A comparison of MR based segmentation methods for measuring brain atrophy progression

Jeroen de Bresser a,b,⁎, Marileen P. Portegies b, Alexander Leemans a, Geert Jan Biessels b, L. Jaap Kappelle b, Max A. Viergever a

a Image Sciences Institute, University Medical Center Utrecht, The Netherlands
b Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, The Netherlands

Abstract

Automated brain segmentation methods with a good precision and accuracy are required to detect subtle changes in brain volumes over time in clinical applications. However, the ability of established methods such as SIENA, US and kNN to estimate brain volume change have not been compared on the same data, nor been evaluated with ground-truth manual segmentations. We compared measurements of brain volume change between SIENA, US and kNN in terms of precision (repeatability) and accuracy (ground-truth) using one baseline and two repeated follow-up 1.5 T MRI scans after 4 years of 10 subjects. The coefficient of repeatability (brain volume/volume change) was larger for US (29.6 cc/2.84%) than for kNN (4.9 cc/0.31%) and SIENA (−0.92%). In terms of absolute brain volume measurements US and kNN showed good correlation with the manual segmentations and with each other (all Spearman’s correlation coefficients ρ > 0.96; all p < 0.001). Concerning brain volume changes, SIENA showed a good (ρ = 0.82; p = 0.004), kNN a moderate (ρ = 0.60; p = 0.067) and US a weak (ρ = 0.50; p = 0.138) correlation with the manual segmentations. For measurements of volume change, SIENA-US (mean correlation coefficient and p-value: ρ = 0.28; p = 0.442) and US-kNN (ρ = 0.17; p = 0.641) showed a weak correlation, but correlation was fairly good for kNN-SIENA (ρ = 0.65; p = 0.048). In conclusion, US and kNN showed a good precision, accuracy and comparability for brain volume measurements. For measurements of volume change, SIENA showed the best performance. kNN is a good alternative if volume change measurements of other brain structures are required.

© 2010 Elsevier Inc. All rights reserved.

Introduction

Brain volumes can be measured to assess atrophy due to normal ageing, or as a marker of disease progression in clinical studies of different pathologic conditions, like dementia (Fox et al., 1996; Karas et al., 2003), or in other diseases with a less marked loss of brain volume, such as multiple sclerosis and type 2 diabetes mellitus (de Bresser et al., 2010; Filippi et al., 2004; Jasperse et al., 2007b; Jongen et al., 2007).

There are a number of automated methods that can estimate brain atrophy from magnetic resonance images; they can be subdivided in methods measuring brain volume and in methods measuring brain volume change over time. Methods measuring brain volume require only a single scan of a subject and generally measure the absolute brain volume. By comparing two scans such methods can also assess brain volume change. By contrast, methods measuring changes in brain volume typically measure a percentage difference in brain volume between two or more scans of the same subject, through a direct comparison of these scans without the need of calculating the absolute brain volume.

Thus far, methods measuring brain volume showed a lower precision for assessment of volume change than methods specifically designed to measure volume change (Smith et al., 2007). High precision and accuracy are important to detect subtle differences in brain volume changes between patient groups. Studies comparing different methods are few in number (i.e. Lee and Prohovnik, 2008; Smith et al., 2007), and only few methods have been evaluated with the reference standard of manual segmentation to test accuracy in addition to precision (i.e. Anbeek et al., 2005).

Structural Image Evaluation, using Normalization, of Atrophy (SIENA), Unified Segmentation (US) and k-Nearest Neighbor-based probabilistic segmentation (kNN) are well-established and widely used methods to measure brain atrophy (Anbeek et al., 2005; Ashburner and Friston, 2005; Smith et al., 2002). Although precision of SIENA has been assessed by repeated imaging with patient repositioning, accuracy has not been evaluated with manual segmentations (Smith et al., 2001, 2002). In addition, accuracy of brain volume measurements by US was tested only on simulated data (Ashburner and Friston, 2005). Furthermore, kNN has been compared

⁎ Corresponding author. University Medical Center Utrecht, Department of Neurology G03.228, P.O. Box 85500, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands. Fax: +31 30 2513399.
E-mail address: j.deBresser@umcutrecht.nl (J. de Bresser).

