Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Approaching expert results using a hierarchical cerebellum parcellation protocol for multiple inexpert human raters

John A. Bogovic [a,*], Bruno Jedynak [b], Rachel Rigg [e], Annie Du [f], Bennett A. Landman [d], Jerry L. Prince [a,c,e], Sarah H. Ying [c,f,g]

[a] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
[b] Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD, USA
[c] Department of Radiology and Radiological Science, Johns Hopkins Medical Institutions, Baltimore, MD, USA
[d] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA
[e] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[f] Department of Neurology, Johns Hopkins Medical Institutions, Baltimore, MD, USA
[g] Department of Ophthalmology, Johns Hopkins Medical Institutions, Baltimore, MD, USA

ARTICLE INFO

Article history:
Accepted 27 August 2012
Available online 4 September 2012

Keywords:
Human cerebellum
Manual labeling
Delineation
Parcellation
STAPLE
STAPLER
Label fusion

ABSTRACT

Volumetric measurements obtained from image parcellation have been instrumental in uncovering structure–function relationships. However, anatomical study of the cerebellum is a challenging task. Because of its complex structure, expert human raters have been necessary for reliable and accurate segmentation and parcellation. Such delineations are time-consuming and prohibitively expensive for large studies. Therefore, we present a three-part cerebellar parcellation system that utilizes multiple inexpert human raters that can efficiently and expediently produce results nearly on par with those of experts. This system includes a hierarchical delineation protocol, a rapid verification and evaluation process, and statistical fusion of the inexpert rater parcellations. The quality of the raters' and fused parcellations was established by examining their Dice similarity coefficient, region of interest (ROI) volumes, and the intraclass correlation coefficient of region volume. The intra-rater ICC was found to be 0.93 at the finest level of parcellation.

Introduction

The cerebellum has a remarkably complex structure that coordinates numerous vital functions of the human body. It is involved in tasks such as eye-movement (McCormick and Thompson, 1984; Ritchie, 1976), speech (Silveri et al., 1994), balance, fine motor control, motor learning, and cognition (Leiner et al., 1986; Schmahmann, 1991). Like the cerebral cortex, the human cerebellum exhibits functional localization. This is reflected in part by its anatomic structure. There are two macroscopic levels of organization: medial–lateral and rostral–caudal. The medial–lateral anatomical divisions of the cerebellum echo the differences in connectivity between the medially located spinocerebellum, and the lateral cerebrocerebellum, or "neo-cerebellum." The spinocerebellum consists of the wormlike "vermis" and the more lateral paravermis, or "intermediate zone." As its name suggests, these regions receive afferents primarily from the spinal cord. The evolutionarily ancient flocculonodular lobe or "vestibulocerebellum" is intimately associated with the vestibular system, and therefore highly influences spatial orientation and balance. In the rostral–caudal direction, transverse fissures create divisions in the cerebellum called lobules.

The study of cerebellar substructures has been confounded by inconsistencies in nomenclature. This has been remedied in the study of humans by the general acceptance of the standard introduced by Schmahmann et al. (1999, 2000). In their work, the cerebellar lobules are numbered from I to X. These lobules stem from white matter branches rooted in the central mass of cerebellar white matter, called the corpus medullare (CM). Under this convention, lobule I is located most rostrally, with lobule numbering increasing caudally. In this work, we adopt the numbering standard of Schmahmann, but will refer to super-groupings as follows: anterior (I–V), superior posterior, or middle (VI and Crus I and II of VIIA, and VIIB), inferior posterior (VIII, IX), and caudal (VIII, IX, and X). Three of these conventions and super-groupings are illustrated in Fig. 1.

Region-specific changes in the cerebellum have been correlated with a number of diseases and functional deficits. For example, regionally selective degeneration of the vermis and anterior lobe has been observed over the course of aging (Andersen et al., 2003; Raz et al., 1998). A decrease in size of the inferior posterior vermis has been observed in boys with attention-deficit and hyperactivity disorder (ADHD) relative to normals (Berquin et al., 1998; Mostofsky et al., 1998). Several studies have shown changes in volumes of the vermis (Nopoulos et al., 1999; Okugawa et al., 2003) and vermian white matter (Levitt et al., 1999) in patients with schizophrenia. Evidence suggests that the vermis and

* Corresponding author.
E-mail address: bogovic@jhu.edu (J.A. Bogovic).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2012.08.075

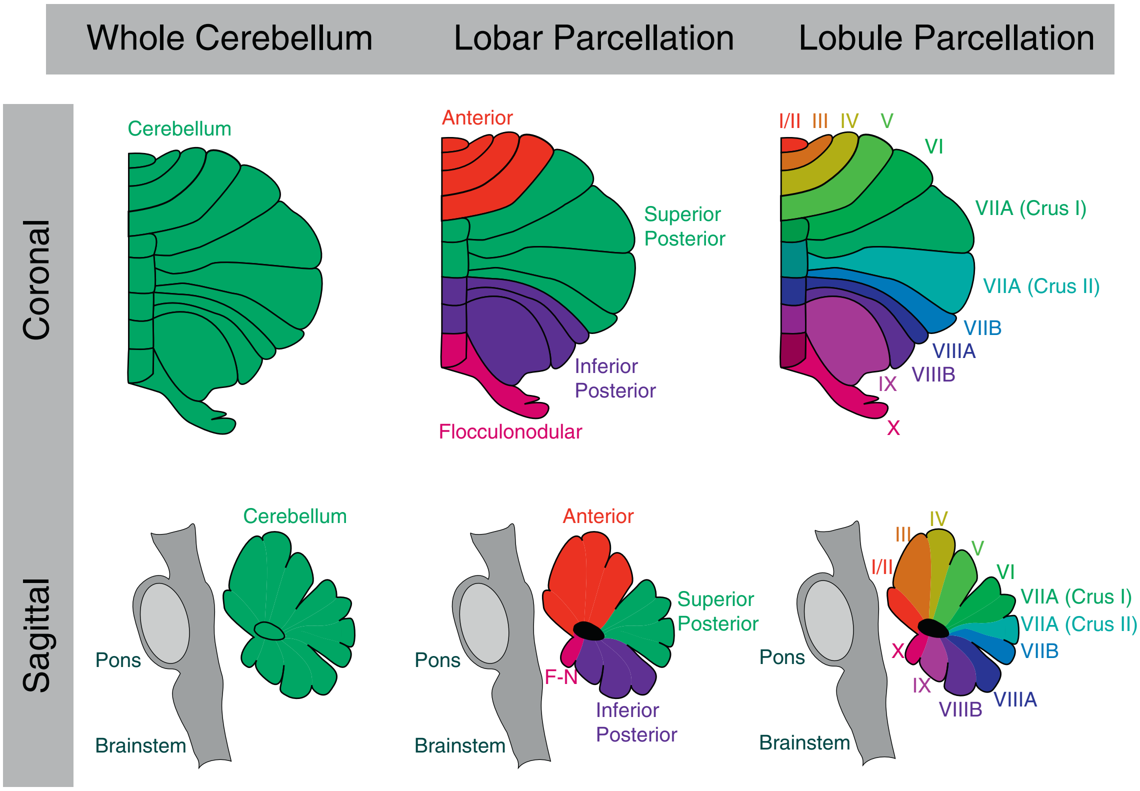

Fig. 1. Illustration of cerebellar anatomy and lobule grouping.

flocculus are targeted by chronic alcoholism (Baker et al., 1999; Cavanagh et al., 1997). Superior posterior lobe volumes have been shown to decrease in patients with Alzheimer's disease relative to controls (Thomann et al., 2008). Several types of cerebellar ataxia have also demonstrated region-specific atrophy within the cerebellum (Brenneis et al., 2003; Jung et al., 2011; Ying et al., 2006).

Clearly, the success of such studies depends on accurate and precise measurements for the structures of interest. Segmentations produced by human experts remain the gold standard despite the progress achieved in automated segmentation algorithms. However, the training of experts is a very long process, typically requiring thousands of hours. The high quality measurements produced by such raters are therefore time-consuming and expensive. Here we briefly review examples of three approaches that have been used to limit the amount of human expertise required: 1) limit the scope of the study, 2) employ automated or semi-automated image analysis methods, or 3) employ inexpert human raters.

In order to make best use of the experts' time, research hypotheses may be tested using a small cohort of subjects, on a small number of subregions, or using a coarse parcellation. For example, one may limit study to the cerebellar vermis. Raz et al. (1998) measured the cross-sectional area of the cerebellar vermis grouping lobules I–V, VI–VII, and VIII–X. A coarse parcellation considered by Levitt et al. (1999), among others, consists of hemispheric white and gray matter, vermian white matter, and gray matter split into three subgroups consisting of lobules I–V, VI–VII, and VIII–X.

A few methods have been introduced that produce a full parcellation of the cerebellar lobules from magnetic resonance (MR) images. A semi-automatic surface-based method presented by Makris et al. (2003, 2005) has the advantage of parcellating the cerebellum into medial-to-lateral subdivisions. Pierson et al. (2002) developed manual and semi-automated methods for delineating the corpus medullare, anterior lobe, superior posterior lobe, and inferior posterior lobe. These manual and semi-automated methods have produced good results, but rater training and delineation time are not discussed. Significant knowledge of cerebellar anatomy is required to follow these protocols, and their use will yield adequate results only when used by expert human raters. As a consequence, they tend also to be time-consuming and expensive.

To our knowledge, there is currently only one publicly available automated method for cerebellar parcellation. Diedrichsen (2006), Diedrichsen et al. (2009) describe an automatic method based on nonlinear registration of a cerebellum label template (SUIT). This methodology was employed in (Donchin et al., 2012) in the case of focal lesions and cerebellar atrophy and showed lobule-specific changes using voxel-based morphometry. However, an explicit evaluation of segmentation performance was not performed. For our data, the SUIT template produced a mean Dice similarity[1] (standard deviation) of 0.55 (0.16) across 15 subjects and 24 labels (see Appendix A for details). Since this method uses a template constructed from control subjects, it is not surprising that results are poor for patients with severe cerebellar degenerative disease (especially those including diverse forms of lobule-specific atrophy). Our data, which include nine subjects with significant degeneration, demonstrate this phenomenon. In any case, testing this approach (or any alternative) would require numerous manual parcellations for validation. The present paper describes a cost-effective approach to produce such manual parcellations on both normal and ataxia subjects.

Recently, the "multi-atlas" segmentation framework has been shown to achieve excellent results for many anatomical segmentation tasks (Heckemann et al., 2006; Isgum et al., 2009). We explored the performance of such a method (see Appendix A for details) for cerebellar parcellation and found that it produced unsatisfactory results: mean Dice similarity (standard deviation) across 15 subjects and 28 labels of

[1] The Dice similarity coefficient (DSC) (Dice, 1945) was computed with the "gold standard" expert rater delineation. For a given label l, the DSC of a parcellation with the expert parcellation is given by: $DSC(l) = 2(|S_{Rl} \cap S_{El}|)/(|S_{El}| + |S_{Rl}|)$, where S_{Rl} and S_{El} denote the sets of voxels assigned to label l by the rater and expert, respectively.

0.66 (0.15). An important factor in explaining this is the high complexity and inter-subject variability of the anatomical shape of the cerebellum.

