Multicenter assessment of reliability of cranial MRI

M. Ewers a, S.J. Teipel a,∗, O. Dietrich b, S.O. Schönberg b, F. Jessen c, R. Hein c, P. Scheltens d, L. van de Pol d, N.R. Freymann c, H.-J. Moeller a, H. Hampel a

a Department of Psychiatry, Dementia and Neuroimaging Section, Alzheimer Memorial Center D2, Ludwig Maximilian University of Munich, Nussbaumstr. 7, 80336 Munich, Germany
b Institute of Clinical Radiology-Großhadern, Ludwig Maximilian University of Munich, Marchioninistr. 15, 81377 Munich, Germany
c Department of Psychiatry, Rheinische Friedrich-Wilhelms-University Bonn, Sigmund-Freud-Str. 25, 53105 Bonn, Germany
d Department of Neurology, VUMedical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands

Received 24 September 2004; received in revised form 6 April 2005; accepted 23 May 2005

Abstract

Clinical utility of magnetic resonance imaging (MRI) for the diagnosis and assessment of neurodegenerative diseases may depend upon the reliability of MRI measurements, especially when applied within a multicenter context. In the present study, we assessed the reliability of MRI through a phantom test at a total of eleven clinics. Performance and entry criteria were defined liberally in order to support generalizability of the results. For manual hippocampal volumetry, automatic segmentation of brain compartments and voxel-based morphometry, multicenter variability was assessed on the basis of MRIs of a single subject scanned at ten of the eleven sites. In addition, cranial MRI scans obtained from 73 patients with Alzheimer’s disease (AD) and 76 patients with mild cognitive impairment were collected at a subset of six centers to assess differences in grey matter volume. Results show that nine out of eleven centers tested met the reliability criteria of the phantom test, where two centers showed aberrations in spatial resolution, slice thickness and slice position. The coefficient of variation was 3.55% for hippocampus volumetry, 5.02% for grey matter, 4.87% for white matter and 4.66% for cerebrospinal fluid (CSF). The coefficient of variation was 12.81% (S.D. = 9.06) for the voxel intensities within grey matter and 8.19% (S.D. = 6.9) within white matter. Power analysis for the detection of a difference in the volumes of grey matter between AD and MCI patients across centers (d = 0.42) showed that the total sample size needed is N = 180. In conclusion, despite minimal inclusion criteria, the reliability of MRI across centers was relatively good.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Multicenter, Reliability, Phantom test, Magnetic resonance imaging, Competence network dementia

1. Introduction

MRI-based assessment of cerebral atrophic changes has been proposed as a useful tool to aid in the early detection and tracking of Alzheimer’s disease (AD) [22]. Numerous studies have examined the diagnostic accuracy of volumetric measurement and rating of selected brain regions such as the medial temporal lobe structures, corpus callosum, and grey matter [4,9,10,13,17,24,25,29,31,32,34]. Recently, in order to assess the clinical utility of neuranatomical markers within a large sample across different clinical sites, several multicenter studies have been reported [12,16,20]. A reliable application of MRI-based diagnostic criteria is pivotal for the utility of MRI for diagnostic purposes. Variability in MRI-based measurement between clinical sites may potentially influence the accuracy of biological measures [6,7,33], and thus compromise applicability of MRI-based diagnostic criteria across sites. The assessment of MRI reliability with regard to volumetric measures is therefore not only important for the internal quality assurance within a multicenter network but also with regard to the applicability of MRI volumetry for clinical diagnostics. In the present study, we aimed to assess the center-induced variability in MRI-based volumetric measures, such as grey matter volume and hippocampal volumetry, that are relevant for the diagnosis of neurodegenerative disorders. First, accuracy and reliability of MRI was assessed via a phantom test. Secondly, the variability of volumetric measures, including manual volumetry of...
the hippocampus and segmentation of brain compartments, was tested.

