the individual human cerebellum. Neuron 100 (4), 977–993.e7. doi:10.1016/j.neuron.2018.10.010.

McLoughlin, H.S., Moore, L.R., Chopra, R., Komlo, R., McKenzie, M., Blumenstein, K.G., Zhao, H., Kordasiewicz, H.B., Shakkottai, V.G., Paulson, H.L., 2018. Oligonucleotide therapy mitigates disease in spinocerebellar ataxia type 3 mice. Ann. Neurol. 84 (1), 64–77. doi:10.1002/ana.25264.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., Beckett, L., 2005. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI).

Okugawa, G., Sedvall, G.C., Agartz, I., 2003. Smaller cerebellar vermis but not hemisphere volumes in patients with chronic schizophrenia. Am. J. Psychiatry 160 (9), 1614–1617. doi:10.1176/appi.ajp.160.9.1614.

Park, M.T.M., Pipitone, J., Baer, L.H., Winterburn, J.L., Shah, Y., Chavez, S., Schira, M.M., Lobaugh, N.J., Lerch, J.P., Voineskos, A.N., Chakravarty, M.M., 2014. Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates. NeuroImage 95, 217–231. doi:10.1016/j.neuroimage.2014.03.037.

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.

Pierson, R., Corson, P.W., Sears, L.L., Alicata, D., Magnotta, V., O'Leary, D., Andreasen, N.C., 2002. Manual and semiautomated measurement of cerebellar subregions on MR images. NeuroImage 17 (1).

Plassard, A.J., Yang, Z., Rane, S., Prince, J.L., Claassen, D.O., Landman, B.A., 2016. Improving cerebellar segmentation with statistical fusion. In: Medical Imaging 2016: Image Processing, Vol. 9784, p. 97842R.

Poldrack, R.A., Congdon, E., Triplett, W., Gorgolewski, K.J., Karlsgodt, K.H., Mumford, J.A., Sabb, F.W., Freimer, N.B., London, E.D., Cannon, T.D., Bilder, R.M., 2016. A phenome-wide examination of neural and cognitive function. Sci.. Data 3 (1), 1–12.

Powell, S., Magnotta, V.A., Johnson, H., Jammalamadaka, V.K., Pierson, R., Andreasen, N.C., 2008. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. NeuroImage 39 (1), 238–247. doi:10.1016/j.neuroimage.2007.05.063.

Price, M., Cardenas, V.A., Fein, G., 2014. Automated MRI cerebellar size measurements using active appearance modeling. NeuroImage 103, 511–521.

Rezende, T.J.R., de Paiva, J.L.R., Martinez, A.R.M., Lopes-Cendes, I., Pedroso, J.L., Barsottini, O.G.P., Cendes, F., França Jr., M.C., 2018. Structural signature of SCA3: from presymptomatic to late disease stages. Ann. Neurol. 84 (3), 401–408. doi:10.1002/ana.25297.

Romero, J.E., Coupé, P., Giraud, R., Ta, V.T., Fonov, V., Park, M.T.M., Chakravarty, M.M., Voineskos, A.N., Manjón, J.V., 2017. CERES: A new cerebellum lobule segmentation method. NeuroImage 147, 916–924.

Schmahmann, J.D., Doyon, J., McDonald, D., Holmes, C., Lavoie, K., Hurwitz, A.S., Kabani, N., Toga, A., Evans, A., Petrides, M., 1999. Three-dimensional MRI atlas of the human cerebellum in proportional stereotaxic space. NeuroImage 10 (3), 233–260.

Schmitz-Hübsch, T., du Montcel, S.T., Baliko, L., Berciano, J., Boesch, S., Depondt, C., Giunti, P., Globas, C., Infante, J., Kang, J.-S., Kremer, B., Mariotti, C., Melegh, B., Pandolfo, M., Rakowicz, M., Ribai, P., Rola, R., Schöls, L., Szymanski, S., van de Warrenburg, B.P., Dürr, A., Klockgether, T., 2006. Scale for the assessment and rating of ataxia. Neurology 66 (11), 1717–1720.

Shrout, P., Fleiss, J., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 2, 420–428.

Sörös, P., Wölk, L., Bantel, C., Bräuer, A., Klawonn, F., Witt, K., 2021. Replicability, repeatability, and long-term reproducibility of cerebellar morphometry. Cerebellum.

Sørensen, Julius, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons 5, 1–34.

Ta, V.-T., Giraud, R., Collins, D.L., Coupé, P., 2014. Optimized patchmatch for near real time and accurate label fusion. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, pp. 105–112.

Toniolo, S., Serra, L., Olivito, G., Marra, C., Bozzali, M., Cercignani, M., 2018. Patterns of cerebellar gray matter atrophy across alzheimers disease progression. Front. Cell Neurosci. 12, 430. doi:10.3389/fncel.2018.00430.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., 2012. The human connectome project: a data acquisition perspective. NeuroImage 62 (4), 2222–2231.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 897–908.

Webb, S.J., Sparks, B.-F., Friedman, S.D., Shaw, D.W.W., Giedd, J., Dawson, G., Dager, S.R., 2009. Cerebellar vermal volumes and behavioral correlates in children with autism spectrum disorder. Psychiatry Res. 172 (1), 61–67. doi:10.1016/j.pscychresns.2008.06.001.

Weier, K., Fonov, V., Lavoie, K., Doyon, J., Louis Collins, D., 2014. Rapid automatic segmentation of the human cerebellum and its lobules (RASCAL)-Implementation and application of the patch-based label-fusion technique with a template library to segment the human cerebellum. Hum. Brain Mapp. 35 (10), 5026–5039.

Womer, F.Y., Tang, Y., Harms, M.P., Bai, C., Chang, M., Jiang, X., Wei, S., Wang, F., Barch, D.M., 2016. Sexual dimorphism of the cerebellar vermis in schizophrenia. Schizophr. Res. 176 (2), 164–170. doi:10.1016/j.schres.2016.06.028.

Yang, Z., Ye, C., Bogovic, J.A., Carass, A., Jedynak, B.M., Ying, S.H., Prince, J.L., 2016. Automated cerebellar lobule segmentation with application to cerebellar structural analysis in cerebellar disease. NeuroImage 127, 435–444.