In order to address the need for simultaneously accurate and inexpensive ratings, numerous explorations into recruiting members of the public to complete rating tasks ("crowd sourcing") have been performed. These studies have challenges ranging from object identification and annotation in images (Russell et al., 2007; Sorokin and Forsyth, 2008) to language translation (Callison-Burch, 2009). The results suggest that good results can be obtained once enough data from the crowd has been collected. Of course, the quality of the final results depends on the difficulty of the task, the extent to which the "crowd" is trained, and the method for combining their ratings. Tasks that require little to no training (e.g., object identification or naming) often produce excellent results, while the results for more challenging tasks are usually sub-par. In a previous work, a large number of minimally trained raters performed an easy version of the cerebellar parcellation task with acceptable accuracy (Landman et al., 2012a). In particular, the labeling task was two-dimensional with only four labeled substructures whose boundaries are readily discernible by underlying image intensities, while here we seek three-dimensional labeling with as many as 29 substructures, some of which are only separable by knowledge of cerebellar anatomy. Minimally trained "crowd" raters, having no training in cerebellar anatomy, are incapable of producing acceptable, detailed three-dimensional parcellations, even when statistically combined.

In this work, we sought to develop a delineation scheme that yields the accuracy of expert raters with the efficiency (with regard to time and cost) of crowd raters. The anatomical complexity and subtlety of image features in the cerebellum necessitate a large time commitment be devoted to training and practice in order to become an expert cerebellum delineator. In our experience, approximately 2000 h were devoted in order to achieve excellent reliability (>0.99 ICC) (Jung et al., 2011). Note that this figure depends on many factors such as the difficulty of the parcellation problem or previous experience or qualifications of the rater (e.g., a seasoned radiologist will require less training for a particular task than a novice). We recommend that researchers take these factors into account and perform "on-the-fly" evaluation during the training process to determine when a rater achieves the desired performance. On the other hand, raters from the crowd (with less than 20 h of experience), struggle to produce acceptable delineations due to the anatomical knowledge required. We have found that individuals with a moderate amount of training (about 150 h), whom we call "inexpert raters", are markedly superior to "crowd" raters, and after review and statistical fusion, their results can yield delineations approaching the quality of that produced by an expert.

This paper describes a system, involving three stages, for the manual parcellation of the cerebellum by multiple inexpert raters. The first stage in this system is a hierarchical delineation protocol which enables a newcomer to label the cerebellum with relative ease and accuracy. The hierarchical scheme involves first delineating the whole cerebellum, and then outlining finer structures until all sub-structures of the cerebellum have been obtained. A hierarchical approach reduces rater variability since the larger, more easily located anatomical landmarks are delineated at early stages in the hierarchy. At later stages, the raters need only concentrate on further delineation of the regions of the cerebellum that were previously established. An additional advantage of this approach is that a coarser parcellation can be achieved in less time than that required for the full parcellation, and coarser parcellations may be ideal for some studies. Our proposed methodology achieves a mean (standard deviation) Dice similarity of 0.84 (0.10) and outperforms all automated methods that we have been able to test (see Appendix A for further details).

Our protocol is specifically designed and tested on both normal and highly atrophied cerebella. Subjects diagnosed with spinocerebellar ataxia types 2 and 6, and subjects with non-genetic cerebellar ataxia were among those used for rater training and testing. Our protocol is likely to be applicable for studies involving ADHD, schizophrenia, or chronic alcoholism and others, because the extent and patterns of atrophy are comparable in these conditions. The proposed methodology may not be successful in more extreme conditions such as Dandy–Walker syndrome. Because of the potential for large inter-subject variations in cerebellar size and shape, raters can make significant errors, as illustrated in Fig. 2. Therefore, the second stage of our delineation system is a method for rapid rater review, evaluation, and correction with emphasis on the large errors that our inexpert raters sometimes make.

The third stage of our delineation system is the application of a robust statistical fusion method to the results of three or more inexpert raters (Landman et al., 2012b). This stage removes errors made by the raters, since the consensus of these imperfect delineations tend to be accurate. Fig. 2 shows an example of a fused label result that is similar to that of an expert rater. While the excellent outcome of the fusion despite high variability in raters may not occur in all cases, this example serves to show that combining raters can produce a more robust and often more accurate result than the inexpert raters alone.

Cerebellum delineation/parcellation protocol

Our protocol took a hierarchical approach to delineation: early steps focus on large, clear boundaries, and subsequent levels of hierarchy sub-parcellate previously defined regions. This scheme is easily adapted to a fast, abridged protocol, in which coarse structures are delineated but the finest structures are not. The choice of the hierarchical level at which to stop involves a tradeoff between the detail of a fine parcellation, and the reliability of a coarse parcellation (larger structures are able to be delineated more consistently). The NIH software, "Medical Image Processing, Analysis, and Visualization" (MIPAV) (McAuliffe et al., 2001) was used throughout the delineation process. The entire protocol, complete with snapshots of each stage is found online at http://iacl.ece.jhu.edu/Cerebellum_Protocol. The description we provide here is intended to convey a general sense of the strategies we employ.

Image acquisition, processing, and interaction

Acquisition protocol

For each subject, two magnetization prepared rapid gradient echo (MP-RAGE) images were obtained using a 3.0 T MR scanner (Intera, Phillips Medical Systems, Netherlands). The first MP-RAGE was acquired with the following parameters: 132 slices, axial orientation, 1.1 mm slice thickness, 8° flip angle, TE = 3.9 ms, TR = 8.43 ms, FOV 21.2 × 21.2 cm, matrix 256 × 256. The second was acquired with the following parameters: 162 slices, axial orientation, 0.9 mm slice thickness, 8° flip angle, TE = 3.9 ms, TR = 8.35 ms, FOV 23 × 18.328 cm, matrix 256 × 256. The two acquired images were co-registered and averaged to produce a volume with 0.8 mm isotropic voxels in order to improve SNR. If a different image acquisition protocol is used, the delineation process can still be carried out, but specific results may vary.

Registration

The registration was performed by using MIPAV's (McAuliffe et al., 2001) Optimized Automatic Registration tool, which is based on (Jenkinson and Smith, 2001) using a rigid (6 degrees of freedom) transformation. The average image was spatially normalized to a stereotactic space by registration to the ICBM atlas (Mazziotta et al., 1995).

Histogram adjustment

In order to increase contrast among the tissue types of interest, raters were instructed to adjust the window and level to optimize apparent contrast between CSF, gray matter and white matter. Low intensity background, and high intensity regions such as blood vessels can reduce contrast between the cerebellar features when the contrast window covers the min and max intensities. Raters adjusted the window based on the image histogram only and did not change

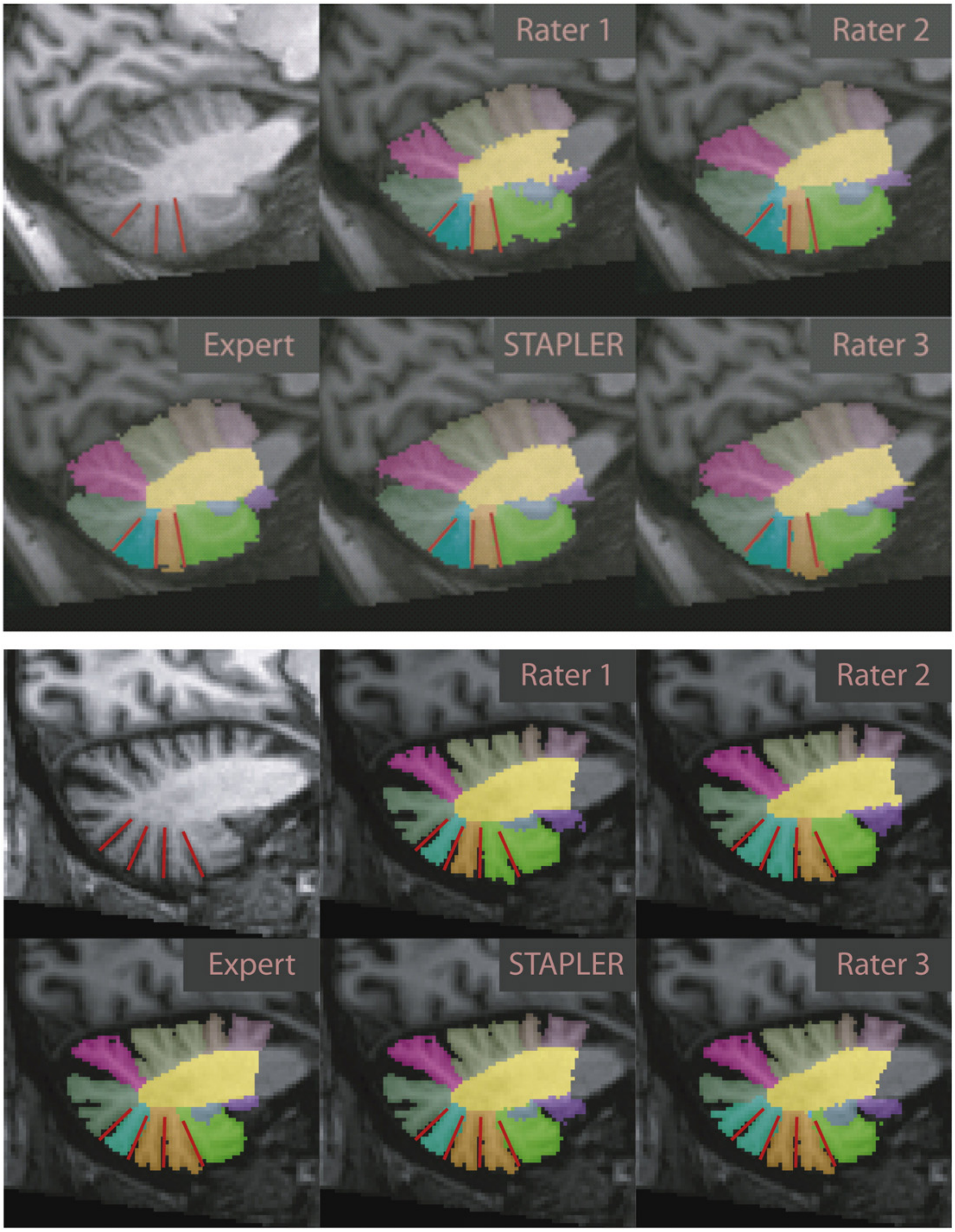

Fig. 2. An illustration of the lobules VIIB through VIIIB being consistently (above) and inconsistently delineated (below). Fissures are indicated with red lines. The anatomical differences between the two subjects were a significant source of error. In particular, the lower subject had an "extra" fissure that caused confusion among the inexpert raters.

it during the delineation process and thereby avoided a potential influence on their boundary choices (a concern described in (Steenbakkers et al., 2005)).

Locking labels

After delineating the whole cerebellum (see below) it is further subdivided in subsequent steps. Therefore, users are instructed to "lock" labels that are in the background (complement) to the object that they are currently subdividing. This simple step greatly expedites the delineation process because 1) it prevents boundaries that have already been established from changing, 2) it prevents accidental mouse clicks from creating isolated, random labels from occurring in remote regions, and 3) it permits larger brushes to be used when delineating near existing boundaries.

Coarse delineation

Fig. 3 outlines the levels of the hierarchy our raters traverse as part of the delineation protocol, the details of which are described below.