2. Methods

The inclusion criteria for the clinical centers consisted of the presence of a 3.5 T scanner and the opportunity to establish at least a 3D T1 weighted (Philips scanners) or MP-RAGE (Siemens scanners) acquisition sequence. For standardization of MRI acquisition across centers, acquisition parameters were provided as a guideline to all centers. The phantom test of the American College of Radiology MRI Accreditation Program \[1\] was conducted at 11 German clinical centers of the Competence network on dementia. The phantom consisted of a 148 mm long cylinder filled with a solution of 10 mM nickel chloride and 75 mM sodium chloride in 3.81 distilled water (Fig. 1). The MRI scans were acquired according to a standardized protocol including an axial T1-weighted 3D sequence and an axial T2-weighted double echo sequence on a 1.5 T scanner, where the hardware and sequence parameters varied between centers. Siemens scanners were used (Siemens Sonata or Siemens Magnetom Vision) at eight centers and Philips scanners (Philips Gyroscan and Philips Intera) at the remaining centers. For the T1 3D sequence, repetition time (TR) varied between 450 and 662 ms and echo time (TE) between 13 and 20 ms. The slice thickness was 5 mm and the slice gap 5 or 10 mm across centers. For the T2-weighted double echo sequence, TR was fixed at 2000 ms and TE at 20/80 ms for all centers, with the slice thickness being 5 mm and slice gap 5 or 10 mm. Geometric accuracy was, in addition, assessed on the basis of a sagittal single slice spin-echo acquisition.

Accuracy of MRI was assessed according to the following seven ACR-criteria \[1\]: (1) geometric accuracy pertaining to the measurement of the length (criterion = 148 ± 2 mm) and diameter (criterion = 190 ± 2 mm) of the phantom. (2) High contrast spatial resolution was assessed to measure the scanner’s ability to resolve small objects when the signal-to-noise ratio is high enough not to be the limiting factor (criterion: objects of size <1 mm must be discernable). (3) Slice thickness accuracy included the measurement of the actual slice thickness, with the prescribed slice thickness being 5 mm (criterion = 5 ± 0.7 mm). (4) Slice position accuracy tested the actual location of the slice. A shift in the slice position was determined by the difference in the length between two adjacent bars, which are of equal length if the prescribed slice position is accurate. The criterion is a maximal difference in the bar length as large as 5 mm, corresponding to an actual shift in the slice position of 5 mm. (5) Image intensity uniformity refers to homogeneity of the voxel intensities across a water-only region of the phantom. The image uniformity was assessed according to the following formula:

\[
\text{Percent integral uniformity (PIU)} = 100 \times \left(1 - \frac{\text{low}}{\text{high} + \text{low}}\right),
\]

where low and high correspond the low and high extreme voxel intensities, which had to be present in at least 100 voxels (size: 1 mm²) within the phantom image. The criterion was that PIU must be greater than or equal to 87.5%. (6) Signal ghosting refers to the artifact in which a faint copy (ghost) of the imaged object is visible next to the object. For the computation of the ghosting ratio, four small regions of interest (ROI = 10,000 mm²) were placed next to and above the phantom within the air region of the MRI image. A large ROI (about 20,000 mm²) was placed within a water-only region of the phantom. The percent signal-to-ghosting ratio was computed according to:

\[
\text{Percent ghosting ratio} = \frac{\text{PIU} - \text{PIU}_{\text{ROI}}}{\text{PIU} - \text{PIU}_{\text{ROI}}},
\]

where the variables left, right, top, bottom and large ROI correspond to the mean intensity value of the respective ROIs. The criterion was a ghosting ratio ≤0.025. (7) Low contrast object detectability refers to the detectability of a maximum of 40 objects displayed at varying size and contrast values. The criterion of the minimum number of discernable objects was nine.

2.1. In vivo MRI

A single healthy male volunteer (31 years old, MMSE = 30) was scanned at a subset of 10 centers. MRI scans were conducted with a sagittal magnetization-prepared rapid gradient echo (MP-RAGE) sequence on Siemens scanners (n = 8) and a 3D-fast-T1-weighted-gradient-echo sequence on Philips scanners (n = 2). Among all centers the TR varied between 9.3 and 20 ms and TE between 3.93 and 4.38 ms.

