NeuroImage 51 (2010) 1047–1056

Contents lists available at ScienceDirect

NeuroImage

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

Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods

Renske de Boer [a,b,*], Henri A. Vrooman [a], M. Arfan Ikram [b,c], Meike W. Vernooij [b,c], Monique M.B. Breteler [b], Aad van der Lugt [c], Wiro J. Niessen [a,d]

[a] *Biomedical Imaging Group Rotterdam, Departments of Radiology & Medical Informatics, Erasmus MC, Rotterdam, The Netherlands*
[b] *Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands*
[c] *Department of Radiology, Erasmus MC, Rotterdam, The Netherlands*
[d] *Imaging Science & Technology, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands*

ARTICLE INFO

Article history:
Received 9 October 2009
Revised 2 March 2010
Accepted 3 March 2010
Available online 10 March 2010

Keywords:
MRI
Brain tissue segmentation
Reproducibility
Reliability
Accuracy
Sample size estimates

ABSTRACT

The ability to study changes in brain morphometry in longitudinal studies majorly depends on the accuracy and reproducibility of the brain tissue quantification. We evaluate the accuracy and reproducibility of four previously proposed automatic brain tissue segmentation methods: FAST, SPM5, an automatically trained *k*-nearest neighbor (kNN) classifier, and a conventional kNN classifier based on a prior training set. The intensity nonuniformity correction and skull-stripping mask were the same for all methods. Evaluations were performed on MRI scans of elderly subjects derived from the general population. Accuracy was evaluated by comparison to two manual segmentations of MRI scans of six subjects (mean age 65.9 ± 4.4 years). Reproducibility was assessed by comparing the automatic segmentations of 30 subjects (mean age 57.0 ± 3.7 years) who were scanned twice within a short time interval. All methods showed good accuracy and reproducibility, with only small differences between methods. The conventional kNN classifier was the most accurate method with similarity indices of 0.82/0.90/0.94 for cerebrospinal fluid/gray matter/white matter, but it showed the lowest reproducibility. FAST yielded the most reproducible segmentation volumes with volume difference standard deviations of 0.55/0.49/0.38 (percentage of intracranial volume) respectively. The results of the reproducibility experiment can be used to calculate the required number of subjects in the design of a longitudinal study with sufficient power to detect changes over time in brain (tissue) volume. Example sample size calculations demonstrate a rather large effect of the choice of segmentation method on the required number of subjects.

Introduction

Longitudinal MR brain imaging studies provide unique insight into morphometric changes in the brain owing to the aging brain in general, neurodegenerative diseases such as Alzheimer's disease, or the effect of possible treatments. The assessment of relevant changes in brain morphometry is affected by the accuracy and reproducibility of the image acquisition and the subsequent image analysis with brain tissue segmentation tools. When designing a longitudinal study with sufficient power to detect changes in brain (tissue) volume, this reproducibility should be taken into account, as it affects the required sample size for a specific interscan interval. In a limited number of previous studies, e.g., Fox et al. (2000) and Schott et al. (2006), this type of sample size calculations have been performed for clinical trials in Alzheimer's disease.

In literature, most brain tissue segmentation methods have been evaluated on their accuracy using real data (e.g., Anbeek et al., 2005; de Boer et al., 2009; Vrooman et al., 2007), simulated data (e.g., Ashburner and Friston, 2000, 2005; Kovacevic et al., 2002), or both (e.g., Amato et al., 2003; Awate et al., 2006; Cocosco et al., 2003; Song et al., 2006; van Leemput et al., 1999). Ten studies have reported the reproducibility of brain tissue segmentation methods (Cardenas et al., 2001; Chard et al., 2002; Clark et al., 2006; Fotenos et al., 2005; Harris et al., 1999; Kovacevic et al., 2002; Lemieux et al., 1999; Lemieux et al., 2003; Shuter et al., 2008; Wang et al., 1998). Most of these studies did not compare the reproducibility of several tissue segmentation methods applied to the same data sets. This makes it difficult to compare the results, as reproducibility is also influenced by imaging hardware and acquisition parameter settings. A few studies have investigated the impact of factors related to image acquisition, including scan sessions, acquisition sequences, data analyses, scanner upgrades, scanner vendors, field strengths, signal-to-noise ratio, and scanner software (Jovicich et al., 2009; Shuter et al., 2008) on the segmented volumes. The study by Clark et al. (2006) is the only one to compare reproducibility of several tissue segmentation methods on the same data sets. By evaluating these on 20 MR images of the same subject, they did not, however, take into account the robustness of these methods to anatomical variation.

* Corresponding author.
E-mail address: renske.deboer@erasmusmc.nl (R. de Boer).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.03.012

In this study, we compare both the accuracy and reproducibility in segmenting cerebrospinal fluid (CSF), gray matter (GM), white matter (WM), and total brain (GM + WM) of four well-known segmentation methods. One of the methods also incorporates white matter lesion (WML) segmentation and is also evaluated on this aspect. Accuracy is measured by comparing each automatic segmentation result to manual segmentations performed by two observers on six data sets. The reproducibility of all methods is assessed based on the automatic brain tissue segmentations of MRI scans of 30 subjects who were scanned twice within a short time interval. All subjects were scanned on the same scanner with the same acquisition protocol and no scanner (software) updates were performed during the course of this study. The resulting reproducibility measures can be used to estimate the required minimal number of subjects to find a certain effect in a longitudinal study with sufficient power.

Materials and methods

Imaging data

Imaging data from the population-based Rotterdam Scan Study (Hofman et al., 2009) acquired in 2005–2008 were used for the evaluation of the accuracy and reproducibility. Scans were obtained on a 1.5 T GE scanner using an 8-channel head coil. The protocol included three high-resolution axial MRI sequences, i.e., a T1-weighted 3D fast RF spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse (T1w) sequence (TR = 13.8 ms, TE = 2.8 ms, TI = 400 ms, FOV = 25 × 25 cm^2, matrix = 416 × 256 (interpolated to 512 × 512 resulting in voxel sizes of 0.49 × 0.49 mm^2), flip angle = 20°, NEX = 1, bandwidth (BW) = 12.50 kHz, 96 slices with a slice thickness of 1.6 mm zero-padded in the frequency domain to 0.8 mm), a proton density-weighted (PDw) sequence (TR = 12,300 ms, TE = 17.3 ms, FOV = 25 × 25 cm^2 , matrix = 416 × 256, NEX = 1, BW = 17.86 kHz, 90 slices with a slice thickness of 1.6 mm), and a fluid-attenuated inversion recovery (FLAIR) sequence (TR = 8000 ms, TE = 120 ms, TI = 2000 ms, FOV = 25 × 25 cm^2, matrix = 320 × 224, NEX = 1, BW = 31.25 kHz, 64 slices with a slice thickness of 2.5 mm).

Brain MRI scans of six subjects were used to assess the accuracy of the different methods (one woman, five men, average age 65.9 ± 4.4 years when scanned). Two physicians independently performed manual segmentations of the brain tissues in the cerebrum on all slices of these six data sets, using a paintbrush method with a locally adapted threshold in the MNI-tool 'Display.' Scans were manually segmented into CSF, GM, and WM on the T1w volumes, and WMLs were manually segmented on the FLAIR volumes. To evaluate the reproducibility, 30 different subjects were scanned twice within an average interval of 18.5 days (median 14.5 days, interquartile range 10–23 days). This group of subjects consisted of 16 women and 14 men, and the average age was 57.0 ± 3.7 years at the time of the first scan.