Whole cerebellum

For each scan a threshold is determined and employed when delineating the whole cerebellum. This threshold is determined by examining the intensity of the scan and selecting a threshold that includes the gray

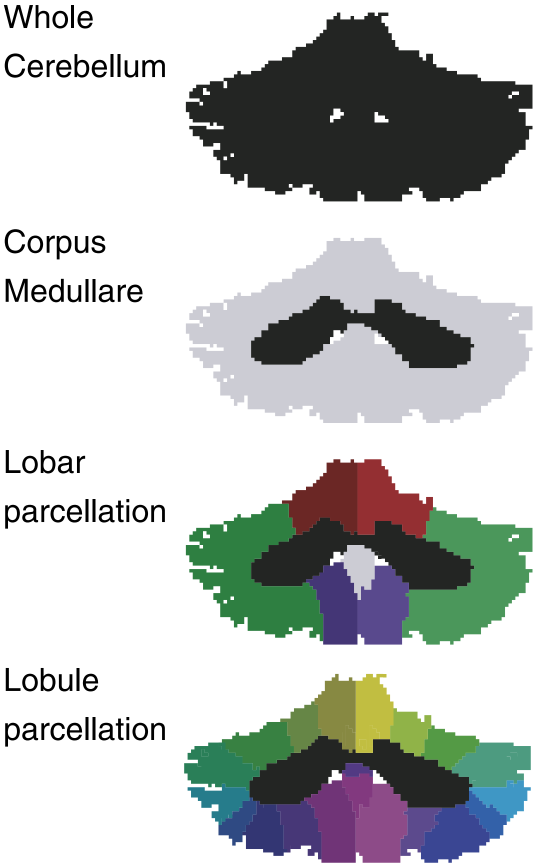

Fig. 3. A schematic showing the steps of the hierarchical delineation protocol.

matter but not the cerebrospinal fluid (CSF). In our experience, this semi-automated thresholding allows raters to more consistently and efficiently separate cerebellar gray matter from the surrounding CSF.

The tentorium cerebelli (the thin layer of dura mater separating the occipital lobe and cerebellum) is not included in the delineation, nor is any other portion of the dura. The pattern and texture of white matter and gray matter in the cerebellum are used to delineate the boundary between the cerebellum and cerebrum. A layer of CSF between the cerebellar and cortical gray matter is often visible. The boundary between the white matter of the cerebellum and the spinal cord is defined as the line in the sagittal plane connecting the most anterior of the dorsal and ventral gray matter in the cerebellum.

Throughout the volume, the sagittal, coronal, and axial views are routinely consulted in order to ensure the highest level of certainty in the delineations and to produce smooth parcellations. In addition, all boundaries are examined and corrected in each of the cardinal planes to remove extraneously labeled voxels.

Corpus medullare

The corpus medullare (CM) is defined as the white matter on the interior of the cerebellum that contains the deep cerebellar nuclei and connects through the cerebellar peduncles to the brain stem. It is the portion of white matter that cannot reasonably be assigned to any particular lobule. Delineation begins with a sagittal view near the midline and proceeds laterally. The label for the CM does not extend into the white matter branches in the lobules but rather connects the bottoms of the deepest fissures. It includes all the white matter within these limits and extends to the anterior boundary of the cerebellum and brain stem. This volume is then adjusted in the sagittal, coronal, and axial planes until it appears smooth in each.

Lobar delineation

The second level of the hierarchy parcellates the cerebellar gray matter into three large structurally significant components: the anterior lobules (I–V), the middle (superior–posterior) lobules (VI–VII), and caudal (inferior–posterior and flocculonodular) lobules (VIII–X). These are chosen because the fissures separating them are prominent and reliably delineated. These divisions also correspond roughly to important functional specialization in the cerebellum, as the anterior lobules have been identified as contributing to sensorimotor function (Ito, 1984; Nitschke et al., 1996), the middle lobules seem to be involved in cognition (Middleton and Strick, 1994; Schmahmann, 1991), while the caudal lobules (especially IX and X) are essential for ocular motor tasks (Ito, 1984).

Primary and prebiventer fissures

At the midline of the sagittal plane, the primary (dorsal) and prebiventer (ventral) fissures are readily identifiable. Raters define these fissures by referencing examples in our protocol and the Schmahmann et al. (1999, 2000) atlas. The primary fissure divides lobules V and VI while the prebiventer fissure separates lobules VIIB and VIIIA, creating a total of four regions: the anterior lobules (I–V), the middle lobules (VI–VII), the caudal lobules (VIII–X), and the corpus medullare. The remainder of the cerebellum is delineated by following these anatomical boundaries through the hemispheres. These divisions are also examined and refined in the axial and coronal planes.

Vermis

Vermal lobules VIIIA through X are delineated as a group. The vermis is defined as a roughly symmetrical shape in the center of the cerebellum and is best identifiable in the coronal plane. It is clearest in the most posterior and most anterior planes. In the posterior planes of the coronal view it appears as a circular shape between the two hemispheres inferior to the CM. In the anterior planes it is more triangular and appears to be "wedged" between the two hemispheres. There is typically uncertainty near the midline of the volume which can be clarified using the sagittal plane to connect the posterior and anterior regions delineated in the coronal view.

Lobule delineation

The finest level of the hierarchy is the delineation of lobules. We define all lobule boundaries using the Schmahmann et al. (2000) atlas. All unchanging paint/labels remain locked and efforts are made to ensure continuity of labels between hemispheres when possible. Raters typically proceed "clockwise" (rostral-to-caudal) in the sagittal view in order to best take advantage of the label "locking" capability and to improve their efficiency.

Lobules I/II

This is the smallest lobule and is visible only in several slices at the midline of the sagittal plane. It is most readily identified in the sagittal plane and appears as a thin curve of grey matter bordering the superior cerebellar peduncle.

Lobule III

This lobule is adjacent to I/II and usually consists of only one major white matter branch. In the axial view it terminates near the boundaries of the spinal cord. It is typically shorter and smaller than both lobules IV and V.

Lobule IV

This lobule is adjacent to Lobule III and is best recognized by locating the next fissure (when proceeding clockwise from the fissure defining the boundary between III and IV) that remains uninterrupted throughout the entirety of the hemisphere. This boundary is somewhat unreliable because such a boundary between the anterior edge and the primary fissure is not present in every subject. In such cases the most prominent fissure in the region to be sub-parcellated is chosen.

Lobule V

This lobule is trivially delineated once the previous four labels have been applied as it lies between lobule IV and the primary fissure. It is usually similar in size to lobule IV and is best labeled in the sagittal view.

Lobule VI

This lobule is bounded anteriorly by the primary fissure and posteriorly by Crus I through the majority of the hemisphere. Near the midline, Crus I is diminished, and lobule VI may directly abut Crus II. This boundary is most readily identified roughly 15 mm from the midline in the sagittal view. Lateral to the midline, Crus I is identified as the growing branch and can be used to identify the boundary of lobule VI. We begin labeling the hemisphere of the cerebellum and proceed laterally. Next, we return to our starting point (15 mm from the midline) and proceed toward the midline, creating a boundary following the anatomy.

VIIA-Crus I

This lobule is not present in the midline sagittal plane and is best identified by examining concurrent sagittal slices between the midline and a slice roughly 15 mm lateral to the midline. Crus I is a large lobule with only one branch, which appears several slices from the midline and continues to grow, remaining present for the remainder of the hemisphere.

VIIA-Crus II

This lobule is adjacent to VI and Crus I. We define the second boundary as the most prominent fissure between Crus I and the prebiventer fissure. This fissure must be present through the entirety of the hemisphere. However, occasionally there exist two candidate fissures, neither of which spans the hemisphere. In this case the fissure spanning the greater portion of the hemisphere is selected.

Lobule VIIB

This lobule is trivially delineated as the remaining superior posterior lobe after the previous three lobules are identified.

Lobule VIIIA

This lobule is adjacent to lobule VIIB and is distinguishable from lobule VIIIB because it does not abruptly curve in the axial view. It sometimes consists of two smaller branches that meet the point of the vermis (see the axial view). It is best recognized in the axial view and delineated in the sagittal view.

Lobule VIIIB

This lobule usually begins as one white matter branch at the center of the cerebellum and becomes two branches as it extends laterally; this behavior is best seen in the axial view. In the sagittal view it appears to hook or curve around lobule IX, and it also borders lobule X. This boundary is a smaller fissure and can be distinguished by the difference in orientation between lobules VIIIB and X. It is best delineated in the sagittal and axial views.

Lobule IX

Lobule IX appears as a round structure at the anterior portion of the cerebellum in the axial view. It is present only near the midline of the cerebellum. In the coronal view it is the most medial lobule and appears to extend horizontally towards the center of the cerebellum. It is easiest to delineate in this view as the boundary is a straight line. It borders primarily the vermis, lobule VIIIB, lobule IX of the other hemisphere, and the spinal cord.

Lobule X

Lobule X is the anterior most lobule of the cerebellum. It is present primarily half way in between the center and the outer edge. It is smaller than most other lobes and is near the spinal cord. It appears as a bump in the sagittal view, not prominent enough to be part of lobule VIIIB.

Vermis VIIIA

This is the small posterior most part of the vermis and seems to take only half of the posterior most white matter trunk. It is the lobule that is present in the most lateral slices of the vermis and appears to meet vermis VIIIA in the axial plane. It is best delineated in the sagittal plane.

Vermis VIIIB

This region consists of the other half of the posterior most white matter branch and is present, laterally, nearly as long as the vermis of VIIIA. It is also best delineated in the sagittal plane.

Vermis IX

This lobule is a triangular shape and is bordered on one side by lobule VIIIB and on the other by X. It is the largest lobe in the sagittal view of the midline. The border between IX and X is a fissure that reaches the CM and can be determined by considering the difference in the branches on either side. It is best delineated in the sagittal plane.

Vermis X

Lobule X is the anterior most region of the vermis and contains no fissures. It is usually slightly wider (laterally) than lobule IX. The fissure separating vermis lobules IX and X is often wider in the more lateral sagittal slices than in the medial. These lobules are sometimes separated by lobule IX of the hemisphere. Lobule X of the vermis is disjoint from lobule X of the hemisphere.

Rapid rater review

The consistency of the rater delineations for each subject was established by manual examination using a method of "rapid rater review". An automated routine located all delineations of a single subject completed by any rater and produced a series of mosaic images. Each image displays, side-by-side, a slice of a single subject from each rater's delineation (including repeat delineations by a single rater) as shown in Fig. 4. Because errors tended to be large, representative slices were sufficient to identify regions of disagreement. Note that this type of verification depends on the fact that several human raters delineate each subject. Specifically, this allows "rapid-reviewers" to assess delineations for similarity to each other, rather than for accuracy. In this way, rapid-reviewers could manually identify raters who were visibly inconsistent with either themselves or the other raters. Though the process is unable to correct some inconsistent delineations found, it does flag scans for redelineation or exclusion from further analysis. This was done at the last stage of the hierarchy because most errors were found there. When a particular label is found to be in inconsistent, only that label is excluded while any other valid regions produced by that rater for that subject remain.