2.2. Hippocampus volumetry

One scan was manually reformatted (using GE Radworks) coronally in a plane perpendicular to the long axis of the
The hippocampus was measured twice on a second occasion. The hippocampal volume (in mm$^3$) was computed for an estimate of the variability of the voxel intensities within the hippocampus. For an estimate of the variability of voxel intensities within the grey and white matter regions, the brain map of covariance coefficients was multiplied with a white and respectively grey matter mask, using SPM2. In a final step, the mean covariance coefficient within the white and grey matter regions were computed, using Matlab 5.3. Extreme values (>60%) that resulted from effects of misregistration and could be visually located within a discrete band along the border of the brain were excluded from the analysis.

2.3. Reliability of voxel-based analysis

The in vivo MRI scans of the single subject scanned at 10 of the 11 centers were spatially coregistered, using an affine transformation with 12 parameters (SPM2, Welcome Department, London, UK). The brain volume was isolated from the skull, using the computer program “Brain Extraction Tool” (BET) included in the “Oxford Center of Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library 3.1”[30]. Subsequently, the brain was segmented into grey matter, white matter and cerebrospinal fluid (CSF), using “FMRIB’s Automated Segmentation Tool” (FAST). For the segmentation, we chose not to apply a priori probabilities for the voxel classification to increase the sensitivity for potential variation in the volumes of the brain compartments between centers, and thus to achieve a conservative test of the multi-center reliability.

For the quantification of the variability in the voxel intensities between centers, voxel intensity values were first normalized to the mean voxel intensity per image and scaled by a factor of 100. Subsequently, coefficients of variation for each voxel of the coregistered and extracted brains across all centers were computed. For the visualisation of the spatial distribution of the voxel-based coefficients of variation, the coefficients of variation were mapped onto a brain with a customized program written in Matlab 5.3 (The Mathworks Inc., 1999). The resulting brain map of coefficients of variation was subsequently used for manual tracing of the hippocampus within one slice that allowed for an easy identification of the borders of the hippocampus. The mean value of the coefficients of variation within the outlined region of the hippocampus was computed for an estimate of the variability of the voxel intensities within the hippocampus. For an estimate of the variability of voxel intensities within the grey and white matter regions, the brain map of covariance coefficients was multiplied with a white and respectively grey matter mask, using SPM2. In a final step, the mean covariance coefficient within the white and grey matter regions were computed, using Matlab 5.3. Extreme values (>60%) that resulted from effects of misregistration and could be visually located within a discrete band along the border of the brain were excluded from the analysis.

2.3.1. Statistics

For the phantom test, a bias in the measurements across centers was determined using a two-sided one-sample t-test (α=0.05) to test for a directional deviation of the observed measurement from the true measurement of the phantom. In addition, the 95% confidence intervals of the deviations of the observed measurements were computed. For the measurement of the image intensity homogeneity and ghosting artefact, the 95% confidence intervals of the deviations between the observed measurement and the criterion value (instead of the “true” value) were employed, since no actual “true” measurement of the phantom was known for these two variables.

The coefficient of variation (V) was computed by dividing the standard deviation of the measurements (S.D.) by the mean value (X) of the measurements averaged across all centers (X), multiplied by 100:

$$V = \left( \frac{S.D.}{X} \right) \times 100$$

The coefficient of variation was computed for the hippocampus volumetry, segmentation of each brain compartment, and for the intensities of each voxel of the in vivo MRI brain scans.

2.4. Grey matter volume in AD and MCI patients

A total of 73 AD and 76 amnestic MCI patients were available from six of the 11 centers where each center contributed at least n = 10 per group (one center had n = 9, Table 2). AD patients were diagnosed according to the NINCDS-ADRDA criteria [18] and the MCI patients according to the Petersen criteria [26]. For the AD patients (39 of female gender), mean MMSE score was 24.2 (S.D. = 3.1). For the MCI patients (37 of female gender), mean MMSE score was 26.7 (S.D. = 2.5). The difference in MMSE between groups was statistically significant (p < 0.0001).

