A Semi-Automatic Algorithm for Determining the Demyelination Load in Metachromatic Leukodystrophy

Philipp Clas, Samuel Groeschel, MD, Marko Wilke, MD, PhD

Rationale and Objectives: Metachromatic leukodystrophy is a lysosomal storage disorder leading to progressive demyelination of brain white matter. This is sensitively detected using magnetic resonance imaging. The volume of demyelination, the "demyelination load," could serve as a useful parameter for assessing both the natural course of the disease and treatment effects. The aim of this study was to develop and validate a semiautomated approach for determining the demyelination load to achieve reliable and time-efficient segmentation results.

Materials and Methods: The demyelination load was determined in 77 magnetic resonance imaging data sets from 35 patients both manually and semiautomatically. For manual segmentation, regarded as the gold standard, the software ITK-Snap was used. For semiautomatic segmentation, a new algorithm called Clusterize was developed and implemented in MATLAB, consisting of automatic iterative region growing followed by the interactive selection of clusters. Results were compared in terms of the obtained volumes, spatial overlap, and time taken to conduct the segmentation.

Results: Performance of the semiautomatic algorithm was excellent, with the volumes generated by the new algorithm showing good agreement with the ones generated by the gold standard (93.4 ± 45.5 vs 96.1 ± 49.0 mL, *P* = NS) with high spatial overlap (Dice's similarity coefficient = 0.7861 ± 0.0697). The semiautomatic algorithm was significantly faster than the gold standard (8.2 vs 27.0 min, *P* < .001). Intrarater and interrater reliability determined high reproducibility of the method.

Conclusion: The demyelination load in metachromatic leukodystrophy can be determined in a time-efficient manner using a semiautomatic algorithm, showing high agreement with the current gold standard.

Key Words: Metachromatic leukodystrophy; demyelination load; semiautomatic segmentation; magnetic resonance imaging; lesion load.

Metachromatic leukodystrophy (MLD) is a rare inherited lysosomal storage disorder leading to the degradation of the myelin sheath in both the central and peripheral nervous system (1). Three forms are distinguished: a late infantile (onset before 3 years of age), a juvenile (16 years), and an adult form. Patients develop progressive neurologic symptoms with different rates of disease progression and different initial manifestations (1,2). Currently, no curative treatment is available, though promising new therapeutic approaches are under investigation (1,3).

The imaging hallmark of this disease is a rapidly progressive leukodystrophy, which can be detected sensitively using magnetic resonance (MR) imaging. Demyelinated white matter (WM) presents as hyperintensities on T2-weighted images. Only recently, Eichler et al (4) developed a score for visually assessing these hyperintensities. This score ranges from 0 to 34 points and has been applied to describe the natural course of the disorder in children (5). Here, a rater scores a number of predefined brain regions as "normal" (healthy), "faint" (1 point), or "dense hyperintensity" (2 points). Although this allows the assessment of global disease progression in a structured way, it would be beneficial to have a more sensitive and quantitative measure of the affected tissue (5). This is particularly important when evaluating new therapeutic approaches, for which an objective validation of putative treatment-induced changes is of high relevance.

Therefore, the basis for our work is a measure that has been used to describe disease progression in multiple sclerosis but has not yet been applied to leukodystrophies, the "demyelination load" (termed "lesion load," "disease burden," or "burden of disease" in multiple sclerosis [6–9]). The idea is to compute the volume of demyelinated WM in the central nervous system. This requires demyelinated WM to be delineated from normal-appearing WM, for which the current gold standard is manual tracing (10,11). However,

Acad Radiol 2012; 19:26–34

From the Department of Pediatric Neurology & Developmental Medicine, Children's Hospital, and Experimental Pediatric Neuroimaging, Children's Hospital and Neuroradiological Clinic, University of Tübingen, Hoppe-Seyler-Straße 1, 72076 Tübingen, Germany. Received July 27, 2011; accepted September 7, 2011. Mr Clas was supported by the IZKF-Promotionskolleg PK2011-6, Medical Faculty, University of Tübingen. Dr Groeschel has received institutional research support from the German Federal Ministry of Education and Research and from Shire plc. Dr Wilke was supported by DFG WI3630/1-1 from the German Research Council. **Address correspondence to:** S.G. e-mail: samuel.groeschel@med.uni-tuebingen.de

doi:10.1016/j.acra.2011.09.008

delineating areas of demyelination (AODs) slice by slice on current high-resolution MR images is very time consuming and may lack objectivity (see below for more details).

The aim of the present study was therefore to develop and evaluate a semiautomated approach for determining the demyelination load in patients with MLD. Validation was performed by comparing the method to the current gold standard, manual delineation. Secondary aims were shorter processing times and a robust procedure that would allow the use of standard clinical MR images.

MATERIALS AND METHODS

MR Imaging Data

Overall, 77 data sets from 35 patients were used: 45 MR images for the late infantile form (18 patients; mean age, 3.5 ± 1.0 years; 7 male), 26 of which were acquired under therapy (intravenous enzyme replacement therapy), and 32 MR images for the juvenile form (17 patients; mean age, 13.9 ± 8.0 years; 7 male), 13 of which were acquired under therapy (hematopoietic stem cell transplantation). The images were collected in a nationwide leukodystrophy network and in the context of a phase I therapeutic trial (12); consequently, they were acquired using different MR scanners at several hospitals. Raters were blind with regard to clinical details, in particular whether images were obtained under therapy or not. Data collection was approved by the local ethics committee. Informed written consent was given by the parents.

All 77 data sets consisted of at least an axial T2-weighted sequence. In-plane resolution was 0.64 ± 0.18 × 0.64 ± 0.18 mm, with a slice thickness of 4.46 ± 0.92 mm (29 ± 6 slices). For 49 data sets, an additional high-resolution T1-weighted data set was available, with a resolution of 1 ± 0.09 × 0.94 ± 0.15 × 0.96 ± 0.14 mm (272 ± 67 slices). For the remaining 28 data sets, either a low-resolution T1-weighted sequence (resolution, 0.61 ± 0.21 × 0.61 ± 0.21 × 4.98 ± 0.91 mm; 27 ± 6 slices) or a fluid-attenuated inversion recovery (FLAIR) sequence (resolution 0.77 ± 0.2 × 0.77 ± 0.2 × 5.71 ± 1.1 mm; 24 ± 7 slices) were available.