1053-8119/– see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.09.060
with manual segmentations, but only evaluating accuracy for absolute brain volume measurements, not for brain volume change assessment (Anbeek et al., 2005). Finally, no studies have compared these methods on the same data.

In this study, we compared measurements of brain volume change between SIENA, US and kNN, and assessed potential differences in precision (via repeatability) and accuracy (compared with ground-truth manual segmentations). The methods were applied as described in the respective validation studies in order not to favor one method (Anbeek et al., 2005; Ashburner and Friston, 2005; Smith et al., 2002).

Materials and methods

Data

Ten subjects were included in this study (mean age ± SD = 70 years ± 6). Subjects were part of the Utrecht Diabetic Encephalopathy Study (de Bresser et al., 2010; van den Berg et al., 2010). This study was approved by the local ethics committee of the University Medical Center Utrecht and all participants signed an informed consent form. Baseline (BL) magnetic resonance imaging (MRI) scans and follow-up scans 4 years later were acquired on the same 1.5 T Philips MR system. At follow-up, scanning was performed twice (FU1/FU2) with patient repositioning in between. All three scans were made by a standardized scanning protocol which consisted of an axial T1 (TR/TE = 234/2 ms), T2 (TR/TE = 2200/100 ms), proton density (PD) (TR/TE = 2200/11 ms), inversion recovery (IR) (TR/TE/T1 = 2919/22/410 ms) and fluid attenuated inversion recovery (FLAIR) (TR/TE/T1 = 6000/100/2000 ms); all scans with 38 contiguous slices and 0.90 × 0.90 × 4.00 mm voxels.

SIENA

SIENA is a fully automated method that calculates brain volume change between two scans (Smith et al., 2001, 2002). SIENA as part of the FMRIB Software Library (FSL) 4.0 was used on the T1 images to calculate differences between BL and FU1, BL and FU2, and FU1 and FU2 (Smith et al., 2004). Brain masks were created with the Brain Extraction Tool (BET) (Smith, 2002), the resulting masked brain images were aligned to each other (Jenkinson et al., 2002; Jenkinson and Smith, 2001) and both data sets were resampled into the space halfway between the two. Brain/non-brain edge points were detected by tissue type segmentation (Zhang et al., 2001), and used to calculate perpendicular edge displacement between the two scans. The resulting mean edge displacement was converted into a global estimate of percentage brain volume change between the scans. BET parameters and other parameters were optimized for our data by qualitatively evaluating output. Resulting BET masks were of good quality and no problems due to non-uniformities were detected. All resulting output images were visually inspected and all output was considered to be of good quality.

kNN

kNN is a fully automated method that can calculate absolute brain volume (Anbeek et al., 2005). This method is based on manually segmented training data, which consisted of IR and FLAIR images of 10 subjects from another cohort without known neurodegenerative disease, comparable in age to the subjects in our study and made by an identical scanning protocol to the images in our study (Anbeek et al., 2005; de Bresser et al., 2010). A percentage of 40% training samples was first randomly selected from the training data, and then fixed for use in this study. For one scan, the T1, T2, PD and IR were rigidly registered to the FLAIR image with the Elastix tool (Klein et al., 2010). Scan inhomogeneities were corrected by a shading correction algorithm (Likar et al., 2001). A BL brain mask was created by using a k-means clustering algorithm with eight clusters (Jongen et al., 2007). FU masks were created by rigidly registering the BL to the FU FLAIR of the same subject and by using the resulting transform parameters to transform the BL brain mask (de Bresser et al., 2010). The uncorrected FLAIR images were multiplied voxelwise by the binary brain mask, followed by a shading correction (Likar et al., 2001). IR and FLAIR images were used for measurements of brain volume, by applying the k-nearest neighbor classification technique that builds a feature space from spatial information and voxel intensities of manually segmented training data (Anbeek et al., 2005). For each voxel to be classified, it determines the k (=100) nearest neighbors from the training data and calculates a probability that a voxel is of a certain tissue type. Parameters were optimized for our data by qualitative evaluation of the output. All resulting segmentations of brain volumes were visually inspected and considered to be of good quality. Brain volumes for each subject were calculated by multiplying the probabilities by the voxel volume, and volume changes were calculated by taking the difference between normalized BL and FU brain volumes. Depending on the types of tissues classified in the training data other volumes can also be determined separately, including gray matter, white matter, CSF, lateral ventricular, and white matter hyperintensity volume.