Rapid reviewers also classified error into three categories: "sloppiness," "disagreement," and "misattribution." An inconsistency was labeled "sloppy," when a raters delineation failed to follow either anatomical or image based cues. Examples of sloppiness included: isolated pixels assigned to a label, unsmooth label boundaries, or label boundaries that did not follow either tissue or fissure boundaries (e.g., bias caused by not properly following the correct intensity value). Errors were assigned to "disagreement," when appropriate cues were followed, but a rater chose a different feature to follow relative to the others (Fig. 2 shows a typical example of this). Finally, a "misattribution" error occurred when a rater correctly delineated boundaries, but assigned an

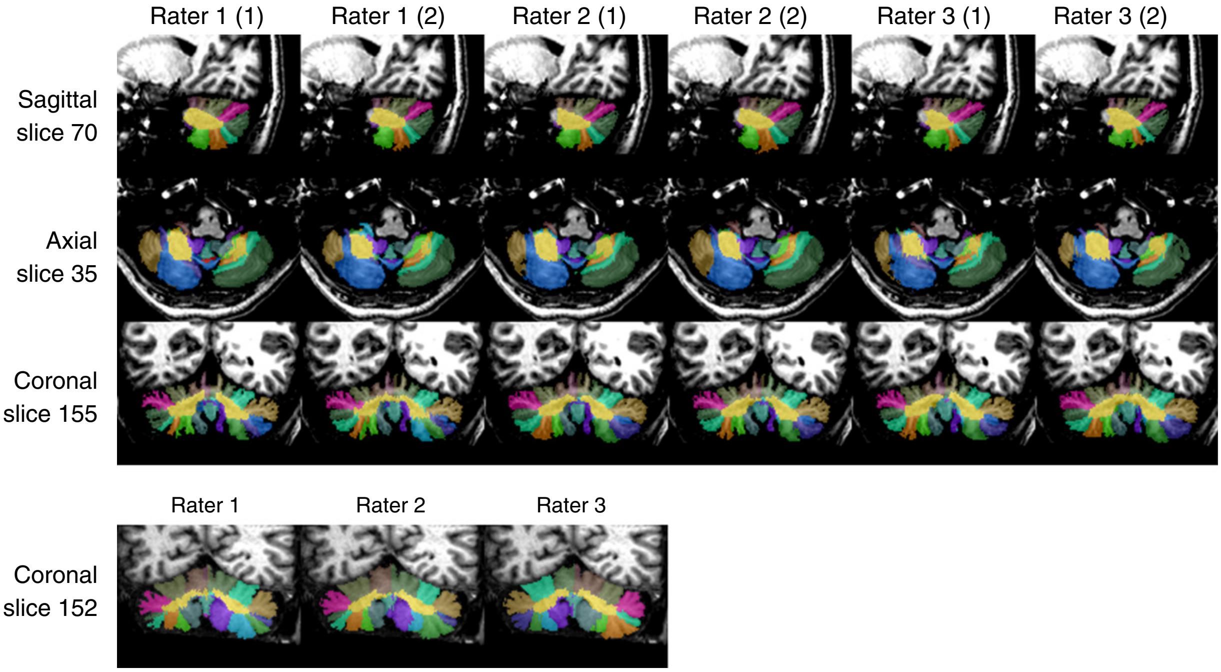

Fig. 4. Example of an automatically generated montage of rater delineations used for rapid rater review. The first example shows examples of axial, coronal, and sagittal slices for one subject delineated twice by three raters. All raters were consistent with themselves and with each other. The second example shows a "misattribution" error, in which Rater 3 delineated the boundaries correctly but switched the labels for the left and right.

incorrect label. For example, assigning the lobule Crus1-right label to Crus1-left was considered a misattribution error. An example is shown in Fig. 4, where Rater 3 exchanged the right and left labels. Misattribution errors were corrected by reassigning the correct labels to the rater delineations.

Rater fusion

Statistical label fusion methods can produce more accurate delineations by combining the labels produced by multiple raters. We report results using the STAPLER algorithm (Landman et al., 2012b), a robust extension of "Simultaneous Truth and Performance Level Estimation" (STAPLE) (Warfield et al., 2004). STAPLE models the performance of each rater as a confusion matrix, where element (i,j)indicates the probability that the rater assigned object i when the true label was object j. A "consensus" labeling can then be estimated using the probability of every label for each voxel given the performances of all raters. The estimation gives more weight to raters that were deemed more accurate for particular labels. The rater performances are then re-estimated using the consensus label probabilities from all voxels, and the whole process is iterated to convergence. This scheme differs from majority vote in that raters are weighted unequally according to how well they agree with the consensus, where this agreement is estimated globally. Furthermore, it can resolve situations in which all raters assign different labels by selecting the label assigned by the best (most accurate) rater.

STAPLER follows a similar procedure but includes prior knowledge that regulates the performance estimation in order to avoid unlikely label configurations that result from poor estimates of the rater performance. It can also handle cases where there are irregularities in the numbers of raters or their labels, but this feature was not used here. We performed experiments comparing STAPLER with the STAPLE algorithm and a majority vote fusion scheme, and found that the STAPLER method performed best among these three. Note that while it did not uniformly improve upon the other methods, its performance was superior across all labels and several measures (see the Methods and results section for the evaluation metrics). A discussion of the types and extent of errors that can be corrected by statistical fusion can be found in the Discussion section.

We used the publicly available implementation of the multi-compartment STAPLER algorithm (http://www.nitrc.org/projects/masi-fusion), initialized with the empirical label probability map. Convergence was declared when the normalized trace of the confusion matrix changed less than 10^{-4} between iterations. The value of the rater performance bias parameter was set to 0.5I for all raters, where I is the identity matrix.

Methods and results

Three human raters completed the protocol on 48 subjects. These raters were undergraduate students at Johns Hopkins University, Baltimore, MD. Each student underwent approximately 150 h of training in the protocol before their results were considered valid. The training consisted of a study of the protocol document and delineation of training subjects followed by expert feedback and subsequent correction by the student. The cohort consisted of 18 healthy control subjects, six subjects diagnosed with spinocerebellar ataxia (SCA) type 6, 22 subjects with a non-genetic cerebellar ataxia, and two subjects with other diagnoses. One expert rater (with approximately 5000 h of experience) delineated 15 subjects from the cohort, and was considered the "gold standard." This expert has been shown to have exceptional reliability (inter-rater ICC with another expert of 0.991) based on delineation of 23 regions for 22 subjects (Jung et al., 2011). The amount of time required for the finest parcellation of a single subject is roughly equivalent for multiple inexpert raters and one expert rater: three raters each taking about 15 h per subject, versus the expert rater taking approximately 50 h per subject.

The inexpert raters underwent rapid review by two independent human judges. They found that approximately 4% of the regions produced by the three raters we used in this study had an error of some sort detected (216 errors out of 5376 rater-regions) by this process. These included both "sloppiness" and "disagreement" errors. Over half (113) of these errors came from lobules VIIB through VIIIA. Examples of a "disagreement" and "misattribution" errors discovered by this review process are shown in Figs. 2 and 3, respectively. See the Rapid rater review section for definitions of these error types.

Label fusion using STAPLER was run on the delineations provided by three inexpert raters. Since our raters repeated the delineations of some subjects there were often many unique groups of three ratings that could be used to produce a label fusion result. We therefore fused all groups of three raters available to us and were able to provide an indication of the variability in a label fusion result given inexpert raters.

The reliability of the volume measurements obtained from the delineations were measured using the intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979; McGraw and Wong, 1996). We used the icc function in the irr package for the R software to compute the ICC. We selected the "one way" option because the lobule volumes for each subject are considered random effects, but not the raters. All other options were left to their default values. The mathematical details of this ICC can be found in Table 4 in McGraw and Wong (1996), under the designation "ICC(1)". This type of ICC measures the absolute agreement between different raters' volumes, as we are interested in determining whether the volumes produced by different raters were identical (rather than consistent). These and all measures of delineation quality were computed at all levels of the label hierarchy. Note that the reliability measures tend to be greater at coarse parcellations and lesser for finer parcellations. Specific explanations for this behavior are given in the Discussion section.

Table 1 shows computed intra-rater ICC values for a coarse grouping of labels that was described in Makris et al. (2005), and for the full lobule parcellation. The 95% confidence interval for the estimate is given in parentheses. For the coarse grouping, the inexpert raters achieve good intra-rater ICCs for the corpus medullare and anterior and posterior lobes (estimates range from 0.62 to 0.99). By this measure, STAPLER performed well on the flocculonodular lobe (intra-rater ICC range from 0.89 to 0.94), despite the fact that inexpert raters were less reliable (−0.67 to 0.82). There is a similar trend for the lobule reliabilities; the STAPLER results tend to be more reliable than individual raters. Intra-rater ICCs for the inexpert raters ranged from 0.26 to 0.96 for rater one, −0.07 to 0.87 for rater two, and −0.66 to 0.98 for rater three. The STAPLER fused results had ICCs ranging from 0.71 to 0.99, and indicate that the fused results are much more reliable than the individual raters.

Table 2 shows the inter-rater ICCs for the lobar and lobule parcellation of the cerebellum. This measure was always computed relative to the expert rater. For the corpus medullare and anterior and posterior lobes, inter-rater ICCs ranged from 0.49 to 0.98 for the inexpert raters, and from 0.94 to 0.99 for STAPLER. For the remaining lobules, ICCs ranged from −0.04 to 0.98 for the inexpert raters and from −0.19 to 0.99 for STAPLER.

In contrast with the ICC, the Dice similarity coefficient (DSC) measures the degree of spatial overlap of two delineations and for this reason is a stricter measure of similarity. Fig. 5 shows the boxplots of the DSC of the inexpert raters and the fusion results with the expert rater. Note that the DSC will give a result equal to sensitivity and positive predictive value when the volumes of expert and rater are equal. The fusion results tend to have higher similarity coefficients than the inexpert raters, as well as more compact distributions.

We also computed the signed volume difference given by: SVD $= (V_R - V_E)/V_E$, where V_R and V_E are the volumes of a region delineated by the rater and expert, respectively. This quantity measures the fractional difference of the rater's volume in comparison to the expert. A delineator that agrees with the expert would achieve an SVD of zero, while raters that consistently overestimate or underestimate the volumes of certain regions would have positive and negative SVDs, respectively. We report this measure to show that the volumes of the structures found are similar in absolute terms in addition to being consistent (as measured by the ICC). Fig. 6 shows the boxplots of the SVD score for all the inexpert raters and the fused results. In general, the results of label fusion have compact distributions around zero, whereas the inexpert raters' volume measurements are more spread out. This indicates that fusion produces volume measurements similar to the expert more often than the raw inexpert raters. We further report in Table 3 the nominal volumes (mean and standard deviation) for normal subjects of the structures described here obtained by expert rater delineation and STAPLER fused inexpert delineation.

Discussion

This is the first study to demonstrate that relatively inexperienced human raters, when given a hierarchical protocol, can produce delineations close to those of an expert rater. As inexpert raters are easier to train and less expensive to employ than expert raters, our methodology could facilitate the accurate labeling of very large data sets in a practical fashion. Given the approximate training and delineation times we have compiled during this study, we can report a rough estimate of the time and cost savings achievable using this methodology. Training time for the inexperts was reduced by about 75% because the three raters could be trained in parallel and because their training was faster, consisting simply of learning the protocol. This suggests that having this hierarchical protocol enables raters to be quickly mobilized for a particular study. After training, the three inexperts can complete a subject in about 70% less time than the expert, primarily because they are able to work in parallel. Furthermore, we observed cost savings of 95% for training and 60% per subject delineation after taking into account the amount of time spent on training and delineation and typical hourly wages for researchers in metropolitan areas and for undergraduate interns. Of course, it is important that this methodology produce results as close as possible to those of the expert.