Manual Segmentation

An AOD within brain WM on a T2-weighted image was defined as consisting of connected voxels with higher signal intensity than normal WM and gray matter and comparable signal intensity to cerebrospinal fluid (CSF).

We specifically abstained from restricting the search volume, because demyelination in MLD may occur throughout supratentorial and infratentorial WM (5). The main problem for the determination of the demyelination is the fact that an AOD sometimes lacks a clear boundary because of gradual disease progression and partial volume effects. Therefore, the transition from clearly pathologic and hyperintense tissue to clearly healthy and nonhyperintense tissue may sometimes be gradual. This is the main reason why it can be very difficult to delineate the “true” extent of an AOD, making the decision to some extent subjective.

In recognition of that fact, a training session with two experienced raters was instituted to have the two raters come to a similar appreciation regarding the identification of an AOD and its extent. Manual tracings were then completed by each rater independently.

For manual tracings, ITK-Snap (13) was used to analyze T2-weighted images. As a first step, contrast settings were manually adjusted for each image individually to optimize the contrast between normal WM and demyelination. In the axial view, a manual tracing tool and a tool to zoom were used to delineate AODs as precisely as possible on all slices. The resulting volume was computed by counting the voxels labeled as an AOD and by multiplying this count with the respective voxel volume.

All manual tracings were conducted by one rater (P.C.). To assess interrater and intrarater reliability, a subset of 18 data sets was selected from the original data set, which was reanalyzed by the main rater and was also analyzed by a second rater (S.G.). Care was taken to ensure that the subset was similar to the overall image data set with respect to the patient’s age, gender, disease variant, and demyelination load.

Semiautomatic Segmentation

A new algorithm, called Clusterize, was developed and implemented in MATLAB (The MathWorks, Inc, Natick, MA) to conduct semiautomatic segmentation. The general approach is to search each slice for local intensity maxima, which are then defined to be cluster cores. Using iterative region growing (clustering) and applying meaningful constraints (see the next section), every eligible voxel in the brain is ultimately assigned to a cluster. This initial automated preprocessing is followed by a manual interactive selection of clusters corresponding to AODs. In detail, the algorithm proceeds as follows:

Automated preprocessing. First, images are scaled to an intensity range of 0 to 1000, which serves to standardize the following calculations.

Each slice is considered separately and smoothed using a Gaussian kernel with full width at half maximum of 2 mm in each dimension (x, y). This leads to more regularly shaped clusters, avoiding biologically implausible edges.

Each slice is then searched for local intensity maxima of at least the number of voxels specified in the extent threshold, t_E. To prevent oversegmented results (14), this value was set to 100 voxels, setting a minimum size for the initiation of a cluster. By iteratively adapting the intensity threshold in steps of 1%, starting at the highest and proceeding to the lowest value, each voxel is consequently assigned to a cluster core. This is similar to a region-growing approach as, with lower thresholds, the cluster will contain more and more voxels. An illustration of this approach is shown in Figure 1.

Every voxel can be part of only one cluster. The following rules were used to ensure that this would be the case: (1) a voxel

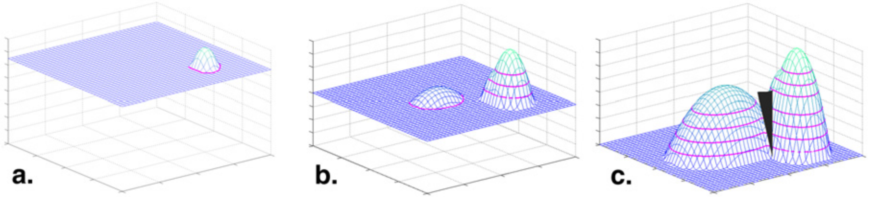

Figure 1. Approach of the algorithm. **(a-c)** The approach of processing an idealized slice dimensioned $x \times y$, with the z axis showing the intensity. The moving grid represents the changing intensity threshold, from high **(a)** to low **(c)**. The first core is found in **(a)** and the second in **(b)**. **(c)**The final result, where each voxel above the lower threshold, t_L, is assigned to a cluster core. The gray rectangle separates the two clusters (cf Fig 2).

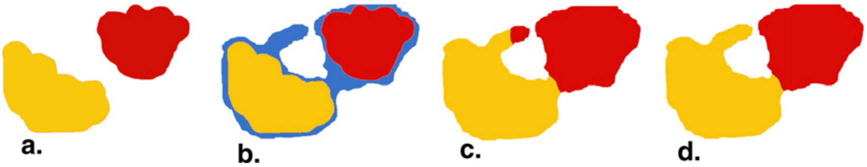

Figure 2. Competing clusters. **(a)** Two clusters (*yellow* and *red*). In the next iteration **(b)**, the threshold is lower, so voxels emerge (*blue*) that must be assigned to either of these two clusters. They are then assigned according to the minimum distance **(c)**, which causes an implausible result (*red tip* at the extension of the *yellow cluster*). **(d)** This is corrected by assigning it to the cluster with which it is connected (*yellow*).

belongs to the cluster core with which it is connected, and (2) if a voxel is connected with two cluster cores, it is assigned to the cluster to which it is closer (on the basis of the last iteration). This prevents implausible assignments of voxels to competing clusters (see Fig 2 for an illustration). "Connection between two voxels" was defined as side to side connection, resulting in a maximum of four connected voxels in one plane ("four-connected" [14]).

For each intensity threshold, clusters are stored in a two-dimensional matrix, ultimately resulting in a three-dimensional cluster matrix with one plane for each iteration. For each image slice, one such cluster matrix is stored, representing the result of the automated preprocessing (see Fig 3 for an illustration).