Segmentation methods

The accuracy and reproducibility of four fully automated brain tissue segmentation methods were compared. These methods are listed below with specific details on the parameter settings that were used in the experiments.

Preprocessing

As the methods incorporate different nonuniformity correction and skull-stripping/masking procedures, it is difficult to compare the effects of the tissue segmentation. We therefore performed our experiments using the same preprocessing for all methods. In addition, we reported the results obtained with the default preprocessing in Appendix A.

The scans are corrected for intensity nonuniformity using the N3 method (Sled et al., 1998) within a mask. The skull-stripping mask was obtained by nonrigid registration of a manual segmented brain mask to the T1w image using Elastix[1] (Klein et al., 2010). This mask excludes the cerebellum since this structure was not always completely included in the PDw image.

(1) FAST (Zhang et al., 2001) is a brain tissue segmentation method, which is part of FSL[2] (Smith et al., 2004). This method is based on a hidden Markov random field model and an associated expectation–maximization algorithm. Both probabilistic and deterministic segmentations are given as output. We ran FAST version 4.1 without bias field correction and with otherwise default parameter settings, using the nonuniformity corrected and masked T1w image as input. Tissue volumes were calculated from the probabilistic images.

(2) SPM5[3] contains a probabilistic brain tissue segmentation method (Ashburner and Friston, 2005). A model, based on a mixture of Gaussians and tissue probability maps as deformable spatial priors, is fitted in an iterative procedure. This model combines image registration, tissue classification and (if applicable) bias correction. The output images are probabilistic images per tissue class. Default parameter settings were used, except for switching off the bias field correction. The nonuniformity corrected T1w image and the mask were given as input. Subsequently, the mask is applied to all output images, since the GM and WM images are masked by default but the CSF image is not skull-stripped and includes non-CSF components. Tissue volumes were calculated from the probabilistic images. If an experiment required a deterministic segmentation result, majority voting was used by classifying a voxel as the tissue type with the highest probability.

(3) A *k*-nearest neighbor (kNN) brain tissue segmentation method, automatically trained on the subject itself using atlas registration, and extended with white matter lesion segmentation (Cocosco et al., 2003; de Boer et al., 2009) is the third method considered. This method uses 12 nonrigidly registered atlases to obtain locations where training samples for the kNN classifier are extracted using the T1w and PDw intensities as features. The features are normalized by a simple range matching procedure. The resulting brain tissue segmentation is used to automatically derive a subject-specific intensity threshold for white matter lesions in the FLAIR image. Before segmentation, the PDw and FLAIR images are rigidly registered to the T1w image. This method was also evaluated without the PDw input image.

(4) The conventional kNN brain tissue classifier is constructed from a prior training set of atlases using the T1w and PDw intensities as features (Vrooman et al., 2007). The features are normalized by a simple range matching procedure within a brain mask. Contrary to the previous method, this kNN classifier is not trained on the subject itself. The training set is obtained from the six subjects with manual segmentations by two observers. Since the accuracy experiment uses these same six subjects, the accuracy of this method was assessed in a leave-one-out experiment. The same coregistered and nonuniformity corrected T1w and PDw images were used as for the automatically trained kNN method. Similar to automatically trained kNN classifier, this method was also evaluated using only the T1w intensities.

Experiments

Accuracy

Segmentation accuracy was assessed by comparing the automatically obtained results to manual segmentations. The automatic segmentation methods that do not classify WML were only compared

[1] Elastix is available at http://elastix.isi.uu.nl/.

[2] FSL is available at http://www.fmrib.ox.ac.uk/fsl/.

[3] SPM5 is available at http://www.fil.ion.ucl.ac.uk/spm/.

to the manual segmentations made on the T1w images (these segmentations also did not include WML). Accuracy is reported using four measures. The true positive fraction (TPF) and extra fraction (EF) are reported to express sensitivity and oversegmentation, respectively:

$$\mathrm{TPF} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad (1)$$

$$\mathrm{EF} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FN}} \qquad (2)$$

based on true positives (TP), false negatives (FN), and false positives (FP). Third, the similarity index (SI), or Dice coefficient (Dice, 1945; Zijdenbos et al., 1994), is used to express overlap between segmentations:

$$\mathrm{SI} = \frac{2(S_1 \cap S_2)}{S_1 + S_2} \qquad (3)$$

where S_1 and S_2 denote the segmented volumes and $(S_1 \cap S_2)$ is the overlap of S_1 and S_2. In addition, the overlap measure conformity (C) (Chang et al., 2009) is used:

$$C = 1 - \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}} \qquad (4)$$

Reproducibility

The results of the reproducibility experiments are presented in two ways. Firstly, based on the segmentations of both MRI scans, the resulting volumes for CSF, GM, and WM (and WML) were compared. These volumes were expressed as percentages of intracranial volume (ICV = CSF + GM + WM + WML) to correct for differences in head size. The differences between the two sequential scans were calculated from these fractional volumes by subtracting the volume of the first scan from the volume of the second scan. The fractional volumes of the sequential scans were also used to compute the coefficient of variation (CoV). The CoV is defined as the ratio of the standard deviation to the mean, and is expressed in percentages:

$$\mathrm{CoV} = \frac{1}{N}\sum_i \frac{\sigma_i}{\mu_i} \times 100\% \qquad (5)$$

where N is the number of subjects, i indexes subjects, σ_i is the standard deviation of subject i, and μ_i is the mean of subject i.

Secondly, the segmentation obtained from the second scan was transformed to the first scan by rigid registration of the T1w images using the Image Registration Toolkit (IRTK)[4] (Rueckert et al., 1999). The overlap of the transformed segmentation and the segmentation of the first scan is represented by the similarity index and the conformity measure. Similarly, the overlap of the skull-stripping masks was calculated, and their SI and C are given as indicators of the error caused by the registration. A two-tailed paired t-test compared the methods based on the SI values.

Sample size calculations for longitudinal studies

Reproducibility influences the number of subjects needed in a longitudinal study. In a study, two types of statistical errors can be made. The probability of a type I error, or level of significance (α), is the probability of rejecting the null hypothesis when the null hypothesis is true. The probability of a type II error (β) is the probability of accepting the null hypothesis when the null hypothesis is false. The power of a study is the probability of rejecting the null hypothesis when the alternative hypothesis is true ($1-\beta$). The number of subjects required to find an effect depends on the power, level of significance, and the effect size. The effect size is defined as follows.

