Comparing fully automated state-of-the-art cerebellum parcellation from magnetic resonance images

Aaron Carass[a,b,*], Jennifer L. Cuzzocreo[c,*], Shuo Han[d,e,*],
Carlos R. Hernandez-Castillo[f], Paul E. Rasser[g], Melanie Ganz[h,i],
Vincent Beliveau[h,j], Jose Dolz[k], Ismail Ben Ayed[k], Christian Desrosiers[k],
Benjamin Thyreau[l], José E. Romero[m], Pierrick Coupé[n,o], José V. Manjón[m],
Vladimir S. Fonov[p], D. Louis Collins[p], Sarah H. Ying[r,*], Deana Crocetti[q,*],
Bennett A. Landman[t,*], Stewart H. Mostofsky[q,r,s,*], Paul M. Thompson[u,v,*],
and Jerry L. Prince[a,b,*]

[a] *Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218, USA*
[b] *Department of Computer Science, The Johns Hopkins University, Baltimore, MD 21218, USA*
[c] *Department of Radiology, The Johns Hopkins School of Medicine, Baltimore, MD 21287, USA*
[d] *Department of Biomedical Engineering, The Johns Hopkins University, Baltimore, MD 21218, USA*
[e] *Laboratory of Behavioral Neuroscience, National Institute on Aging, National Institutes of Health, Baltimore, MD 20892, USA*
[f] *Consejo Nacional de Ciencia y Tecnología, Instituto de Neuroetología, Universidad Veracruzana, Xalapa, Mexico*
[g] *Priority Research Centre for Brain & Mental Health and Stroke & Brain Injury, University of Newcastle, Callaghan NSW, Australia*
[h] *Neurobiology Research Unit, Rigshospitalet, Copenhagen, Denmark*
[i] *Department of Computer Science, University of Copenhagen, Copenhagen, Denmark*
[j] *Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark*
[k] *Laboratory for Imagery, Vision, and Artificial Intelligence, École de Technologie Supérieure, Montreal, QC, Canada*
[l] *Institute of Development, Aging and Cancer, Tohoku University, Japan*
[m] *Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, España*
[n] *University of Bordeaux, LaBRI, UMR 5800, PICTURA, Talence, F-33400, France*
[o] *CNRS, LaBRI, UMR 5800, PICTURA, Talence, F-33400, France*
[p] *Image Processing Laboratory, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada*
[q] *Center for Neurodevelopmental Medicine and Imaging Research, Kennedy Krieger Institute, Baltimore, MD 21205, USA*
[r] *Department of Neurology, The Johns Hopkins School of Medicine, Baltimore, MD 21287, USA*
[s] *Department of Psychiatry, The Johns Hopkins School of Medicine, Baltimore, MD 21287, USA*
[t] *Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235, USA*
[u] *Imaging Genetics Center, Mark and Mary Stevens Institute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Marina del Rey, CA 90292, USA*
[v] *Departments of Neurology, Pediatrics, Psychiatry, Radiology, Engineering, and Ophthalmology, University of Southern California, Los Angeles, CA 90033, USA*

[*] *These authors curated the data and organized the comparison, all others contributed results.*
Please address correspondence to:

Aaron Carass,
Department of Electrical and Computer Engineering, The Johns Hopkins University,
105 Barton Hall, 3400 N. Charles St., Baltimore, MD 21218, USA.

Email address: `aaron_carass@jhu.edu` (Aaron Carass)

Preprint submitted to NeuroImage *April 23, 2018*

Abstract

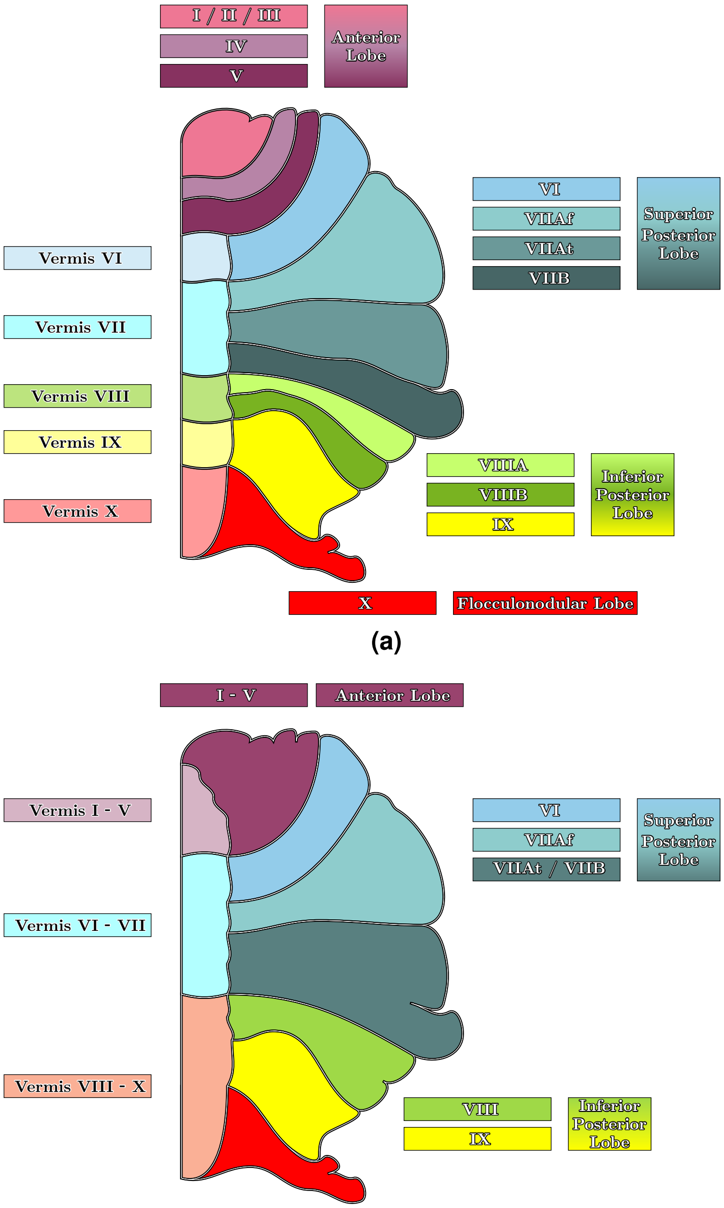
The human cerebellum plays an essential role in motor control, is involved in cognitive function (i.e., attention, working memory, and language), and helps to regulate emotional responses. Quantitative in-vivo assessment of the cerebellum is important in the study of several neurological diseases including cerebellar ataxia, autism, and schizophrenia. Different structural subdivisions of the cerebellum have been shown to correlate with differing pathologies. To further understand these pathologies, it is helpful to automatically parcellate the cerebellum at the highest fidelity possible. In this paper, we coordinated with colleagues around the world to evaluate automated cerebellum parcellation algorithms on two clinical cohorts showing that the cerebellum can be parcellated to a high accuracy by newer methods. We characterize these various methods at four hierarchical levels: coarse (i.e., whole cerebellum and gross structures), lobe, subdivisions of the vermis, and the lobules. Due to the number of labels, the hierarchy of labels, the number of algorithms, and the two cohorts we have restricted our analyses to the Dice measure of overlap. Under these conditions, and in conjunction with the rank-sum computation we identified the most effective strategies and an overall winning method.

Keywords: Magnetic resonance imaging, cerebellar ataxia, attention deficit hyperactivity disorder, autism.

1. Introduction

The cerebellum is a structure of great importance in the neuroanatomy of humans. It plays an essential role in motor coordination (Ito, 1984; Manto et al., 2013), as well as cognitive function such as attention (Schmahmann, 1991, 2004), working memory (Desmond and Fiez, 1998), and language (Silveri et al., 1994; Desmond and Fiez, 1998), regulates emotional responses (Schutter and Van Honk, 2005) including fear (Schmahmann and Caplan, 2006), and there is increasing understanding of perceptual processes in the cerebellum (Baumann et al., 2015). Anatomically, the cerebellum is nestled underneath the cerebral hemispheres behind the brainstem in the posterior cranial fossa. It is separated from the cerebrum by the tentorium cerebelli, a dura structure, and is connected to the brainstem at the pons. The cerebellum is divided into two hemispheres, like the cerebrum, and also has a midline zone which is known as the vermis. The cortical surface of the cerebellum is made up of finely spaced branches that radiate outwards from the cerebellar white matter (WM), which is known as the corpus medullare (CM). These WM branches conceal that the volume of the cerebellum is a tightly folded layer of gray matter (GM). Anatomists differentiate regions of the cerebellum hierarchically into groups of folds, known as lobes, and then into individual folds, referred to as lobules. The lobes are the anterior, superior posterior, inferior posterior, and the flocculonodular. The lobules are identified by Roman Numerals I through X (Schmahmann et al., 2000), however Lobules VII and VIII are further differentiated. This nomenclature comes from Schmahmann et al. (2000), derived in part from Larsell (1952); we refer to it as the Schmahmann nomenclature and note the differences between it and the classical nomenclature (Malacarne, 1776; Henle, 1879) in Table 1. Figure 1 shows the anatomical structure of the cerebellum, including the hierarchical breakdown of the lobes and lobules. Due to the importance of the cerebellum, any pathology can have serious consequences; however, the tightly folded structure of the cerebellum makes identifying specific structures challenging. Below we outline the clinical relevance of understanding the structure of the cerebellum and the various effects of cerebellar pathologies; we then provide an overview of the fully automated parcellation tools that exist in the literature.

Cerebellum centric disorders, such as spinocerebellar ataxia (SCA), have been previously shown to have cerebellar shape (Yang et al., 2016a), clinical disability scores (Ying et al., 2006), and functional scores (Yang et al., 2014; Kansal et al., 2016) that correlate with SCA subtype in a region specific manner. More importantly, the cerebellum has been shown to be affected in diseases ranging from attention-deficit and hyperactivity disorder (ADHD) (Mostofsky et al., 1998b), schizophrenia (Nopoulos et al., 1999; Parker et al., 2014), Alzheimer's disease (Thomann et al., 2008; Colloby et al., 2014), to chronic alcoholism (Victor et al., 1959; Torvik and Torp, 1986; Cavanagh et al., 1997; Baker et al., 1999; Fitzpatrick et al., 2008). In patients with schizophrenia, a reduction in the volume of the vermis has been observed in multiple studies (Nopoulos et al., 1999; Okugawa et al., 2002, 2003) based on the manual parcellation of the cerebellum. Moreover, when the vermis has been further subdivided into the anterior and posterior portions, the volume differences are driven by changes in the posterior vermis (Womer et al., 2016) with a significant diagnosis-by-sex interaction. Several types of dementia exhibit correlations with the cerebellum; Alzheimer's disease (AD) has shown a reduction in the volume of the posterior lobes (Thomann et al., 2008), whereas dementia with Lewy bodies has shown greater GM loss in Lobule VII than AD (Colloby et al., 2014). Several recent voxel based morphometry (VBM) studies have shown regional patterns of atrophy between AD and cerebellar GM and WM (Möller et al., 2013) and correlations between GM loss and the constructional praxis and constructional praxis recall test in the CERAD test battery (Dos Santos et al., 2011). However, older studies (Karas et al., 2003) that relied upon studying large regions—due to the FWHM size used in the VBM—showed no significant GM loss in the cerebellum suggesting that the effects of cerebellum/AD interaction can only be identified when smaller regions of interest are used. These and other studies are summarized in Table 2. There are two key points to take from this past work: 1) in-vivo assessment of the cerebellum through magnetic resonance (MR) imaging (MRI) is imperative to further our understanding and 2) manual parcellation or delineation remains a widely used approach for studying the cerebellum.



(a)

(b)

Figure 1: An illustration of a coronal view of one hemisphere of the human cerebellum. Shown are the lobule labels for our **(a)** Adult and **(b)** Pediatric Cohort with their corresponding lobe groupings, based on the Schmahmann nomenclature (Schmahmann et al., 2000). Table 4 has a complete list of the provided labels for both cohorts. It is widely acknowledged that there is no true vermis for the Anterior Lobe (Lobules I-V). Thus the distinction between vermis and body in the Anterior Lobe differentiates the midline portion from the body of the lobe. Our Adult Cohort does not use this differentiation, whereas our Pediatric Cohort does.

Table 1: A key to convert between the nomenclature of Schmahmann (Schmahmann et al., 2000), derived from Larsell (1952), and the classical nomenclature (Malacarne, 1776; Henle, 1879) of common cerebellar structures.

Vermal Nomenclature		**Hemisphere Nomenclature**	
Schmahmann	**Classical**	**Schmahmann**	**Classical**
Vermis I / II[†]	Lingula	L/R Lobule I/II	L/R Lingula (or Lingulae)
Vermis III[†]	Centralis	L/R Lobule III	L/R Centralis
Vermis IV[†]	Culmen I	L/R Lobule IV	L/R Quadrangularis
Vermis V[†]	Culmen II	L/R Lobule V	L/R Quadrangularis
Vermis VI	Declive	L/R Lobule VI	L/R Quadrangularis
Vermis VIIAf	Folium	L/R Lobule VIIAf (Crus I)	L/R Semi-Lunaris Superior
Vermis VIIAt	Tuber I	L/R Lobule VIIAt (Crus II)	L/R Semi-Lunaris Inferior
Vermis VIIB	Tuber II	L/R Lobule VIIB	L/R Semi-Lunaris Inferior
Vermis VIIIA	Pyramis I	L/R Lobule VIIIA	L/R Biventer I
Vermis VIIIB	Pyramis II	L/R Lobule VIIIB	L/R Biventer II
Vermis IX	Uvula	L/R Lobule IX	L/R Tonsilla (or Tonsil)
Vermis X	Nodulus	L/R Lobule X	L/R Flocculus

[†] It is widely acknowledged that there is no true vermis for the Anterior Lobe (Lobules I-V). The division in our Pediatric Cohort differentiates the midline portion of the Anterior Lobe from the body of the lobe.

Table 2: A summary of some cerebellar focused imaging studies exploring various pathologies. We include whether the study used manual delineations (MD) and the key cerebellar related findings. $N\ (M/F)$ denotes the number of patients and the male/female ratio. Abbreviations: ADHD - Attention-deficit and hyperactivity disorder; AD - Alzheimer's disease.

Disease	**Citation**	$N\ (M/F)$	**MD**	**Observations**
ADHD	Mostofsky et al. (1998b)	35 (35/0)	Y	Decreased inferior posterior vermis
Alcoholism	Torvik and Torp (1986)	65 (65/0)	Y	Decreased vermis segments
	Baker et al. (1999)	19 (14/5)	Y	Non-significant loss in vermis and flocculus
AD	Thomann et al. (2008)	60 (29/31)	Y	Decreased superior and inferior posterior lobes
	Möller et al. (2013)	344 (175/169)	–†	Reduced GM throughout the cerebellum
	Colloby et al. (2014)	127 (84/43)	–†	Bilateral reduction of Lobule VI
Autism	Courchesne et al. (1994)	103 (84/19)	Y	Reduced area in the vermis of Lobule VI and VII
	D'Mello et al. (2015)	70 (51/19)	N	Reduced GM in Lobule VII
Fragile X Syndrome	Mostofsky et al. (1998a)	188 (98/90)	Y	Decreased posterior vermis in males and females, though less significant in females.
Schizophrenia	Nopoulos et al. (1999)	130 (130/0)	N‡	Smaller vermis area and smaller anterior lobe
	Okugawa et al. (2002)	30 (30/0)	N‡	Reduced posterior superior vermis
	Okugawa et al. (2003)	116 (73/43)	N‡	Reduced anterior vermis, posterior superior vermis, and posterior inferior vermis volumes
	Womer et al. (2016)	104 (48/56)	Y	Decreased posterior vermis volumes in males

† The studies did not differentiate regions of the cerebellum and based assessment on an anatomists interpretation of the areas of change.