Manual segmentations

In the implementation of the methods described in the respective validation studies, SIENA and US use T1 images and kNN mainly uses IR images to determine the brain CSF border (Anbeek et al., 2005; Ashburner and Friston, 2005; Smith et al., 2002). Therefore, the intra-rater reliability of both sequences for manual segmentation of brain volume was first determined voxelwise and as a mean difference, to select the optimal sequence for manual segmentation. Brain volumes were manually segmented twice on IR images and twice on T1 images by a trained researcher (MP) blinded for subject number, on ten randomly selected half slices from the data in this study. The intra-rater reliability of these manual segmentations for the T1 and IR images showed a sensitivity of 0.98 and 0.97, specificity of 0.99 and 0.99, similarity index of 0.96 and 0.96 and a mean difference ± SD of 0.53 cc ± 0.53 and 0.19 cc ± 0.69, respectively (Dice, 1945). Because of the slightly superior intra-rater reliability, the IR images were used for further manual segmentations. Brain volume was manually segmented by two trained researchers (JB, MP) on six additional randomly selected slices, which showed a good inter-rater reliability with a sensitivity of 0.95, specificity of 0.99 and similarity index of 0.96 (Dice, 1945).

Finally, total brain volumes of all BL and FU1 scans were manually segmented by the same trained researcher (MP). Intra-cranial nerves, such as the optic nerve, and large intra-cranial vessels were not included in the brain segmentations. All segmentations were inspected by a neurologist experienced in neuroimaging (GB) and segmentations were adjusted accordingly. Brain volume change was calculated by taking the difference of the BL and FU1 measurements.
**Statistical analysis**

Nonparametric statistics were used because of the limited number of subjects.

The first step in the analysis was to determine precision for each of the methods, i.e. the extent to which repeated measurements under unchanged conditions show the same results. Precision was determined for each method by testing for differences of brain volume and volume change measurements between repeated imaging, with Wilcoxon signed rank tests. A coefficient of repeatability (i.e. the variation in measurements performed by the same method under the same conditions) was calculated to evaluate variation from the mean in scan–rescan measurements (Bland and Altman, 1986).

The second step was to determine accuracy of the methods, i.e. the extent to which the measurements correspond with the reference standard of manual segmentation. To determine the accuracy of brain volume and volume change measurements, differences between manual segmentations and the methods were tested with Wilcoxon signed rank tests, correlations were tested with Spearman’s rank correlation tests, and Bland–Altman plots were used for detailed comparison of brain volume change measurements (Bland and Altman, 1986).

The third step was to compare the measurements between the different methods. Wilcoxon signed rank tests and Spearman’s rank correlation tests were used to determine differences and correlations in brain volume and volume change measurements between the methods. Repeated–measurements Bland–Altman plots with corrected limits of agreement were used to compare measurements of brain volume change between methods, taking into account the repeated measurements (Bland and Altman, 1986).

**Results**

**Precision of the methods**

The individual brain volume and volume change measurements for all subjects and each method are shown in Fig. 1. Examples of the measurements by US, kNN and SIENA for one subject are shown in Fig. 2. In Table 1, the precision of brain volume and volume change measurements for the different methods is shown. The mean difference between the FU1 and FU2 scans was close to 0 for kNN (mean ± SD = 0.01 ± 0.16), but was larger for SIENA (-0.10 ± 0.25) and US (-0.11 ± 1.48). Between the FU1 and FU2 scan the mean absolute difference was small for SIENA (0.20% ± 0.17) and kNN (0.12% ± 0.10) and larger for US (1.06% ± 0.99). Because the FU1 and FU2 scans were made after each other, the mean differences should be 0 if there is no error in precision of the methods. The mean absolute difference was used to give an indication of the size of the error in precision.