In the original implementation, every voxel in the image volume was assigned to a cluster, which is neither meaningful nor efficient. We therefore implemented two approaches to eliminate noncontributing voxels. The first option is to introduce a lower intensity threshold, t_L. Only voxels with intensities above t_L are considered, effectively excluding background voxels. This threshold t_L can be set to zero, such that every voxel will be labeled, and a lower value does not change the labeling of voxels above that threshold, but using it substantially speeds up computation. On the basis of simulations in the current data set, t_L was set to 20% of the maximum intensity. Removing the ventricles is also recommended, because this improves the separation of ventricles and small nearby AODs. If only T2-weighted images are available, clusters corresponding to the ventricles can be identified and removed by the user, upon which the cluster matrix is updated. Alternatively, the exclusion of the ventricles can be done using an accompanying T1-weighted or FLAIR image from the same subject. In this case, routines available within SPM8 (The Wellcome Trust Centre for NeuroImaging, University College London, London, United Kingdom) are used to coregister the images, and as a first approximation, an Otsu threshold (15) is suggested to separate CSF from AODs, taking advantage of the fact that CSF spaces will have low intensity in these images; the user must set the final intensity threshold interactively. This excludes CSF spaces very effectively.

On a standard computer workstation, the time needed for the automated preprocessing of one MR image was on the order of several minutes. The limiting factor is high image resolution. For very high resolution images, an optional step was included that resizes images to a lower resolution, speeding up processing. This option was not used for any of the analyses in this study.

User interaction. Following automated processing, the user is presented with an axial slice with clusters (initially using the results from the final iteration at the lowest threshold t_L) on the left and the corresponding original slice on the right (Fig 4). The user can then left-click to select a cluster that corresponds to an AOD on the left, upon which the outline of the cluster is shown on the right. The outline is used because it allows for a better estimate of the correspondence of the cluster and the AOD. To achieve an optimal overlap, the user can then adapt the size of the chosen cluster by holding down the right mouse button while moving the mouse, which effectively walks through the slices from the corresponding cluster matrix. Further clusters can be added and adapted until all AODs are appropriately described. When the user is done with one slice, the clusters chosen from that slice are combined and stored (Fig 4).

This process is repeated for every slice. Upon moving to the next slice, clusters showing a substantial overlap (>50%) with the clusters from the previous slice are preselected, significantly speeding up the process. Final selection of clusters, however, rests with the user. Manual editing of clusters to remove or add parts is also possible using a freehand "scissors" tool. Upon completion of the last slice, the generated volume

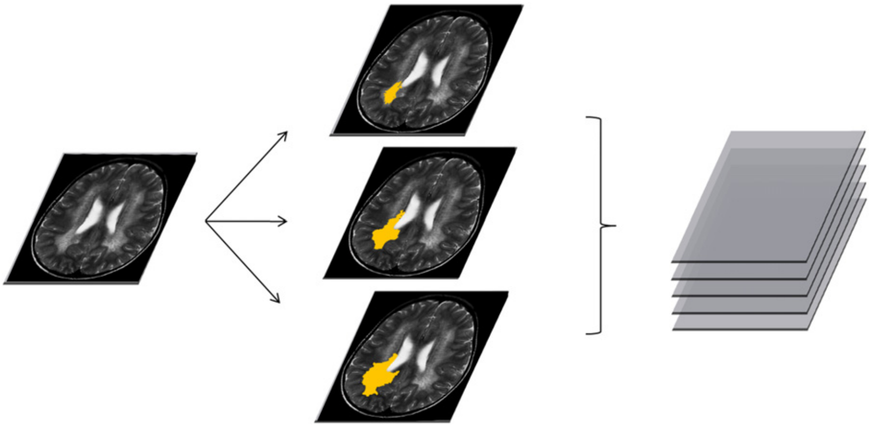

Figure 3. A cluster matrix is calculated for each slice of a magnetic resonance image, storing different versions of the same cluster (as derived by changing the intensity threshold; cf Fig 1), which the user can then interactively choose from. The maximum number of slices in a cluster matrix is 80 (if t_L = 20% and intensity step = 1%).

of interest is stored as an image file, and the volume (in milliliters) of the combined cluster is calculated. As automated processing is separate from the interactive part, the user can apply automated preprocessing to a whole stack of images and begin user interaction at a later point in time.

Paralleling the procedure established for the manual rating, all semiautomated processing was conducted by one rater (P.C.). To assess interrater and intrarater reliability, the same subset of 18 data sets selected above was used and, to this effect, also analyzed by a second rater (S.G.). For the current data set, T1-weighted or FLAIR images were available for all subjects and were used to remove the ventricles. To assess the impact of using this optional removal of the ventricles, the same 18 subjects were also analyzed without using the accompanying T1-weighted or FLAIR images to remove the ventricles ($T2_{only}$). The intraclass correlation coefficient (ICC) was used to evaluate interrater and intrarater reliability both for the manual and the semiautomated approach.

The semiautomated approach was compared with the manual approach, using the following parameters: agreement of the resulting demyelination volume, voxelwise agreement (ie, spatial overlap), and time taken to complete the rating.

To calculate the agreement of the resulting demyelination volume between the manual and semiautomatic approach, the volumes of the two approaches were evaluated with respect to their deviation from the mean, and a confidence interval was calculated. Results are visualized using Bland-Altman diagrams (16).

To calculate the voxelwise overlap between the manually and semiautomatically derived demyelination load, Dice's similarity coefficient (DSC) was calculated according to

$$\mathrm{DSC} = 2 \times \frac{\mathrm{MAN} \cap \mathrm{SAM}}{(\mathrm{MAN} + \mathrm{SAM})},$$

where MAN denotes the voxels segmented manually and SAM the voxels segmented semiautomatically. A DSC of 1 means 100% congruency; no congruency will result in a DSC of 0. There is no common agreement on a minimum value for an acceptable DSC, but coefficients > 0.7 have been considered good or high in previous publications (10,17,18).