‡ Automated processing for cerebellar volumes based on registration to a Talairach Atlas, augmented by manual tracings of the vermis.

Despite the continued use of manual delineation of the cerebellum in various studies (Womer et al., 2016) there has been work on both semi-automated (Pierson et al., 2002) and fully automated segmentation and parcellation of the cerebellum. We are only concerned with those methods that provide at a minimum the lobes of the cerebellum; hence methods like FreeSurfer (Fischl, 2012), TOADS (Bazin and Pham, 2008), and MA-CRUISE (Huo et al., 2016) that only provide tissue classes are not directly relevant unless used in combination with other tools. The first published method that provided a fully automated parcellation of the cerebellar lobules was SUIT (Diedrichsen, 2006); the method used a spatially unbiased template of the human cerebellum that when registered with a subject image provided the parcellation. The method was later updated (Diedrichsen et al., 2009) to include a probabilistic atlas. As powerful as SUIT is in identifying the subdivisions of the cerebellum, it has primarily been used only for identifying cerebellar GM as a normalizing factor in functional MRI analysis. Prior to the introduction of the probabilistic version of SUIT, Powell et al. (2008) presented machine learning approaches for cerebellar parcellation that identified the lobes and vermis of the cerebellum. Bogovic et al. (2013a) presented ACCLAIM, a multi-object geometric deformable model (Bogovic et al., 2013c; Carass and Prince, 2016) approach that provides a parcellation of 28 labels of the cerebellum and included a comparison to SUIT. Price et al. (2014) presented the Cerebellar Analysis Toolkit (CATK) which used a Bayesian Appearance Modeling (Patenaude et al., 2011) with prior knowledge of shape, image intensity, and inter-shape relationships to provide five cerebellar labels. Weier et al. (2014) described the Rapid Automatic Segmentation of the human Cerebellum And its Lobules (RASCAL) which is a patch matching based approach that improved on the multi-atlas segmentation fusion technique presented in Coupe et al. (2011). Romero et al. (2017) presented CERES another patch-matching technique, that uses OPAL (Giraud et al., 2016; Ta et al., 2014) for its label fusion. Yang et al. (2016b) presented a multi-atlas labeling approach that used a graph-cut to help regularize the final segmentation. A more detailed description of these methods is provided in Section 3 to help describe the approaches presented in this paper. To summarize, the previous work in this area includes: single and multi atlas registration, level sets, graph methods, a Bayesian framework, neural networks, support vector machines, and patch matching. Table 5 presents an overview of the methods presented and evaluated in this paper. It can be seen that deep learning, an important new class of algorithms in medical imaging, are represented among the methods tested in this paper.

There has been an increasing movement towards *Grand Challenges* (Styner et al., 2008; Schaap et al., 2009; Heimann et al., 2009; Menze et al., 2015; Mendrik et al., 2015; Maier et al., 2017; Carass et al., 2017) in the medical imaging community in recent years. These challenges have helped to develop standards for evaluating the performance of different categories of medical imaging problems and for helping those on the peripheral of the community to understand the state-of-the-art and the general direction in which the technology is moving. In particular, the 2008 MICCAI MS Lesion challenge (Styner et al., 2008) was a significant step forward in the sharing of clinically relevant data. More recently, the 2015 Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) (Menze et al., 2015) has been a disruptive step forward, having allowed groups without access to high-quality data with delineations to contribute innovative new solutions for segmenting brain tumors (Sauwen et al., 2015; Banerjee et al., 2016; Kamnitsas et al., 2017). Thus in the spring of 2017, we invited colleagues from around the world to participate in a Cerebellum Parcellation Challenge as part of MICCAI 2017. As only eight groups responded to this call, it was decided that the workshop itself would not go forward due to lack of broad interest. Coming out of the discussions for this Cerebellum Parcellation Challenge (hereafter the Comparison), it was agreed that we would present the performance findings from seven of the research teams who participated in the Comparison (hereafter the Participants). In Section 2, we outline the two cohorts of data that were provided to the Participants and the evaluation used in comparing the submitted results from each of the Participants. One of the Participants submitted two methods, however two of the Participants contributed no results for one of the cohorts. Thus, our first cohort has results from six algorithms, while our second cohort was processed by eight algorithms. Both cohorts are imaged using standard clinical protocols with an approximately 1 mm isotropic resolution, with complete details of the acquasition in Section 2. In our examination of these data and methods, we restrict our analyses to the Dice overlap; we outline our rationale behind this decision in Section 2.2. Section 3 provides a complete description of the methods contributed by the Participants for the Comparison. Section 4 includes the Comparison between the manual delineations for our two cohorts and the algorithms; it is broken down into hierarchical levels: 1) Coarse level including the whole cerebellum, whole vermis, and CM (3 labels); 2) Lobe level including the left and right of the four lobes (8 labels); 3) Vermis level which included the vermal subdivisions of the vermis (5 labels for our Adult Cohort, 3 labels for our Pediatric Cohort); 4) Lobule level (22 labels for our Adult Cohort, 14 labels for our Pediatric Cohort); and a 5) Consolidated level, with further details in Sec. 4. In general the methods show agreement with the manual delineations of the cerebellar structures. However, the size of our cohorts restricted our statistical analyses, with rank-sum computations being used to determine an overall highest ranked method.

2. Materials and Metrics

2.1. Data

The Participants were given data from our Adult and Pediatric Cohorts, described below. The Participants were also encouraged to take advantage of other available data sets; in particular, they were made aware of data provided by Jörn Diedrichsen of the University of Western Ontario[1]. The Diedrichsen data comprises 20 normal adult subjects, each of which have 30 labeled cerebellar components. Our Adult Cohort is an expertly labeled data set collected by the Image Analysis and Communications Laboratory (IACL) at Johns Hopkins University (PI: J.L. Prince) (Bogovic et al., 2013b). It contains 20 subjects, a mix of healthy controls and ataxia

[1] Available from: http://www.diedrichsenlab.org/imaging/propatlas.htm

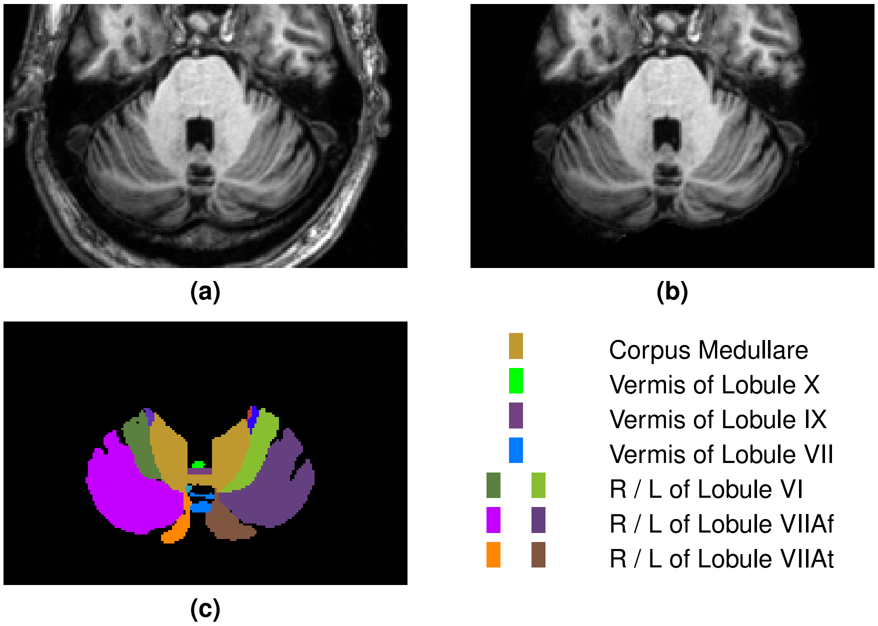


Figure 2: For our Adult Cohort, we show a cropped portion of a typical axial slice of **(a)** the defaced MP-RAGE, **(b)** the skull-stripped MP-RAGE, and **(c)** the manual labels with a corresponding color key for the prominent labels. The images are shown in radiological convention. A complete list of all the labels for the Adult Cohort is provided in Table 4. Results of the methods on the same data are shown in Fig. 4.

patients, each with 28 labeled cerebellar components (complete demographic information is provided in Table 3; see Fig. 2 for an example image and corresponding manual labels). Fifteen training examples were provided to the Participants, and the remaining five data sets were used for testing, with the goal being to label the cerebella of the test subjects to best agree with the expert labels. Magnetization prepared rapid gradient echo (MP-RAGE) images using a 3.0 T MR scanner (Intera, Phillips Medical Systems, Netherlands) were acquired with the following parameters: 1.1 mm slice thickness, 8° flip angle, TE = 3.9 ms, TR = 8.43 ms, FOV 21.2×21.2 cm, image matrix of 256×256. The images were resampled to have a 1.0 mm isotropic voxel; subsequently they were defaced using `mri_deface` from FreeSurfer (v5.3) (Fischl, 2012), a skull stripping mask was generated using SPECTRE (Carass et al., 2010), and the skull-stripped image was white matter (WM) peak normalized so that all images have a consistent WM peak intensity (Nyúl and Udupa, 1999a). For the training data, the defaced MR image, the WM peak skull stripped image, and the expert manual cerebellar parcellations were provided to the Participants. For the test subjects only the defaced MR image and the WM peak skull-stripped image were provided. All images in this cohort were acquired in an axial orientation. An example of both the defaced and WM peak skull stripped image for a data set are shown in Fig. 2 with the corresponding manual delineation.

Our Pediatric Cohort comprises data collected at the Center for Neurodevelop-

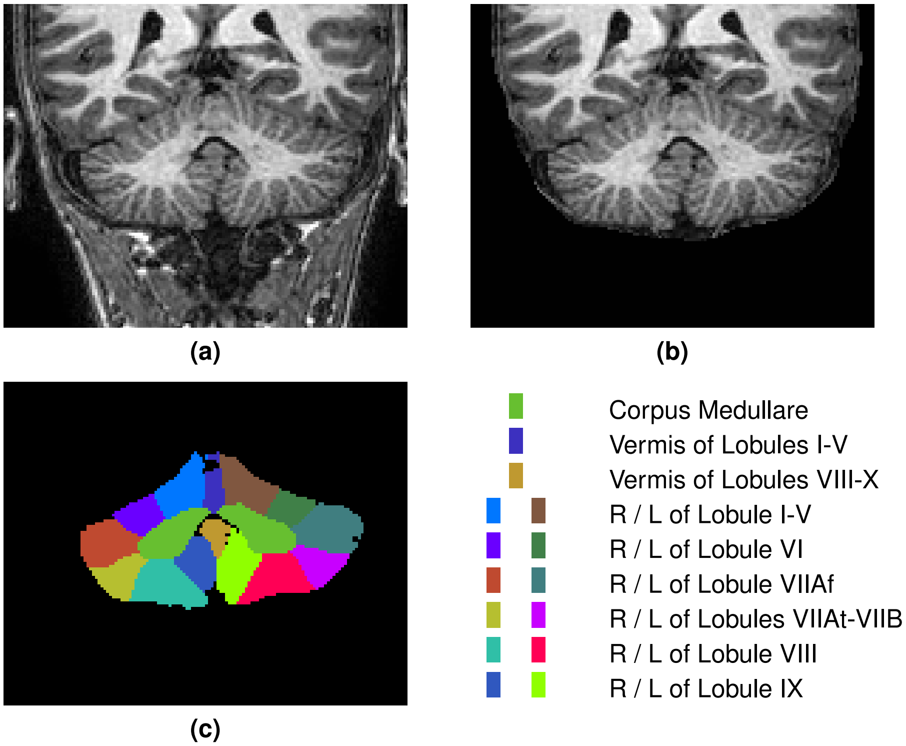


Figure 3: For our Pediatric Cohort, we show a cropped portion of a typical coronal slice of **(a)** the defaced MP-RAGE, **(b)** the skull-stripped MP-RAGE, and **(c)** the manual labels with a corresponding color key for the prominent labels. The images are shown in radiological convention. A complete list of all the labels for the Pediatric Cohort is provided in Table 4. Results of the methods on the same data are shown in Fig. 9.

mental and Imaging Research at the Kennedy Krieger Institute (PI: S.H. Mostofsky). These 30 expertly labeled data sets, with 18 labeled cerebellar components, are from 8-12 year old boys and girls with a mix of healthy controls, ADHD and high-functioning Autism (HFA) patients (complete demographic information is provided in Table 3). 20 of these were provided as training and 10 were reserved for testing. The objective was to label these cerebella to best agree with the expert labels. The provided MR images were MP-RAGE, acquired on a 3T Philips Gyroscan NT (Royal Philips Electronics) system with the following parameters: 1 mm slice thickness, 8° flip angle, TE = 3.0 ms TR = 7.0 ms, image matrix of 256 × 256. The Pediatric Cohort was preprocessed in an identical manner to our Adult Cohort; specifically, the images were defaced using `mri_deface`, skull-stripped using SPECTRE, and the skull-stripped image was WM peak normalized. For each of the 20 training images, the defaced MR image, the WM peak skull stripped image, and the expert manual cerebellar parcellation were provided to the Participants. For the test subjects only the defaced MR image and the WM peak skull stripped image were provided. All images in this cohort were acquired in a coronal orientation. An example of both the

Table 3: Demographic details for the training and test data for both cohorts. The top line is the information of the entire data set, while subsequent lines within a section are specific to the patient diagnoses. N (M/F) denotes the number of patients and the male/female ratio, respectively. The Age column lists the mean, standard deviation, min, and max, in years, at scan time. The codes for the patient groups are: HC – Healthy controls; CB – Symptoms of cerebellar dysfunction without genetic diagnosis; SCA6 – Spinocerebellar ataxia type 6; ADHD – Attention-deficit and hyperactivity disorder; HFA – High-functioning Autism.

Data Set	**N (M/F)**	**Age**	
		Mean (SD)	**[Min, Max]**
Adult Cohort			
Training	15 (5/10)	54.7(±11.97)	[30.0, 71.0]
HC	6 (2/4)	54.3(±14.69)	[30.0, 71.0]
CB	3 (1/2)	54.3(±8.02)	[46.0, 62.0]
SCA6	6 (2/4)	55.3(±12.60)	[35.0, 70.0]
Testing	5 (5/0)	69.2(±5.81)	[62.0, 78.0]
CB	5 (5/0)	69.2(±5.81)	[62.0, 78.0]
Pediatric Cohort			
Training	20 (7/13)	10.1(±1.36)	[8.3, 13.2]
HC	10 (4/6)	10.2(±1.33)	[8.4, 13.2]
ADHD	7 (0/7)	10.4(±1.61)	[8.3, 12.2]
HFA	3 (3/0)	9.2(±0.65)	[8.5, 9.7]
Testing	10 (3/7)	10.1(±1.29)	[8.4, 12.6]
HC	5 (1/4)	9.9(±1.04)	[8.4, 11.2]
ADHD	3 (0/3)	10.2(±1.06)	[9.2, 11.3]
HFA	2 (2/0)	10.6(±2.76)	[8.7, 12.6]

defaced and WM peak skull stripped image for a training data set are shown in Fig. 3 with the corresponding manual delineation. A complete list of the labels provided for the two cohorts is available in Table 4 and a key is provided in Table 1 to convert between the Schmahmann (Schmahmann et al., 2000) and classical (Malacarne, 1776; Henle, 1879) nomenclature.

Table 4: The labeled cerebellar structures of both cohorts. For reference, we include a key to convert between the Schmahmann and classical nomenclature in Table 1.