Brain volume (FU1 versus FU2) and volume change measurements (BL–FU1 versus BL–FU2) were not significantly different for each of the methods (all p > 0.05). The coefficient of repeatability for brain volume measurements was larger for US (29.6 cc) than kNN (4.9 cc), and the coefficient of repeatability for brain volume change measurements for US (2.84%) was also larger than SIENA (0.92%) and kNN (0.31%). This coefficient should be small, indicating a small variation from the mean in scan–rescan measurements of the methods.

The mean and SD of baseline brain volume measurements by manual segmentations (1094.9 ± 137.3 cc) were relatively large compared with the coefficient of repeatability for both US and kNN. In contrast, the mean and SD of volume change measurements by manual segmentations (-1.60 ± 1.26%) were relatively smaller compared with the coefficient of repeatability of the methods for brain volume change measurements, particularly for US. This implies that the error in precision of the measurements for the methods had a relatively larger impact on the measurement of brain volume change than on brain volume.

**Accuracy of the methods evaluated with manual segmentations**

In Table 2, the comparison of brain volume and volume change measurements of the different methods with manual segmentations is shown. For the measurements of brain volume at BL and FU1, US and kNN showed a good correlation with manual segmentations (all correlation coefficients ≥ 0.96; all p < 0.001). However, the volumes as determined with US at FU1 were significantly smaller than manual segmentations (mean difference ± SD = -17.1 cc ± 21.9; p = 0.037).

For measurements of brain volume change, SIENA showed the best correlation with manual segmentation (ρ = 0.82; p = 0.004), followed by kNN (ρ = 0.60; p = 0.067), and US (ρ = 0.50; p = 0.138). No significant differences between the methods and manual segmentations for change in brain volume from BL to FU1 were found (p > 0.05). In Fig. 3, Bland–Altman plots comparing brain volume change measurements of SIENA, US and kNN with manual segmentations are shown. The mean difference between manual segmentations and SIENA (-0.22%) and kNN (0.27%) were close to 0, unlike US (0.76%). The limits of agreement, which should be as small as possible, relative to the mean of each method and manual segmentation (x-axis of Bland–Altman plot), were small for SIENA (-1.82% to 1.37%) and kNN (-1.66% to 2.20%), and larger for US (-3.30% to 4.83%).

**Comparison between the methods**

In Table 3, the comparisons between the different methods on brain volume and volume change measurements are shown. The correlation between US–kNN for brain volume measurements was good (all correlation coefficients ≥ 0.96; all p < 0.001), but US measured smaller volumes at BL and FU1 (mean difference ± SD = -15.5 cc ± 14.7; p = 0.013 (BL); -14.5 cc ± 21.0; p = 0.047 (FU1)).

For measurements of brain volume change, the correlations between SIENA and US (mean correlation coefficient and p-value of both volume change measurements ρ = 0.28; p = 0.442) and US–kNN (ρ = 0.17; p = 0.641) were weak. For kNN–SIENA (ρ = 0.65; p = 0.048) the correlation was fairly good, but the volume change over time was larger for kNN than for SIENA (mean difference ± SD = -0.50% ± 0.55, p = 0.022 (BL–FU1); -0.45% ± 0.56, p = 0.047 (BL–FU2)). In Fig. 4, Bland–Altman plots comparing brain volume change measurements between SIENA, US and kNN are shown. The mean differences between SIENA–kNN (0.47%) and kNN–US (0.56%) were close to 0, unlike the difference between US and SIENA (-1.03%). The limits of agreement, which should be as small as possible, related to the mean of the two methods (x-axis of Bland–Altman plot), were small for SIENA–kNN (-0.63% to 1.57%), and larger for kNN–US (-3.79% to 4.90%) and US–SIENA (-5.05% to 2.99%).