To compare the time taken for the manual and the semiautomatic approach, Student's *t* tests were conducted. Significance was assumed at $P \leq .05$.

RESULTS

Manual Segmentation

The average demyelination load as generated by manual segmentation was 96.1 ± 49.0 mL. Over the whole data set, the average time required for manual segmentation was 27.0 ± 13.0 minutes per data set.

For the subset selected to test intrarater reliability, an average demyelination load of 87.1 ± 44.9 mL was determined in a first round of manual segmentation, while an average demyelination load of 90.2 ± 48.0 mL was determined in a second round. In a Bland-Altman plot, a mean deviation of +3.1 mL was seen (95% confidence interval [CI], −11.0 to 17.2 mL; Fig 5a). The ICC was 0.9873 (95% CI, 0.9665 to 0.9952). The mean DSC was 0.8328 ± 0.0586. The second round of manual segmentation was not significantly faster than the first (26.2 vs 23.7 minutes, P = .54).

In the interrater-test, an average demyelination load of 84.4 ± 43.1 mL was determined by the second rater in the subset. Compared to the first manual segmentation of the first rater, a mean deviation of −2.7 mL was seen (95% CI, −19.3 to 13.9 mL; Fig 5c) in a Bland-Altman plot. The ICC was 0.9804 (95% CI, 0.9484 to 0.9926). The mean DSC was 0.8026 ± 0.0808. The second rater needed on average 23.7 ± 14.1 minutes for one data set.

Semiautomatic Segmentation

The average demyelination load as generated by semiautomatic segmentation was 93.4 ± 45.5 mL. Over the whole data set, the average time required for semiautomatic segmentation was 8.2 ± 2.8 minutes per data set.

For the subset selected to test intrarater reliability, an average demyelination load of 87.8 ± 42.9 mL was determined in a first round of semiautomatic segmentation, while an average demyelination load of 86.9 ± 44.7 mL was

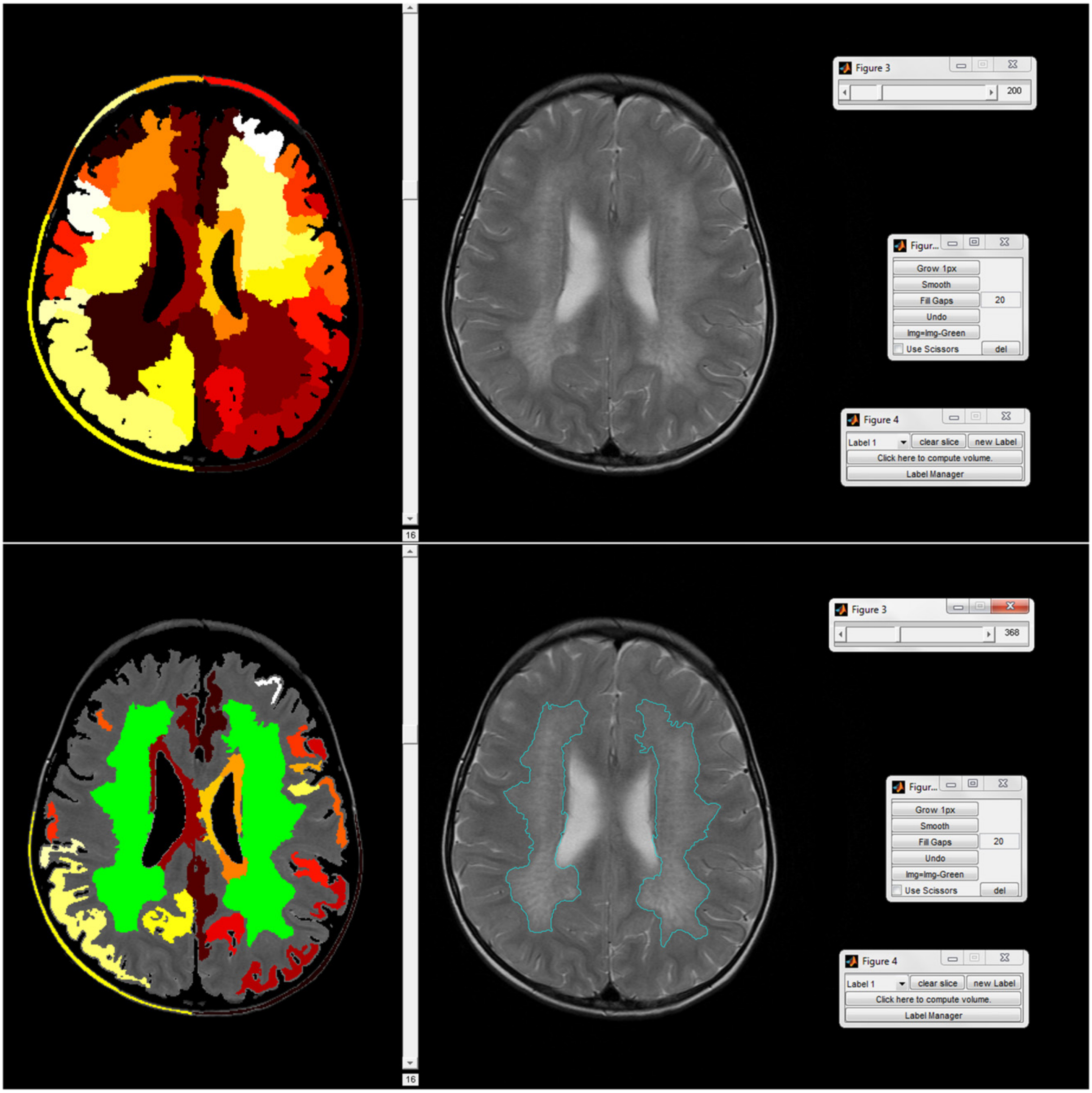

Figure 4. The *upper panel* shows a screenshot of the software toolbox in the initial view after calculating the clusters (following removal of cerebrospinal fluid using a T1-weighted image). Clusters are displayed on the left in different colors, while the original T2-weighted image is shown on the right. The *lower panel* shows the process of selecting and adjusting clusters representing areas of demyelination (*green*). The resulting contours of the selected clusters are updated in real time and are shown on the original T2-weighted image on the right.

determined in a second round. In a Bland-Altman plot, a mean deviation of −0.9 mL was seen (95% CI, −16.9 to 15.1 mL; Fig 5b). The ICC was 0.9817 (95% CI, 0.9517 to 0.9931). The mean DSC was 0.8683 ± 0.0686. The second round of semiautomatic segmentation was significantly faster than the first (7.6 ± 2.2 vs 5.7 ± 1.7 minutes, $P \leq .01$).