Agreement with an expert rater can be achieved by employing a three-stage approach. A hierarchical delineation protocol reduces rater variability by limiting the number of decisions a rater needs to make at any given moment. Gross errors are isolated using rapid review and verification of consensus. Raters tend to make "sloppiness" errors independently (in a statistical sense), and so often fit the error model of statistical fusion algorithms. This explains the improvement of accuracy and reliability gained from using the fused results.

This leads to an important discussion regarding the types of rater errors that can be discovered and corrected using our methodology. As our methodology does not employ expert raters, we rely on the "consensus" of the inexperts to determine locations of possible errors. As a result, boundaries that have been delineated incorrectly (relative to the truth) but identically (i.e., a consensus was reached) by all raters are not detectable as errors. Nevertheless, the quality of our results suggests that this situation does not occur very often, and when a rater makes an error, there is usually disagreement regarding the boundary's location.

We present data defining the inherent tradeoff between granularity and accuracy. Understanding this choice could facilitate the statistical design of hypothesis-driven cerebellar parcellation protocols in the future. Our results show that the overlap and volume repeatability measures degrade as raters complete finer stages of the labeling hierarchy (see Figs. 5 and 6 and Tables 1 and 2). This can be caused by reduced or lack of contrast between small structures or inter-subject differences in anatomy. This in turn results in differences in rater judgment in those regions, and denoted a "disagreement" error by rapid reviewer. In particular, the most prominent fissures give rise to significant image contrast, are stable from subject to subject, and therefore are reliably delineated. On the other hand, the boundaries between some lobules are small fissures between the corresponding folia that produce only a small amount of image contrast. As was noted in the Cerebellum delineation/parcellation protocol section, when fissures are not visible in the image, their presence had to be inferred from the positions of

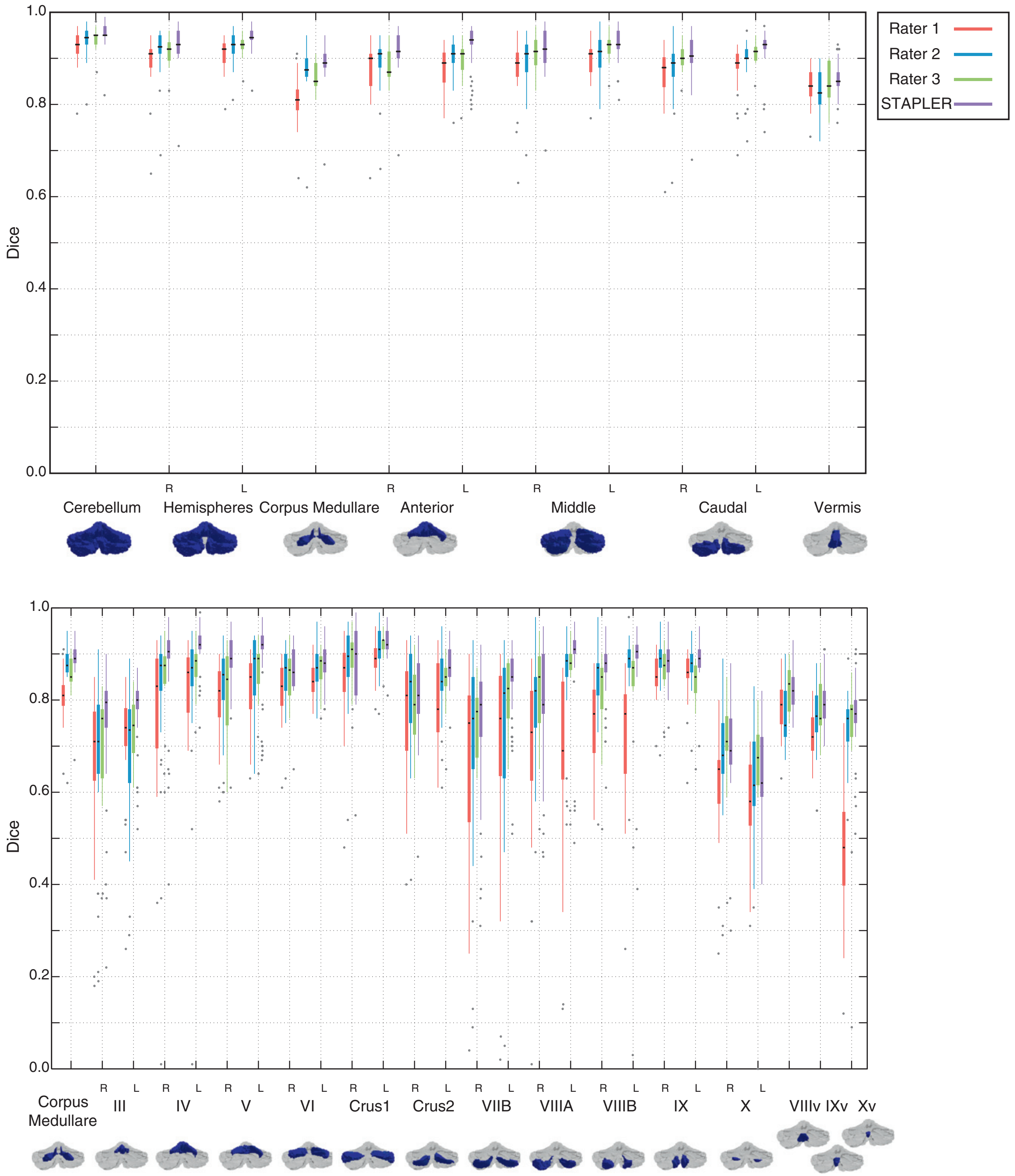

Fig. 5. Box plots of Dice similarity coefficient with the expert rater for the human raters and label fusion by STAPLER. The upper and lower plots show overlaps for the lobar parcellation and lobule parcellations of Fig. 1, respectively.

white matter branches. This judgment is largely dependent on experience and anatomical knowledge.

Other situations in this labeling task arise in which image contrast gives little or no information. It is not uncommon for two lobules to sprout from a single white matter branch emanating from the corpus medullare. Since white matter branches were included as part of the lobule volume, the shared white matter branches were split between the two lobules approximately equally. A related challenge is the appearance of marked fissures between different parts of the same lobule (lobules often consist of multiple folia). Consequently, raters must distinguish between fissures or other image features corresponding to lobule boundaries and fissures that separate different parts of a single lobule. To make matters worse, the size and depth of the fissure may or may not be a reliable feature in determining which fissures separate

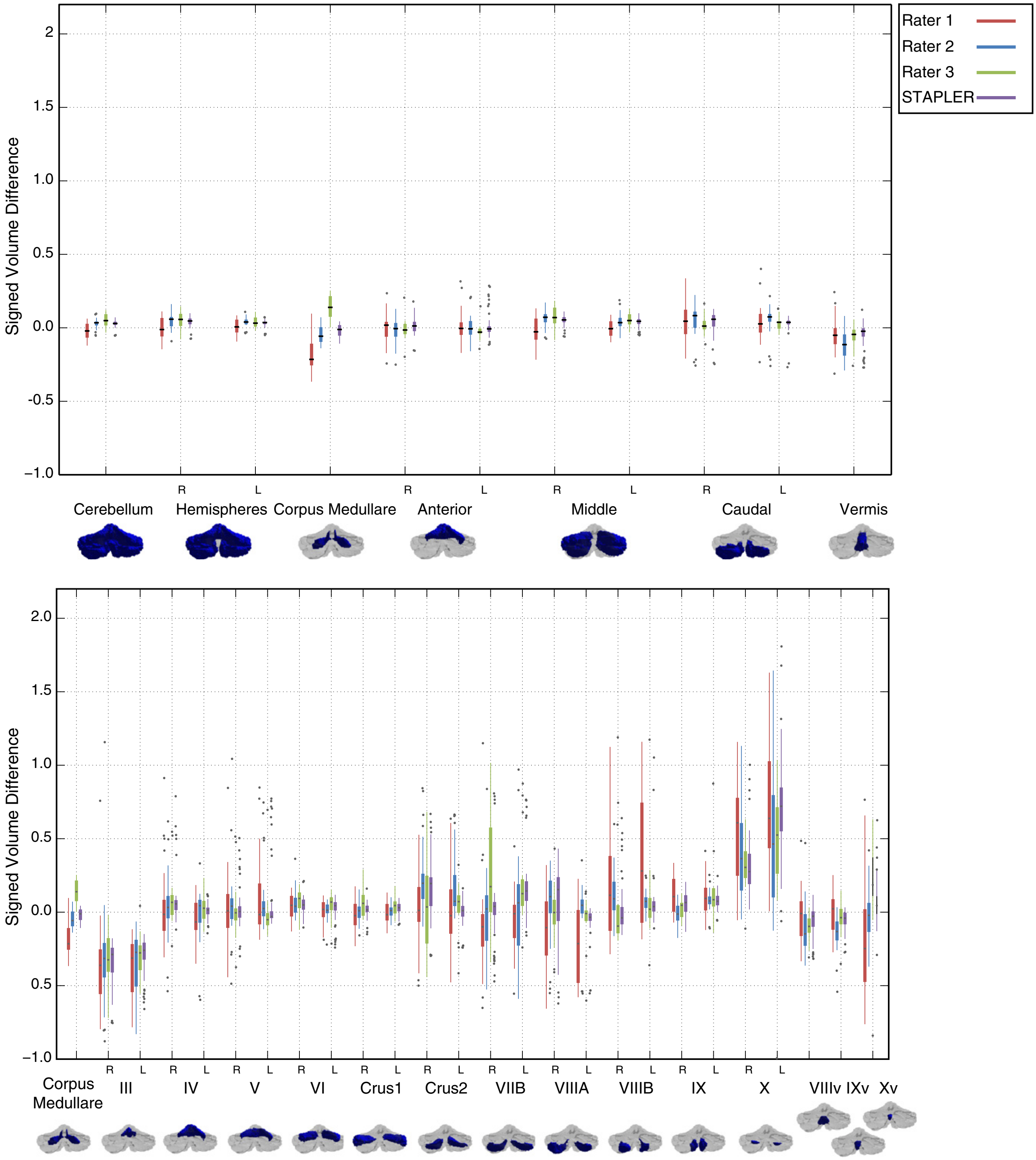

Fig. 6. Box plots of signed volume difference with the expert rater for the human raters and label fusion by STAPLER. The upper and lower plots show comparisons for the lobar parcellation and lobule parcellations of Fig. 1, respectively.

lobules. (i.e., fissures within a single lobule can appear larger/deeper than fissures separating two lobules). As a result, raters occasionally choose to follow *different image features* in delineating the same lobule boundary. This phenomenon explains some large discrepancies between raters that we observed, and the surprisingly wide confidence intervals of ICC for some lobules. Future work in improving the rapid review process could help alleviate the effect of these outliers.