**Discussion**

Few published studies have compared different methods of brain volume or volume change measurements (i.e. Lee and Prohovnik, 2008; Smith et al., 2007), and only a proportion of available methods have been evaluated with manual segmentations to test accuracy in addition to precision (i.e. Anbeek et al., 2005). High precision and accuracy are important to be able to detect subtle differences in brain volume changes between patient groups. Therefore, in this work brain volumes (US, kNN) and brain volume changes (SIENA, US, kNN) were compared between different methods, focusing not only on precision but also on accuracy by comparing with ground-truth manual segmentations. Crucial for our study is that the methods were implemented as described in the respective validation studies, to prevent favoring one method (Anbeek et al., 2005; Ashburner and Friston, 2005; Smith et al., 2002).
Fig. 1. Individual brain volume and brain volume change measurements for all subjects and each method. Brain volume or volume change ($y$) are plotted against individual subjects ($x$). Brain volume measurements (US, kNN, and manual segmentation) and volume change measurements (SIENA, US, kNN, and manual segmentation) are all shown in a separate figure. BL: Baseline; FU1: First follow-up; FU2: Second follow-up; N/A: Not applicable.
Precision

Precision of automated methods is often the first feature that is tested in validation studies. Without good precision, there is no need to check accuracy. Our results for precision of brain volume change measurements by SIENA (absolute mean = 0.20%) were even slightly better than the results described in a comparison study by the developer of the method (0.27%) (Smith et al., 2007). The precision of US for our scans could not be compared with the precision found by the developer, because only the voxelwise similarity index for simulated data was reported in the validation study (Ashburner and Friston, 2005; Dice, 1945). The precision for kNN was also only reported as a voxelwise similarity index by the developer (Anbeek et al., 2005).

An important factor in automated measurements is the actual variation between subjects on the measure that is being assessed relative to the error of the measurements. This variation in individual measurements, which is much larger for brain volume (1049.4±137.3 cc in our subjects) than for volume change (−1.60±1.26% in our subjects) relative to the coefficients of repeatability of the methods. This implies that the precision of volume change measurements has to be good and even better than required for brain volume.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>SIENA</th>
<th>US</th>
<th>kNN</th>
<th>Manual segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL brain volume (cc)</td>
<td>N/A</td>
<td>1040.2±135.1</td>
<td>1055.7±143.2</td>
<td>1049.4±137.3</td>
</tr>
<tr>
<td>Volume change</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL–FU1 (%)</td>
<td>−1.37±0.74</td>
<td>−2.36±2.30</td>
<td>−1.87±0.82</td>
<td>−1.60±1.26</td>
</tr>
<tr>
<td>Volume change</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL–FU2 (%)</td>
<td>−1.41±0.96</td>
<td>−2.48±1.81</td>
<td>−1.86±0.82</td>
<td>N/A</td>
</tr>
<tr>
<td>Difference</td>
<td>−0.10±0.25</td>
<td>−0.11±1.48</td>
<td>0.01±0.16</td>
<td>N/A</td>
</tr>
<tr>
<td>Absolute difference</td>
<td>0.20±0.17</td>
<td>1.06±0.99</td>
<td>0.12±0.10</td>
<td>N/A</td>
</tr>
<tr>
<td>Coefficient of repeatability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brain volume (cc)</td>
<td>N/A</td>
<td>29.6</td>
<td>4.92</td>
<td>N/A</td>
</tr>
<tr>
<td>Volume change (%)</td>
<td>0.92</td>
<td>2.84</td>
<td>0.31</td>
<td>N/A</td>
</tr>
</tbody>
</table>

All volumes and percentages are means±SD. The coefficient of repeatability should be small, indicating a small variation from the mean in scan–rescan measurements of the methods.

BL: Baseline; FU1: First follow-up; FU2: Second follow-up; N/A: Not applicable.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>SIENA</th>
<th>US</th>
<th>kNN</th>
<th>Manual segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean difference</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIENA</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.22±0.80</td>
</tr>
<tr>
<td>US</td>
<td>−9.2 cc±13.5</td>
<td>−17.1 cc±21.9</td>
<td>−0.76±2.03</td>
<td></td>
</tr>
<tr>
<td>kNN</td>
<td>6.4 cc±15.4</td>
<td>−2.6 cc±15.9</td>
<td>−0.27±0.96</td>
<td></td>
</tr>
<tr>
<td>Correlation coefficients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIENA</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.82†</td>
</tr>
<tr>
<td>US</td>
<td>0.99‡</td>
<td>0.50</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>kNN</td>
<td>1.00‡</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data in the top section are mean differences between manual segmentations and automatic segmentation by each of the methods. Differences were tested by Wilcoxon signed rank tests; p-values indicate the significance of differences between manual segmentation and the respective method.