MRI scans were preprocessed by setting the origin of the image to the anterior commissure. Subsequently, the MRI scans positioned in native space were segmented, using the SPM2 standard T1 template, grey matter, white matter and CSF priors. This segmentation step involves an affine transformation of each scan to the template and subsequent retransformation into the native space.
The effect size $d$ of the difference in the volume of grey matter between AD and MCI patients was estimated according to the following formula:

$$d = \frac{(X - Y) \delta_{xy}}{\delta_{xy}}$$

where $X - Y$ is the difference in the mean volume of grey matter between AD and respectively MCI patients, and $\delta_{xy}$ the pooled standard deviation of volume of grey matter obtained from AD and MCI patients. The two-tailed independent samples $t$-test of the group comparison and corresponding effect size was determined for each center separately (henceforth, referred to as “monocenter effect size”). In addition, the effect size of the group difference on the basis of the grey matter volumes pooled across centers was estimated (henceforth, referred to as “pooled effect size”). In order to compare the monocenter effect size to a “multicenter effect size”, $n=10$ data points were randomly sampled each from the MCI group ($N=76$) and AD group ($N=73$) pooled among centers and the effect size of the group difference was calculated. The random sampling of grey matter volumes from each group and the calculation of the respective effect size was iterated 1000 times. The mean effect size of the resulting sampling distribution along with the 95% confidence interval (95% CI) was calculated, using a customized program written in Matlab 5.6. We chose to sample $n=10$ in each sampling trial applied to the pooled data since this sample size approximates the one collected at each single center, thus allowing for a comparison between the monocenter and multicenter effect size. A one-sample $t$-test was used to test whether the multicenter effect size significantly deviated from the value 0. The sample size required to detect the pooled effect size with a statistical power of 0.80, $\alpha=0.05$ for a two tailed $t$-test, was estimated, using the computer program GPOWER [5].

The intraclass correlation coefficient adjusted for the mean was computed with SPSS 11.5.1, using a two-way random model.

3. Results

3.1. Phantom test

Descriptive statistics are presented in Table 1. Two out of nine centers did not meet all criteria. For these two centers, the criterion of slice thickness was exceeded by more than 2.69 S.D. At one of these two centers, slice position exceeded the criterion (0.7 mm) by more than 2.53 S.D. and the low contrast object detectability by 0.95 S.D. (criterion >9 objects) in addition. The criterion of high contrast resolution was not met at the other center (Table 1).

Across centers, the 95% confidence intervals of deviations in the measurement of the length and diameter (Fig. 2), and low contrast object detectability (Fig. 3) were within the limits of the ACR criteria. In contrast, the upper boundary of the confidence intervals of deviations in the slice position and thickness exceeded the limits of the ACR criteria (Fig. 2). After exclusion of those two centers that did not meet the criteria, the upper boundary of the confidence interval of measurement deviations were confined within the limits of the criteria of slice position (0.65 mm for T1 and DE; criterion =±0.7 mm) but still exceeded slightly the upper limit of the criterion of slice thickness (5.58 mm for T1, 5.37 mm for DE; criterion = 5 mm). The confidence interval of the absolute measurements of the ghosting artifact and intensity uniformity (Fig. 4) fell within the limits of the ACR criteria.

Fig. 2. Mean values (filled symbols) and 95% confidence intervals (bars) of the deviations in the measurement of the length, diameter, slice thickness and position for the T1 and the DE sequence. For all variables, the lower and upper cut-off values of the criteria are indicated (blank circles), where for the slice position only the upper limit is defined.