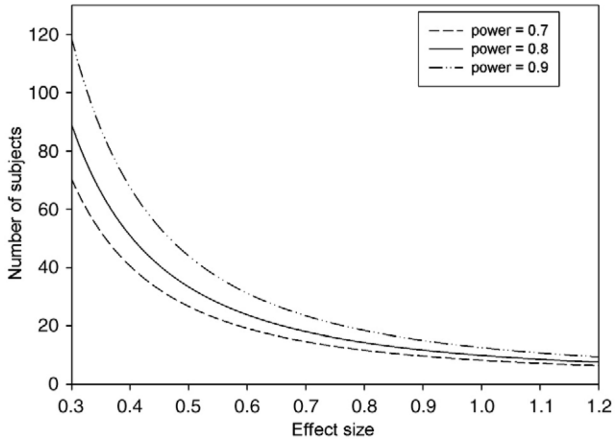

Fig. 1. Effect size versus required number of subjects for the design of a longitudinal study (two-tailed paired t-test, $\alpha = 0.05$).

$$\text{effect size} = \frac{\text{mean difference over time period}}{\text{standard deviation}} \qquad (6)$$

The standard deviation depends on the segmentation method and results from the reproducibility experiment as the standard deviation of the volume differences. Given the effect size (e.g., rates of tissue atrophy), power, and significance level, the required number of subjects can be calculated using the program G*Power 3[5] (Faul et al., 2007). Fig. 1 shows a graph relating the required number of subjects to effect size. This graph is obtained with G*Power 3, using varying powers and a level of significance of $\alpha = 0.05$ in a paired two-tailed t-test. We give several example calculations using the results from the reproducibility experiment.

The sample size calculations presented in this paper assume a longitudinal study of a single group of subjects. If a longitudinal study is designed to compare the atrophy rates of two groups, e.g., patients with Alzheimer's disease and a control group, the required number of subjects also depends on the standard deviations and the difference in sizes of these groups.

Results

Segmentation

Fig. 2 shows a representative result of an axial slice of the MR images and their corresponding manual and automatic segmentations of a subject with low WML load.

One scan of 1 of the 30 subjects used in the reproducibility experiments was not masked properly by the nonrigidly registered skull-stripping mask as it included part of both eyes. This resulted in an erroneous segmentation by the conventional kNN classifier, as it performs its feature normalization within the brain mask. A new brain mask was created by an alternative registration strategy, and a new conventional kNN segmentation was obtained. The new mask was also used for all other segmentations of this subject.

Processing time of the FAST method was approximately 26 min on a 64-bit Linux cluster node. SPM5 processing took approximately 8 min on a 64-bit Linux system. The nonrigid atlas registration for the automatically trained kNN classifier took approximately 6–7 hours per atlas using IRTK and a control point spacing of 2.5 mm on a 64-bit Linux system. Comparable nonrigid registration of the same atlas using Elastix

[4] IRTK is available at http://www.doc.ic.ac.uk/~dr/software/.

[5] G*Power 3 is available at http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/.

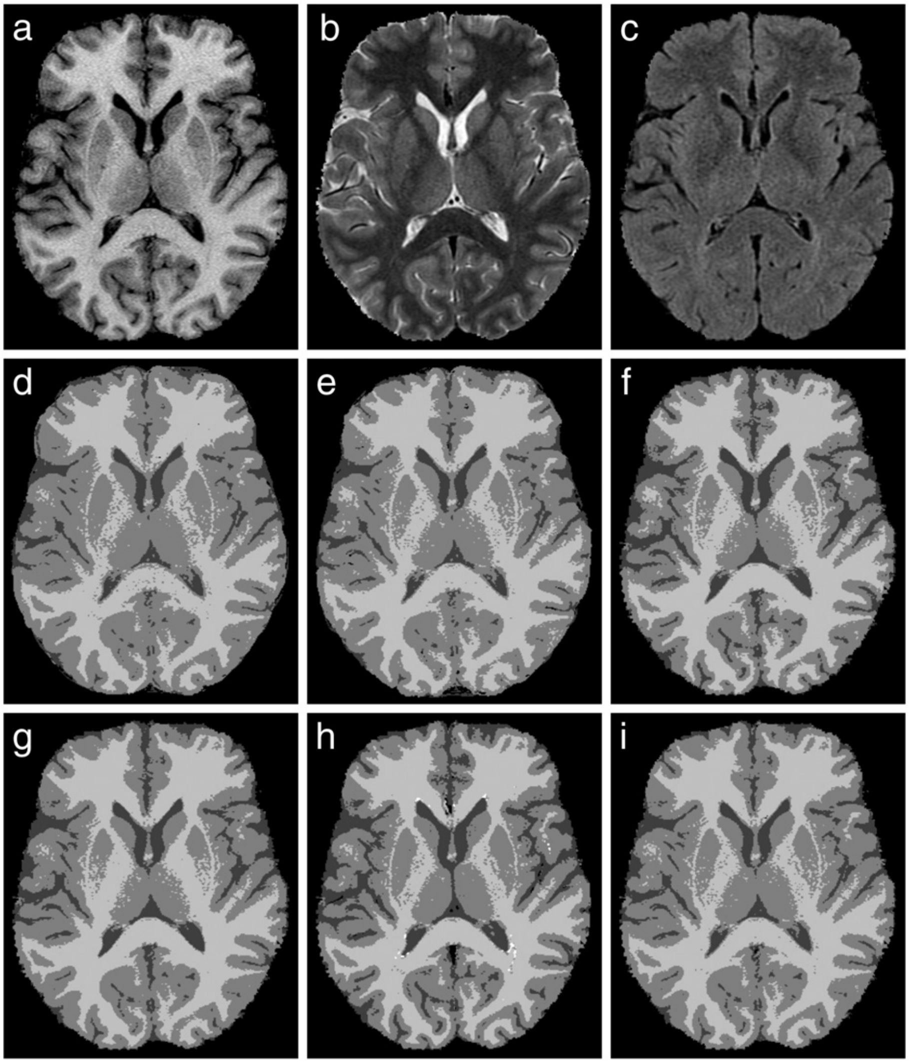

Fig. 2. Axial slice of a subject with low WML load included in the accuracy study: (a) T1w image, (b) PDw image, (c) FLAIR image, (d) manual segmentation based on T1w by observer 1, (e) manual segmentation based on T1w by observer 2, (f) FAST segmentation based on T1w, (g) SPM5 segmentation based on T1w, (h) automatically trained kNN classifier segmentation (WML in white) based on T1w/PDw/FLAIR, and (i) conventional kNN classifier segmentation based on T1w/PDw.

took approximately 15 min on a 64-bit Linux system yielding a final kNN segmentation with similar accuracy. After registering the atlases and creating tissue probability maps, the remaining processing time of the automatically trained kNN method (including WML segmentation) was 22 min on a 32-bit Windows desktop machine. The conventional kNN classifier took approximately 23 min on a 64-bit Linux system.

Accuracy

The results for the accuracy measures are shown in Table 1. The interobserver measures did not differ for the manual segmentation including or excluding WML. Overall, the different segmentation methods showed only small differences in accuracy. Most methods showed high accuracy for all tissue classes, and the SIs were close to the interobserver SI of the manual segmentations. As conformity and SI are closely related, they show the same trends. However, for tissue types with less overlap, the conformity measure shows a better distinction between the segmentation methods. The conventional kNN classifier showed the highest overlap with the manual segmentation for all tissues. SPM5 and the automatically trained classifier using PDw and T1w input images showed the lowest accuracy. The accuracy of the automatically trained classifier improved if the PDw image was left out. For the conventional kNN classifier, there was no clear improvement or decline in accuracy if only the T1w image was used.