Data in the bottom section are Spearman’s rank correlation coefficients of the methods with manual segmentation.

BL: Baseline; FU1: First follow-up; N/A: Not applicable.

† p<0.01.
‡ p<0.05.
* p<0.001.
Fig. 3. Bland–Altman plots comparing SIENA, US and kNN with manual segmentation on measurements of brain volume change. Differences (y) are plotted against means (x), comparing the automatic methods with manual segmentation in measurements of brain volume change. The solid lines represent means of all subjects. The dashed lines represent lower and upper limits of agreement.
methods. Differences were tested by Wilcoxon signed rank tests; (Smith et al., 2007). However, in our study precision of kNN was even slightly better than that of SIENA.

Accuracy

Accuracy is also very important for automated measurements, because it is an indication how the measurements compare to a reference standard. In our study, US and kNN showed a good accuracy for brain volume measurements compared to the reference standard of manual segmentation. For volume change measurements, SIENA, kNN and US, showed a good, moderate and weak accuracy, respectively. In contrast to precision, there is no uniform approach for determining accuracy across different studies. Manual segmentation can be regarded as the reference standard to determine accuracy. However, accuracy by comparing measurements with manual segmentations has not been checked for SIENA (Smith et al., 2001, 2002), and accuracy of brain volume measurements by US was tested only on simulated data (Ashburner and Friston, 2005), kNN had thus far only been compared with manual segmentation to evaluate accuracy for brain volume measurements (Anbeek et al., 2005), but not for the assessment of brain volume changes. Even in studies comparing different methods, differences in accuracy are not always addressed (i.e. Smith et al., 2007). We identified only one study that compared different methods also on accuracy by manual segmentation, but only for brain volume measurements, and not for volume change measurements (de Boer et al., 2010).

Comparison between the methods

Our study showed a good correlation between US and kNN for brain volume measurements, but US measured lower volumes. For brain volume change measurements SIENA–US and US–kNN showed a weak correlation, but the correlation was fairly good for SIENA–kNN. As discussed before, for US the good correlation for brain volume measurements, but weak correlation for volume change assessment can be attributed to a lower precision compared to SIENA and kNN.

Although SIENA and kNN showed a fairly good accuracy and comparability for brain volume change measurements, it was not perfect. Because brain atrophy rates are increasingly used as biomarkers in etiological and treatment studies, it is important to acknowledge the effects of suboptimal comparability. For studies comparing groups, the error in accuracy will be acceptable when comparing larger groups, because of the good precision. However, when analyzing data at a subject level, for example in identifying risk factors for increased change or protective factors, this might be a larger problem.

Choice of a method

In choosing a method for brain volume or volume change measurements, practical issues such as availability and experience with a certain method often play an important role. As the present study illustrates, the choice for a method has the largest impact on detectable variability, followed by the choice of a pulse sequence (Clark et al., 2006). Therefore, based on the research questions within a study, choices should be made for the most appropriate pulse sequence and method. This method should be validated for the intended use, and have sufficient precision and accuracy.

There are also other factors or problems that could influence the choice for a certain method. A general problem, particular in studies with a long duration, is scanner software updates, which could influence precision and accuracy of brain volume change measurements by increasing signal-to-noise ratio. Changes in signal-to-noise ratio have already shown to significantly influence measurements of US (Shuter et al., 2008). However, for most methods it is still unclear what the effects of changes in signal-to-noise ratio for the automated measurements are, but the increased signal-to-noise ratio could potentially affect methods in different ways. It is recommended to acknowledge this problem and check the influence it has on the measurements by routinely assessing signal-to-noise ratio or adding an additional covariate in the analysis (Shuter et al., 2008). Another factor that should be taken into account in choosing a method is if the method is used for a multi-center study. A study based on another automated method, FreeSurfer, showed that using scanners from the same vendor, even after software upgrades resulted in reliable measurements (Jovicich et al., 2009). On the other hand, using scanners from different vendors and with different field strengths does introduce a bias which should be considered (Jovicich et al., 2009). Application of the same software on the scans of the same subjects, but in different centers, also influences precision and accuracy. Manual editing in SIENA showed to influence measurements only to a minor extent (Jasperse et al., 2007a), however, varying parameters in US could potential alter the results of the measurements in the same order of magnitude as the studied diseases (Henley et al., 2010).