We now will present several examples of anatomical features of the cerebellum that confound accurate parcellation. Fig. 2 shows an example of this phenomenon focusing on the labeling of lobules Crus II, VIIB, VIIIA, and VIIIB. Notice that the first subject has three fissures separating four regions, and the second subject has four fissures separating five regions. All inexpert raters agree with the expert rater on the first subject since the image features indicate the anatomical divisions. The labeling task for the second subject is much more difficult, since raters needed to determine which of the image features correspond to anatomical divisions. In this case, the expert determined that VIIIA (pictured in orange) contained two branches. All of the inexpert raters correctly

Table 1
Intra-rater ICCs for coarse (lobe) and fine (lobules) parcellations (see Fig. 1).

Lobule		Intra-rater ICC			
		Rater 1	Rater 2	Rater 3	STAPLER
CM		0.62 (0.33, 0.81)	0.79 (0.60, 0.89)	0.93 (0.65, 0.99)	0.99 (0.96, 1.00)
Anterior	R + L	0.95 (0.90, 0.98)	0.79 (0.61, 0.90)	0.99 (0.93, 1.00)	1.00 (0.99, 1.00)
	L	0.95 (0.89, 0.97)	0.84 (0.70, 0.92)	0.94 (0.69, 0.99)	0.99 (0.97, 1.00)
	R	0.95 (0.89, 0.98)	0.75 (0.53, 0.87)	0.89 (0.49, 0.98)	0.99 (0.96, 1.00)
Posterior	R + L	0.84 (0.69, 0.92)	0.71 (0.48, 0.85)	0.86 (0.37, 0.98)	0.97 (0.92, 0.99)
	L	0.90 (0.80, 0.95)	0.74 (0.52, 0.87)	0.88 (0.45, 0.98)	0.99 (0.96, 1.00)
	R	0.77 (0.55, 0.89)	0.69 (0.44, 0.84)	0.76 (0.11, 0.96)	0.95 (0.86, 0.98)
I–III	L	0.72 (0.47, 0.86)	0.67 (0.41, 0.82)	0.97 (0.85, 1.00)	0.98 (0.95, 0.99)
	R	0.61 (0.31, 0.80)	0.41 (0.07, 0.67)	0.98 (0.89, 1.00)	0.97 (0.92, 0.99)
IV	L	0.76 (0.54, 0.88)	0.79 (0.60, 0.89)	0.98 (0.90, 1.00)	0.98 (0.93, 0.99)
	R	0.74 (0.51, 0.87)	0.66 (0.40, 0.82)	0.73 (0.03, 0.96)	0.97 (0.92, 0.99)
V	L	0.77 (0.56, 0.89)	0.86 (0.72, 0.93)	0.81 (0.24, 0.97)	0.95 (0.85, 0.98)
	R	0.72 (0.48, 0.86)	0.86 (0.73, 0.93)	0.93 (0.63, 0.99)	0.98 (0.93, 0.99)
VI	L	0.96 (0.92, 0.98)	0.82 (0.66, 0.91)	0.92 (0.62, 0.99)	0.99 (0.97, 1.00)
	R	0.95 (0.90, 0.98)	0.73 (0.50, 0.86)	0.89 (0.47, 0.98)	0.97 (0.91, 0.99)
VIIA	L	0.88 (0.76, 0.94)	0.76 (0.56, 0.88)	0.91 (0.54, 0.99)	0.99 (0.98, 1.00)
Crus1	R	0.66 (0.39, 0.83)	0.70 (0.46, 0.84)	0.62 (−0.17, 0.94)	0.98 (0.93, 0.99)
VIIA	L	0.46 (0.11, 0.71)	0.67 (0.41, 0.83)	0.30 (−0.53, 0.86)	0.80 (0.51, 0.93)
Crus2	R	0.18 (−0.21, 0.52)	0.63 (0.36, 0.80)	0.89 (0.47, 0.98)	0.89 (0.72, 0.96)
VIIB	L	0.28 (−0.10, 0.59)	0.53 (0.22, 0.75)	0.34 (−0.49, 0.87)	0.97 (0.90, 0.99)
	R	0.38 (0.01, 0.66)	0.54 (0.23, 0.75)	0.89 (0.47, 0.98)	0.86 (0.65, 0.95)
VIIIA	L	0.57 (0.25, 0.78)	0.44 (0.11, 0.69)	0.79 (0.18, 0.97)	0.72 (0.35, 0.89)
	R	0.36 (−0.01, 0.65)	0.51 (0.19, 0.73)	0.89 (0.47, 0.98)	0.89 (0.70, 0.96)
VIIIB	L	0.62 (0.33, 0.81)	0.56 (0.25, 0.76)	0.58 (−0.22, 0.93)	0.85 (0.62, 0.95)
	R	0.63 (0.35, 0.81)	0.73 (0.52, 0.86)	0.95 (0.76, 0.99)	0.97 (0.93, 0.99)
VIII	v	0.87 (0.73, 0.94)	0.64 (0.38, 0.81)	0.83 (0.30, 0.97)	0.98 (0.95, 0.99)
IX	L	0.78 (0.57, 0.89)	0.87 (0.74, 0.93)	0.87 (0.42, 0.98)	0.96 (0.89, 0.99)
	R	0.76 (0.55, 0.88)	0.68 (0.43, 0.83)	0.49 (−0.34, 0.91)	0.97 (0.92, 0.99)
	v	0.56 (0.24, 0.77)	0.42 (0.08, 0.67)	0.71 (−0.01, 0.95)	0.93 (0.80, 0.97)
X	R + L	0.74 (0.51, 0.87)	0.52 (0.20, 0.74)	−0.33 (−0.84, 0.56)	0.91 (0.77, 0.97)
	L	0.63 (0.34, 0.81)	0.52 (0.21, 0.74)	0.25 (−0.57, 0.84)	0.89 (0.72, 0.96)
	R	0.73 (0.50, 0.87)	0.50 (0.18, 0.72)	−0.67 (−0.94, 0.17)	0.94 (0.83, 0.98)
	v	0.28 (−0.10, 0.59)	0.48 (0.16, 0.72)	0.82 (0.26, 0.97)	0.85 (0.62, 0.95)

Table 2
Inter-rater ICCs for coarse (lobe) and fine (lobules) parcellations (see Fig. 1). This measure compares a particular human rater or fusion method to the "gold standard" expert rater.

Lobule		Inter-rater ICC with ground truth			
		Rater 1	Rater 2	Rater 3	STAPLER
CM		0.49 (−0.01, 0.80)	0.92 (0.79, 0.97)	0.76 (0.38, 0.92)	0.94 (0.81, 0.99)
Anterior	R + L	0.98 (0.96, 0.99)	0.78 (0.59, 0.89)	0.79 (0.62, 0.89)	0.96 (0.85, 0.99)
	L	0.98 (0.96, 0.99)	0.76 (0.55, 0.88)	0.80 (0.63, 0.89)	0.96 (0.86, 0.99)
	R	0.95 (0.90, 0.97)	0.79 (0.61, 0.89)	0.74 (0.54, 0.86)	0.95 (0.83, 0.99)
Posterior	R + L	0.93 (0.88, 0.96)	0.56 (0.26, 0.76)	0.63 (0.37, 0.80)	0.99 (0.95, 1.00)
	L	0.96 (0.92, 0.98)	0.57 (0.27, 0.77)	0.64 (0.39, 0.81)	0.99 (0.96, 1.00)
	R	0.90 (0.83, 0.94)	0.55 (0.24, 0.76)	0.60 (0.34, 0.78)	0.98 (0.92, 0.99)
I–III	L	0.75 (0.40, 0.91)	0.45 (−0.04, 0.77)	0.67 (0.20, 0.89)	0.56 (−0.03, 0.87)
	R	0.47 (−0.04, 0.79)	0.21 (−0.31, 0.64)	0.21 (−0.37, 0.68)	0.27 (−0.37, 0.75)
IV	L	0.85 (0.61, 0.95)	0.82 (0.55, 0.93)	0.88 (0.64, 0.96)	0.96 (0.85, 0.99)
	R	0.87 (0.66, 0.96)	0.91 (0.76, 0.97)	0.84 (0.56, 0.95)	0.96 (0.84, 0.99)
V	L	0.78 (0.45, 0.92)	0.77 (0.46, 0.92)	0.67 (0.21, 0.89)	0.86 (0.57, 0.96)
	R	0.86 (0.63, 0.95)	0.84 (0.60, 0.94)	0.73 (0.32, 0.91)	0.83 (0.48, 0.95)
VI	L	0.96 (0.88, 0.99)	0.99 (0.96, 1.00)	0.94 (0.81, 0.98)	0.97 (0.89, 0.99)
	R	0.96 (0.89, 0.99)	0.97 (0.92, 0.99)	0.91 (0.73, 0.97)	0.95 (0.83, 0.99)
VIIA	L	0.95 (0.85, 0.98)	0.98 (0.96, 0.99)	0.95 (0.84, 0.99)	0.99 (0.97, 1.00)
Crus1	R	0.92 (0.77, 0.97)	0.97 (0.93, 0.99)	0.88 (0.66, 0.96)	0.96 (0.86, 0.99)
VIIA	L	0.90 (0.72, 0.97)	0.68 (0.29, 0.88)	0.75 (0.36, 0.92)	0.85 (0.53, 0.96)
Crus2	R	0.60 (0.14, 0.85)	0.65 (0.23, 0.86)	0.53 (−0.01, 0.84)	0.56 (−0.03, 0.87)
VIIB	L	0.82 (0.54, 0.94)	0.11 (−0.40, 0.57)	0.60 (0.09, 0.86)	0.73 (0.27, 0.93)
	R	0.51 (0.01, 0.81)	0.40 (−0.10, 0.75)	0.04 (−0.51, 0.58)	0.54 (−0.05, 0.86)
VIIIA	L	0.57 (0.10, 0.84)	0.86 (0.63, 0.95)	0.84 (0.54, 0.95)	0.81 (0.44, 0.95)
	R	0.58 (0.11, 0.84)	0.64 (0.22, 0.86)	0.75 (0.37, 0.92)	0.33 (−0.31, 0.78)
VIIIB	L	0.39 (−0.14, 0.75)	0.96 (0.90, 0.99)	0.80 (0.46, 0.94)	0.61 (0.04, 0.88)
	R	0.73 (0.37, 0.91)	0.90 (0.75, 0.97)	0.57 (0.05, 0.85)	0.84 (0.50, 0.96)
VIII	v	0.82 (0.53, 0.94)	0.70 (0.32, 0.89)	0.87 (0.61, 0.96)	0.80 (0.40, 0.94)
IX	L	0.77 (0.45, 0.92)	0.91 (0.76, 0.97)	0.49 (−0.07, 0.82)	0.90 (0.67, 0.97)
	R	0.83 (0.57, 0.94)	0.95 (0.87, 0.98)	0.94 (0.82, 0.98)	0.94 (0.78, 0.98)
	v	0.85 (0.61, 0.95)	0.73 (0.38, 0.90)	0.85 (0.57, 0.95)	0.87 (0.59, 0.97)
X	R + L	0.74 (0.57, 0.85)	0.57 (0.27, 0.77)	0.62 (0.35, 0.79)	−0.15 (−0.68, 0.49)
	L	−0.04 (−0.53, 0.48)	0.52 (0.04, 0.81)	0.46 (−0.10, 0.81)	−0.15 (−0.67, 0.49)
	R	0.02 (−0.48, 0.53)	0.36 (−0.15, 0.73)	0.25 (−0.33, 0.70)	−0.19 (−0.69, 0.46)
	v	0.11 (−0.41, 0.59)	0.54 (0.08, 0.82)	0.46 (−0.11, 0.80)	0.41 (−0.22 0.81)