Table 2 shows the accuracy of the white matter lesion segmentation by the automatically trained classifier. White matter lesion segmentation is a difficult task and interobserver variability is caused by both differences in detection and in segmented volumes. Due to the small volumes of the lesions, small differences in segmentations have a relatively large effect on the evaluation measures. The SI of the automatically trained kNN method was, however, close to the interobserver SI of the manual segmentations. Other accuracy evaluation measures of the WML segmentation of this method are available in de Boer et al. (2009).

Reproducibility

Table 3 shows the results for the first reproducibility experiment. In general, the reproducibility results of the different segmentation methods only showed small differences. The segmentation methods showed small volume differences and standard deviations and low CoV for all tissue classes indicating high reproducibility. Since the volume differences are expressed as percentage of ICV, the CSF volume differences are equal to the brain volume differences. SPM5 had the lowest reproducibility for brain segmentation, as its standard

Table 1
Accuracy of segmentation methods.

	CSF				Gray matter				White matter				Brain[a]			
	TPF	EF	SI	*C*	TPF	EF	SI	*C*	TPF	EF	SI	*C*	TPF	EF	SI	*C*
FAST	0.88 ± 0.03	0.47 ± 0.20	0.75 ± 0.07	0.33 ± 0.23	0.82 ± 0.02	0.05 ± 0.02	0.88 ± 0.01	0.72 ± 0.03	0.97 ± 0.02	0.10 ± 0.05	0.94 ± 0.02	0.87 ± 0.04	0.93 ± 0.02	0.02 ± 0.01	0.96 ± 0.01	0.91 ± 0.02
SPM5	0.87 ± 0.03	0.49 ± 0.22	0.75 ± 0.07	0.30 ± 0.24	0.83 ± 0.03	0.07 ± 0.03	0.87 ± 0.02	0.70 ± 0.05	0.93 ± 0.03	0.08 ± 0.04	0.93 ± 0.01	0.84 ± 0.03	0.93 ± 0.02	0.01 ± 0.01	0.96 ± 0.01	0.91 ± 0.02
Auto. trained kNN PDw + T1w + FLAIR	0.82 ± 0.07	0.22 ± 0.15	0.81 ± 0.03	0.52 ± 0.09	0.85 ± 0.02	0.10 ± 0.06	0.87 ± 0.03	0.70 ± 0.07	0.97 ± 0.02	0.13 ± 0.06	0.92 ± 0.02	0.83 ± 0.04	0.96 ± 0.01	0.04 ± 0.03	0.96 ± 0.01	0.91 ± 0.02
Auto. trained kNN T1w (+ FLAIR)	0.85 ± 0.04	0.26 ± 0.15	0.81 ± 0.04	0.51 ± 0.13	0.85 ± 0.03	0.07 ± 0.04	0.89 ± 0.02	0.74 ± 0.05	0.98 ± 0.01	0.14 ± 0.06	0.93 ± 0.02	0.84 ± 0.05	0.96 ± 0.01	0.03 ± 0.02	0.96 ± 0.01	0.93 ± 0.01
Conventional kNN[b] PDw + T1w	0.86 ± 0.03	0.24 ± 0.09	0.82 ± 0.04	0.55 ± 0.12	0.87 ± 0.03	0.07 ± 0.03	0.90 ± 0.01	0.78 ± 0.03	0.96 ± 0.02	0.09 ± 0.04	0.94 ± 0.01	0.87 ± 0.02	0.96 ± 0.02	0.03 ± 0.01	0.97 ± 0.01	0.93 ± 0.01
Conventional kNN[b] T1w	0.82 ± 0.03	0.22 ± 0.08	0.81 ± 0.03	0.51 ± 0.10	0.88 ± 0.03	0.09 ± 0.04	0.90 ± 0.01	0.77 ± 0.03	0.95 ± 0.02	0.08 ± 0.04	0.94 ± 0.01	0.87 ± 0.02	0.96 ± 0.01	0.03 ± 0.02	0.96 ± 0.00	0.93 ± 0.01
Interobserver			0.89 ± 0.05	0.74 ± 0.14			0.93 ± 0.02	0.86 ± 0.04			0.95 ± 0.02	0.90 ± 0.04			0.98 ± 0.01	0.96 ± 0.01

Reported values are mean ± standard deviation.
TPF, true positive fraction; EF, extra fraction; SI, similarity index; *C*, conformity.
[a] Brain segmentation is defined as the combined segmentations of gray matter and white matter.
[b] Leave-one-out evaluation.

deviations and CoV were highest. The conventional kNN classifier with PDw and T1w input images had the lowest reproducibility for the other classes. FAST showed the best reproducibility for all tissues. Using only the T1w image as input improved the reproducibility of the conventional kNN classifier segmentation for all tissue types. The reproducibility of the GM and WM segmentations from the automatically trained kNN classifier was slightly worsened by leaving out the PDw image. Table 2 shows the results for both reproducibility experiments for WML.

Table 4 shows the SI and *C* values of the second reproducibility experiment where the segmentation of the second scan is rigidly transformed to the first scan. The mean SI and *C* (± standard deviation) of the rigidly transformed skull-stripping mask of the second scan and the mask of the first scan equaled 0.973 (±0.005) and 0.945 (±0.011), respectively, and can be used as a reference. Although the differences between the segmentation methods were small, the *p*-values in Table 5 show that there is a significant difference between most methods at a $p<0.05$ level. For the automatically trained kNN classifier and conventional kNN classifier results, the paired *t*-tests were performed only with the T1w and PDw input images as these resulted in higher SI than with only the T1w input image. Most methods differed significantly for (almost) all tissue types. Only SPM5 and the automatically trained kNN classifier showed no significant difference in the overlap of their WM segmentations. In addition, the overlap between the brain segmentations of the automatically trained kNN and SPM5 and the automatically trained kNN and the conventional kNN had no significant differences. The automatically trained kNN classifier and SPM5 had the highest overlap between the registered segmentations and FAST, and the conventional kNN classifier, the lowest.

Sample size calculations for longitudinal studies

Finally, we performed an analysis to relate the estimated reproducibilities to sample size calculations for longitudinal studies into brain atrophy. The number of subjects required to find a given effect size (rate of atrophy of brain or tissue type), given a certain power and a level of significance of $\alpha = 0.05$, can be deduced for each method from Fig. 1. The effect size can be calculated using the standard deviation of the volume differences mentioned in Table 3. An example is the design of a longitudinal study with the aim to find a brain atrophy rate of −0.45% per year (Fotenos et al., 2005). If FAST is used for the brain segmentation, the reproducibility experiment shows that the standard deviation of the brain volume difference is 0.55% (Table 3). With an interscan interval of 1 year, the effect size will be $0.45/0.55 = 0.82$. Fig. 1 indicates that the required number of subjects included in the analysis will be 18 for a power of 0.9 or 14 for a power of 0.8. If the conventional kNN classifier method is used the effect size will be $0.45/0.98 = 0.46$. In this case, the required number of subjects will be 52 for a power of 0.9 or 40 for a power of 0.8, as indicated by the graph.