Adult Cohort (Healthy Controls and Ataxia Patients)	
Major Structure	**Cerebellar Sub-components**
Corpus Medullare	
Vermis	Vermis of Lobule VI
	Vermis of Lobule VII
	Vermis of Lobule VIII
	Vermis of Lobule IX
	Vermis of Lobule X
L/R Anterior	L/R Lobule I / II / III
	L/R Lobule IV
	L/R Lobule V
L/R Superior Posterior	L/R Lobule VI
	L/R Lobule VIIAf (Crus I)
	L/R Lobule VIIAt (Crus II)
	L/R Lobule VIIB
L/R Inferior Posterior	L/R Lobule VIIIA
	L/R Lobule VIIIB
	L/R Lobule IX
L/R Flocculonodular	L/R Lobule X

Pediatric Cohort (Healthy Controls, ADHD & HFA Patients)	
Major Structure	**Cerebellar Sub-components**
Corpus Medullare	
Vermis	Vermis of Lobule I-V
	Vermis of Lobule VI-VII
	Vermis of Lobule VIII-X
L/R Anterior	L/R Lobule I-V
L/R Superior Posterior	L/R Lobule VI
	L/R Lobule VIIAf (Crus I)
	L/R Lobule VIIAt (Crus II) & VIIB
L/R Inferior Posterior	L/R Lobule VIII
	L/R Lobule IX
L/R Flocculonodular	L/R Lobule X

2.2. Comparison Metric

To compare the results from the available methods with our expert delineations, we used the Dice overlap (Dice, 1945). The Dice overlap is a commonly used volume metric for comparing labels masks. If $\mathcal{M}_G$ is the *gold standard* mask of a human rater and $\mathcal{M}_A$ is the mask generated by a particular algorithm, then the Dice overlap for binary objects is computed as

$$\text{Dice}(\mathcal{M}_G, \mathcal{M}_A) = 2\frac{|\mathcal{M}_G \cap \mathcal{M}_A|}{|\mathcal{M}_G| + |\mathcal{M}_A|},$$

where $|\cdot|$ is the cardinality (number of voxels). This overlap measure has values in the range $[0, 1]$, with 0 indicating no agreement between the two masks, and 1 meaning the two masks are identical. We have chosen to explicitly restrict our analysis to the Dice overlap for two reasons: 1) it is a widely reported and understood measure; 2) due to the large number of labels, the hierarchy of labels (from coarse to fine), the number of algorithms, and the two cohorts that we report on would make reporting multiple measures very lengthy. We note that in two recent challenge papers (Carass et al., 2017; Maier et al., 2017) the final rankings of the methods—which used multiple metrics—were well correlated with the Dice overlap; see Table 7 in Maier et al. (2017) for example. A benefit of using a single measure in this manner is the clarity that is afforded in declaring a *best* method. We comment more on the pros and cons of this evaluation in Section 5.

3. Methods

Here we provide descriptions of all the methods used in the Comparison along with the names of the specific contributors. A brief summary of each of the methods is provided in Table 5.

🟥 **SUIT**

Default SUIT v3.2

(Carlos H. Castillo)

Data analysis were performed using MATLAB R2015b (The Mathworks Inc. Natick, MA), SPM12 (Ashburner et al., 2000), and the spatially unbiased atlas template of the human cerebellum (SUIT) toolbox v3.2 (Diedrichsen et al., 2009). To achieve the best performance from SUIT, all anatomical images were first reoriented into LPI (Neurological) orientation and then the origin of each T1-w image was assigned to the manually selected anterior commissure.

To ensure the correct normalization of the cerebellar cortex into the atlas template, SUIT first isolates the infra-tentorial structures from the rest of the brain. This is important because the occipital cortex has a similar intensity as the cerebellum and in most cases there is not a clearly visible separation between these two structures. SUIT v3.2 achieves this separation by using the unified segmentation (Ashburner and Friston, 2005) of SPM12; this segmentation procedure combines tissue classification and registration by means of both a mixture of Gaussians and tissue

Table 5: An overview of the methods used in our comparison, with details of each method listed in the remainder of this Section.

Name	Approach
🟥 **SUIT****	Default SUIT v3.2
🟧 **C-SUIT***	C-SUIT is a customized SUIT, with Correction and Customized Atlas based on the Pediatric Cohort
🟨 **FS-SUIT**	FreeSurfer and SUIT in collaboration
🟩 **FCN**	A thirteen layer fully convolution network (FCN)
🟦 **ConvNet**	Convolution neural network
🟪 **CERES2**	Updated version of CERES with improved intensity normalization and a new error correction method based on an ensemble of boosted patch-based neural networks
🟪 **RASCAL**	Updated patch-matching technique with cohort specific templates, improved intensity normalization, and non-linear registration
◼ **DeepNet**	A U-net based FCN with ten layers

$*$ – Denotes methods that only contributed results for the Pediatric Cohort.

probability maps. Using this technique, the brain is segmented into eight tissue types: cerebral GM, cerebral WM, cerebellar GM, cerebellar WM, cerebrospinal fluid (CSF), bone, fat/skin, and air. Finally, a binary cerebellar mask is created by combining the cerebellar GM and WM segmentation maps including voxels with a tissue probability of greater than or equal to 90% of coming from either of those classes.

After the cerebellar isolation, SUIT uses a fast-diffeomorphic normalization algorithm (DARTEL) (Ashburner, 2007). DARTEL uses the probabilistic GM and WM segmentation maps to align the anatomy of the cerebellum of each participant to the SUIT atlas template. To increase the speed of the process, the non-linear registration is solved using a Levenberg-Marquardt strategy and a multigrid method; see Ashburner (2007) for complete details. The result is a non-linear deformed image coregistered to the SUIT atlas template and its respective deformation field.

To identify the cerebellar lobular boundaries, the probabilistic atlas of the cerebellum included in the SUIT toolbox was used. The SUIT atlas consists of a set of 34 probabilistic maps that indicates the likelihood that a certain voxel in the reference space belongs to each lobule. The SUIT atlas includes the cerebellar left and right lobules (I-IV, V, VI, Crus I, Crus II, VIIb, VIIIa, VIIIb, IX, X), vermis (VI, Crus I, Crus II, VIIb, VIIIa, VIIIb, IX, and X) and deep cerebellar nuclei. For this work, these compartments were combined to have only 18 labels (I-V, VI, Crus I, Crus II-VIIb, VIII, IX, X, Vermis I-V, Vermis VI-VII, Vermis VIII-X, and corpus medullare). For each subject, the inverse warp deformation field was calculated and then applied to the SUIT atlas using a nearest neighbor approach, so the values of each label were preserved. For each voxel one label was assigned depending on the maximum probability of the SUIT atlas, resulting in a lobular segmentation of the subject's native space.

🟧 C-SUIT

Customized-SUIT (C-SUIT) with Corrections and Customized Atlas based on the Pediatric Cohort

(Paul Rasser)

FreeSurfer (v5.3) (FS) (Fischl, 2012) was used to preprocess the images from the Pediatric Cohort training set by performing bias correction, intensity normalization, and skull stripping on the provided defaced MP-RAGE images. A 6-paramerter transformation from their original defaced MP-RAGE to the conformed FreeSurfer space was found using `minctracc` (Collins et al., 1994) and was applied to both the defaced and the manually parcellated volumes. The parcellated volumes were used to correct differences in cerebellum GM as defined by the FreeSurfer subcortical segmentation output. SPM (Ashburner et al., 2000) was used to correct the coordinate system of the data sets to match the requirements of SUIT (Diedrichsen et al., 2009). The correction was followed by application of SUIT to provide an initial SUIT cerebellum mask that was subsequently corrected to conform to the cerebellar volume as defined by the corrected subcortical segmentation of FreeSurfer.

ANTs (Avants et al., 2008)—using the fast cross correlation metric—was used to find a symmetric diffeomorphic transformation between the normalized and bounded MP-RAGE to the SUIT space. All subject images (MP-RAGE, parcellated cerebellum, cerebellum binary mask) were transformed to the SUIT space. FreeSurfer was then used to create a bounded normalized atlas in SUIT space as well as a cerebellum mask in SUIT space by combining the output from 19 of the 20 training subjects from the Pediatric Cohort. The 20th subject was excluded from this atlas construction step due to its poor registration with the SUIT space. In the SUIT space, the 18 labels from the 19 subjects in the Pediatric Cohort were combined to create a customized parcellation using `mri_concat` from FreeSurfer.

To apply the customized atlas to the remaining ten subjects in the Pediatric Cohort, C-SUIT first preprocesses the ten subjects using FreeSurfer to perform bias correction, intensity normalization, and skull stripping. Preprocessing the data in this manner provided a consistency between the volumetric measures based on the cerebellar parcellation and the existing cerebral measures derived from FreeSurfer. This feature was included to provide a supplementary cerebellar parcellation for projects with existing measures of the cerebrum derived from FreeSurfer, such as the ENIGMA Project[2].

Using SUIT the preprocessed MP-RAGED is bounded, then ANTs estimates a symmetric diffeomorphic transformation into the customized atlas. This is then applied to the normalized and bounded MP-RAGE. The inverse of this symmetric diffeomorphic transformation was applied to the average cerebellum hemisphere mask in SUIT space and used to create a binary SUIT style cerebellum mask. The SUIT

[2] `http://enigma.ini.usc.edu/`

command `suit_normalize` was applied to the bounded and normalized volume and its mask, with the inverse of this transformation applied to the customized cerebellum parcellation. As the cerebellum parcellation was required in the native space, the test subject's cerebellum parcellation was transformed from the bounded space to their native space using `fsl_rigid_register` supplied with the FreeSurfer package. Finally, to remove any potential outliers in the cerebellum parcellation, FSL's FAST tissue segmentation algorithm (Zhang et al., 2001) was used to create GM and WM masks that were applied to the final cerebellum parcellation.

🟨 FS-SUIT

FreeSurfer and SUIT in collaboration for Cerebellar Segmentation
(Melanie Ganz & Vincent Beliveau)

The approach considered in this work combines FreeSurfer (v5.3) (Fischl, 2012) and SUIT (Diedrichsen et al., 2009), to overcome some of the limitations of these algorithms when used independently. Whole brain and cerebellar GM and WM segmentation of structural MRI data was performed with FreeSurfer. FreeSurfer processing included motion correction, removal of non-brain tissue, automated Talairach transformation, segmentation of subcortical WM and deep GM volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles) (Fischl et al., 2002, 2004), intensity normalization, and further cortical surface processing. FreeSurfer's cerebellum segmentation is driven by a probabilistic atlas segmentation (Fischl et al., 2002). The procedure maintains tissue class statistics (e.g., means and variances of the MRI intensities of a given neuroanatomical structure) on a per-location per-class basis throughout an atlas space. In addition, local spatial relationships between structures are encoded in a Markov random field (MRF). FreeSurfer has been demonstrated to have good test-retest reliability across scanner manufacturers and field strengths (Han et al., 2006).

As FreeSurfer only provides a segmentation of the cerebellar GM and WM, FS-SUIT augments it with SUIT (v2.7) (Diedrichsen et al., 2009) to identify cerebellar lobules. SUIT is available as a compatible toolbox for SPM12 (Ashburner et al., 2000) which allows for the creation of a segmentation of the cerebellum into different lobules. SUIT consists of three steps: 1) cerebellum isolation; 2) normalization to SUIT atlas space; 3) reslicing into SUIT atlas space/individual subject space. SUIT uses SPM to perform the cerebellum isolation by segmenting the brain into tissue-types. The volume is cropped to include anything inferior to the tentorium cerebelli. The tissue-types are used to compute posterior probability for each voxel. The normalization to SUIT atlas space is performed by a nonlinear deformation map to the SUIT template using the cosine-basis function approach (Ashburner and Friston, 1999). Finally, SUIT applies the estimated deformation to map the subject into the SUIT atlas space. In SUIT the atlas is spatially unbiased.

To unify the results from FreeSurfer and SUIT into a coherent cerebellum lobule parcellation, a final segmentation is created by limiting the SUIT lobule parcellation to their intersection with the FreeSurfer cerebellar GM. While SUIT yields a segmentation of the cerebellar lobules, it largely ignores the individual WM and GM intensities. Whereas FreeSurfer provides a more accurate representation of the cerebellar WM and GM tissue classes, but it is also more sensitive in regions of low contrast between tissue types and tends to over-label peripheral tissue. Thus, the intersection provides a refinement of the GM and WM segmentation while reducing the over-labeling of peripheral tissue. Correspondingly, due to the restriction of the SUIT labels to FreeSurfer cerebellar GM, FS-SUIT uses the FreeSurfer WM to label the CM; otherwise there would be gaps between the lobule labels and the CM.

🟩 FCN

Cerebellum parcellation from a deep learning perspective
(Jose Dolz, Ismail Ben Ayed, & Christian Desrosiers)

Originally designed for image recognition and classification tasks, convolutional neural networks (CNNs) are now commonly employed for semantic segmentation. The most naive approach follows a sliding-window strategy where regions defined by the window are processed one-by-one. This technique presents two main drawbacks: 1) processing image regions independently provides non-structured output, which reduces segmentation accuracy; and 2) due to many redundant convolution and pooling operations, the process is inefficient. To mitigate these limitations, the spatial map of class probabilities can be obtained in a single, dense inference step. This approach, known as a fully convolutional neural network (FCN) (Long et al., 2015), represents the network as a single non-linear convolution, which is trained end-to-end. Unlike the sliding-window approach, FCNs can avoid redundant convolution and pooling operations, making them computationally more efficient. Additionally, fully convolutional networks have an order of magnitude fewer coefficients, which makes them easier to train with fewer training samples.

The proposed method, which is built on top of DeepMedic (Kamnitsas et al., 2017), is based on the FCN described in Dolz et al. (2018) had state-of-the-art performances for subcortical brain segmentation. This network is composed of 13 layers in total: 9 convolutional layers, followed by 3 fully-connected layers, and the classification layer. The number of kernels in each convolutional layer—from shallow to deep—is 25, 25, 25, 50, 50, 50, 75, 75, and 75, respectively. The kernel sizes are equal to $3 \times 3 \times 3$ in all the convolutional layers. Three fully-connected layers, composed of 400, 200, and 150 hidden units each, are added after these convolutional layers for encoding semantic information. To ensure that the network contains only convolutional layers, fully-connected layers are converted to a collection of $1 \times 1 \times 1$ convolutions (Kamnitsas et al., 2017). Dolz et al. (2018) described two intermediate-layer outputs (i.e., feature maps) that were embedded in the final predictions, encouraging consistency between features extracted at different scales, while injecting fine-grained information directly in the segmentation process. As the structures in the cerebellum are often thinner than subcortical structures, to avoid losing small details when passing the target structures through several convolutional blocks, FCN embeds the feature maps from all layers into the fully-connected layers.

Due to computation and memory limitations, the FCN network cannot apply dense training over the whole 3D input volume. Instead, FCN sub-samples this volume into S smaller sub-volumes, which are then fed into the network. In this way: 1) FCN avoids memory issues when the input is not down-sampled (as in this work), and 2) FCN has a high number of samples from each image, removing the need for data augmentation. A Parametric Rectified Linear Unit (PReLU) (He et al., 2015), which applies an element-wise activation function, follows each convolutional layer. Let θ be the network trainable parameters, and $\mathcal{L}$ as the set of ground-truth labels such that $L_s^v \in \mathcal{L}$ represents the label of voxel v in the s^{th} sampled sub-volume for all the predicted voxels V. The cost function is

$$J(\theta, \mathcal{L}) = -\frac{1}{S\,V} \sum_{s=1}^{S} \sum_{v=1}^{V} \log p_{L_s^v}(X_v)\,,$$

where $p_{L_s^v}(X_v)$ is the output of the classification layer for voxel v in the segment s (i.e. softmax output) given their input feature maps, X_v. This cost function corresponds to the mean cross-entropy computed over all voxels and sub-volumes. Sample sizes were set to be larger for inference than in training (Dolz et al., 2018). In this particular application, the following combination was found to give satisfactory results: 27 × 27 × 27 for training and 35 × 35 × 35 for testing.