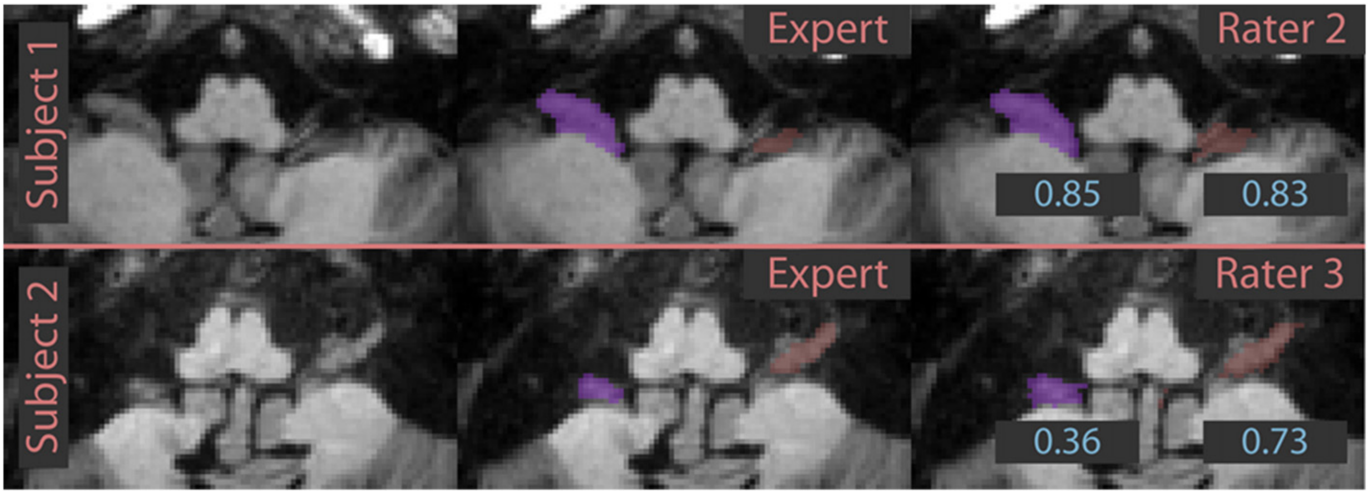

Fig. 7. Examples of lobule X delineations. The inexpert rater delineations are labeled with Dice similarity coefficients with the expert delineation. Note that small absolute errors translate to large relative errors due to the small size of the structure.

identified the fissures, but chose to split the five branches among the four labels in ways that did not agree with the expert. This highlights the importance of anatomical knowledge for certain aspects of this delineation task. We found that a consensus labeling, determined by statistical fusion performs well and compensates, in part, for the relative inexperience of raters.

A different challenge was faced by our inexpert raters when delineating lobule X (the flocculonodular lobe). These rather small regions on the anterior part of the cerebellum are in very close proximity to other nerves and connective tissue. The inexpert raters had difficulty in distinguishing these structures and consistently over-estimated the size of lobule X relative to the expert rater. The details can be observed in Fig. 6; the signed volume difference is positive for lobule X for all inexpert raters and fused results. A specific example of this phenomenon is given in Fig. 7. Despite the qualitative similarity of the inexpert and expert delineations, the DSC (displayed in blue) is small due to the small size of the lobule. Even though their errors were small on an absolute scale, the small size of these structures caused unacceptably large relative errors in volume. We plan to address this impediment in a subsequent revision of our protocol, though this structure may be too

Table 3
Means and standard deviations (in parentheses) for the lobar and lobule parcellations of control subjects.

Lobule		Expert	Rater 1	Rater 2	Rater 3	STAPLER
CM		12.14 (0.57)	10.79 (2.16)	12.44 (1.45)	14.42 (1.36)	12.18 (1.84)
Anterior	R + L	14.54 (1.71)	15.66 (2.97)	15.47 (2.80)	14.25 (2.81)	15.78 (2.46)
	L	6.89 (0.92)	7.52 (1.36)	7.36 (1.58)	6.98 (1.51)	7.59 (1.30)
	R	7.65 (1.27)	8.14 (1.86)	8.11 (1.67)	7.27 (1.80)	8.19 (1.55)
Posterior	R + L	88.93 (7.46)	92.65 (13.63)	97.80 (12.61)	94.67 (10.55)	92.55 (13.62)
	L	46.18 (3.75)	48.17 (6.39)	50.50 (5.88)	48.78 (4.72)	48.20 (6.31)
	R	42.75 (3.73)	44.48 (7.48)	47.54 (6.71)	46.01 (5.75)	44.34 (7.48)
I–III	L	0.77 (0.21)	0.80 (0.27)	0.57 (0.23)	0.66 (0.29)	0.78 (0.26)
	R	0.74 (0.13)	0.91 (0.42)	0.71 (0.40)	0.67 (0.28)	1.13 (0.42)
IV	L	2.93 (0.58)	2.71 (0.64)	2.85 (0.90)	3.07 (0.83)	3.01 (0.91)
	R	3.52 (1.14)	3.10 (1.01)	3.43 (0.97)	3.35 (1.26)	3.11 (1.01)
V	L	3.20 (0.61)	4.00 (0.75)	3.94 (0.84)	3.25 (0.83)	3.81 (0.67)
	R	3.39 (0.77)	4.12 (0.85)	3.97 (1.03)	3.25 (0.83)	3.94 (0.76)
VI	L	8.26 (1.71)	9.15 (1.86)	9.03 (2.18)	9.30 (2.05)	9.18 (2.12)
	R	7.49 (1.73)	8.67 (1.93)	8.74 (1.76)	8.66 (1.45)	8.07 (1.76)
VIIA	L	13.49 (1.77)	13.11 (2.25)	13.89 (2.51)	13.74 (2.34)	13.15 (1.82)
Crus1	R	12.49 (1.36)	11.98 (2.42)	12.99 (2.69)	12.92 (2.43)	12.03 (3.17)
VIIA	L	6.35 (0.97)	8.42 (2.70)	8.37 (1.62)	7.15 (1.60)	8.53 (1.50)
Crus2	R	7.71 (1.93)	8.02 (1.82)	9.72 (1.74)	7.60 (2.15)	6.96 (2.48)
VIIB	L	4.64 (0.60)	4.49 (0.94)	5.46 (1.62)	5.81 (1.83)	4.56 (2.36)
	R	4.82 (0.68)	4.59 (1.74)	5.46 (1.61)	6.07 (1.60)	6.92 (1.41)
VIIIA	L	6.31 (0.55)	4.26 (1.57)	5.91 (1.77)	5.64 (1.67)	5.47 (2.07)
	R	3.62 (1.20)	3.54 (0.86)	4.01 (0.95)	4.06 (1.20)	2.88 (1.04)
VIIIB	L	4.05 (0.71)	5.26 (1.71)	4.32 (0.94)	3.69 (0.77)	3.58 (1.06)
	R	3.44 (0.70)	4.37 (1.27)	3.87 (0.83)	3.28 (0.83)	4.36 (1.09)
VIII	v	1.83 (0.41)	0.64 (0.35)	0.74 (0.20)	0.55 (0.21)	0.91 (0.28)
IX	L	3.08 (0.25)	3.60 (0.76)	3.74 (0.94)	3.45 (0.75)	3.74 (0.81)
	R	3.18 (0.18)	3.54 (0.85)	3.27 (0.92)	3.34 (0.85)	3.12 (0.86)
	v	0.99 (0.07)	1.09 (0.20)	0.92 (0.15)	0.95 (0.15)	1.15 (0.34)
X	R + L	0.91 (0.19)	1.29 (0.17)	1.31 (0.30)	1.20 (0.29)	1.72 (0.31)
	L	0.44 (0.10)	0.66 (0.08)	0.69 (0.18)	0.60 (0.19)	0.93 (0.27)
	R	0.47 (0.10)	0.63 (0.11)	0.62 (0.14)	0.61 (0.13)	0.79 (0.10)
	v	0.33 (0.02)	0.23 (0.09)	0.33 (0.07)	0.33 (0.06)	0.34 (0.08)

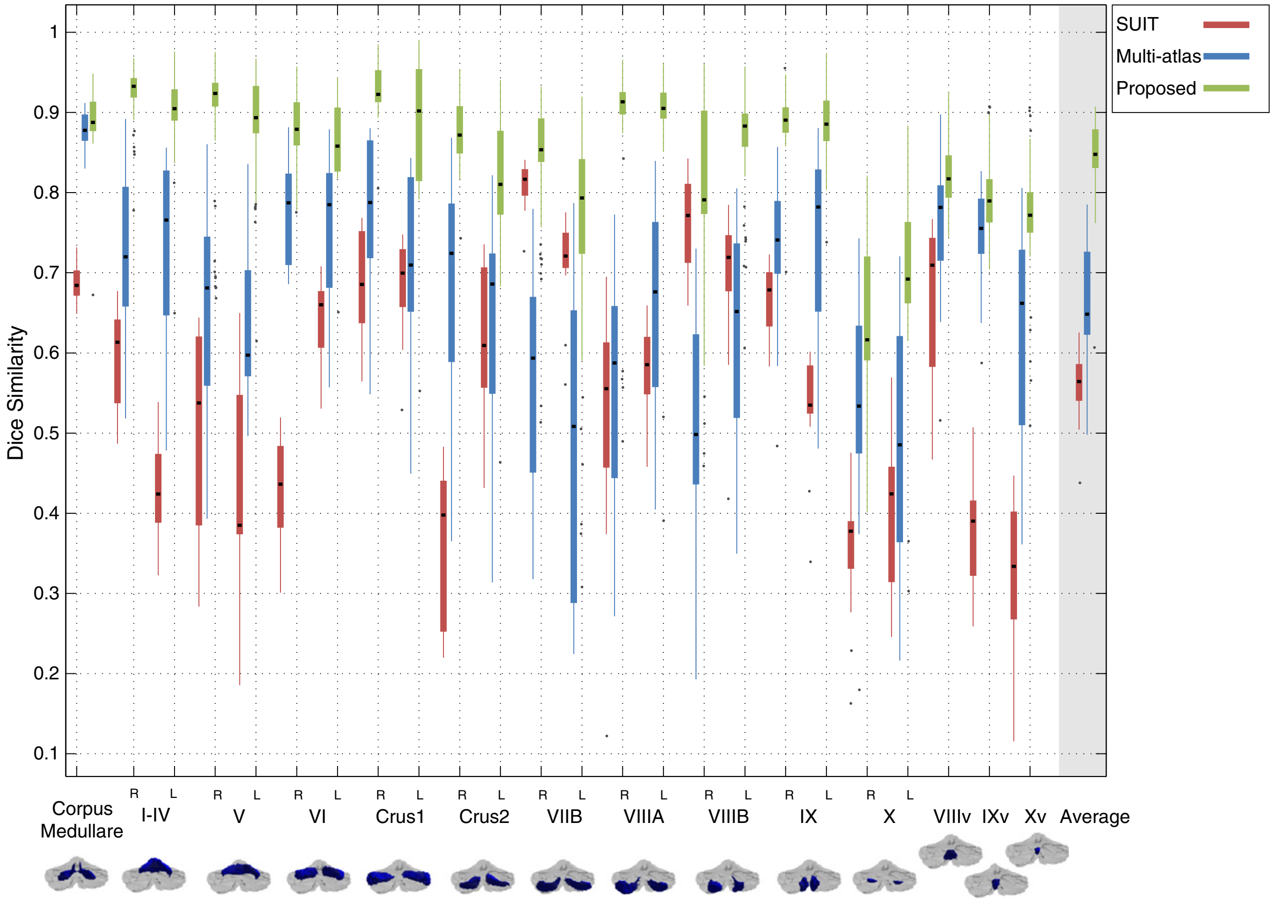

Fig. 8. Box plots of Dice similarity coefficient for SUIT, multi-atlas segmentation and the proposed manual method using inexpert raters. The rightmost set of boxes shows the of the average Dice similarity across all labels.

variable and challenging for inexpert raters to delineate reliably. Dedicated training for this lobule may be necessary. To summarize, the situations described above rely on rater judgment and experience rather than observing image features. Therefore, differences between raters in this estimation account for errors in the smaller lobules. Incorporating different image features (for example diffusion weighted image contrasts) may be useful in future efforts to provide image cues not present in the T1 images used here. Diffusion imaging in particular, could play a role in future studies in sub-parcellating the cerebellar white matter and identifying the deep nuclei.