Similar calculations can be made for GM, WM, or WML. Table 6 shows the required number of subjects for several example volume differences of different tissues for a longitudinal study with a 1-year interscan interval, a power of 0.9 and a level of significance of 0.05. Although the segmentation methods only showed small differences in reproducibility, the effect on the required number of subjects can be large. If the aimed volume difference is relatively small compared to the reproducibility standard deviation, the required number of subjects increases dramatically, as is the case for the example gray matter atrophy rates. Since FAST has the lowest reproducibility standard deviation, it requires the least subjects for longitudinal studies on volume differences of all tissue types.

Discussion and conclusion

Comparing accuracy and reproducibility of segmentation methods based on literature can be difficult due to the use of different evaluation

Table 2
Results for white matter lesion segmentations.

	Accuracy experiment				Reproducibility experiment 1		Reproducibility experiment 2	
	TPF	EF	SI	*C*	Δ (%)	CoV (%)	SI	*C*
Auto. trained kNN PDw + T1w + FLAIR	0.79 ± 0.14	0.48 ± 0.60	0.73 ± 0.16	0.09 ± 0.74	0.05 ± 0.12	7.45	0.569 ± 0.137	−0.762 ± 1.106
Auto. trained kNN T1w + FLAIR	0.80 ± 0.15	0.50 ± 0.59	0.72 ± 0.17	0.06 ± 0.77	0.01 ± 0.05	5.87	0.567 ± 0.140	−0.791 ± 1.166
Interobserver			0.75 ± 0.15	0.22 ± 0.58				

TPF, true positive fraction (mean ± standard deviation); EF, extra fraction (mean ± standard deviation); SI, similarity index (mean ± standard deviation); *C*, conformity (mean ± standard deviation); Δ, volume difference (mean ± standard deviation) as percentage of intracranial volume; CoV, coefficient of variation.

Table 3
Reproducibility of segmentation methods.

	CSF		Gray matter		White matter		Brain[a]	
	Δ (%)	CoV (%)	Δ (%)	CoV (%)	Δ (%)	CoV (%)	Δ (%)	CoV (%)
FAST	0.09 ± 0.55	1.48	−0.22 ± 0.49	0.69	0.13 ± 0.38	0.53	−0.09 ± 0.55	0.39
SPM5	−0.20 ± 1.05	2.98	−0.05 ± 1.16	1.50	0.25 ± 0.86	1.29	0.20 ± 1.05	0.72
Auto. trained kNN PDw + T1w + FLAIR	−0.14 ± 0.69	2.26	−0.03 ± 1.02	1.30	0.14 ± 0.86	1.42	0.14 ± 0.69	0.52
Auto. trained kNN T1w (+ FLAIR)	0.29 ± 0.64	2.27	−0.57 ± 1.17	1.50	0.27 ± 1.02	1.53	−0.29 ± 0.64	0.50
Conventional kNN PDw + T1w	0.22 ± 0.98	3.53	−0.33 ± 1.61	2.13	0.11 ± 1.53	2.29	−0.22 ± 0.98	0.62
Conventional kNN T1w	0.07 ± 0.55	2.16	−0.03 ± 0.67	0.74	−0.04 ± 0.67	0.92	−0.07 ± 0.55	0.37

Δ, volume difference (mean ± standard deviation) as percentage of intracranial volume; CoV, coefficient of variation.
[a] Brain volume is defined as the combined volumes of gray matter and white matter.

measures, different manual segmentation protocols, and, most importantly, different imaging data. We compared both the accuracy and the reproducibility of several previously proposed brain tissue segmentation methods on the same data sets. All scans were made with the same acquisition protocol on the same scanner, without any scanner updates during the course of the study. The rescans, for the reproducibility experiment, were made on average 18.5 days after the first scan. This period ensures that the natural variations in, for example, fluid balance in the brain are captured, while no significant brain changes take place. In general, all segmentation methods showed good accuracy and reproducibility. There were, however, small differences between the various methods. The conventional kNN classifier method performed best in the accuracy experiment and worst in the reproducibility experiment. FAST showed the best reproducibility, but its accuracy was relatively low for CSF and GM.

Despite its high accuracy, the conventional kNN classifier method has several weaknesses. Due to the fixed training set, any changes in image acquisition require a laborious training stage. Furthermore, the MRI contrast between GM and WM tissues changes with age (Cho et al., 1997) and might therefore result in an age-related bias in tissue volumes obtained with the conventional kNN classifier.

The accuracy experiment compared the automatic segmentations to manual segmentations based on the T1w scan. This might induce a bias for the methods, as they are based on the same T1w image. The automatically trained kNN classifier results support this hypothesis as the accuracy increased when the PDw image was left out of the analysis.

Table 4
Measured overlap of the rigidly registered segmentation of scan 2 and the segmentation of scan 1 of the reproducibility data set.

	CSF		Gray matter		White matter		Brain[a]	
	SI	*C*	SI	*C*	SI	*C*	SI	*C*
FAST	0.796 ± 0.024	0.486 ± 0.077	0.845 ± 0.015	0.631 ± 0.044	0.904 ± 0.011	0.788 ± 0.027	0.960 ± 0.005	0.916 ± 0.012
SPM5	0.802 ± 0.024	0.504 ± 0.076	0.868 ± 0.013	0.695 ± 0.036	0.909 ± 0.009	0.800 ± 0.022	0.964 ± 0.005	0.925 ± 0.011
Auto. trained kNN PDw + T1w (+ FLAIR)	0.820 ± 0.022	0.559 ± 0.067	0.865 ± 0.012	0.687 ± 0.032	0.909 ± 0.010	0.800 ± 0.025	0.961 ± 0.005	0.918 ± 0.011
Auto. trained kNN T1w (+ FLAIR)	0.773 ± 0.026	0.409 ± 0.088	0.844 ± 0.016	0.628 ± 0.047	0.894 ± 0.013	0.762 ± 0.033	0.958 ± 0.006	0.912 ± 0.012
Conventional kNN PDw + T1w	0.783 ± 0.029	0.443 ± 0.097	0.853 ± 0.013	0.654 ± 0.036	0.896 ± 0.013	0.768 ± 0.032	0.961 ± 0.005	0.919 ± 0.011
Conventional kNN T1w	0.752 ± 0.029	0.337 ± 0.106	0.846 ± 0.013	0.636 ± 0.035	0.890 ± 0.013	0.752 ± 0.034	0.959 ± 0.005	0.915 ± 0.011

Reported values are mean ± standard deviation.
SI, similarity index; *C*, conformity.
[a] Brain segmentation is defined as the combined segmentations of gray matter and white matter.

Table 5
Results of two-tailed paired *t*-tests ($df = 29$) comparing the SI values of the registered segmentation of scan 2 and the segmentation of scan 1 of the reproducibility data set.