Optimization of network parameters was performed via the RMSprop optimizer (Tieleman and Hinton, 2015). Momentum was set to 0.6 and the initial learning rate was set to 0.001; the latter was reduced by a factor of 2 after every 5 epochs (starting from epoch 10). Weights in layer l were initialized based on a zero-mean Gaussian distribution with standard deviation $\sqrt{2/n_l}$, where n_l denotes the number of connections to units in that layer. FCN was trained for 35 epochs, with each epoch composed of 20 sub-epochs. At each sub-epoch, a total of 1000 samples were randomly selected from the training images, and processed in batches of size 10. The network architecture was developed using Theano (Bergstra et al., 2010), with modifications being made through LiviaNET [3]. FCN was used on a server with a NVIDIA Tesla P100 GPU and 16 GB of RAM; training took approximately 15 hours for the FCN (taking 25 minutes per epoch).

🟦 ConvNet

Cerebellum segmentation using convolutional neural networks
(Benjamin Thyreau)

The basis of this work is a convolutional neural network (ConvNet) that learns to segment the MRI using the expert labels as training data. ConvNet was intended to investagate whether whole-image input, as opposed to patch-based, could better capture high-level structure and human-expert variation. ConvNet usually requires a large data set, so increasing the number of training examples available was important. As there is overlap between the labeling schemes in both cohorts, ConvNet uses a merged set of each provided expert delineation to have a common labeling scheme across the two cohorts. ConvNet is trained on this common set, followed by duplication and refinement of the ConvNets for each of the two cohorts.

A template image was created from all provided delineations using ANTs (Avants

[3] https://github.com/josedolz/LiviaNET

et al., 2008). A cerebellum mask is defined in this template space as the bounding box of the union of all labels in this template space. This cerebellum mask reduced the size of the computational domain for the processing. The general left/right symmetry was used to double the training-set size and further reduce the input FOV. Then each image was augmented 200× using random elastic distortions focused especially at label borders. The images were also intensity-normalized. Due to the memory constraints of the GPU platform (NVidia GTX 1080, 8Gb), the left/right resolution had to be halved which diminished some image details.

ConvNet was created with computational and time constraints in mind. Thus ConvNet structure has an alternating stack of 40- and 24-wide convolutional layers, with batch-normalization layers. No max-pooling nor other resolution changes were incorporated to reduce the model complexity. ConvNet also refrained from recalling low-level features within the final layers—usually a good practice in deep learning segmentation systems—as the goal is not pixel-perfect accuracy but rather expert imitation.

The loss function was a standard cross-entropy metric, weighted by each label size. When a label did not exist, which is a possibility since different label sets were merged, its weight was not accounted for in the loss function. Stochastic gradient decent (SGD) with momentum was used for the optimization of the initial ConvNet—as it tends to find better-quality minima with less over-fitting. The initial ConvNet model fitting took two days to train. The fine tuning of the separate ConvNets for each cohort used ADAM (Kingma and Ba, 2014) instead of SGD. This allowed for faster fine-tuning of the two ConvNets, taking approximately four hours per model. Mini-batches of size 6 were used. During development, one subject was left out of every data set for internal evaluation purpose, and later, as a final step the models were refitted on all available training data.

This ConvNet has limitations and room for improvement. For instance incorporation of multi-resolution features, tuning the number of parameters, and a different dropout scheme to ease convergence should all help improve the results. Additionally, as each of the ConvNets were trained on half-cerebella, the central voxels are arbitrary labeled according to their bounding-box side, which cause inaccuracies when the vermis is absent. Some post-processing, such as the use of an MRF, could be employed to enhance the results.

■ CERES2

Cerebellum multi-atlas patch-based segmentation with a patch-based boosted neural network error corrector

(José E. Romero, Pierrick Coupé, & José V. Manjón)

A new version of CERES (Romero et al., 2017), which is a cerebellum lobule segmentation algorithm that is based on a recent method called Optimized PatchMatch Label fusion (OPAL) (Giraud et al., 2016; Ta et al., 2014). The method consists of a multi-atlas patch-based (Rousseau et al., 2011; Coupe et al., 2011) non-local label fusion technique that produces segmentations using a library of manually annotated cases. CERES2 improves on CERES by using a different intensity normalization method and by adding a systematic error correction step based on an ensemble of patch-based boosted neural networks.

CERES2 preprocesses all the imaging data by denoising it using the spatially adaptive non-local means filter (Manjón et al., 2010) and corrects for intensity inhomogeneity using the N4 bias correction method (Tustison et al., 2010). The images were registered to the Montreal Neurological Institute (MNI) space using an affine transformation estimated using ANTs (Avants et al., 2008) and the MNI152 T1-w template; this is followed by intensity normalization (Nyúl and Udupa, 1999b) with the MNI152 images as the reference template. The images are then cropped to the cerebellum based on the manual segmentations of the subjects in the MNI space. Non-linear deformation was estimated using ANTs (Avants et al., 2008) using the cropped MNI152 template as reference. CERES2 completes the preprocessing by again applying intensity normalization (Nyúl and Udupa, 1999b) to the cropped images, further improving the intensity matching. After preprocessing, CERES2 has a library consisting of a set of cropped images (and segmentations) and their non-linear transformations to the cropped MNI space. Similar to CERES, when given a new subject to segment a custom library is created online in the subject's linear MNI space to avoid unwanted interpolation artifacts. This is done by concatenating the direct non-linear transformations of the library templates with the inverse non-linear transformation of the target case.

CERES2 employs a non-local patch-based label fusion, which is a multi-atlas segmentation fusion technique (Coupe et al., 2011). The resultant label for each image voxel is calculated by a weighted label fusion of multiple sample patches from a specific search area surrounding the target voxel for all the cases in the library, computed as

$$v(x_i) = \frac{\sum_s \sum_{j \in V_i} w(x_i, x_{s,j}) y_{s,j}}{\sum_s \sum_{j \in V_i} w(x_i, x_{s,j})}.$$

Here V_i is the search area around the i^{th} voxel of the subject image, s iterates over the number of subjects in the library, and $y_{s,j}$ is the candidate label from voxel $x_{s,j}$, the j^{th} voxel in the s^{th} subject. The function $w(x_i, x_{s,j})$ is the similarity between patches, which is defined as,

$$w(x_i, x_{s,j}) = \exp\left(- \frac{\left\| P(x_i) - P(x_{s,j}) \right\|_2^2}{h^2} \right).$$

where $P(\cdot)$ is the patch around the respective voxel, $\|\cdot\|_2$ is the normalized L_2 norm (normalized by the number of elements), and h is a normalization factor calculated as the minimum of all patch distances from the search area.

The PatchMatch algorithm (Barnes et al., 2010) is an efficient way to find patch-wise correspondences between 2D images based on the approximated nearest neighbor field (ANNF) method. The core idea is that if two patches are a good match, then adjacent patches are likely to be a good match as well. Ta et al. (2014) presented the Optimized PatchMatch Label fusion (OPAL), an adaptation to 3D MR images, establishing correspondences between the input subject image and the library of L templates; the major benefit of OPAL is its run time is independent of the size of the library. A key difference between Barnes et al. (2010) and OPAL is the assumption, within OPAL, that both subject and library templates are located in the same coordinate frame (MNI space). A more complete description is available in Ta et al. (2014). Giraud et al. (2016) presented a multi-scale extension to OPAL, thus avoiding issues relating to fixed size patches. In this extension multiple label probability maps are computed and combined equally for each label before the label fusion step. Like CERES, CERES2 uses two scales and a label dependent weight as follows,

$$p(l) = \alpha(l)p_1(l) + (1 - \alpha(l))\, p_2(l),$$

where $p.(l)$ is the corresponding probability map for label l and $\alpha(l) \in [0, 1]$ is the mixing coefficient for label l. Details about the coefficient optimization are available in Romero et al. (2017). The final label for a voxel is simply the label with the highest probability.

Automatic segmentation methods suffer from random and systematic errors. Despite the fact that random error can be mitigated with aggregation techniques (such as multi-atlas techniques), systematic errors cannot be reduced using this strategy as they are not random. Nevertheless, their bias can be learned and used to correct the segmentations. Inspired by the work of Wang et al. (2013), in CERES2 a systematic error corrector using a patch-based ensemble of boosted neural networks to improve the segmentation accuracy is incorporated. The neural network ensemble is trained using samples from a region of interest of the label to correct as done in Wang et al. (2013). The feature vector was created by concatenating the following data: image patches of sizes 3 × 3 × 3 voxels (fully sampled); 7 × 7 × 7 and 11 × 11 × 11 voxels (subsampled by skipping two and four voxels at each dimension, respectively) from the T1-w image; the corresponding patches from the automatic segmentations; the Euclidean distance value of the voxel to the edge of the structure, and the coordinates in MNI space. Thus CERES2 has a feature vector of length 166 that is mapped to the corresponding manual segmentation patch of size 3 × 3 × 3 voxels. Features were extracted using an overcomplete scheme as done in Manjón et al. (2016). Such a structured prediction not only provides more accurate results than the voxel-wise version but also produces a more regular correction.

The neural network model used in CERES2 consists of a multilayer perceptron with two hidden layers of size 83 and 55 neurons resulting in a topology of 166 × 83 × 55 × 27 weights. CERES2 uses an ensemble comprising of 10 neural networks trained using a boosting strategy were wrongly classified training data samples were selected with higher probability than correctly classified ones. One ensemble per label is trained. At test time, the trained ensembles are used to correct the automatic segmentations produced by OPAL. The correction process takes a few seconds.

🟪 RASCAL

Patch-based label fusion

(Vladimir S. Fonov and D. Louis Collins)

The previously published RASCAL (Rapid Automatic Segmentation of the Human Cerebellum and its Lobules) (Weier et al., 2014) was adapted for use with the two cohorts. The data was preprocessed as follows: 1) linear registration to MNI-ICBM152 2009c stereotaxic space (Fonov et al., 2010); 2) linear intensity normalization based on quantile matching to normalize the intensity range to the MNI-ICBM152 2009c template; 3) extracted brain mask using thresholding of the provided SPECTRE brainmask; 4) created an unbiased population specific template (Fonov et al., 2010), the resultant template was used as a reference template for RASCAL.

The RASCAL segmentation algorithm is an improved version of Coupe et al. (2011), which is also used by CERES2 and described above. The key differences are a change in how intensity normalization is done, majority voting to account for multiple labels, and the use of nonlinear registration. RASCAL (Weier et al., 2014) was then fine-tuned for the proposed data sets by employing a leave-one-out cross-validation. Linear registration, localized to the cerebellum, is performed using the affine registration mode of ANTs (Avants et al., 2008) with the Mattes cost function; which is followed by non-linear registration to the reference template using ANTs with the cross-correlation cost-function. After the registration steps are complete, all images are resampled to have the same resolution as the subject. Segmentations are then fused using the non-local patch-based algorithm.

■ DeepNet

U-Net Parcellation of the Cerebellum

(Vladimir S. Fonov and D. Louis Collins)

DeepNet is an exploration of the potential of using an FCN based on U-net (Ronneberger et al., 2015; Çiçek et al., 2016) to parcellate the cerebellum. Recall that FCNs are CNNs designed for semantic segmentation. U-net has demonstrated state-of-the-art performance in several tasks (Chen et al., 2016; He et al., 2017; Zhao et al., 2017) while preserving the high resolution information throughout the contraction-expansion layers of the network. In brief, 3D U-net consists of a contracting (analysis) path and an expanding (synthesis) path. Each layer of the contracting portion consists of two 3×3×3 unpadded convolutions, followed by a ReLU and a 2×2×2 max pooling operation with a stride of 2 for downsampling in each dimension. Every step of the expansion consists of an upsampling convolution of 2×2×2 with a stride of 2 in each dimension followed by two 3×3×3 convolutions each followed by a ReLU. Shortcut connections from layers of equal resolution in the contracting path provide the high-resolution features to the expansion portion of the network. The last layer consists of a 1×1×1 convolution to reduce the number of output channels to the number of labels. Before each ReLU layer, batch normalization is performed during training with the mean, standard deviation, and global statistics updated using these values. This is followed by a layer to learn the scale and bias explicitly. At test time, normalization is done via the computed global statistics and the learned scale and bias.

As originally presented, 3D U-net used four analysis/synthesis steps; for the cerebellum parcellation task DeepNet uses five analysis/synthesis steps. DeepNet also modifies the default convolution kernels on a per-layer basis, with the details

Table 6: Layer parameters for ■ **DeepNet**.

U-net	Input	Output	Convolution Kernels		Upsampling
Layer	Channels	Channels	#1	#2	Kernel
1	4	256	5×5×5	5×5×5	5×5×5
2	16	128	5×5×5	5×5×5	3×3×3
3	16	64	3×3×3	3×3×3	3×3×3
4	16	64	3×3×3	3×3×3	3×3×3
5	32	64	1×1×1	3×3×3	3×3×3

listed in Table 6. The final layer of DeepNet contains two fully connected convolutional layers with 256 and 128 channels and a Dropout Layer. Thus the final layer creates a mapping from the 128 features into the segmentation labels—28 for the Adult Cohort and 18 for the Pediatric Cohort. Log Soft Max was used to calculate the negative log-likelihood error function, generalized kappa overlap metric was used to track performance on out-of-sample validation data. The total number of trainable parameters input into the model is 9,479,573.