In the interrater test, an average demyelination load of 78.3 ± 40.3 mL was determined by the second rater in the subset. Compared to the first semiautomatic segmentation of the first rater, a mean deviation of −9.5 mL was seen (95% CI, −24.6 to 5.6 mL; Fig 5d) in a Bland-Altman plot. The ICC was 0.9820 (95% CI, 0.9525 to 0.9932). The mean DSC was 0.8433 ± 0.0649. The second rater needed on average 8.1 ± 2.1 minutes for one MR image.

When assessing the influence of using a T1-weighted or FLAIR image to remove the ventricles, the average time needed for segmentation (time to remove the ventricles included) was 6.9 ± 2.1 minutes when not using the T1-weighted or FLAIR image ($T2_{only}$) compared to 7.6 ± 2.2 when using the T1-weighted or FLAIR image ($P = .35$).

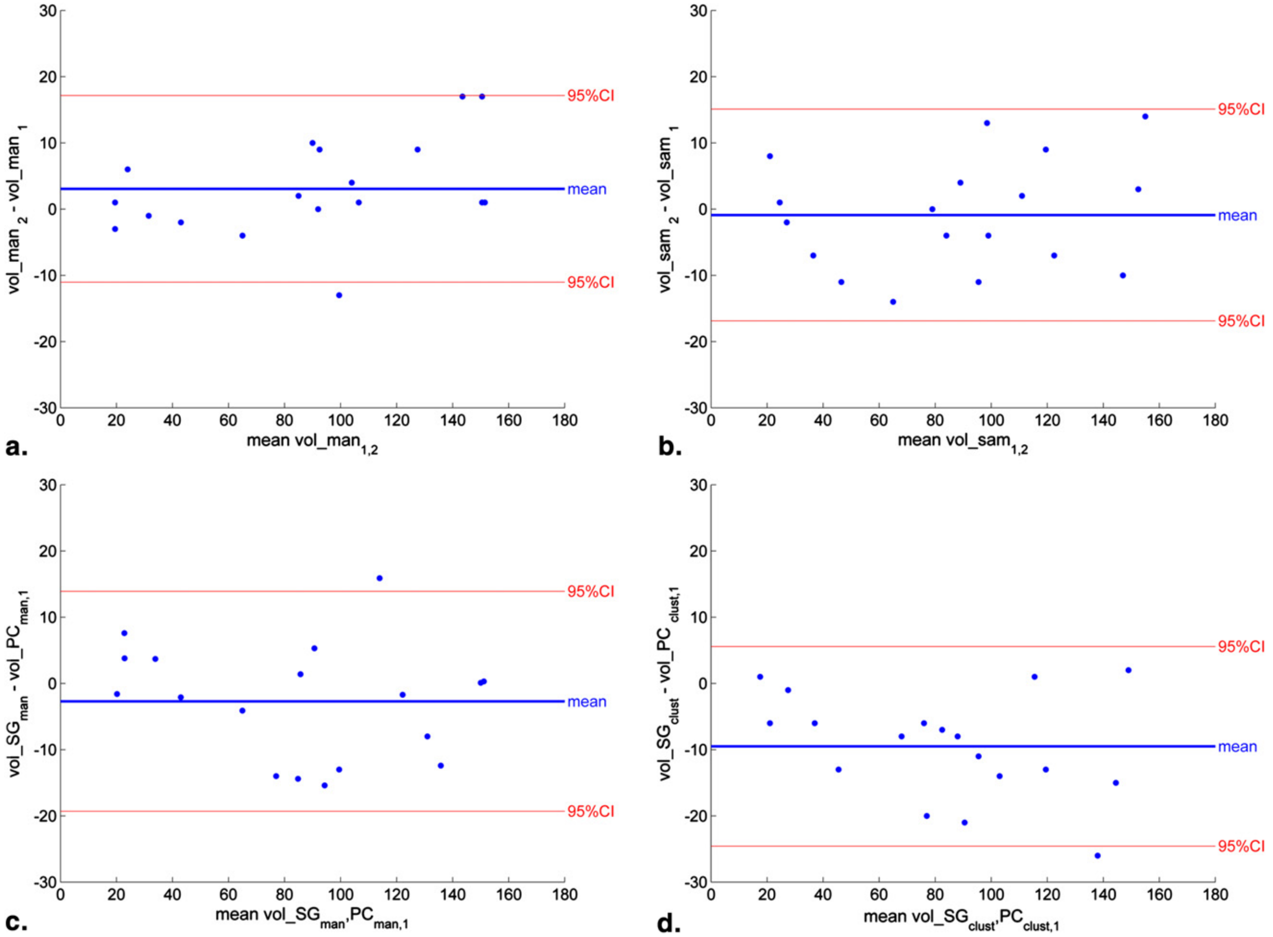

Figure 5. **(a)** Intrarater reliability for manual segmentation in a subset of *n* = 18. Each point represents one magnetic resonance (MR) image in which the demyelination load was determined manually twice by the same rater (P.C.), showing in a Bland-Altman plot the mean volume of the two measures on the *x* axis and the difference between them on the *y* axis. **(b)** Intrarater reliability for semiautomated segmentation in a subset of *n* = 18 (same images as in [a]). Each point represents one MR image in which the demyelination load was determined semiautomatically twice by the same rater (P.C.). **(c)** Interrater reliability for manual segmentation in a subset of *n* = 18 (same images as in [a]). Each point represents one MR image in which the demyelination load was determined manually twice, once by P.C. and once by S.G. **(d)** Interrater reliability for semiautomated segmentation in a subset of *n* = 18 (same images as in [a]). Each point represents one MR image in which the demyelination load was determined semiautomatically twice, once by P.C. and once by S.G.