Fig. 3. Mean value of measurement deviation in the low contrast object detectability (filled circles) and corresponding 95% confidence interval (bars) for the T1 and DE sequences. The cut-off value of the criterion indicates the maximally acceptable deviation (31 objects) from the total amount of detectable objects (40), i.e. at least nine objects need to be detectable in order to meet the criterion.
Table 1

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sequence</th>
<th>Criterion</th>
<th>Arithmetic mean (S.D.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (in mm)</td>
<td>T1</td>
<td>148 ± 2</td>
<td>147.08 (0.78)</td>
</tr>
<tr>
<td>Diameter (in mm)</td>
<td>T1</td>
<td>190 ± 2</td>
<td>189.8 (1.14)</td>
</tr>
<tr>
<td>Spatial resolution (in mm)</td>
<td>T1</td>
<td>≤1</td>
<td>HD: 0.95 (0.05)</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>≤1</td>
<td>VD: 0.95 (0.05)</td>
</tr>
<tr>
<td>Slice thickness (in mm)</td>
<td>T1</td>
<td>5 ± 0.7</td>
<td>5.86 (1.3)</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>5 ± 0.7</td>
<td>5.92 (1.29)</td>
</tr>
<tr>
<td>Slice position (in mm)</td>
<td>T1</td>
<td>≤5</td>
<td>3.07 (4.39)</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>≤5</td>
<td>3.16 (4.52)</td>
</tr>
<tr>
<td>Intensity uniformity index (in %)</td>
<td>T1</td>
<td>≥87.5</td>
<td>94.08 (1.9)</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>≥87.5</td>
<td>93.01 (2.05)</td>
</tr>
<tr>
<td>Ghosting (in %)</td>
<td>T1</td>
<td>≤0.025</td>
<td>0.003 (0.002)</td>
</tr>
<tr>
<td>Low contrast object (number)</td>
<td>T1</td>
<td>≥9</td>
<td>27 (10.63)</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>≥9</td>
<td>20.55 (9.43)</td>
</tr>
</tbody>
</table>

* HD: objects to be discerned were arrayed in horizontal direction.
  * VD: objects to be discerned were arrayed in vertical direction.

Across all centers, a bias was observed in the geometric measurement, with the length (p = 0.003) but not the diameter (p = 0.57) being underestimated. The slice thickness was significantly higher than the prescribed slice thickness (p = 0.04 for T1 and p = 0.05 for DE). No bias was determined for the other phantom measurements due to the fact that by definition only unidirectional deviation in these measurements was possible (e.g. the spatial resolution could only be too low but not too high).

3.2. Hippocampus volumetry

The volume of the right hippocampus was measured twice by the same rater. The mean hippocampus volume was 3362.05 mm³ (S.D. = 138.94) for the first measurement, 3273.14 mm³ (S.D. = 102.99) for the second measurement. The coefficient of variation was 4.13% and 3.15% for the first and second measurement, respectively. For the mean hippocampal volume averaged across both time points of measurement, the coefficient of variation was 3.55%. The intraclass correlation coefficient was 0.86, indicating a high reliability of the manual hippocampus measurement across centers.

3.3. Segmentation of brain compartments and voxel-based morphometry

The mean volume was 758.94 cm³ (S.D. = 38.07) for cortical grey matter, 599.93 cm³ (S.D. = 29.20) for white matter and 248.99 cm³ (S.D. = 11.61) for CSF (Fig. 5). The coefficient of variation was 5.02% for grey matter, 4.87% for white matter, and 4.66% for CSF.

For the voxel-based morphometric analysis, the mean coefficient of variation averaged across voxels was 12.81% (S.D. = 9.07) within the grey matter, 8.19% (S.D. = 6.9) within the white matter, and 7.97% (S.D. = 3.28) within the hippocampus (Fig. 6).

3.4. Multicenter assessment of difference in grey matter volume between MCI and AD patients

Mean volume of grey matter in AD and MCI patients and the moncenter effect sizes of group differences for each center are displayed in Table 2. The mean of the sampling distribution of the moncenter effect sizes was $d = -0.41$ (p < 0.0001, Fig. 7; negative effect size indicates larger grey matter volume in MCI patients). All moncenter effect sizes, except for two centers ($d = -1.31$ and $d = 0.27$), were within the 95% CI of the moncenter sampling distribution (95% CI = $-1.03 - 0.25$). Those two centers that showed effect sizes exceeding the 95% CI did not correspond to those centers who deviated from the reliability criteria of the phantom test reported above. A t-test confirmed that the moncenter effect sizes did not differ in either direction from the mean
Table 2
Comparison of grey matter volume (ml) between AD and MCI patients at each center