4. Results

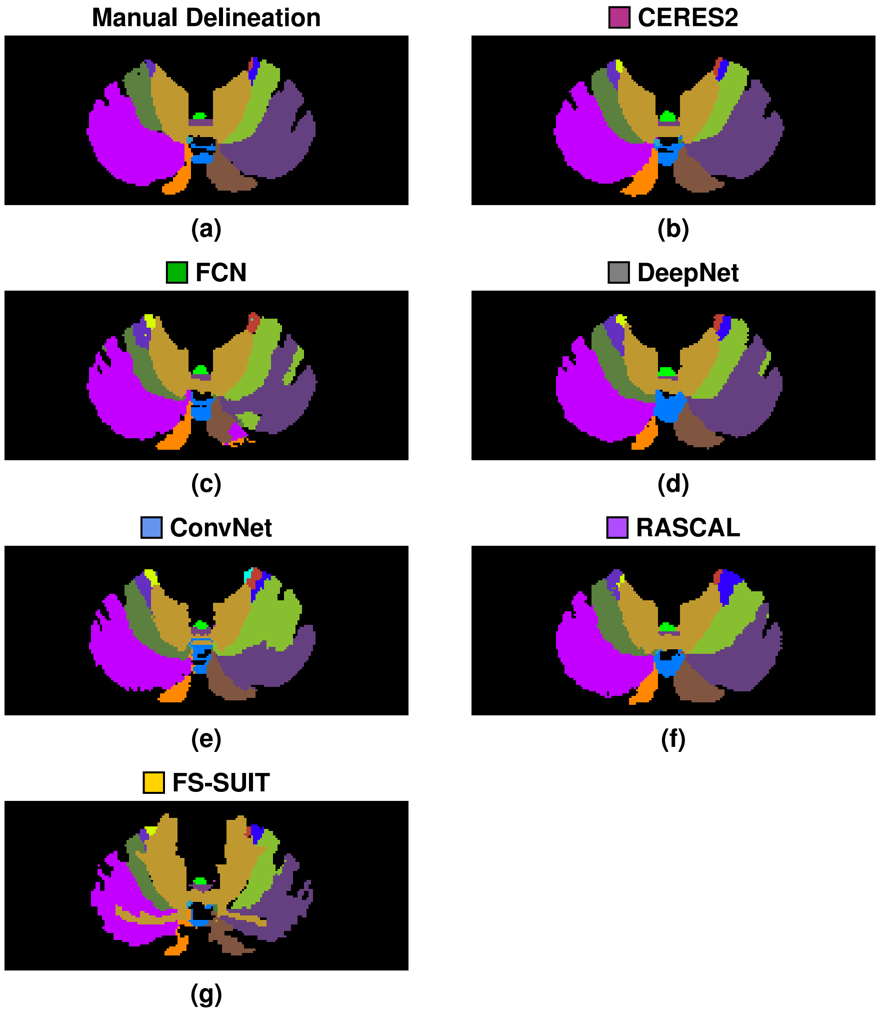
We present results using the Dice overlap measure to characterize the performance of the methods applied to both cohorts in our Comparison. Each Participating group provided a parcellation of the test data sets into lobules respecting the labeling scheme used in the respective cohort. To better characterize performance, we broke down the analysis using a hierarchical scheme. At the coarsest level we have the gross structures of the whole cerebellum, the whole vermis, and the corpus medullare (CM). We then have the subdivisions of the cerebellum into its left and right lobes; see Table 4 for the definitions of these structures for each cohort. The final two levels are the subdivisions of the vermis and the individual lobules, these are different for both cohorts—as the delineations draw distinctions between the vermis and the granularity with which the cerebellum compartments are identified. Specifically, for the Adult Cohort there are five subdivisions of the vermis and 22 lobule labels (11 per hemisphere), whereas for the Pediatric Cohort there are three vermal subdivisions and 14 lobule labels (seven per hemisphere). These levels are identified and defined as: 1) Coarse level which includes the whole cerebellum, whole vermis, and CM (3 labels); 2) Lobe level including the left and right of the four lobes (8 labels); 3) Vermis level which includes the vermal subdivisions of the vermis (5 labels for our Adult Cohort, 3 labels for our Pediatric Cohort); 4) Lobule level (22 labels for our Adult Cohort, 14 labels for our Pediatric Cohort); and a grouping listed as 5) Consolidated, which is a union of all the available labels (38 labels for the Adult Cohort, 28 labels for the Pediatric Cohort). These hierarchies have been generated (where necessary) based on the supplied parcellation of each algorithm by merging the appropriate labels; for example, the whole cerebellum label is given by merging all the labels. In Subsection 4.3, we summarize the Dice overlap results using the rank-sum to compare the performance of the various methods in a succinct manner. The rank-sum scoring assigns a score of 1 to the method with the highest mean Dice overlap measure, 2 to the second highest mean Dice overlap measure, et cetera, for each label. Table 7 provides a summary of the rank-sums for each of the hierarchies. The supplemental material includes details of the rank-sum calculation.

4.1. Adult Cohort

Figure 4 shows the results of the six methods on a typical axial slice from a test data set in the Adult Cohort: Fig. 2 shows the underlying MR data. Figures 5–8 show the Dice overlap for each of the methods across the various hierarchies; these plots show the individual data point for each of the five test data sets as well as showing the mean Dice overlap as a horizontal bar. Specifically, Fig. 5 shows the Dice overlap for the whole cerebellum, the whole vermis, and the CM. The mean Dice overlap of the methods on whole cerebellum was used to order the methods in Fig. 4. We can see that CERES2 has the highest mean Dice overlap for each of the Coarse labels; however, for the whole cerebellum label the difference between CERES2 and FCN is quite small (0.950 vs. 0.949), though this is not the case for the other two Coarse Labels. This result sets the tone for many of the other labels in the Adult Cohort; in general for a given label the mean Dice overlap of CERES2 is the highest of the methods, with FCN typically coming in second and on occasion the difference is negligible. Typical examples of this behavior are the Left and Right Anterior Lobe (Fig. 6), the Left and Right Superior Posterior Lobe (Fig. 6), Vermis of Lobules VIII through X (Fig. 7), and several cases in the Lobule hierarchy shown in Fig. 8. There are of course example of labels on which CERES2 does not achieve the maximum mean Dice overlap. See the Left and Right Inferior Posterior Lobe in Fig. 6, and the Vermis of Lobule VI in Fig. 7 for examples. In all 38 labels under consideration, there are 11 labels on which CERES2 is not ranked first; these 11 cases are evenly split between FCN (3 times), ConvNet (5 times), and DeepNet (3 times); see the supplemental material for complete details. We also observe in Figs. 6 and 8 that each algorithm has similar performance on both the left and right for each label. We make the observation that most of the methods have a mean Dice overlap above 0.8 for all the lobes except the Flocculonodular Lobe. For the vermal subdivisions, we see a slight degradation in results (mean Dice overlap in the range 0.7 to 0.9). Of course we see a further drop in performance when considering the lobe subdivisons, particularly for Lobules V, VIIB, and VIIIA. In fact, these lobules appear to be the most difficult to parcellate for all the methods; as each method has a large range of Dice overlap values for these regions.

4.2. Pediatric Cohort

Figure 9 shows the results of the eight methods on a typical coronal slice from a test data set in the Pediatric Cohort, Fig. 3 shows the underlying MR data. Figures 10–13 show the Dice overlap for each of the methods across the various hierarchies; these plots show the individual data point for each of the ten test data sets as well as showing the mean Dice overlap as a horizontal bar. Specifically,



Manual Delineation

(a)

🟪 CERES2

(b)

🟩 FCN

(c)

⬛ DeepNet

(d)

🟦 ConvNet

(e)

🟪 RASCAL

(f)

🟨 FS-SUIT

(g)

Figure 4: Shown for a test data set in the Adult Cohort are the **(a)** manual delineation, and the results for each of the methods: **(b)** 🟪 **CERES2**; **(c)** 🟩 **FCN**; **(d)** ⬛ **DeepNet**; **(e)** 🟦 **ConvNet**; **(f)** 🟪 **RASCAL**; and **(g)** 🟨 **FS-SUIT**, for the same axial slice shown in Fig. 2. The methods are ranked based on their mean whole cerebellum parcellation, see Fig. 5 for details.

Fig. 10 shows the Dice overlap for the whole cerebellum, the whole vermis, and the CM. The mean Dice overlap of the methods on the whole cerebellum was used to order the methods in Fig. 9. We can see that FCN has the highest mean Dice overlap for the whole cerebellum and CM labels with CERES2 in second place; however, for the other coarse label the order of these two methods is reversed. In fact, unlike the Adult Cohort, where CERES2 was on top but definitely not unopposed, in the Pediatric Cohort CERES2 is quite dominant. The only labels for which it is not ranked first are the whole cerebellum and the CM. Similar to the Adult Cohort,

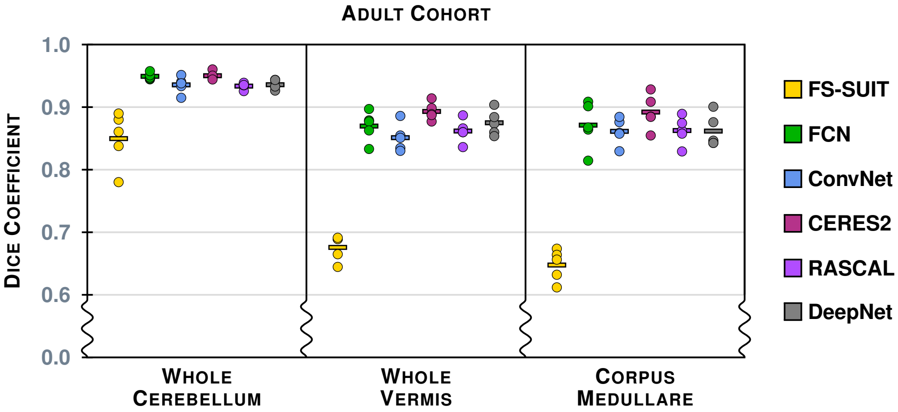


Figure 5: The Dice overlap for the three labels associated with the Coarse hierarchy is shown for the Adult Cohort. Each column includes five data points, for the five test data sets in the Adult Cohort, showing the Dice overlap for a method-label pair (some of the data points are *on top* of one another and are thus occluded from view). The horizontal line in each column shows the mean Dice overlap for that particular method and label. We note that the scale has been zoomed to help appreciate the differences between the algorithms.

we observe in Figs. 11 and 13 for the Pediatric Cohort that each algorithm performs consistently on both the left and right for each label.

4.3. Summary

To create a readily interpretable representation of these results we computed the rank-sum for each method over the various hierarchies and both cohorts. These rank-sum results are presented in Table 7, with the details of the computation included in the supplemental material. Over both cohorts, we can easily discern some patterns in Table 7: clearly CERES2 is the overall winner, with FCN and DeepNet trading back and forth between second and third place. We also see RASCAL is quite consistently fourth in both cohorts. Given the outcome of our rank-sum analysis, we identify the top three methods as CERES2, FCN, and DeepNet. We next want to determine if there is a statistically significant difference between these top three methods. To this end, we use a two-sided Wilcoxon paired signed-rank test (Wilcoxon, 1945) between CERES2 & FCN, and between CERES2 & DeepNet, to establish statistical significance. The Wilcoxon test is a nonparametric test of the null hypothesis that the two samples come from the same population against an alternative hypothesis. We tested using all the available Dice overlap values for a particular hierarchy; thus for the Coarse level on the Adult Cohort there are 15 values for each method (3 labels × 5 data sets). For the statistical comparisons we use an α level of 0.001 to note *weak statistical significance* and an α level of 0.0001 to denote *strong statistical significance*; we use these α values as we do not employ any multiple comparison correction techniques. The p-values for the Wilcoxon test and the mean values for the Dice overlap (for our top three methods) are shown in Table 8 for the Adult Cohort and Table 9 for the Pediatric Cohort. For the five

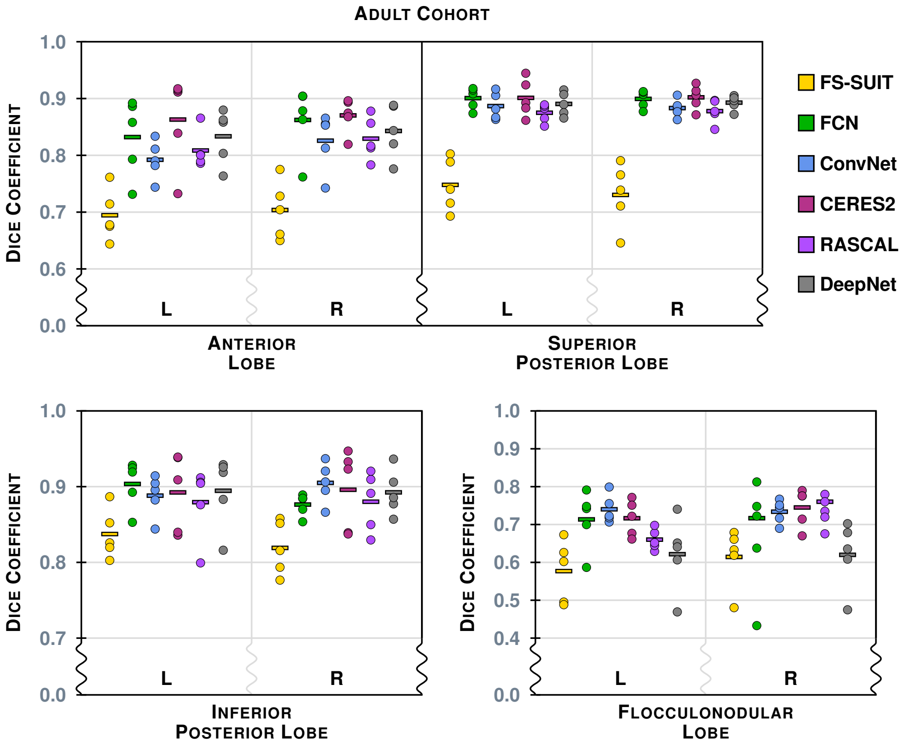


Figure 6: The Dice overlap for the eight labels associated with the Lobe hierarchy is shown for the Adult Cohort, see Table 4 for the list of lobe labels. We note that the scale has been zoomed to help appreciate the differences between the algorithms.

670 hierarchies (Coarse, Lobe, Vermis, Lobule, and Consolidated) on the Adult Cohort CERES2 has the highest mean Dice overlap on all five hierarchies and is statistically significantly different on eight of the ten comparisons (with strong significance in five instances). The two cases where there is no statistically significant difference are between CERES2 and FCN for the Lobe and Vermis hierarchies. For the Pediatric 675 Cohort CERES2 has the highest mean Dice overlap on all five hierarchies and is statistically significantly different on nine of the ten comparisons (strong significance in all nine cases). The single comparison for which there is not significance is between CERES2 and FCN for the Coarse hierarchy.

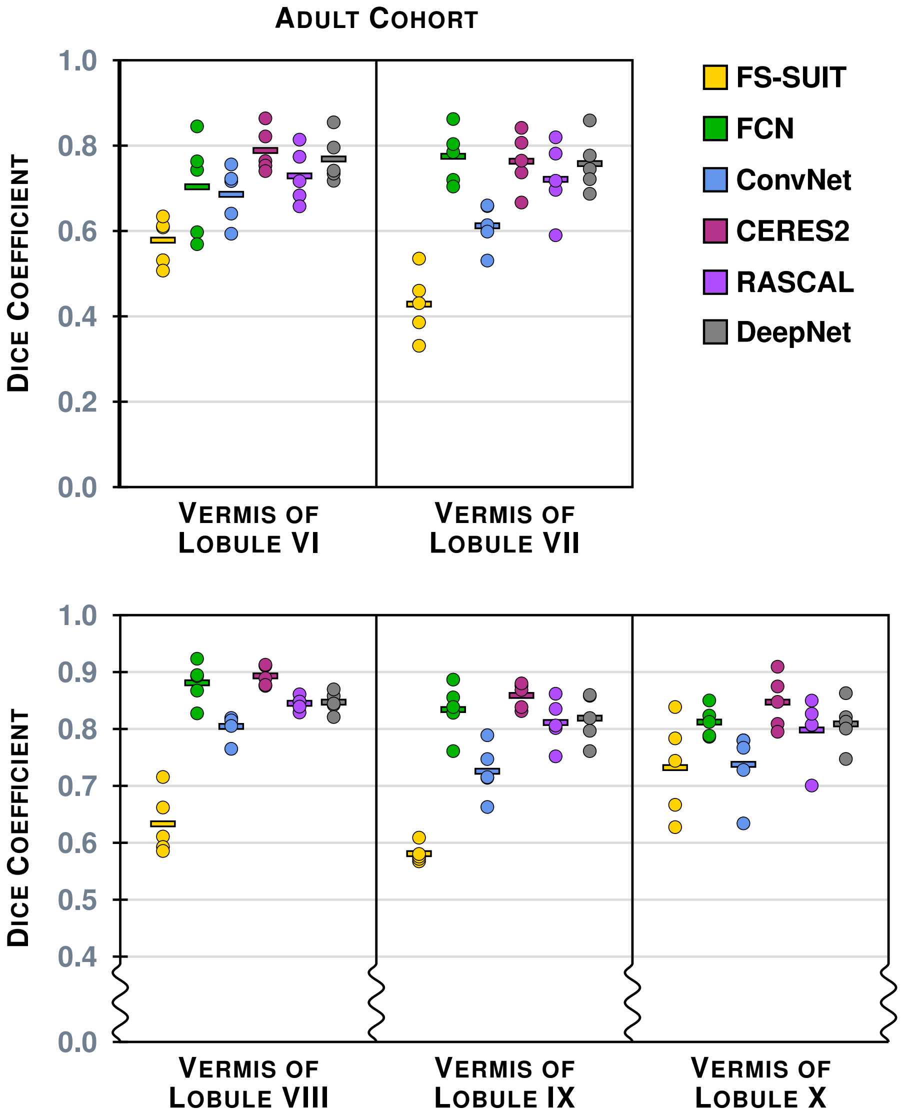


Figure 7: The Dice overlap for the five labels associated with the Vermis hierarchy is shown for the Adult Cohort, see Table 4 for the list of vermis labels. See Fig. 5 for instructions on interpreting the plots. We note that some of the scale has been zoomed to help appreciate the differences between the algorithms.