The average demyelination load was determined as 81.9 ± 41.4 mL. In a Bland-Altman plot comparing this segmentation to the first round of semiautomatic segmentation (T1-weighted or FLAIR image, 87.8 ± 42.9 mL), a mean deviation of −5.9 mL was seen (95% CI, −22.3 to 10.4 mL). The mean DSC (between $T2_{only}$ and T1-weighted or FLAIR image) was 0.8365 ± 0.0618.

Comparison between Manual and Semiautomatic Segmentation

Comparing manual and semiautomatic segmentation, an average demyelination load of 96.1 ± 49.0 mL was determined in manual segmentation, while an average demyelination load of 93.4 ± 45.5 mL was determined in semiautomatic segmentation. In a Bland-Altman plot, a mean deviation of −2.7 mL was seen (95% CI, −18.8 to 13.4; Fig 6). The mean DSC was 0.7861 ± 0.0697. Semiautomatic segmentation was significantly faster than manual segmentation (8.2 ± 2.8 vs 27.0 ± 13.0 minutes, $P < .001$).

DISCUSSION

MLD is a rare, genetically determined, and clinically devastating neurodegenerative disorder, mainly manifesting in childhood (1). Demyelination is the imaging hallmark of MLD, and a better description and quantification of this phenomenon would contribute to an improved understanding of the disease. In this work, a method to compute the volume of demyelination in MLD, the demyelination load, is described and evaluated. A volumetric analysis of demyelination has not been conducted in leukodystrophies before, although the value of such an approach had been recognized (19). The demyelination load is a novel measure that in the future could serve as a tool to assess the natural

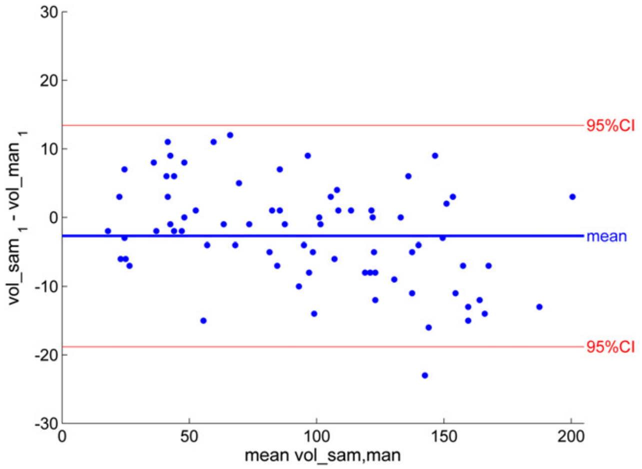

Figure 6. Comparison of the output from the semiautomated algorithm to the results from manual segmentation (n = 77), showing in a Bland-Altman plot the mean volume of the two measures on the x axis and the difference between them on the y axis.

course of the disease as well as to investigate the effects of new therapeutic strategies. Although it can be delineated manually, this is both subjective and time consuming. Our goal therefore was to develop an algorithm that allows to determine the demyelination load as precisely as, but faster than, the current gold standard.

Manual segmentation must be considered the current gold standard in the determination of structural abnormalities in the brain (10,11). Expert raters have a maximum of control over the labeling of voxels, but the process of manual segmentation is a very time-consuming task: in this study, the rater spent 27.0 ± 13.0 minutes on each image volume. When conducted twice on the same images, manual tracing produced reliable results, but there was still considerable variance between repeated ratings (Fig 5a) as well as between two raters. This discrepancy is also evident in the spatial overlap measure of 0.8328 for intrarater and 0.8026 for interrater reliability, which can be considered high (10,17,18) but not perfect agreement, demonstrating that perfect congruency is unattainable even within the current gold standard. The most likely reason for this is the lack of clear boundaries of AODs as well as partial volume effects, making it difficult to define a definitive border.

When comparing the new method with the current gold standard (Fig 6), good agreement of both approaches was evident. Our approach had a tendency to produce smaller demyelination loads, but this difference was not significant. The good agreement was also apparent in the spatial overlap, as determined by the DSC of 0.7861, which again can be considered high (10,17,18). Compared to the intrarater reliability of manual segmentation (Fig 5a), it is apparent that the variance of the current gold standard is on the same order of magnitude as when comparing the manual and the semiautomated approaches. Therefore, although perfect agreement with manual segmentation was not achieved, the discrepancy between the new method and the current gold standard was similar to the discrepancy that was seen if manual delineation was done twice. Furthermore, it may not only be questioned whether perfect agreement is possible, but also whether it would be desirable if one cannot be sure that manual segmentation is perfect (20).

Our results demonstrate that the reproducibility of the semiautomatic approach is comparable to the reproducibility that is obtained when performing manual tracing twice, in terms of the obtained volumes, the spatial overlap measure, and the variance (95% CI width, 28.2 [manual] vs 32.0 [semiautomated]; Figs 5a and 5b).

For the sample used here, matching T1-weighted images were available for all subjects, which is helpful because it allows very quick removal of the ventricles, speeding up later calculations. To assess whether this systematically changes the results, we investigated a subset of images by using only the T2-weighted images ($T2_{only}$). Compared to the approach using an accompanying T1-weighted or FLAIR image, there was a tendency to produce smaller demyelination loads. The difference, however, was not significant (P = .68). Furthermore, we conducted a post hoc analysis comparing the overlap between the $T2_{only}$ method and the gold standard on one hand and the T2-weighted plus T1-weighted or FLAIR image approach and the gold standard on the other hand. The respective DSCs were nearly identical ($T2_{only}$, 0.7834 [95% CI, 0.7502–0.8165] vs T2 plus T1, 0.7843 [95% CI, 0.7542–0.8144]; P = .97). This indicates that although processing is faster when using an accompanying T1-weighted image, the procedure can also be used with a T2-weighted image only, with similar accuracy.

