Neuroradiology (2012) 54:1321–1330
DOI 10.1007/s00234-012-1058-0

DIAGNOSTIC NEURORADIOLOGY

Is radiological evaluation as good as computer-based volumetry to assess hippocampal atrophy in Alzheimer's disease?

Claire Boutet · Marie Chupin · Olivier Colliot · Marie Sarazin · Gurkan Mutlu · Aurélie Drier · Audrey Pellot · Didier Dormont · Stéphane Lehéricy · And the Alzheimer's Disease Neuroimaging Initiative

Received: 6 April 2012 / Accepted: 14 June 2012 / Published online: 11 July 2012

Abstract
Introduction Hippocampus volumetry is a useful surrogate marker for the diagnosis of Alzheimer's disease (AD). Our purpose was to compare visual assessment of medial temporal lobe atrophy made by radiologists with automatic hippocampal volume and to compare their performances for the classification of AD, mild cognitive impairment (MCI) and cognitively normal (CN).
Methods We studied 30 CN, 30 MCI and 30 AD subjects. Six radiologists with two levels of expertise performed two

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://www.loni.ucla.edu/ADNI/Data/ADNI_Authorship_List.pdf.

C. Boutet (✉) · A. Drier · A. Pellot · D. Dormont · S. Lehéricy
Department of Neuroradiology, AP-HP,
Groupe Hospitalier Pitié-Salpêtrière,
47-83, Boulevard de l'Hôpital,
75651 Paris Cedex 13, France
e-mail: claireboutet@yahoo.fr

C. Boutet · M. Chupin · O. Colliot · M. Sarazin · A. Drier · D. Dormont · S. Lehéricy
Université Pierre et Marie Curie-Paris 6, Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière,
UMR-S975, Paris, France

C. Boutet · M. Chupin · O. Colliot · M. Sarazin · A. Drier · D. Dormont · S. Lehéricy
Inserm,
U975 Paris, France

C. Boutet · M. Chupin · O. Colliot · M. Sarazin · A. Drier · D. Dormont · S. Lehéricy
CNRS,
UMR 7225 Paris, France

C. Boutet · M. Chupin · O. Colliot · M. Sarazin · A. Drier · D. Dormont · S. Lehéricy
ICM-Institut du Cerveau et de la Moelle épinière,
Paris, France

C. Boutet · A. Drier · D. Dormont · S. Lehéricy
Centre de Neuroimagerie de Recherche-CENIR,
Paris, France

M. Chupin · O. Colliot
Equipe Cogimage-CRICM,
47, Boulevard de l'Hôpital,
75651 Paris Cedex 13, France

M. Sarazin
Department of Neurology,
Institut de la Mémoire et de la Maladie d'Alzheimer-IM2A,
Groupe Hospitalier Pitié-Salpêtrière,
47-83, Boulevard de l'Hôpital,
75651 Paris Cedex 13, France

G. Mutlu
Urgences Cérébro-Vasculaires,
Université Pierre et Marie Curie-Paris 6,
Groupe Hospitalier Pitié-Salpêtrière,
47-83, Boulevard de l'Hôpital,
75651 Paris Cedex 13, France

G. Mutlu
Inserm, Université Paris 7-Denis Diderot, Hôpital Saint-Louis,
U717 Paris, France

readings of medial temporal lobe atrophy. Medial temporal lobe atrophy was evaluated on coronal three-dimensional T1-weighted images using Scheltens scale and compared with hippocampal volume obtained using a fully automatic segmentation method (Spearman's rank coefficient).
Results Visual assessment of medial temporal lobe atrophy was correlated with hippocampal volume ($p<0.01$). Classification performances between MCI converter and CN was better using volumetry than visual assessment of non-expert readers whereas classification of AD and CN did not differ between visual assessment and volumetry except for the first reading of one non-expert ($p=0.03$).
Conclusions Visual assessment of medial temporal lobe atrophy by radiologists was well correlated with hippocampal volume. Radiological assessment is as good as computer-based volumetry for the classification of AD, MCI non-converter and CN and less good for the classification of MCI converter versus CN. Use of Scheltens scale for assessing hippocampal atrophy in AD seems thus justified in clinical routine.

Keywords Alzheimer's disease · Mild cognitive impairment (MCI) · Volumetric MRI · Visual scale

Introduction

Consensus exist that current criteria for the clinical diagnosis of Alzheimer's disease (AD) should be revised [1]. New criteria have thus been proposed recently, which incorporate research results on biomarkers of the underlying disease state, including cerebrospinal fluid and brain imaging [2, 3]. These new criteria stipulate that in addition to the core memory impairment, there must also be at least one or more abnormal biomarkers among structural MRI, PET molecular neuroimaging and cerebrospinal fluid analysis of β-amyloid or tau proteins [2].