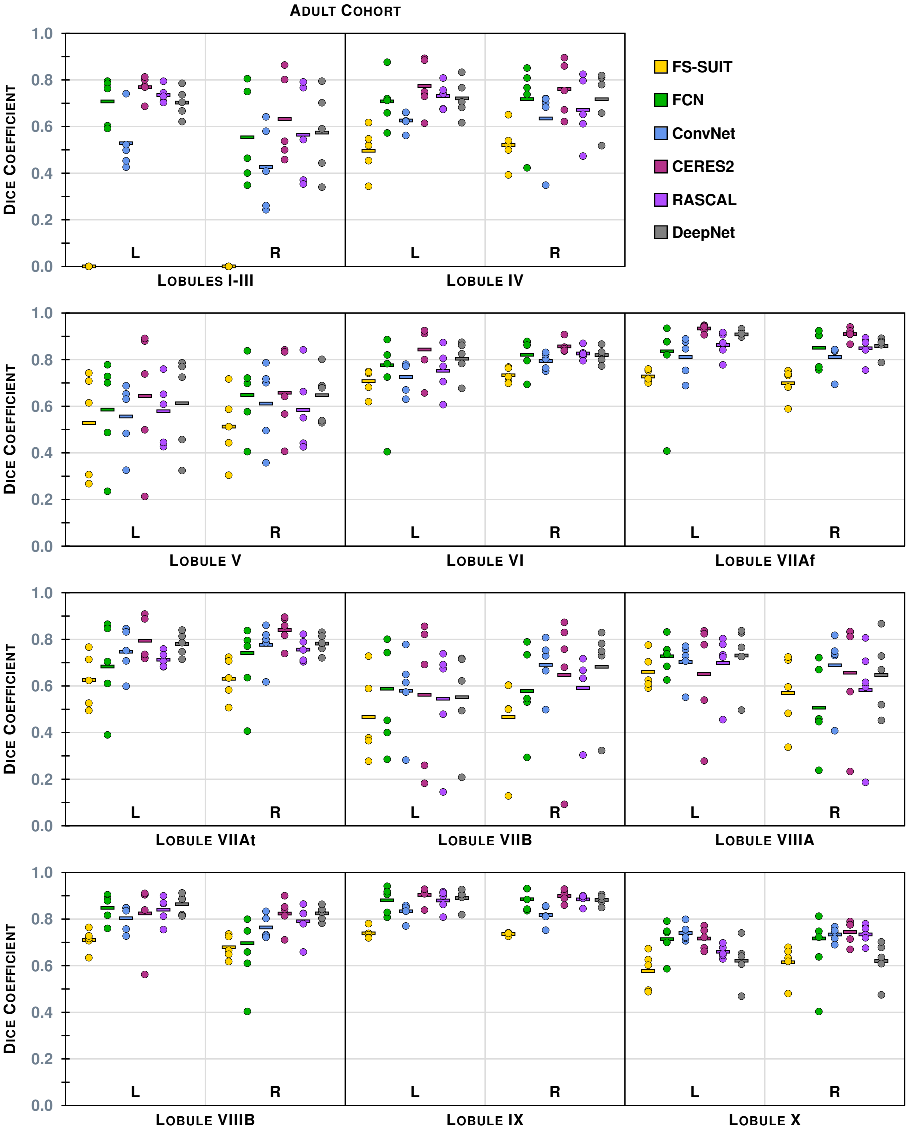


Figure 8: The Dice overlap for the 22 labels (11 per hemisphere) associated with the Lobule hierarchy is shown for the Adult Cohort, see Table 4 for the list of lobule labels. See Fig. 5 for instructions on interpreting the plots.

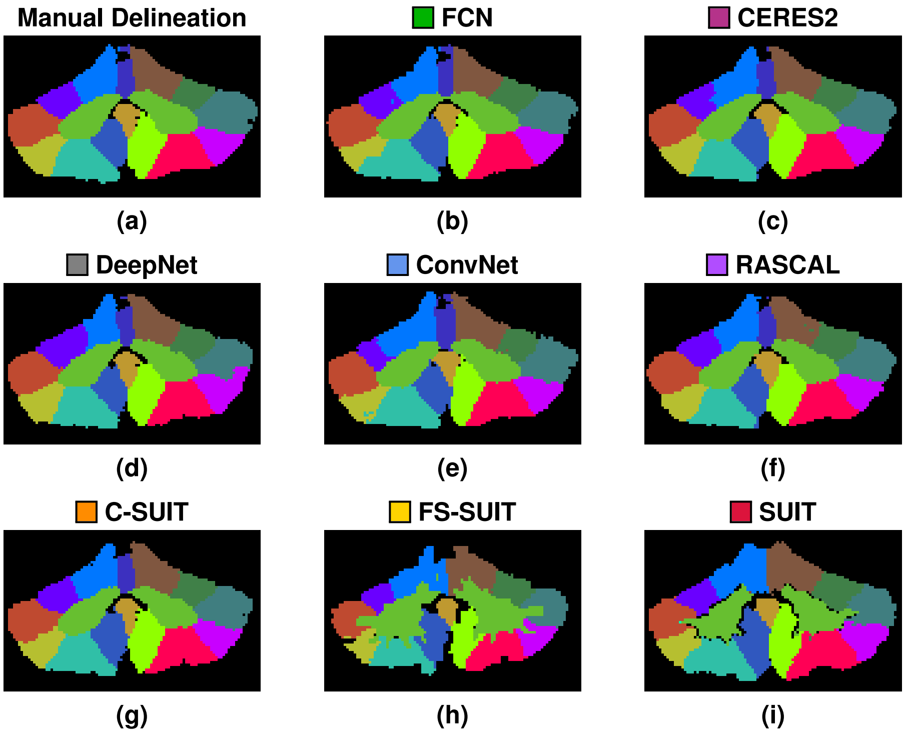


Figure 9: Shown for a test data set in the Pediatric Cohort are the **(a)** manual delineation, and the results for each of the methods: **(b)** 🟩 **FCN**; **(c)** 🟪 **CERES2**; **(d)** ⬛ **DeepNet**; **(e)** 🟦 **ConvNet**; **(f)** 🟪 **RASCAL**; **(g)** 🟧 **C-SUIT**; **(h)** 🟨 **FS-SUIT**; and **(i)** 🟥 **SUIT**, for the same coronal slice shown in Fig. 3. The methods are ranked based on their mean whole cerebellum parcellation (see Fig. 10 for details).

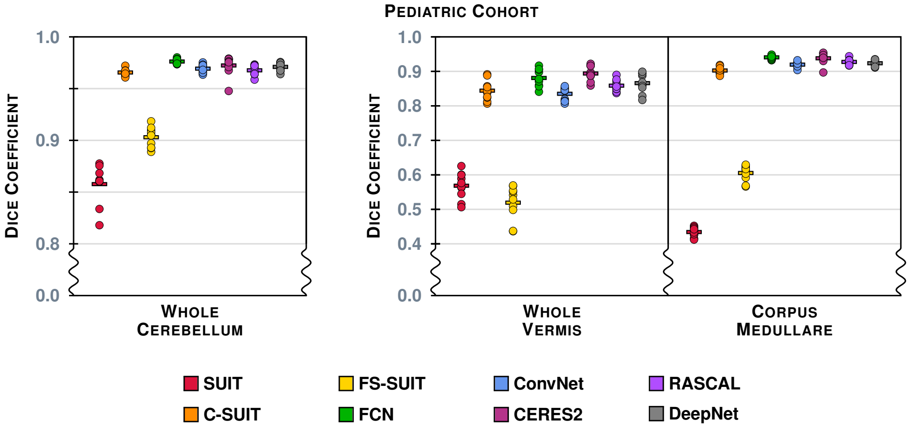


Figure 10: The Dice overlap for the three labels associated with the Coarse hierarchy are shown for the Pediatric Cohort. Each column includes ten data points, for the ten test data sets in the Pediatric Cohort, showing the Dice overlap for a method-label pair (some of the data points are *on top* of one another and are thus occluded from view). The horizontal line in each column shows the mean Dice overlap for that particular method and label. We note that the scale has been zoomed to help appreciate the differences between the algorithms.

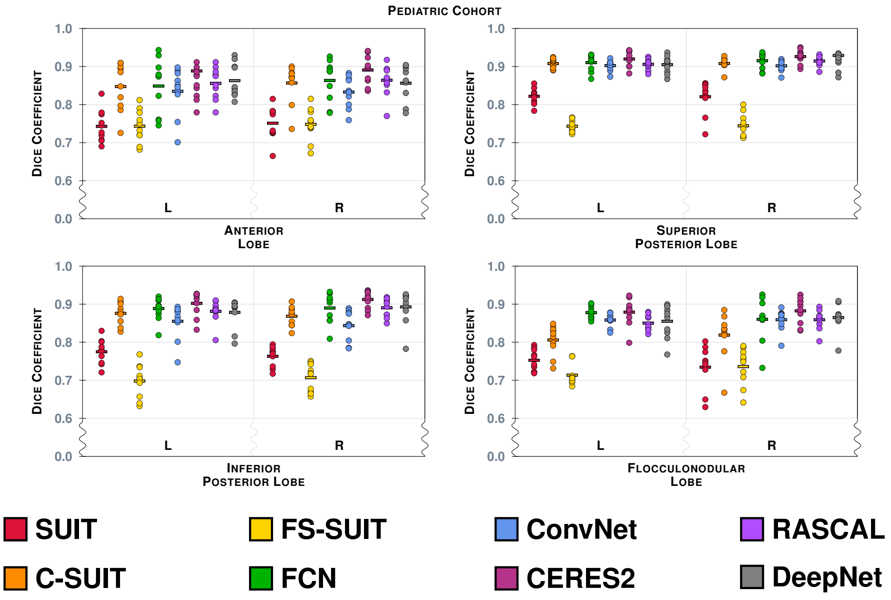


Figure 11: The Dice overlap for the eight labels associated with the Lobe hierarchy are shown for the Pediatric Cohort, see Table 4 for the list of lobe labels. See Fig. 10 for instructions on interpreting the plots. We note that the scale has been zoomed to help appreciate the differences between the algorithms.

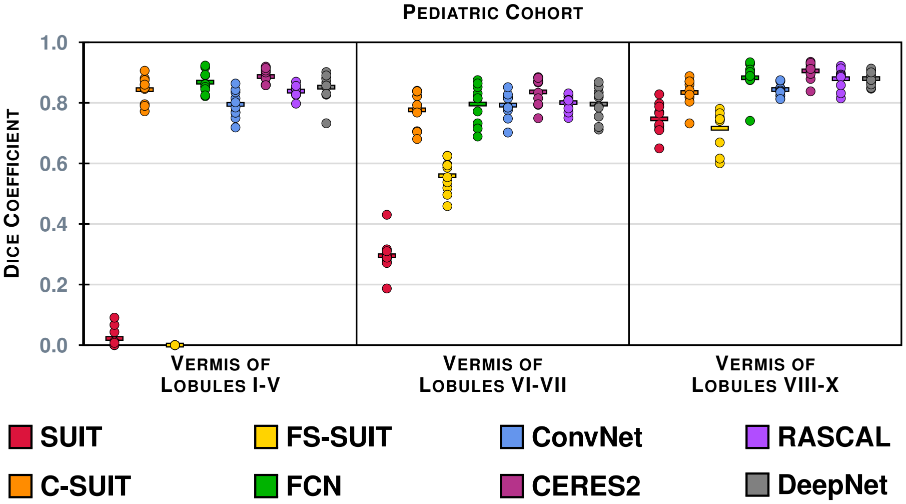


Figure 12: The Dice overlap for the three labels associated with the Vermis hierarchy are shown for the Pediatric Cohort, see Table 4 for the list of vermis labels. See Fig. 10 for instructions on interpreting the plots.

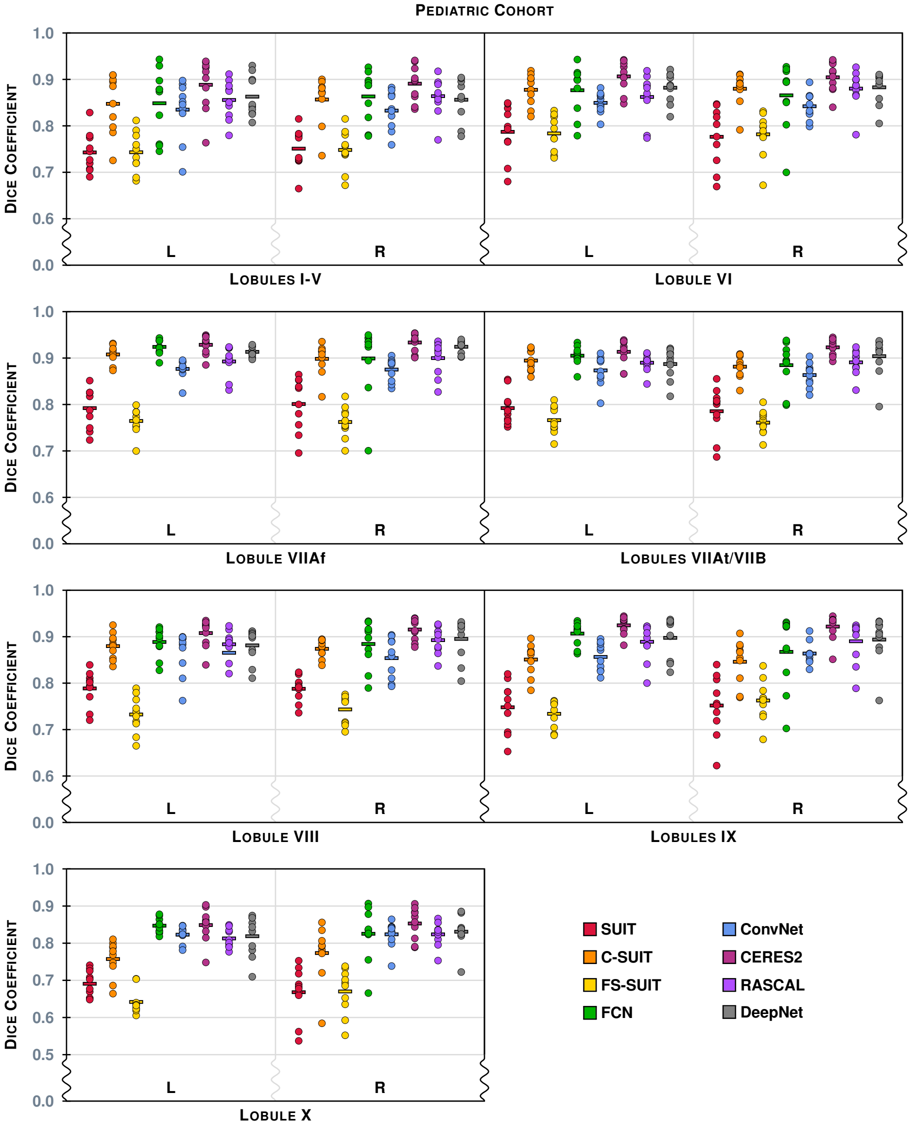


Figure 13: The Dice overlap for the 14 labels (7 per hemisphere) associated with the Lobule hierarchy are shown for the Pediatric Cohort, see Table 4 for the list of lobule labels. See Fig. 10 for instructions on interpreting the plots. We note that the scale has been zoomed to help appreciate the differences between the algorithms.