As hypothesized, the semiautomated method was significantly faster than the manual method (8.2 ± 2.8 vs 27.0 ± 13.0 minutes per data set). Although this may seem like a trivial difference, it amounts to several weeks of concentrated manual processing when analyzing larger data sets. Considering the inherent variability in brain structure, especially in children (21), large-scale studies may be necessary to find smaller effects. Depending on the resources, such large-scale studies are effectively made possible only when using semiautomated or fully automated approaches. Furthermore, the immediate visual feedback of the interactive cluster manipulation allows for a critical appraisal of the user's choice as each modification is updated in real time. In our experience, the semiautomated approach suggested here requires less training, making it easier to learn than exact, expert manual tracing.

Regarding the approach taken here to "cluster" the input image, numerous approaches for tissue classification are described (for an excellent overview, see Sharma and Aggarwal [22]). For example, in a widely used software package such as SPM8, segmentation is based both on tissue probability maps and a mixture of Gaussians (23). Watershed transforms (24,25) constitute another approach; they do not require a priori knowledge but work on a topographic representation of the image with voxel intensity or gradient magnitude constituting the third dimension. Algorithms based on active contours ("snakes") (13,26) usually require the user to set seeds, which are then expanded, usually implementing criteria derived from differential equations. The application of simple intensity thresholds (27) requires the thresholds to be defined, which can either be done automatically (15) or manually. For manual fine-tuning, intelligent scissors or "live wire" (28) can be used. Although more advanced methods may be applicable in settings in which high, consistent data quality is guaranteed, we believe that the approach taken here, combining iterative intensity thresholding, morphologic operations, and easy manual fine-tuning, is appropriate for the typical clinical setting of possibly suboptimal data quality and images stemming from different scanners.

It has recently been shown (29) that a simple intensity thresholding approach alone is prone to generate false-positive findings. Moreover, in the context of AODs, using a single threshold can be problematic to determine the demyelination load, because abnormal WM intensity may be very similar to normal CSF intensity. Consequently, AODs and CSF may be connected, requiring further constraints. This leakage can also pose a problem for methods based on active contours. We addressed this issue here by searching for cluster cores in each slice and then assigned every voxel to the local maximum to which it belongs (using the criteria stated above), consequently achieving a successful separation of structures despite their similar intensity values, and by removing CSF in a first step.

The algorithm we developed resembles a watershed transform and can, in the terminology of Sharma and Aggarwal (22), be called "region-based segmentation" in which "region splitting" is used to find cluster cores and "region merging" to assign the voxels to them. We decided not to use a gradient representation of the image (in contrast to Letteboer et al [25]), because AODs very often lack clear boundaries, so the intensity-based approach seemed more applicable (a gradient-based approach would lead to a less gradual transition between healthy and affected tissue). A further difference to the watershed as conducted by Wang et al (24) is that our algorithm enables for user interaction: in the approach taken here, several different iterations are computed, resulting in a matrix of possible results, and the user can select the one that best fits the present lesion. Thus, AODs can be delineated differently for different regions in the brain, which allows the user to define the border of an AOD individually and flexibly; this is also helpful in the presence of inhomogeneities of the magnetic field (30), leading to different contrasts in different parts of the brain.

The current algorithm has potential for future modifications. MLD can be accompanied by cerebral atrophy (5). One potential advantage in using a semiautomated algorithm over the manual tracing method is the possibility of relating the demyelination load to brain atrophy, which could also be determined (semi)automatically from the input images. Apart from serving as a (perhaps more sensitive) measure for disease progression, the demyelination load contains additional information about the precise spatial location of AODs, as (in addition to yielding a volume in milliliters) it can be written out as a three-dimensional image volume. Further analyzing these may allow addressing a wide range of questions using approaches, such as lesion-symptom mapping (11,31) or even voxelwise analyses (akin to voxel-based morphometry [32]). Ultimately, the effect of therapeutic interventions should also be measurable using such an approach, which is likely to be of high relevance in the context of assessing and comparing current and future therapeutic approaches.

The current study had several limitations. As an inherent disadvantage of semiautomation, it must be noted that it is never completely objective and that for examining very large samples, even minimal user interaction can be too time consuming. On the one hand, it could be argued that the ultimate aim would be a fully automated approach to determine the demyelination load. On the other hand, the fact that the rater interactively defines AODs and thus controls the process can be seen as a general advantage of semiautomated methods over fully automated methods. Moreover, such an automated segmentation approach will likely require high-quality, ideally multispectral imaging data (33) from the same scanner (34). In contrast to this, the approach suggested here is very robust, making the algorithm applicable to images acquired with differing parameters in the clinical routine setting and from different scanners.

It should also be noted that it is necessary to set a value for the extent threshold t_E to prevent the image from being over-segmented: the smaller the value is, the more clusters will be obtained. The value of 100 voxels used here was based on simulations run on several images and was found to work well, but it was not systematically explored. Depending on the image resolution, this may correspond to a different volume (in our cases, ranging from 61 to 244 μL). In future versions of the software, the user will have the option to adjust the value of t_E both in units of voxels and microliters before clustering, extending the applicability possibly to other conditions with lesions on MR imaging.

CONCLUSIONS

With this work, we demonstrate that the demyelination load in MLD can be determined in a time-efficient manner using a semiautomatic algorithm, showing high agreement with the current gold standard. The algorithm presented here seems to be a promising approach to quantifying, in a more objective way, the demyelination occurring in this disease. The resulting overall demyelination load as well as the spatial delineation of the affected tissue may be used to further investigate both the natural course and the effect of therapeutic interventions in leukodystrophies in the future.