	CSF		Gray matter		White matter		Brain[a]	
	t-value	*p*-value	*t*-value	*p*-value	*t*-value	*p*-value	*t*-value	*p*-value
FAST vs. SPM5	−12.6	<**0.001**	−28.6	<**0.001**	−8.4	<**0.001**	−22.6	<**0.001**
FAST vs. Auto. trained kNN	−14.2	<**0.001**	−13.3	<**0.001**	−5.5	<**0.001**	−1.8	0.090
FAST vs. Conventional kNN	10.3	<**0.001**	−7.7	<**0.001**	14.7	<**0.001**	−5.3	<**0.001**
SPM5 vs. Auto. trained kNN	−10.4	<**0.001**	2.2	**0.038**	0.1	0.913	5.6	<**0.001**
SPM5 vs. Conventional kNN	13.6	<**0.001**	13.9	<**0.001**	13.7	<**0.001**	10.0	<**0.001**
Auto. trained kNN vs. Conventional kNN	17.5	<**0.001**	8.0	<**0.001**	11.0	<**0.001**	−0.6	0.539

Bold indicates significance at a $p < 0.05$ level.
[a] Brain segmentation is defined as the combined segmentations of gray matter and white matter.

Table 6
Example estimations of the required number of subjects of a longitudinal study (1-year interscan interval, power of 0.9, level of significance 0.05) aimed to find selected volume differences.

	Gray matter			White matter			Brain[a]			White matter lesions		
Volume difference[b] (% of ICV per year)	−0.05	−0.10	−0.15	−0.3	−0.4	−0.5	−0.3	−0.4	−0.5	0.05[c]	0.06[c]	0.07[c]
FAST	1012	255	115	19	12	9	38	22	15			
SPM5	5658	1416	631	89	51	34	131	75	49			
Auto. trained kNN PDw + T1w (+ FLAIR)	4375	1096	488	89	51	34	58	34	23	63	44	33
Auto. trained kNN T1w (+ FLAIR)	5756	1441	642	124	71	46	50	29	20	13	10	8
Conventional kNN PDw + T1w	10,897	2725	1213	276	156	101	115	66	43			
Conventional kNN T1w	1889	474	212	55	32	21	38	22	15			

[a] Brain volume is defined as the combined volumes of gray matter and white matter.

[b] Volume differences are based on cross-sectional and a limited number of longitudinal studies, e.g., DeCarli et al. (2005), Fotenos et al. (2005), Ge et al. (2002), and Ikram et al. (2008).

[c] Natural log-transformed.

In our experiments, we performed segmentations on T1w and PDw data or on T1w data only. It is also possible to segment the brain tissues based on only the PDw image. The contrast between GM and WM, however, is lower on the PDw image than on the T1w image. Furthermore, PDw sequences are less commonly used nowadays. This experiment was therefore not performed in this study.

The reproducibility experiment is influenced by the reproducibility of the image acquisition. A difference between the segmentations of the two sequential scans must therefore be expected, no matter how reproducible the segmentation method is. Since all methods are tested on the same data, the image acquisition variation is the same for every method.

In some studies, the correlation coefficient between segmented volumes at two time points is used as evaluation measure for reproducibility (e.g., Cardenas et al., 2001; Harris et al., 1999; Wang et al., 1998). We decided not to use this measure, as it depends on the dispersion of the segmented volume measurements. Instead, the coefficient of variation was used. Different definitions of CoV are used in the literature. We used the same definition as Cardenas et al. (2001) and Wang et al. (1998), but they used absolute volumes instead of fractional volumes for their calculations. The results, therefore, cannot be compared.

The reproducibility experiment measuring the overlap between the registered segmentation of the second scan and the segmentation of the first scan showed less overlap for CSF and GM compared to WM. As CSF and GM are not as compact as WM, they have more boundary voxels. Since boundary voxels are more difficult to segment due to partial volume effects, a lower overlap can be expected, as is supported by the accuracy evaluation. In addition, this reproducibility evaluation suffers from registration errors. Especially for CSF and GM this is a disadvantage, since they are on the outside of the brain, close to the border of the skull-stripping mask. Errors in mask registration will therefore mainly be reflected in less CSF and GM overlap.

The subjects in this study are participants of a population-based cohort study of the elderly. As white matter lesions are commonly found in elderly subjects (de Leeuw et al., 2001), they are also present in the subjects used for the evaluation. Three of the subjects with manual segmentations had a low WML load with a mean (± standard deviation) of the manual segmentations of 1.54 (± 1.09) ml and a mean (± standard deviation) of the automatic segmentations of 1.96 (± 0.31) ml. The other three accuracy subjects had a high WML of 15.56 (± 4.32) ml according to the manual segmentations and 13.63 (± 3.34) ml according to the automatic segmentations. The 30 subjects for the reproducibility experiment were picked randomly. These subjects had no manual segmentations but according to the automatic segmentations of their first scan, their mean (± standard deviation) WML load was 5.00 (± 6.84) ml with a range of 0.97 – 34.81 ml.

White matter lesions can take up a considerable part of the white matter, so excluding these WMLs might improve the WM segmentation. Furthermore, WMLs may be of interest themselves as they are associated with cognitive decline and increased risk of stroke and dementia. FSL 4.1 and SPM5 have no automatic WML segmentation method included. FAST v4.1 and the conventional kNN classifier are capable of multimodal segmentation and it is possible to add an extra segmentation class. There are, however, no processing steps that ensure that the additional class will contain only WMLs, so other components might get the same label. SPM5 is not capable of multimodal brain tissue segmentation.

Besides the automatic WML segmentation method evaluated in this study, several automatic WML segmentation methods have been proposed. Some of these studies evaluated the reproducibility of their method (Admiraal-Behloul et al., 2005; Jack et al., 2001). Both of these studies performed the rescan with repositioning within several minutes of the first scan. Admiraal-Behloul et al. reported only intraclass correlation coefficients and no CoVs. Jack et al. reported CoVs based on absolute volumes and subjects with larger WML volumes and are therefore not comparable to our results.

Since reproducibility depends on the image acquisition variables, a change in these variables might influence the required number of subjects in the design of a longitudinal study. The sample size estimates in this study can, in that case, be used as example calculations for a new reproducibility study. In addition, they demonstrate the influence of the choice of the segmentation method on the required number of subjects.

The small differences in reproducibility of the different segmentation methods have a rather large effect on the required number of subjects in a study that aims to detect a certain longitudinal change in tissue volume. Especially if the aim of the study is to find a tissue volume difference that is relatively small compared to the reproducibility standard deviation of the segmentation method, the choice of the segmentation method has a large influence on the required number of subjects. Segmentation methods with lower reproducibility standard deviations require fewer subjects to find the same tissue volume difference in a longitudinal study than methods with higher reproducibility standard deviations.

Currently, data on the association between age and gray matter volume are inconsistent. Several studies report a decline in GM volume from early adulthood onwards, e.g., Fotenos et al. (2005) and Ge et al. (2002), while others find no such decline, e.g., Ikram et al. (2008). If there is an association between GM volume and age, its rate is likely to be small. Consequently, our power calculations show that the required number of subjects to find a small GM atrophy rate in a longitudinal study would be very large.

In conclusion, we compared the accuracy and reproducibility of four known brain tissue segmentation methods. Overall, the accuracy and reproducibility were good, and there were only small differences between the methods. The small differences in reproducibility do, however, have a relatively large effect on the required number of subjects in the design of a longitudinal study with sufficient power.