Table 7: A summary of the rank-sum calculation for each of the hierarchies. The Coarse hierarchy includes three labels: whole cerebellum, whole vermis, and CM; the Lobe hierarchy includes eight labels: Left/Right Anterior Lobe, Left/Right Superior Posterior, Left/Right Inferior Posterior, and Left/Right Flocculonodular; the Vermis hierarchy is five labels for the Adult Cohort and three labels for the Pediatric Cohort (see Table 4 for details); the Lobule hierarchy contains 22 labels for the Adult Cohort and 14 labels for the Pediatric Cohort (see Table 4 for details). Complete rank-sum calculation is included in the supplemental material.

		1st	2nd	3rd	4th	5th	6th	7th	8th
Adult Cohort	**Coarse**	🟪 **CERES2**	🟩 **FCN**	⬛ **DeepNet**	🟪 **RASCAL**	🟦 **ConvNet**	🟨 **FS-SUIT**		
	Lobe	🟪 **CERES2**	🟩 **FCN**	🟦 **ConvNet**	⬛ **DeepNet**	🟪 **RASCAL**	🟨 **FS-SUIT**		
	Vermis	🟪 **CERES2**	🟩 **FCN**	⬛ **DeepNet**	🟪 **RASCAL**	🟦 **ConvNet**	🟨 **FS-SUIT**		
	Lobule	🟪 **CERES2**	⬛ **DeepNet**	🟩 **FCN**	🟪 **RASCAL**	🟦 **ConvNet**	🟨 **FS-SUIT**		
	Consolidated	🟪 **CERES2**	⬛ **DeepNet**	🟩 **FCN**	🟪 **RASCAL**	🟦 **ConvNet**	🟨 **FS-SUIT**		
Pediatric Cohort	**Coarse**	🟩 **FCN**	🟪 **CERES2**	⬛ **DeepNet**	🟪 **RASCAL**	🟦 **ConvNet**	🟧 **C-SUIT**	🟨 **FS-SUIT**	🟥 **SUIT**
	Lobe	🟪 **CERES2**	🟩 **FCN**	⬛ **DeepNet**	🟪 **RASCAL**	🟧 **C-SUIT**	🟦 **ConvNet**	🟥 **SUIT**	🟨 **FS-SUIT**
	Vermis	🟪 **CERES2**	🟩 **FCN**	⬛ **DeepNet**	🟪 **RASCAL**	🟦 **ConvNet**	🟧 **C-SUIT**	🟥 **SUIT**	🟨 **FS-SUIT**
	Lobule	🟪 **CERES2**	⬛ **DeepNet**	🟩 **FCN**	🟪 **RASCAL**	🟧 **C-SUIT**	🟦 **ConvNet**	🟥 **SUIT**	🟨 **FS-SUIT**
	Consolidated	🟪 **CERES2**	🟩 **FCN**	⬛ **DeepNet**	🟪 **RASCAL**	🟧 **C-SUIT**	🟦 **ConvNet**	🟥 **SUIT**	🟨 **FS-SUIT**

Table 8: For the Adult Cohort, we show the p-value for the two-sided Wilcoxon paired signed-rank test comparing the second (🟩 **FCN**) and third (⬛ **DeepNet**) placed teams to the top (🟪 **CERES2**) ranked team across the four hierarchies (Coarse, Lobe, Vermis, Lobule) of labeling and also the combination of all 38 labels (Consolidated). The mean Dice overlap for each method, at the respective hierarchy, is shown underneath the method's name.

Hierarchy	**Method** **Mean Dice Overlap**		p**-value**
Coarse	🟪 **CERES2** 0.9118	*vs.* 🟩 **FCN** 0.8967	6.9×10^{-3} †
		vs. ⬛ **DeepNet** 0.8908	6.1×10^{-5} ‡
Lobe	🟪 **CERES2** 0.8395	*vs.* 🟩 **FCN** 0.8289	2.2×10^{-1}
		vs. ⬛ **DeepNet** 0.8021	1.9×10^{-4} †
Vermis	🟪 **CERES2** 0.8302	*vs.* 🟩 **FCN** 0.8012	1.2×10^{-2}
		vs. ⬛ **DeepNet** 0.8003	5.6×10^{-4} †
Lobule	🟪 **CERES2** 0.7657	*vs.* 🟩 **FCN** 0.7168	5.5×10^{-5} ‡
		vs. ⬛ **DeepNet** 0.7382	1.2×10^{-5} ‡
Consolidated	🟪 **CERES2** 0.8013	*vs.* 🟩 **FCN** 0.7657	3.0×10^{-7} ‡
		vs. ⬛ **DeepNet** 0.7719	3.1×10^{-12} ‡

† Denotes weak statistical significance (p-value < 0.001).

‡ Denotes strong statistical significance (p-value < 0.0001).

5. Discussion and Conclusions

5.1. Ranking the Methods

The primary result of this Comparison is a ranking of the state-of-the-art methods for parcellating the cerebellum, which is summarized in Table 7 for both the Adult and Pediatric Cohorts. The different levels of labeling, which we have referred to as hierarchies, allows for some granularity in understanding the ranking of the various methods on our cohorts. Had all the Participants contributed results for the two cohorts it would have been feasible to merge the rankings; regardless of this, there is an obvious stratification that occurs across both cohorts that is almost in-

Table 9: For the Pediatric Cohort, we show the p-value for the two-sided Wilcoxon paired signed-rank test comparing the second (🟩 **FCN**) and third (◼ **DeepNet**) placed teams to the top (🟪 **CERES2**) ranked team across the four hierarchies (Coarse, Lobe, Vermis, Lobule) of labeling and also the combination of all 28 labels (Consolidated). The mean Dice overlap for each method, at the respective hierarchy, is shown underneath the methods name.

Hierarchy	**Method** **Mean Dice Overlap**		**p-value**
Coarse	🟪 **CERES2** 0.9348	*vs.* 🟩 **FCN** 0.9326	2.1×10^{-1}
		vs. ◼ **DeepNet** 0.9201	6.0×10^{-6} ‡
Lobe	🟪 **CERES2** 0.9033	*vs.* 🟩 **FCN** 0.8859	7.4×10^{-6} ‡
		vs. ◼ **DeepNet** 0.8827	4.9×10^{-7} ‡
Vermis	🟪 **CERES2** 0.8763	*vs.* 🟩 **FCN** 0.8491	2.7×10^{-5} ‡
		vs. ◼ **DeepNet** 0.8427	7.5×10^{-5} ‡
Lobule	🟪 **CERES2** 0.9043	*vs.* 🟩 **FCN** 0.8776	1.6×10^{-11} ‡
		vs. ◼ **DeepNet** 0.8808	1.4×10^{-12} ‡
Consolidated	🟪 **CERES2** 0.9043	*vs.* 🟩 **FCN** 0.8828	2.2×10^{-16} ‡
		vs. ◼ **DeepNet** 0.8815	2.2×10^{-16} ‡

† Denotes weak statistical significance (p-value < 0.001).

‡ Denotes strong statistical significance (p-value < 0.0001).

dependent of the hierarchy. We observe that the order of CERES2, FCN, DeepNet, and RASCAL (as first through fourth) is very stable across both cohorts and the hierarchies. This is quite pleasing, as it points to a stability of both the algorithms and the labeling schemes used on both cohorts—even though the cohorts were labeled independently. We observe that these top methods all used spatial and intensity normalization to the MNI space.

5.2. Criticisms

The current work has two major shortcomings: 1) flawed cohorts and 2) exclusive use of Dice overlap. The two cohorts are flawed in different ways. Firstly, the Adult

Cohort while having a rich label set (CM label, five vermal labels, and 22 lobule labels) provided only five test data sets each of which showed signs of cerebellar dysfunction without a genetic diagnosis. In particular, the test data for the Adult Cohort had a mean age of 69.2 years of age, whereas the training data had a mean age of 54.7 years of age (see Table 3). A two-sided Wilcoxon signed-rank test (Wilcoxon, 1945) between the ages of the training and testing portions of the Adult Cohort has a p-value of 0.02, not significant but not a satisfactory situation either. The other issues with the Adult Cohort are its gender bias (all male test data versus training data that is only one third male) and the small size of the test data ($N = 5$). The effects of the gender bias are an unknown and the cohort size limits the statistical power of any tests. The cohort size also reduced the organizers' willingness to report standard deviations for the Dice overlap, with such a small sample any reported standard deviations would be erroneous. In contrast, to the Adult Cohort, the Pediatric Cohort has a slightly smaller label set (CM label, three vermal labels, and 18 lobule labels), a larger training pool of 20 data sets and a larger testing pool with 10 data sets. The gender proportions are consistent throughout the training and testing data sets as well as throughout the disease classifications in both the training and testing data. When using a two-sided Wilcoxon signed-rank test to perform a comparison between the ages of the training and test data, we get a more pleasing p-value of 0.95. The unfortunate drawback of the Pediatric Cohort is that it is pediatric data. The pediatric cerebellum is an area of great potential research and the availability of these automated methods for future work is very promising. However, the pediatric cerebellum remains an understudied portion of the central nervous system. The organizers believe that the pooling of these two cohorts to validate these methods is still a comprehensive test for any cerebellum parcellation method. Moreover, the range of ages (correspondingly head sizes) and cerebellar disease subtypes suggest a robustness of all the methods presented.

The remaining concern is the exclusive use of the Dice overlap measure throughout the paper. If we ignore the hierarchical label evaluation we employed, there were 28 labels in the Adult Cohort and 22 labels in the Pediatric Cohort. Given this many labels it seemed impractical to the organizers to report multiple metrics. Moreover, it would have been quite difficult to develop a consensus as to how to combine such metrics in a meaningful and unbiased manner. We also note that the majority of papers comparing multiple algorithms, as this paper does, are focused on a small number of labels. In fact it is typical for there to be only one label under consideration: white matter lesions, for example (Styner et al., 2008). As organizers, we observed in Maier et al. (2017) (from Table 7) that the final ranking correlated with the mean Dice overlap; in fact, the mean Dice overlap correctly predicts the top three methods and only incorrectly ranks three of the fourteen methods under consideration. This occurs despite the fact that the Dice overlap is only one component of a multi-measure evaluation (Maier et al., 2017). Thus, we believe exclusive use of the Dice overlap is acceptable and that our analysis of this Comparison correctly represents the state-of-the-art in fully automated cerebellum parcellation.

5.3. Comment on Inter-rater Performance

A portion of our Adult Cohort, along with other similarly acquired data, was used as part of an inter-rater comparison (Bogovic et al., 2013b). It is reassuring to see that the performance of the top methods in this Comparison have similar Dice overlap to that reported for the inter-rater analysis. In particular, the mean Dice overlap for CERES2, FCN, and DeepNet, for the whole vermis are larger than those reported for the inter-rater values (Fig. 5 in Bogovic et al. (2013b)).

5.4. Impact of this Work

Identifying the state-of-the-art in cerebellum parcellation is important for improving the robustness and speed with which cerebellum imaging studies can be completed. Although SUIT (Diedrichsen, 2006; Diedrichsen et al., 2009) has been available and widely used for over 10 years, our study clearly reveals that there are emerging methods with significantly better performance (given our performance criteria); we note that the probabilistic lobular segmentation generated by SUIT was meant to be informative and not definitive. As studies begin to emerge relating the volumes of cerebellar lobules to functional brain performance (cf. Kansal et al. (2016)), methods such as CERES2, FCN, and DeepNet may offer a better alternative for identifying these volumes. As well, this study provides a baseline for future work on cerebellar parcellation, both in providing information on the best strategies to date and in providing Dice coefficients for comparison.

Acknowledgments

The data collection and labeling of the cerebellum was supported in part by the NIH/NINDS grant R01 NS056307 (PI: J.L. Prince) and NIH/NIMH grants R01 MH078160 & R01 MH085328 (PI: S.H. Mostofsky). PMT is supported in part by the NIH/NIBIB grant U54 EB020403. CERES2 development was supported by grant UPV2016-0099 from the Universitat Politécnica de Valencia (PI: J.V. Manjón); the French National Research Agency through the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02, HL-MRI Project; PI: P. Coupé) and Cluster of excellence CPU and TRAIL (HR-DTI ANR-10-LABX-57; PI: P. Coupé). Support for the development of FCN was provided by the National Science and Engineering Research Council of Canada (NSERC), discovery grant program, and by the ETS Research Chair on Artificial Intelligence in Medical Imaging. The authors wish to acknowledge the invaluable contributions offered by Dr. George Fein (Dept. of Medicine and Psychology, University of Hawaii) in preparing this manuscript.

References

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95–113.

Ashburner, J., Andersson, J., Friston, K.J., 2000. Image registration using a symmetric prior - in three-dimensions. Human Brain Mapping 9, 212–225.

Ashburner, J., Friston, K.J., 1999. Nonlinear spatial normalization using basis functions. Human Brain Mapping 7, 254–266.

Ashburner, J., Friston, K.J., 2005. Unified Segmentation. NeuroImage 26, 839–851.

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. Medical Image Analysis 12, 26–41.

Baker, K.G., Harding, A.J., Halliday, G.M., Kril, J.J., Harper, C.G., 1999. Neuronal loss in functional zones of the cerebellum of chronic alcoholics with and without Wernicke's encephalopathy. Neuroscience 91, 429–438.

Banerjee, S., Mitra, S., Shankar, B.U., Hayashi, Y., 2016. A Novel GBM Saliency Detection Model Using Multi-Channel MRI. PLOS ONE 11, e0146388.

Barnes, C., Shechtman, E., Golman, D.B., Finkelstein, A., 2010. The generalized patchmatch correspondence algorithm, in: 2010 European Conference on Computer Vision (ECCV 2010), Springer Berlin Heidelberg. pp. 29–43.

Baumann, O., Borra, R.J., Bower, J.M., Cullen, K.E., Habas, C., Ivry, R.B., Leggio, M., Mattingley, J.B., Molinari, M., Moulton, E.A., Paulin, M.G., Pavlova, M.A., Schmahmann, J.D., Sokolov, A.A., 2015. Consensus Paper: The Role of the Cerebellum in Perceptual Processes. The Cerebellum 14, 197–220.

Bazin, P.L., Pham, D.L., 2008. Homeomorphic brain image segmentation with topological and statistical atlases. Medical Image Analysis 12, 616–625.

Bergstra, J., Breulex, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, B., Bengio, Y., 2010. Theano: A CPU and GPU math compiler in Python, in: Proc. 9th Python in Science Conf., pp. 1–7.

Bogovic, J.A., Bazin, P.L., Ying, S.H., Prince, J.L., 2013a. Automated Segmentation of the Cerebellar Lobules Using Boundary Specific Classification and Evolution, in: 23rd Inf. Proc. in Med. Imaging (IPMI 2013), Springer Berlin Heidelberg. pp. 62–73.

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

Bogovic, J.A., Prince, J.L., Bazin, P.L., 2013c. A multiple object geometric deformable model for image segmentation. Comput. Vis. Image Und. 117, 145–157.

Carass, A., Cuzzocreo, J., Wheeler, M.B., Bazin, P.L., Resnick, S.M., Prince, J.L., 2010. Simple paradigm for extra-cerebral tissue removal: Algorithm and analysis. NeuroImage 56, 1982–1992.