ACKNOWLEDGMENTS

We would like to thank the participating study centers, in particular Christine í Dali (Rigshospitalet, Copenhagen, Denmark) and Lars Hanson (Copenhagen University Hospital Hvidovre, Denmark), for contributing image datasets, and Christiane Kehrer (University Children's Hospital Tübingen, Germany) for patient recruitment. We are also grateful to Professor Krägeloh-Mann (University Children's Hospital Tübingen, Germany) and Professor Ernemann (Department of Neuroradiology, University Hospital Tübingen, Germany) for continued support.

REFERENCES

1. Gieselmann V, Krageloh-Mann I. Metachromatic leukodystrophy—an update. Neuropediatrics 2010; 41:1–6.
2. Kehrer C, Blumenstock G, Gieselmann V, et al. The natural course of gross motor deterioration in Metachromatic leukodystrophy. Dev Med Child Neurol 2011; 53:850–855.
3. Biffi A, Lucchini G, Rovelli A, et al. Metachromatic leukodystrophy: an overview of current and prospective treatments. Bone Marrow Transplant 2008; 42(suppl):S2–S6.
4. Eichler F, Grodd W, Grant E, et al. Metachromatic leukodystrophy: a scoring system for brain MR imaging observations. AJNR Am J Neuroradiol 2009; 30:1893–1897.
5. Groeschel S, Kehrer C, Engel C, et al. Metachromatic leukodystrophy: natural course of cerebral MRI changes in relation to clinical course. J Inherit Metab Dis 2011; 34:1095–1102.
6. Filippi M, Horsfield MA, Tofts PS, et al. Quantitative assessment of MRI lesion load in monitoring the evolution of multiple sclerosis. Brain 1995; 118:1601–1612.
7. Paty DW, Li DK, Oger JJ, et al. Magnetic resonance imaging in the evaluation of clinical trials in multiple sclerosis. Ann Neurol 1994; 36(suppl): S95–S96.
8. Paty DW, Li DK, UBC MS/MRI Study Group and the IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial. Neurology 1993; 43:662–667.
9. Fazekas F, Soelberg-Sorensen P, Comi G, et al. MRI to monitor treatment efficacy in multiple sclerosis. J Neuroimaging 2007; 17(suppl):50S–55S.
10. Seghier ML, Ramlackhansingh A, Crinion J, et al. Lesion identification using unified segmentation-normalisation models and fuzzy clustering. Neuroimage 2008; 41:1253–1266.
11. Rorden C, Karnath HO. Using human brain lesions to infer function: a relic from a past era in the fMRI age? Nat Rev Neurosci 2004; 5:813–819.
12. i Dali C, Hanson LG, Barton NW, et al. Brain N-acetylaspartate levels correlate with motor function in metachromatic leukodystrophy. Neurology 2010; 75:1896–1903.
13. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006; 31:1116–1128.
14. Bleau A, Leon LJ. Watershed-based segmentation and region merging. Comput Vis Image Und 2000; 77:317–370.
15. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 1979; 9:62–66.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between 2 methods of clinical measurement. Lancet 1986; 1:307–310.
17. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004; 11:178–189.
18. Garcia-Lorenzo D, Lecoeur J, Arnold DL, et al. Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts. Med Image Comput Comput Assist Interv 2009; 12:584–591.
19. Minamikawa-Tachino R, Maeda Y, Fujishiro I, et al. Three-dimensional brain visualization for metachromatic leukodystrophy. Brain Dev 1996; 18:394–399.
20. Wilke M, de Haan B, Juenger H, et al. Manual, semi-automated, and automated delineation of chronic brain lesions: a comparison of methods. Neuroimage 2011; 56:2038–2046.
21. Wilke M, Holland SK. Variability of gray and white matter during normal development: a voxel-based MRI analysis. Neuroreport 2003; 14:1887–1890.
22. Sharma N, Aggarwal LM. Automated medical image segmentation techniques. J Med Phys 2010; 35:3–14.
23. Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005; 26: 839–851.
24. Wang H, Chen X, Moss RH, et al. Watershed segmentation of dermoscopy images using a watershed technique. Skin Res Technol 2010; 16:378–384.
25. Letteboer MM, Olsen OF, Dam EB, et al. Segmentation of tumors in magnetic resonance brain images using an interactive multiscale watershed algorithm. Acad Radiol 2004; 11:1125–1138.
26. Kass M, Witkin A, Terzopoulos D. Snakes—active contour models. Int J Comput Vision 1987; 1:321–331.
27. Joe BN, Fukui MB, Meltzer CC, et al. Brain tumor volume measurement: comparison of manual and semiautomated methods. Radiology 1999; 212:811–816.
28. Schenk A, Prause G, Peitgen HO. Efficient semiautomatic segmentation of 3D objects in medical images. Lect Notes Comput Sci 2000 1935;186–195.
29. Kloppel S, Abdulkadir A, Hadjidemetriou S, et al. A comparison of different automated methods for the detection of white matter lesions in MRI data. Neuroimage 2011; 57:416–422.
30. Majumdar S, Orphanoudakis SC, Gmitro A, et al. Errors in the measurements of T2 using multiple-echo MRI techniques. II. Effects of static field inhomogeneity. Magn Reson Med 1986; 3:562–574.
31. Rorden C, Karnath HO, Bonilha L. Improving lesion-symptom mapping. J Cogn Neurosci 2007; 19:1081–1088.
32. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage 2000; 11:805–821.
33. Valdés Hernandez MC, Ferguson KJ, Chappell FM, et al. New multispectral MRI data fusion technique for white matter lesion segmentation: method and comparison with thresholding in FLAIR images. Eur Radiol 2010; 20:1684–1691.
34. Schnack HG, van Haren NE, Brouwer RM, et al. Mapping reliability in multicenter MRI: voxel-based morphometry and cortical thickness. Hum Brain Mapp 2010; 31:1967–1982.