<table>
<thead>
<tr>
<th>Center</th>
<th>Sample size</th>
<th>AD Mean volume of grey matter (S.D.)</th>
<th>MCI Mean volume of grey matter (S.D.)</th>
<th>Effect size d</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>13</td>
<td>619.38 (77.29)</td>
<td>635.11 (90.37)</td>
<td>−0.20</td>
<td>0.68</td>
</tr>
<tr>
<td>II</td>
<td>9</td>
<td>629.4 (64.8)</td>
<td>672.57 (66.5)</td>
<td>−0.69</td>
<td>0.16</td>
</tr>
<tr>
<td>III</td>
<td>11</td>
<td>641.71 (58.57)</td>
<td>642.49 (62.31)</td>
<td>−0.51</td>
<td>0.2</td>
</tr>
<tr>
<td>IV</td>
<td>11</td>
<td>578.48 (67.6)</td>
<td>680.29 (93.53)</td>
<td>−1.31</td>
<td>0.008</td>
</tr>
<tr>
<td>V</td>
<td>10</td>
<td>645.50 (82.38)</td>
<td>624.22 (77.15)</td>
<td>+0.27</td>
<td>0.57</td>
</tr>
<tr>
<td>VI</td>
<td>19</td>
<td>559.75 (105.35)</td>
<td>592.51 (78.0)</td>
<td>−0.36</td>
<td>0.28</td>
</tr>
<tr>
<td>Pooled</td>
<td>73</td>
<td>601.33 (84.63)</td>
<td>636.22 (80.85)</td>
<td>−0.42</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Fig. 5. Box-plot of the volumes of the segmented brain compartments including grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Outliers (blank circles) are defined as deviating by 1.5–3 box-lengths (interquartile range) from the upper or lower edge of the box.

Fig. 6. Colour map of coefficient of variation displayed within a sagittal slice of the brain. The colour bar indicates the correspondence between different colours and the coefficients of variation (in percent).

Fig. 7. Sampling distribution of multicenter effect sizes of the group difference in grey matter volume between AD and MCI patients. Samples of n=10 were randomly drawn from a total of 73 AD and 76 MCI patients each, where sampling without replacement was iterated 1000 times. The blue line indicates the mean of the sampling distribution. For purposes of comparison between multicenter and monocenter effect sizes, the effect size of each monocenter analysis (red stars) were projected onto the plot (note that height of the stars was chosen arbitrarily and does not indicate frequency).

The multicenter effect size ($d = 0.8$), suggesting that the monocenter effect sizes were not superior when compared to the multicenter effect sizes. When pooling data of patients with MCI ($N = 76$) and respectively AD ($N = 73$) across centers, the pooled effect size was $d = −0.42$ ($p = 0.01$). The total sample size needed to detect the pooled effect size with sufficient statistical power yielded $N = 180$.

4. Discussion

The aim of the current study was to investigate the reliability of MRI and variability of MRI-based volumetric measures across multiple sites.
The major results of the phantom test showed that two out of 11 centers did not meet the reliability criteria for slice position and thickness, spatial resolution of low contrast objects and spatial resolution at high spatial contrast. Thus, the majority of centers showed accurate measurements across all dimensions assessed.

In order to estimate the variability of the MRI measurements at a follow-up examination, we computed the 95% CIs of the measurement deviations. The 95% CIs exceeded the limits of the criteria on slice position and thickness, primarily due to those two centers that exceeded the criteria for these two measures, suggesting that MRI measurements at a follow-up assessment are likely to be confined within the criteria limits at most centers. However, one has to caution that this conclusion rests on the assumption that the current measurements are representative of subsequent measurements, which is valid only to a limited extent since fluctuation in measurement accuracy over time have been reported in previous studies [19].