In our study, US and kNN showed a good precision, accuracy and comparability for brain volume measurements. Both methods have the advantage of also being able to measure gray matter, white matter, and CSF. Additionally, kNN can make a distinction in peripheral CSF and CSF in the lateral ventricles and can measure white matter hyperintensities.

SIENA and kNN both showed a fairly good precision and accuracy of brain volume change measurements. An additional value of kNN is that it can already determine brain volume on a single scan. By definition SIENA requires two scans acquired over an interval of time to give an indication of "disease severity", although an adapted cross-sectional version of SIENA, called SIENAX, is available to measure brain volume (Smith et al., 2002). Furthermore, depending on the types of tissues classified in the training data of kNN, other brain structures can also be determined separately, like gray matter, white

Table 3

Comparison of brain volumes (US, kNN) and brain volume changes (SIENA, US, kNN) between methods.

<table>
<thead>
<tr>
<th></th>
<th>Mean differencea</th>
<th>Correlation coefficientb</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Brain volume measurements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US–kNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(BL)</td>
<td>−15.5 cc±14.7⁎</td>
<td>1.00†</td>
</tr>
<tr>
<td>(FU1)</td>
<td>−14.5 cc±2.0⁎</td>
<td>0.96†</td>
</tr>
<tr>
<td>(FU2)</td>
<td>−14.9 cc±2.1</td>
<td>0.98†</td>
</tr>
<tr>
<td><strong>Volume change measurements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIENA–US</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(BL–FU1)</td>
<td>0.99%±2.17</td>
<td>0.35</td>
</tr>
<tr>
<td>(BL–FU2)</td>
<td>1.07%±1.75</td>
<td>0.21</td>
</tr>
<tr>
<td>US–kNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(BL–FU1)</td>
<td>−0.49%±2.33</td>
<td>0.20</td>
</tr>
<tr>
<td>(BL–FU2)</td>
<td>−0.62%±1.98</td>
<td>0.14</td>
</tr>
<tr>
<td>kNN–SIENA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(BL–FU1)</td>
<td>−0.50%±0.55</td>
<td>0.69*</td>
</tr>
<tr>
<td>(BL–FU2)</td>
<td>−0.45%±0.56</td>
<td>0.60</td>
</tr>
</tbody>
</table>

BL: Baseline; FU1: First follow-up; FU2: Second follow-up.
a Mean differences between measurements of the different automatic segmentation methods. Differences were tested by Wilcoxon signed rank tests; p-values indicate the significance of differences between methods.
b Spearman's rank correlation coefficients between methods.
⁎ p<0.05.
† p<0.001.
Fig. 4. Repeated-measurements Bland-Altman plots comparing SIENA, US and kNN on measurements of brain volume change. Differences \(y\) are plotted against means \(x\), comparing the different methods in measurements of brain volume change. The solid lines represent means of all subjects. The dashed lines represent lower and upper limits of agreement.
Strengths and limitations

The strength of the present study is the comparison of the different methods on the same data and in this work, in contrast to other comparison studies, accuracy of brain volume and volume change measurements were evaluated with precise manual segmentations. Limitations could be that other scan parameters of the used pulse sequences could have influenced our results, especially the addition of a high resolution T1 image. However, a standardized scanning protocol was used, comparable to that used in patient studies to assess brain volume or volume change measurements (de Bresser et al., Jongen et al., 2007). Nevertheless, adjustment of scan parameters can possibly influence methods to a different extent. Although the limited number of subjects could have influenced our results, paired and nonparametric statistics were used instead of parametric statistics, in order to prevent overestimating the found results. Finally, the current reference standard also does not have a perfect precision or accuracy towards the actual brain volume.

In conclusion, US and kNN showed a good precision, accuracy and comparability for brain volume measurements. For measurements of volume change, SIENA showed the best performance. kNN is a good alternative if volume change measurements of other brain structures are required.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgments

This study was financially supported by grant 2003.01.004 of the Dutch Diabetes Research Foundation and by a high potential grant from Utrecht University. The authors thank P. Anbeek for supplying software.