Carass, A., Prince, J.L., 2016. An Overview of the Multi-Object Geometric Deformable Model Approach in Biomedical Imaging, in: Zhou, S.K. (Ed.), Medical Image Recognition, Segmentation and Parsing. Academic Press, pp. 259–279.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., Cardoso, M.J., Cawley, N., Ciccarelli, O., Wheeler-Kingshott, C.A.M., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick, O., Barillot, C., Tomas-Fernandez, X., Warfield, S.K., Vaidya, S., Chunduru, A., Muthuganapathy, R., Krishnamurthi, G., Jesson, A., Arbel, T., Maier, O., Handels, H., Iheme, L.O., Unay, D., Jain, S., Sima, D.M., Smeets, D., Ghafoorian, M., Platel, B., Birenbaum, A., Greenspan, H., Bazin, P.L., Calabresi, P.A., Crainiceanu, C., Ellingsen, L.M., Reich, D.S., Prince, J.L., Pham, D.L., 2017. Longitudinal multiple sclerosis lesion segmentation: Resource & challenge. NeuroImage 148, 77–102.

Cavanagh, J.B., Holton, J.L., Nolan, C.C., 1997. Selective damage to the cerebellar vermis in chronic alcoholism: A contribution from neurotoxicology to an old problem of selective vulnerability. Neuropathology and applied neurobiology 23, 355–363.

Çiçek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in: 19th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2016), Springer Berlin Heidelberg. pp. 424–432.

Chen, H., Qi, X., Yu, L., Heng, P.A., 2016. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2487–2496.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3-D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assis. Tomogr. 18, 192–205.

Colloby, S.J., O'Brien, J.T., Taylor, J.P., 2014. Patterns of cerebellar volume loss in dementia with Lewy bodies and Alzheimer's disease: A VBM-DARTEL study. Psychiatry Research: Neuroimaging 223, 187–191.

Coupe, P., Manjon, 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, 940–954.

Courchesne, E., Saitoh, O., Yeung-Courchesne, R., Press, G.A., Lincoln, A.J., Haas, R.H., Schreibman, L., 1994. Abnormality of Cerebellar Vermian Lobules VI and VII in Patients with Infantile Autism: Identification of Hypoplastic and Hyperplastic with MR Imaging. Am. J. of Roentgenology 162, 123–130.

Desmond, J.E., Fiez, J.A., 1998. Neuroimaging studies of the cerebellum: language, learning and memory. Trends Cogn. Sci. 2, 355–362.

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26, 297–302.

Diedrichsen, J., 2006. A spatially unbiased atlas template of the human cerebellum. NeuroImage 33, 127–138.

Diedrichsen, J., Balsters, J.H., Flavell, J., Cussans, E., Ramnani, N., 2009. A probabilistic MR atlas of the human cerebellum. NeuroImage 46, 39–46.

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: Clinical 7, 631–639.

Dolz, J., Desrosiers, C., Ayed, I.B., 2018. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. NeuroImage 170, 456–470.

Dos Santos, V., Thomann, P.A., Wüstenberg, T., Seidl, U., Essig, M., Schröder, J., 2011. Morphological cerebral correlates of CERAD test performance in mild cognitive impairment and Alzheimer's disease. Jrnl. Alzheimer's Disease 23, 411–420.

Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781.

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, 341–355.

Fischl, B., Salat, D.H., van der Kouwe, A.J.W., Makris, N., Ségonne, F., Quinn, B.T., Dale, A.M., 2004. Sequence-independent segmentation of magnetic resonance images. NeuroImage 23, S69–s84.

Fitzpatrick, L.E., Jackson, M., Crowe, S.F., 2008. The relationship between alcoholic cerebellar degeneration and cognitive and emotional functioning. Neuroscience & Biobehavioral Reviews 32, 466–485.

Fonov, V.S., Evans, A.C., Botteron, K., McKinstry, R.C., Collins, D.L., the Brain Development Cooperative Group, 2010. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54, 313–327.

Giraud, R., Ta, V.T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P., Alzheimer's Disease Neuroimaging Initiative, 2016. An optimized patchmatch for multi-scale and multi-feature label fusion. NeuroImage 124, 770–782.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., Fischl, B., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. NeuroImage 32, 180–194.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing humanlevel performance on imagenet classification, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034.

He, Y., Carass, A., Yun, Y., Zhao, C., Jedynak, B.M., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L., 2017. Towards Topological Correct Segmentation of Macular OCT from Cascaded FCNs, in: Fetal, Infant and Ophthalmic Medical Image Analysis: International Workshop, FIFI 2017, and 4th International Workshop, OMIA 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings, Springer Berlin Heidelberg. pp. 202–209.

Heimann, T., van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., Binnig, G., Bischof, H., Bornik, A., Cashman, P., Ying, C., Cordova, A., Dawant, B.M., Fidrich, M., Furst, J.D., Furukawa, D., Grenacher, L., Hornegger, J., Kainmuller, D., Kitney, R.I., Kobatake, H., Lamecker, H., Lange, T., Lee, J., Lennon, B., Li, R., Li, S., Meinzer, H.P., Nemeth, G., Raicu, D.S., Rau, A.M., van Rikxoort, E.M., Rousson, M., Rusko, L., Saddi, K.A., Schmidt, G., Seghers, D., Shimizu, A., Slagmolen, P., Sorantin, E., Soza, G., Susomboon, R., Waite, J.M., Wimmer, A., Wolf, I., 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans. Med. Imag. 28, 1251–1265.

Henle, J., 1879. Handbuch der Nervenlehre des Menschen. Friedrich Vieweg und Sohn, Braunschweig.

Huo, Y., Plassard, A.J., Carass, A., Resnick, S.M., Pham, D.L., Prince, J.L., Landmann, B.A., 2016. Consistent Cortical Reconstruction and Multi-atlas Brain Segmentation. NeuroImage 138, 197–210.

Ito, M., 1984. The cerebellum and neural control. Raven, New York.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menton, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36, 61–78.

Kansal, K., Yang, Z., Fishman, A.M., Sair, H.I., Ying, S.H., Jedynak, B.M., Prince, J.L., Onyike, C.U., 2016. Structural cerebellar correlates of cognitive and motor dysfunctions in cerebellar degeneration. Brain 140, 707–720.

Karas, G.B., Burton, E.J., Rombouts, S.A.R.B., van Schijndel, R.A., O'Brien, J.T., Scheltens, P., McKeith, I.G., Williams, D., Ballard, C., Barkhof, F., 2003. A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. NeuroImage 18, 895–907.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980. arXiv:1412.6980.

Larsell, O., 1952. The morphogenesis and adult pattern of the lobules and tissues of the cerebellum of the white rat. J. Comp. Neuol. 97, 281–356.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H.L., Havaei, M., Iftekharuddin, K.M., Jodoin, P.M., Kamnitsas, K., Kellner, E., Korvenoja, A., Larochelle, H., Ledig, C., Lee, J.H., Maes, F., Mahmood, Q., Maier-Hein, K.H., McKinley, R., Muschelli, J., Pal, C., Pei, L., Rangarajan, J.R., Reza, S.M.S., Robben, D., Rueckert, D., Salli, E., Suetens, P., Wang, C.W., Wilms, M., Kirschke, J.S., Krämer, U.M., Münte, T.F., Schramm, P., Wiest, R., Handels, H., Reyes, M., 2017. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269.

Malacarne, V., 1776. Nuova esposizione della vera struttura del cervelletto umano. Briolo, Torino.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. Mag. Reson. Im. 31, 192–203.

Manjón, J.V., Coupé, P., Raniga, P., Xia, Y., Fripp, J., Salvado, O., 2016. HIST: HyperIntensity Segmentation Tool, in: Patch-MI 2016: Patch-Based Techniques in Medical Imaging, Springer Berlin Heidelberg. pp. 92–99.

Manto, M., Schmahmann, J.D., Koibuchi, N., Rossi, F., 2013. Handbook of the Cerebellum and Cerebellar Disorders. Springer, New York.

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W., de Bresser, J., Alansary, A., de Bruijn, M., Carass, A., El-Baz, A., Jog, A., Katyali, R., Khan, A.R., van der Lijn, F., Mahmood, Q., Mukherjee, R., van Opbroek, A., Paneri, S., Pereira, S., Persson, M., Rajchl, M., Sarikayan, D., Smedby, O., Silva, C.A., Vrooman, H.A., Vyas, S., Wang, C., Zhaon, L., Biessels, G.J., Viergever, M.A., 2015. MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans. Computational Intelligence and Neuroscience .

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Riklin-Raviv, T., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Székely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imag. 34, 1993–2024.

Möller, C., Vrenken, H., Jiskoot, L., Versteeg, A., Barkhof, F., Scheltens, P., van der Flier, W.M., 2013. Different patterns of gray matter atrophy in early- and late-onset Alzheimer's disease. Neurobiology of Aging 34, 2014–2022.

Mostofsky, S.H., Mazzocco, M.M.M., Aakalu, G., Warsofsky, I.S., Denckla, M.B., Reiss, A.L., 1998a. Decreased cerebellar posterior vermis size in fragile X syndrome: Correlation with neurocognitive performance. Neurology 50, 121–130.

Mostofsky, S.H., Reiss, A.L., Lockhart, P., Denckla, M.B., 1998b. Evaluation of cerebellar size in attention-deficit hyperactivity disorder. J. Child Neurol. 13, 434–439.

Nopoulos, P.C., Ceilley, J.W., Gailis, E.A., Andreasen, N.C., 1999. An MRI study of cerebellar vermis morphology in patients with schizophrenia: Evidence in support of the cognitive dysmetria concept. Biological psychiatry 46, 703–711.

Nyúl, L.G., Udupa, J.K., 1999a. New Variants of a Method of MRI Scale Normalization, in: 16th Inf. Proc. in Med. Imaging (IPMI 1999), Springer Berlin Heidelberg. pp. 490–495.

Nyúl, L.G., Udupa, J.K., 1999b. On Standardizing the MR Image Intensity Scale. Mag. Reson. Med. 42, 1072–1081.

Okugawa, G., Sedvall, G., Nordström, M., Andreasen, N., Pierson, R., Magnotta, V., Agartz, I., 2002. Selective reduction of the posterior superior vermis in men with chronic schizophrenia. Schizophrenia Research 55, 61–67.

Okugawa, G., Sedvall, G.C., Agartz, I., 2003. Smaller cerebellar vermis but not hemisphere volumes in patients with chronic schizophrenia. Am. J. of Psychiatry 160, 1614–1617.

Parker, K.L., Narayanan, N.S., Andreasen, N.C., 2014. The therapeutic potential of the cerebellum in schizophrenia. Front Syst Neurosci. 8, 163.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 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, 61–76.

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, 238–247.

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

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.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: 18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2015), Springer Berlin Heidelberg. pp. 234–241.

Rousseau, F., Habas, P.A., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imag. 30, 1852–1862.

Sauwen, N., Sima, D.M., Van Cauter, S., Veraart, J., Leemans, A., Maes, F., Himmelreich, U., Van Huffel, S., 2015. Hierarchical non-negative matrix factorization to characterize brain tumor heterogeneity using multi-parametric MRI. NMR in Biomedicine 28, 1599–1624.

Schaap, M., Metz, C.T., van Walsum, T., van der Giessen, A.G., Weustink, A.C., Mollet, N.R., Bauer, C., Bogunović, H., Castro, C., Deng, X., Dikici, E., O'Donnell, T., Frenay, M., Friman, O., Hoyos, M.H., Kitslaar, P.H., Krissian, K., Kühnel, C., Luengo-Oroz, M.A., Orkisz, M., Smedby, Ö., Styner, M., Szymczak, A., Tek, H., Wang, C., Warfield, S.K., Zambal, S., Zhang, Y., Krestin, G.P., Niessen, W.J., 2009. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. Medical Image Analysis 13, 701–714.

Schmahmann, J.D., 1991. An emerging concept. The cerebellar contribution to higher function. Arch. Neurol. 48, 1178–1187.

Schmahmann, J.D., 2004. Disorders of the cerebellum: ataxia, dysmetria of thought, and the cerebellar cognitive affective syndrome. The Journal of neuropsychiatry and clinical neurosciences 16, 367–378.

Schmahmann, J.D., Caplan, D., 2006. Cognition, emotion and the cerebellum. Brain 129, 290–292.

Schmahmann, J.D., Doyon, J., Toga, A.W., Petrides, M., Evans, A.C., 2000. MRI atlas of the human cerebellum. Academic Press, San Diego, CA.

Schutter, D.J.L.G., Van Honk, J., 2005. The cerebellum on the rise in human emotion. The Cerebellum 4, 290–294.

Silveri, M.C., Leggeio, M.G., Molinari, M., 1994. The cerebellum contributes to linguistic production. Neurology 44, 2047.

Styner, M., Lee, J., Chin, B., Chin, M.S., Commowick, O., Tran, H.H., Markovic-Plese, S., Jewells, V., Warfield, S., 2008. 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation, in: 11th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2008) 3D Segmentation in the Clinic: A Grand Challenge II, pp. 1–6.

Ta, V.T., Giraud, R., Collins, D.L., Coupé, P., 2014. Optimized patchMatch for near real time and accurate label fusion, in: 17th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2014), Springer Berlin Heidelberg. pp. 105–112.

Thomann, P.A., Schläfer, C., Seidl, U., Dos Santos, V., Essig, M., Schröder, J., 2008. The cerebellum in mild cognitive impairment and Alzheimer's disease - A structural MRI study. J. Psychiatric Res. 42, 1198–1202.

Tieleman, T., Hinton, G., 2015. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, in: COURSERA: Neural networks for machine learning. volume 4, pp. 26–31.

Torvik, A., Torp, S., 1986. The prevalence of alcoholic cerebellar atrophy: A morphometric and histological study of an autopsy material. Journal of the Neurological Sciences 75, 43–51.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: Improved N3 Bias Correction. IEEE Trans. Med. Imag. 29, 1310–1320.

Victor, M., Adams, R.D., Mancall, E.L., 1959. A Restricted Form of Cerebellar Cortical Degeneration Occurring in Alcoholic Patients. AMA Arch. Neurol. 1, 579–688.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-Atlas Segmentation with Joint Label Fusion. IEEE Trans. Patt. Anal. Mach. Intell. 35, 611–623.

Weier, K., Fonov, V., Lavoie, K., Doyon, J., Collins, D.L., 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. Human Brain Mapping 35, 5026–5039.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.

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. Schizophrenia Research 176, 164–170.

Yang, Z., Abulnaga, S.M., Carass, A., Kansal, K., Jedynak, B.M., Onyike, C.U., Ying, S.H., Prince, J.L., 2016a. Landmark Based Shape Analysis for Cerebellar Ataxia Classification and Structural Change Pattern Visualization, in: Proceedings of SPIE Medical Imaging (SPIE-MI 2016), San Diego, CA, February 27-March 3, 2016, pp. 9784 – 9784 – 8.

Yang, Z., Ye, C., Bogovic, J.A., Carass, A., Jedynak, B.M., Ying, S.H., Prince, J.L., 2016b. Automated Cerebellar Lobule Segmentation with Application to Cerebellar Structural Analysis in Cerebellar Disease. NeuroImage 127, 435–444.

Yang, Z., Zhong, S., Carass, A., Ying, S.H., Prince, J.L., 2014. Deep Learning for Cerebellar Ataxia Classification and Functional Score Regression, in: Machine Learning in Medical Imaging (MLMI 2014), pp. 68–76.

Ying, S.H., Choi, S.I., Perlman, S.L., Baloh, R.W., Zee, D.S., Toga, A.W., 2006. Pontine and cerebellar atrophy correlate with clinical disability in SCA2. Neurology 66, 424–426.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the Expectation-Maximization algorithm. IEEE Trans. Med. Imag. 20, 45–57.

Zhao, C., Carass, A., Lee, J., He, Y., Prince, J.L., 2017. Whole Brain Segmentation and Labeling from CT Using Synthetic MR Images, in: Machine Learning in Medical Imaging (MLMI 2017), Springer Berlin Heidelberg. pp. 291–298.