A bias was observed in the measurement of the length and slice thickness. However, for the measurement of length, the underestimation, although statistically significant, was rather small since the deviation across all centers did not exceed the limits of the ACR criteria. For the bias in the slice thickness, the effect was due to the extreme values of the two outlier centers, since no bias could be detected after removing those two centers from the analysis, suggesting that the elevation of slice thickness above the limit of the criterion was not representative for the whole group.

One potential explanation of the variance in the measurement accuracy between centers may be differences in the routine quality assurance of MRI implemented at the centers. However, the routine quality check provided by the manufacturers (Siemens and Philips) did not seem to vary between the centers. Quality assurance used at the clinics includes a manufacturer-specific phantom test and the calibration of the scanner twice a year. Measures beyond the quality check provided by the manufacturer were not reported by the clinics. Thus, it is unlikely that any difference in the routine quality assurance between clinics may account for the multicenter variability in measurement accuracy. For those two centers that did not meet the phantom test criteria used in the present study it is possible that transient sources of variability may have lead to measurement inaccuracies. This can only be evaluated within a longitudinal assessment of reliability within centers, which is currently under way at all centers of the network.

Although a number of other studies on multicenter MRI studies [19,27] that included volumetry in patients with AD [12,16,20,21] have been reported, only few studies have assessed multicenter reliability [15,19,27]. The comparison with previous studies using a MRI phantom test shows that the deviation in the measurement of length (about 92 mm) observed in the current study is relatively low. Prott et al. [27] found in a European multicenter study a deviation in the measurement of length of about 1.8 mm at 23 out of 27 scanners tested. McRobbie and Quest [19] assessed the reliability of 17 MRI scanners over an observation period of 8 years, finding a mean deviation in the measurement of length of 1.39 mm. In their study, deviations in the measurement of length were one of most frequent problems, in addition to deviations in intensity homogeneity and ghosting. Jack et al. [16] do not report statistics on phantom measurements, but mention that geometric distortion and signal-to-noise ratio were the most frequent problems in their longitudinal MRI-based assessment of AD patients. In the current study, we did not observe any failures for these measures by any center tested, however, we did find failures to meet the criteria regarding the slice position, slice thickness and spatial resolution, which were only observed at a relatively low frequency in the study of McRobbie and Quest. However, a direct comparison between the studies is not possible due to differences in the methods and study design.

Another aim of the study was to assess to what extent the reliability of the MRI measurement determined in the phantom test can be generalized to manual and fully computerized measures of structural brain changes. Results of the automatic segmentation of the brain compartments showed a variability across centers that is in agreement with other studies on quantification of structural brain deficits, such as the semi-automatic assessment of lesion volume in patients with multiple sclerosis [8]. Both the manual measurement of the hippocampus and the automatic volumetry of grey matter had a coefficient of variation below 5%. More recently, voxel-based morphometry has been applied for the assessment of changes in grey matter concentration in patients with AD and MCI in moncenter studies [2,11,28]. To our knowledge, our study presents the first data on the multicenter variability of voxel-based morphometry. The coefficient of variation for the voxel intensities measured across centers were higher than those of the segmented brain compartments. This increase in variability can at least partially be attributed to misregistration effects, which lead to a mismatch of brain voxel positions between centers. Moreover, misregistration of the brains may lead to the overlap of voxel positions at the border of the brains with voxel intensities of the background, resulting in extreme values of coefficients of variation that were observed along the boundary of the brain. Note, that our protocol of preprocessing of the images did not include any correction for intensity non-uniformity or smoothing procedures in order to maximize the sensitivity for the reaction of the variability in voxel intensities between centers. Thus, the current results can be interpreted as a liberal estimate of multicenter variability of voxel intensities. Still, the estimate of the variability of the voxel intensities within the hippocampus was moderate and is comparable to the variability of other biological measurements, e.g. APP in cerebrospinal fluid [23].