Acknowledgments

The Rotterdam Study is supported by the Erasmus MC and Erasmus University Rotterdam; the Netherlands Organization for Scientific Research (NWO); the Netherlands Organization for Health Research and Development (ZonMw); The Research Institute for Diseases in the Elderly (RIDE); the Ministry of Education, Culture and Science, the Ministry of Health, Welfare and Sports; the European Commission (DG XII); and the Municipality of Rotterdam. This study was further financially supported by the Netherlands Organization for Scientific Research (NWO) grants 948-00-010 and 918-46-615.

This work benefited from the use of Elastix, software developed by Stefan Klein and Marius Staring, originally developed at the Image Science Institute (ISI), on funding granted by the Netherlands Organization for Scientific Research (NWO).

Appendix A. Accuracy and reproducibility results using original preprocessing

Since most researchers use the publicly available methods with their default preprocessing, we also report the results of our experiments with the intensity non-uniformity correction and skull-stripping as provided by the methods FAST and SPM5. FAST intrinsically corrects for spatial intensity variations, and the input image is skull-stripping using FSL's Brain Extraction Tool (BET) (Smith, 2002) version 2.1. Contrary to the skull-stripping mask used in this study, the BET mask does include the cerebellum. SPM5 combines the bias field correction with the image registration and tissue classification. As mentioned before, SPM5 applies no skull-stripping to its CSF segmentation. The cerebellum is included in the GM and WM segmentations. In an attempt to obtain an SPM5 CSF segmentation that can be used for our experiments, we thresholded the CSF segmentation. Visual inspection showed that there is no optimal threshold that excludes all non-CSF components while maintaining the actual CSF. We chose a CSF threshold of 50%.

Fig. A1 shows the resulting segmentations of the same axial slice as Fig. 2. Since the manual segmentations do not include the cerebellum, the accuracy measures were only calculated within the manual mask. This manual mask is defined as all voxels with a CSF, GM, or WM label in the manual segmentation. Contrary to the accuracy experiment mentioned before, this evaluation has a bias since the automatic segmentations can have no false positives outside the manual mask. Table A1 shows the results of this accuracy experiment, including the mean interobserver values determined within the manual mask. The low accuracy of the FAST CSF segmentation is partly due to differences in masking. The BET mask

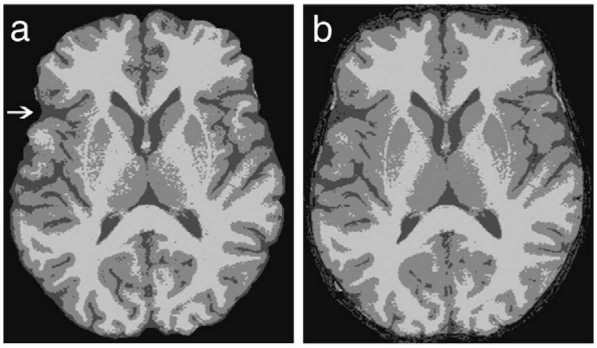

Fig. A1. Axial slice of a subject with low WML load included in the accuracy study: (a) FAST segmentation based on T1w with default preprocessing and (b) SPM5 segmentation based on T1w with default preprocessing and probabilistic CSF threshold of 50%.

Table A1
Accuracy of segmentation methods with default preprocessing.

	CSF				Gray matter				White matter				Brain[a]			
	TPF	EF	SI	C	TPF	EF	SI	C	TPF	EF	SI	C	TPF	EF	SI	C
FAST	0.72 ± 0.08	0.34 ± 0.16	0.70 ± 0.07	0.12 ± 0.31	0.79 ± 0.02	0.01 ± 0.01	0.88 ± 0.01	0.72 ± 0.04	0.98 ± 0.01	0.11 ± 0.04	0.94 ± 0.02	0.87 ± 0.05	0.92 ± 0.02	0.00 ± 0.00	0.96 ± 0.01	0.92 ± 0.02
SPM5 50% CSF-threshold	0.93 ± 0.05	0.19 ± 0.13	0.88 ± 0.04	0.72 ± 0.11	0.88 ± 0.03	0.06 ± 0.04	0.91 ± 0.01	0.80 ± 0.04	0.94 ± 0.02	0.09 ± 0.06	0.93 ± 0.02	0.84 ± 0.05	0.97 ± 0.02	0.01 ± 0.02	0.98 ± 0.00	0.96 ± 0.01
Interobserver			0.93 ± 0.05	0.83 ± 0.13			0.95 ± 0.02	0.89 ± 0.04			0.95 ± 0.02	0.90 ± 0.04			0.99 ± 0.01	0.98 ± 0.01

Evaluation is performed within the manual mask.
Reported values are mean ± standard deviation.
TPF, true positive fraction; EF, extra fraction; SI, similarity index, C: conformity.
[a] Brain segmentation is defined as the combined segmentations of gray matter and white matter.

Table A2
Reproducibility of segmentation methods with default preprocessing (including cerebellum).

	CSF		Gray matter		White matter		Brain[a]	
	Δ (%)	CoV (%)	Δ (%)	CoV (%)	Δ (%)	CoV (%)	Δ (%)	CoV (%)
FAST	0.22 ± 0.52	1.35	−0.36 ± 0.57	0.87	0.14 ± 0.41	0.64	−0.22 ± 0.52	0.32
SPM5 50% CSF-threshold	0.18 ± 1.50	3.09	−0.53 ± 1.34	1.22	0.35 ± 0.99	1.54	−0.18 ± 1.50	0.76

Δ, volume difference (mean ± standard deviation) as percentage of intracranial volume; CoV, coefficient of variation.

[a] Brain volume is defined as the combined volumes of gray matter and white matter.

Table A3
Measured overlap of the rigidly registered segmentation of scan 2 and the segmentation of scan 1.

	CSF		Gray matter		White matter		Brain[a]	
	SI	*C*	SI	*C*	SI	*C*	SI	*C*
FAST	0.765 ± 0.021	0.385 ± 0.073	0.838 ± 0.017	0.614 ± 0.048	0.893 ± 0.011	0.759 ± 0.027	0.961 ± 0.005	0.919 ± 0.010
SPM5 50% CSF-threshold	0.770 ± 0.029	0.399 ± 0.099	0.880 ± 0.017	0.726 ± 0.047	0.895 ± 0.011	0.764 ± 0.028	0.966 ± 0.007	0.929 ± 0.016

Segmentations obtained with default preprocessing (including cerebellum). Reported values are mean ± standard deviation.
SI, similarity index; *C*, conformity.

[a] Brain segmentation is defined as the combined segmentations of gray matter and white matter.

used by FAST excludes more sulcal CSF than the manual segmentations. An example of this is indicated by the arrow in Fig. A1a.

Tables A2 and A3 show the results of both reproducibility experiments. The tissue volume difference standard deviations in Table A2 can be used for sample size calculations.

Keep in mind that due to differences in masking and in the accuracy evaluation, the results in this appendix cannot be compared to the values in the Results section. Also, the results in Tables A2 and A3 cannot be compared between methods due to differences in masking.

References

Admiraal-Behloul, F., van den Heuvel, D.M., Olofsen, H., van Osch, M.J., van der Grond, J., van Buchem, M.A., Reiber, J.H., 2005. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. Neuroimage 28, 607–617.