The rapid-review and fusion aspects of this system are critical as well. The review process was instrumental in identifying gross errors made by inexpert raters. Statistical label fusion techniques (STAPLER) improved the reliability of the final result by combining individual inexpert delineations. We have shown that the consensus labeling estimated using these techniques tends to agree with a parcellation produced by an expert; an example of this is given in Fig. 2. The results suggest that more robust fusion techniques and/or more careful delineation may be necessary when regions are small and/or difficult to delineate (e.g., lobule X).

Finally, our nominal volume measures reported in Table 3 generally agree with those reported in Makris et al. (2005). For example, the mean bilateral (R+L) anterior lobe volume in Makris et al. (2005) was 13.68 (2.4), whereas the STAPLER fusion result here was 12.18 (1.84). The difference observed could be due to the mean age of our cohort (51 years), which is likely larger than the unreported mean age in Makris et al. (2005). In many cases, the mean of the fused results tends to be closer to the mean expert result than any of the individual inexpert raters.

Conclusion

By using this protocol, inexpert raters can produce a parcellation of the human cerebellum that agrees with experts when paired with a system for review/verification and statistical label fusion. The hierarchical nature of this protocol allows for researches to tailor the protocol to a specific hypothesis by balancing reliability and parcellation coarseness. Fusion of multiple inexpert delineations can produce a labeling that approaches the reliability and accuracy of an expert. This is especially useful since employing many inexpert raters may be more time and cost-effective than employing a small number of experts. The efficiency offered by this approach can enable larger scale studies of the cerebellum in the future.

Acknowledgments

This work was supported in part by NIH/NINDS grants R01NS56307 and R01NS37747. The authors would like to thank Neda Jahanshad for her work on initial versions of the protocol, the undergraduate raters, Charles Ang, Lauren Drake, and Javaneh Jabbari, and the rapid-reviewers Aaron Carass and Lorre Atlan.

Appendix A

The evaluation described used leave-one-out cross validation to estimate the performance of the multiple-atlas segmentation framework. Fifteen subjects were used (six controls), meaning that fourteen "atlases" were used to label each subject. Registration was performed using the

SyN algorithm (Avants et al., 2008), which was found to be a leading method for the purposes of labeling the cortex (Klein et al., 2009). SyN was run using the following parameters: cross correlation similarity using a 32 bin square joint histogram, Gaussian regularization with sigma of 3, a three level optimization using 30, 20, and 10 iterations at the coarse-, mid-, and full-resolutions, respectively. Label fusion was accomplished with the multi-category STAPLE algorithm (Warfield et al., 2004) initialized with equal confusion matrices for all raters with diagonal entries equal to 0.9999. Convergence was declared when the difference of the normalized trace of the confusion matrix between two iterations was less than 10^{-5}.

The SUIT algorithm was run using SUIT version 2.4 with SPM8. A corpus medullare label was manually added to the SUIT template. A whole cerebellum mask was generated using the *suit_isolate* command. The SUIT template and subject cerebellum were deformably registered using *suit_normalize*. Next, the lobule labels and cerebellum mask were resampled into the subject space with *suit_reslice_inv*. Finally, we applied the cerebellum mask from the isolation step to the lobule labels to refine the delineation near deep fissures.

The Dice similarity coefficient (DSC) with the expert's lobule labels was computed for each of these 15 subjects for the multi-atlas, and SUIT automated methods, and for the proposed method using inexpert raters, a box plot of which is shown in Fig. 8. The median DSC of the proposed method is higher for all lobules than either of the automatic methods. We also computed the DSC using only the gray matter labels after masking the gray matter labels for all methods with the true (expert rater) gray matter. The mean (standard deviation) DSCs are: 0.63 (0.15) for SUIT, 0.82 (0.13) for multi-atlas, and 0.89 (0.09) for our inexpert rater fusion. This indicates automated methods make errors in assigning different gray matter labels that cannot be solved by incorporating a simple tissue segmentation.

References

Andersen, B.B., Gundersen, H.J.G., Pakkenberg, B., 2003. Aging of the human cerebellum: a stereological study. J. Comp. Neurol. 466, 356–365.

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

Berquin, P.C., Giedd, J.N., Jacobsen, L.K., Hamburger, S.D., Krain, A.L., Rapoport, J.L., Castellanos, F.X., 1998. Cerebellum in attention-deficit hyperactivity disorder: a morphometric MRI study. Neurology 50, 1087–1093.

Brenneis, C., Bosch, S.M., Schocke, M., Wenning, G.K., Poewe, W., 2003. Atrophy pattern in SCA2 determined by voxel-based morphometry. Neuroreport 14, 1799–1802.

Callison-Burch, C., 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. EMNLP, pp. 286–295.

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. Neuropathol. Appl. Neurobiol. 23, 355–363.

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.

Donchin, O., Rabe, K., Diedrichsen, J., Lally, N., Schoch, B., Gizewski, E.R., Timmann, D., 2012. Cerebellar regions involved in adaptation to force field and visuomotor perturbation. J. Neurophysiol. 107, 134–147.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 33, 115–126.

Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M.A., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging 28, 1000–1010.

Ito, M., 1984. The Cerebellum and Neural Control. Raven, New York.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Jung, B.C., Choi, S.I., Du, A.X., Cuzzocreo, J.L., Ying, H.S., Landman, B.A., Perlman, S.L., Baloh, R.W., Zee, D.S., Toga, A.W., Prince, J.L., Ying, S.H., 2011. MRI shows a region-specific pattern of atrophy in spinocerebellar ataxia type 2. Cerebellum 11 (1), 272–279.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46, 786–802.

Landman, B.A., Asman, A.J., Scoggins, A.G., Bogovic, J.A., Stein, J.A., Prince, J.L., 2012a. Foibles, follies, and fusion: web-based collaboration for medical image labeling. Neuroimage 59, 530–539.

Landman, B.A., Asman, A.J., Scoggins, A.G., Bogovic, J.A., Xing, F., Prince, J.L., 2012b. Robust statistical fusion of image labels. IEEE Trans. Med. Imaging 31, 512–522.

Leiner, H.C., Leiner, A.L., Dow, R.S., 1986. Does the cerebellum contribute to mental skills? Behav. Neurosci. 100, 443–454.

Levitt, J.J., Mccarley, R.W., Nestor, P.G., Petrescu, C., Donnino, R., Hirayasu, Y., Kikinis, R., Jolesz, F.A., Shenton, M.E., 1999. Quantitative volumetric MRI study of the cerebellum and vermis in schizophrenia: clinical and cognitive correlates. Am. J. Psychiatry 156, 1105–1107.

Makris, N., Hodge, S.M., Haselgrove, C., Kennedy, D.N., Dale, A., Fischl, B., Rosen, B.R., Harris, G., Caviness, V.S., Schmahmann, J.D., 2003. Human cerebellum: surface-assisted cortical parcellation and volumetry with magnetic resonance imaging. J. Cogn. Neurosci. 15, 584–599.

Makris, N., Schlerf, J.E., Hodge, S.M., Haselgrove, C., Albaugh, M.D., Seidman, L.J., Rauch, S.L., Harris, G., Biederman, J., Caviness, V.S., Kennedy, D.N., Schmahmann, J.D., 2005. MRI-based surface-assisted parcellation of human cerebellar cortex: an anatomically specified method with estimate of reliability. Neuroimage 25, 1146–1160.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development: The International Consortium for Brain Mapping (ICBM). Neuroimage 2, 89–101.

McAuliffe, M.J., Lalonde, F.M., McGarry, D., Gandler, W., Csaky, K., Trus, B.L., 2001. Medical image processing, analysis & visualization in clinical research. IEEE CMBS, p. 381.

McCormick, D.A., Thompson, R.F., 1984. Cerebellum: essential Involvement in the classically conditioned eyelid response. Science 223, 296–299.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46.

Middleton, F.A., Strick, P.L., 1994. Anatomical evidence for cerebellar and basal ganglia involvement in higher cognitive function. Science 266, 458–461.

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

Nitschke, M.F., Kleinschmidt, A., Wessel, K., Frahm, J., 1996. Somatotopic motor representation in the human anterior cerebellum. A high-resolution functional MRI study. Brain 119, 1023–1029.

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. Biol. Psychiatry 46, 703–711.

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

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.

Raz, N., Dupuis, J.H., Briggs, S.D., McGavran, C., Acker, J.D., 1998. Differential effects of age and sex on the cerebellar hemispheres and the vermis: a prospective MR study. AJNR Am. J. Neuroradiol. 19, 65–71.

Ritchie, L., 1976. Effects of cerebellar lesions on saccadic eye movements. J. Neurophysiol. 39, 1246–1256.

Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2007. LabelMe: a database and web-based tool for image annotation. Int. J. Comput. Vision 77, 157–173.

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

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

Schmahmann, J.D., Doyon, J., Toga, A., Petrides, M., Evans, A., 2000. MRI Atlas of the Human Cerebellum. Academic Press, San Diego, CA.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420.

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

Sorokin, A., Forsyth, D., 2008. Utility data annotation with Amazon Mechanical Turk. CVPR. IEEE, Anchorage, AK, pp. 1–8.

Steenbakkers, R.J.H.M., Duppen, J.C., Fitton, I., Deurloo, K.E.I., Zijp, L., Uitterhoeve, A.L.J., Rodrigus, P.T.R., Kramer, G.W.P., Bussink, J., De Jaeger, K., Belderbos, J.S.A., Hart, A.A.M., Nowak, P.J.C.M., van Herk, M., Rasch, C.R.N., 2005. Observer variation in target volume delineation of lung cancer related to radiation oncologist–computer interaction: a "Big Brother" evaluation. Radiother. Oncol. 77, 182–190.

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. Psychiatr. Res. 42, 1198–1202.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903–921.

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.