Based on the observed multicenter variability of volumetric measures, it would be feasible to expect a reduced effect size for multicenter data when compared to moncenter data. However, the present results on the effect size of the differences in grey matter between MCI and AD patients show...
that the multicenter effect size did not seem to systematically differ from the monocenter effect sizes, despite the fact that those two centers that did not meet the criteria of the ACR phantom test were included. In contrast, the results on the multicenter reliability of the measurement of grey matter volume based on the data obtained from the single subject scanned at all centers demonstrated variability between centers. A possible explanation could be that the estimate of the variability between centers based on the single subject observation may partially reflect variability that may occur within centers as well, e.g. positioning effects in the scanner or different radiologists operating the system. Thus, the estimate of the multicenter variability may have been inflated. On the other hand, for the interpretation of the comparison of monocenter and multicenter effect sizes based upon the large sample of MCI and AD patients some caveats need to be considered as well. Note that the estimation of the monocenter effect sizes was based on a rather small sample size per center. Although the effect of diagnosis based upon the pooled multicenter data was statistically significant, this was true for none of the within-center effect sizes of group differences. Thus, it is possible that the rather large variability within centers due to low sample size may have occluded variability between centers. The multicenter effect may have become detectable if larger sample sizes per center had been included.

It should be noted that acquisition parameters and scanner manufacturer did vary between centers to some extent. The competence network pursued a standardization of the MRI acquisition by providing guidelines on acquisition sequences to be used by each center. However, a liberal policy of keeping acquisition parameters and manufacturer of hardware uniform across centers was adopted in order to allow for reasonable variability in sequences. A rigid standardization of MRI acquisition within the network would compromise the conclusiveness regarding the reliability of diagnostic criteria at clinical sites outside the network where acquisition parameters and manufacturers of hardware are likely to vary. Thus, our strategy was to define liberal standards of acquisition and to assess the variability in measurement accuracy by phantom test and measures of interest, i.e. volume of hippocampus and grey matter that are relevant for the detection of AD. Our current results suggest that although measurement accuracy varies between centers, the extent appears to be within limits that allow for a reasonable multicenter assessment of volumetric differences between groups.

From a practical point of view the current results are encouraging since our analysis shows that a multicenter data analysis is sensitive towards the effects of AD on the grey matter volume when compared to MCI patients. Sample size estimation showed that a total of 180 patients are needed in order to reliably detect the expected effect size, which appears to be feasible for multicenter projects. On the basis of these results, we suggest that the ACR reliability criteria may be used in order to assess the extent of measurement accuracy across sites. However, the meaning of the violation of the criteria needs to be determined relative to the specific clinical measures to be studied. In the present study, violation of the criteria did not seem to lead to detectable deviations in measures of interest, i.e. manual tracing of the hippocampus as well as automatic analysis of grey matter volume. Although no direct test on the influence of measurement accuracy on volumetric measures was possible, results of the in vivo MRI of a single subject scanned at 10 out of 11 sites do not suggest that failure of the phantom test necessarily leads to abnormal measurement of the volume of the hippocampus or grey matter. Moreover, the analysis of the multicenter comparison between MCI and AD patients showed that the multicenter analysis produced similar effect sizes of group difference in grey matter volume when compared to the monocenter analysis. This was true even though those two centers that failed the phantom test were included in the analysis, suggesting that the error variance was not increased within a multicenter analysis. Together, the current results suggest that those measurement dimensions in which two centers showed significant deviations on the phantom test may not exert a significant effect on volumetric measures. Future studies may explore systematically measurement dimensions that may be of importance for volumetric measures.

Acknowledgement

The study was funded by a grant from the Bundesministeriums für Bildung und Forschung (BMBF 01 GI 0102) awarded to the dementia network "Kompetenznetz Demenz". We thank Dr. Jens Heinzenreich (Charité Berlin), Dr. Christian Luckhaus (Rheinische Kliniken Düsseldorf), Drs. Thomas Kucinski and Thomas Müller-Thomsen (Universitätsklinikum Hamburg-Eppendorf), Dr. Marco Essig (University of Heidelberg) and Dr. Anke Hensel (University of Leipzig) and the centers of Bonn, Erlangen, Frankfurt, Freiburg, Mannheim for contributing data and helpful comments on the manuscript.

References