Amato, U., Larobina, M., Antoniadis, A., Alfano, B., 2003. Segmentation of magnetic resonance brain images through discriminant analysis. J. Neurosci. Methods 131, 65–74.

Anbeek, P., Vincken, K.L., van Bochove, G.S., van Osch, M.J., van der Grond, J., 2005. Probabilistic segmentation of brain tissue in MR imaging. Neuroimage 27, 795–804.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. Neuroimage 11, 805–821.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Awate, S.P., Tasdizen, T., Foster, N., Whitaker, R.T., 2006. Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. Med. Image Anal. 10, 726–739.

Cardenas, V.A., Ezekiel, F., Sclafani, V.D., Gomberg, B., Fein, G., 2001. Reliability of tissue volumes and their spatial distribution for segmented magnetic resonance images. Psychiatry Res. 106, 193–205.

Chang, H.H., Zhuang, A.H., Valentino, D.J., Chu, W.C., 2009. Performance measure characterization for evaluating neuroimage segmentation algorithms. Neuroimage 47, 122–135.

Chard, D.T., Parker, G.J.M., Griffin, C.M.B., Thompson, A.J., Miller, D.H., 2002. The reproducibility and sensitivity of brain tissue volume measurements derived from an SPM-based segmentation methodology. J. Magn. Reson. Imaging 15, 259–267.

Cho, S., Jones, D., Reddick, W.E., Ogg, R.J., Steen, R.G., 1997. Establishing norms for age-related changes in proton T1 of human brain tissue in vivo. Magn. Reson. Imaging 15, 1133–1143.

Clark, K.A., Woods, R.P., Rottenberg, D.A., Toga, A.W., Mazziotta, J.C., 2006. Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images. Neuroimage 29, 185–202.

Cocosco, C.A., Zijdenbos, A.P., Evans, A.C., 2003. A fully automatic and robust brain MRI tissue classification method. Med. Image Anal. 7, 513–527.

de Boer, R., Vrooman, H.A., van der Lijn, F., Vernooij, M.W., Ikram, M.A., van der Lugt, A., Breteler, M.M., Niessen, W.J., 2009. White matter lesion extension to automatic brain tissue segmentation on MRI. Neuroimage 45, 1151–1161.

de Leeuw, F.E., de Groot, J.C., Achten, E., Oudkerk, M., Ramos, L.M., Heijboer, R., Hofman, A., Jolles, J., van Gijn, J., Breteler, M.M., 2001. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. The Rotterdam Scan Study. J. Neurol. Neurosurg. Psychiatry 70, 9–14.

DeCarli, C., Massaro, J., Harvey, D., Hald, J., Tullberg, M., Au, R., Beiser, A., D'Agostino, R., Wolf, P.A., 2005. Measures of brain morphology and infarction in the framingham heart study: establishing what is normal. Neurobiol. Aging 26, 491–510.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Faul, F., Erdfelder, E., Lang, A.G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39, 175–191.

Fotenos, A.F., Snyder, A.Z., Girton, L.E., Morris, J.C., Buckner, R.L., 2005. Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology 64, 1032–1039.

Fox, N.C., Cousens, S., Scahill, R., Harvey, R.J., Rossor, M.N., 2000. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: power calculations and estimates of sample size to detect treatment effects. Arch. Neurol. 57, 339–344.

Ge, Y., Grossman, R.I., Babb, J.S., Rabin, M.L., Mannon, L.J., Kolson, D.L., 2002. Age-related total gray matter and white matter changes in normal adult brain: Part I. Volumetric MR imaging analysis. AJNR Am. J. Neuroradiol. 23, 1327–1333.

Harris, G., Andreasen, N.C., Cizadlo, T., Bailey, J.M., Bockholt, H.J., Magnotta, V.A., Arndt, S., 1999. Improving tissue classification in MRI: a three-dimensional multispectral discriminant analysis method with automated training class selection. J. Comput. Assist. Tomogr. 23, 144–154.

Hofman, A., Breteler, M.M., van Duijn, C.M., Janssen, H.L., Krestin, G.P., Kuipers, E.J., Stricker, B.H., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R., Witteman, J.C., 2009. The Rotterdam Study: 2010 objectives and design update. Eur. J. Epidemiol. 24, 553–572.

Ikram, M.A., Vrooman, H.A., Vernooij, M.W., van der Lijn, F., Hofman, A., van der Lugt, A., Niessen, W.J., Breteler, M.M., 2008. Brain tissue volumes in the general elderly population. The Rotterdam Scan Study. Neurobiol. Aging 29, 882–890.

Jack Jr., C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., Erickson, B.J., 2001. FLAIR histogram segmentation for measurement of leukoaraiosis volume. J. Magn. Reson. Imaging 14, 668–676.

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., Pacheco, J., Albert, M., Killiany, R., Blacker, D., Maguire, P., Rosas, D., Makris, N., Gollub, R., Dale, A., Dickerson, B.C., Fischl, B., 2009. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. Neuroimage 46, 177–192.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29, 196–205.

Kovacevic, N., Lobaugh, N.J., Bronskill, M.J., Levine, B., Feinstein, A., Black, S.E., 2002. A robust method for extraction and automatic segmentation of brain images. Neuroimage 17, 1087–1100.

Lemieux, L., Hagemann, G., Krakow, K., Woermann, F.G., 1999. Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. Magn. Reson. Med. 42, 127–135.

Lemieux, L., Hammers, A., Mackinnon, T., Liu, R.S.N., 2003. Automatic segmentation of the brain and intracranial cerebrospinal fluid in T1-weighted volume MRI scans of the head, and its application to serial cerebral and intracranial volumetry. Magn. Reson. Med. 49, 872–884.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging 18, 712–721.

Schott, J.M., Frost, C., Whitwell, J.L., Macmanus, D.G., Boyes, R.G., Rossor, M.N., Fox, N.C., 2006. Combining short interval MRI in Alzheimer's disease: implications for therapeutic trials. J. Neurol. 253, 1147–1153.

Shuter, B., Yeh, I.B., Graham, S., Au, C., Wang, S.-C., 2008. Reproducibility of brain tissue volumes in longitudinal studies: effects of changes in signal-to-noise ratio and scanner software. Neuroimage 41, 371–379.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (Suppl 1), S208–219.

Song, Z., Tustison, N., Avants, B., Gee, J.C., 2006. Integrated graph cuts for brain MRI segmentation. Med. Image Comput. Comput. Assist. Interv. 9, 831–838.

van Leemput, K., a.M. F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Transact. Med. Imaging 18, 897–908.

Vrooman, H.A., Cocosco, C.A., van der Lijn, F., Stokking, R., Ikram, M.A., Vernooij, M.W., Breteler, M.M., Niessen, W.J., 2007. Multi-spectral brain tissue segmentation using automatically trained *k*-nearest-neighbor classification. Neuroimage 37, 71–81.

Wang, D., Galloway, G.J., de Zubicaray, G.I., Rose, S.E., Chalk, J.B., Doddrell, D.M., Semple, J., 1998. A reproducible method for automated extraction of brain volumes from 3D human head MR images. J. Magn. Reson. Imaging 8, 480–486.

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

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imaging 13, 716–724.