Decades of research using MRI have shown that atrophy of medial temporal structures is a core imaging feature of AD [4]. Hippocampal volume can be quantified using MR volumetry with manual [5, 6] or automated segmentation [7]. Clinicians mostly rely on visual assessment of medial temporal lobe atrophy. Visual assessment of atrophy is easily applicable in clinical practice [8–10] but subjective and does not provide a true quantitative assessment of hippocampal volumes.

There is also a recent interest in automated techniques for radiological diagnosis. Computer-based methods have been applied to assist the diagnosis in a variety of brain diseases including AD and mild cognitive impairment (MCI) versus controls [7, 11–14]. These approaches rely on automated classification algorithms such as support vector machines (SVM). Various brain features have been used by these algorithms to classify subjects, including whole brain atrophy [15], gray matter volume using VBM [16] or medial temporal lobe atrophy [7, 13, 14]. Recent comparison between radiological reading and whole brain SVM-based techniques has suggested that well-trained neuroradiologists classify typical AD scans as well as the automated techniques but less experienced radiologists reach poorer performances [17]. Before considering automated techniques as a tool for clinical practice, it is therefore crucial to evaluate the actual improvement that they provide for radiological reading.

In this context, the purpose of our study was twofold. We first evaluated the accuracy of visual assessment of medial temporal lobe atrophy made by radiologists with two levels of expertise as compared with automatic hippocampal volumetry. Secondly, we evaluated whether a computer-based classification method of AD and MCI subjects versus healthy cognitively normal (CN) subjects would achieve better performances when using automated hippocampal volumes as compared with visual assessment as classification features.

Methods

Alzheimer's Disease Neuroimaging Initiative

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organisations, as a $60 million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers and the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California—San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the USA and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research—approximately 200 CN older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information,

see www.adni-info.org. The study was conducted with institutional review board approval and in compliance with HIPAA regulations. All experiments on human subjects were conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from each participant or their legal representatives prior to their inclusion in the study.

Subjects

The ADNI eligibility criteria are described at http://www.adni-info.org/index.php?option=com_content&task=view&id=9&Itemid=43. Briefly, subjects were 55–90 years old, not depressed and had a study partner able to provide an independent evaluation of functioning. CN subjects had mini-mental state examination (MMSE) [18] scores between 24 and 30, Clinical Dementia Rating (CDR) [19] of 0. MCI subjects had MMSE scores between 24 and 30, subjective memory complaint, objective memory loss measured by using education-adjusted scores on the Logical Memory II (Delayed Recall) subscale of the Wechsler Memory Scale, a CDR of 0.5, preserved activities of daily living and an absence of dementia. AD subjects had MMSE scores between 20 and 26, CDR of 0.5 or 1.0 and met the National Institute of Neurological Disorders and Stroke and Alzheimer's Disease and Related Disorders Association criteria for probable AD [1]. The Protocol Summary is available from the ADNI Protocol page of the ADNI-Info Web site at http://www.adni-info.org/Scientists/ADNIGrant/ProtocolSummary.aspx.

We randomly selected 30 AD subjects of the ADNI cohort (16 men and 14 women; mean age±SD, 74 years±6.8; MMSE score, 23.0±1.9; and CDR score, 0.5 for 14 patients and 1 for 16 patients), then 30 MCI (18 men and 12 women; mean age±SD, 74.6 years±6.3; MMSE score, 26.4±1.7; and CDR score, 0.5) and 30 CN (19 men, 11 women; mean age±standard deviation (SD), 74.2 years±4.3; MMSE score, 29.2±1.0; and CDR score, 0), matched for age. They were followed up during 18 months.

MRI acquisition

MRI acquisition was done according to the ADNI acquisition protocol, with protocols individualised for each scanner, as defined at http://www.loni.ucla.edu/ADNI/Research/Cores/index.shtml and described in [20].

Different levels of pre-processing correction above are available at ADNI. We used T1-weighted volumes with 3D gradwarp [21] and B1 non-uniformity corrections [22], selected at the available T0 scanning sessions (baseline or screening), as being the best images that can be obtained in clinical routine.

Fully automated segmentation of the hippocampus

Hippocampal segmentation was carried out with a fully automatic method using probabilistic and anatomical priors [23–25]. Briefly, it segments the hippocampus and the amygdala simultaneously on the basis of competitive region-deformation between these structures. It includes prior knowledge of the relative positions of these structures with respect to 11 sets of anatomic landmarks, which are automatically identified at the border of the deforming objects.

For each subjects, both hippocampi were segmented. The time needed for hippocampal segmentation was 15 min/subject. Segmentation quality was controlled by two observers, blinded to clinical data and diagnosis, using a visual scale of 0–4, 0 stands for completely wrong segmentation results and four for perfect segmentation results, as previously described [25]. All the segmentations were used regardless of their quality grades. Volumes were then normalised by the total intracranial volume (TIV), computed by using SPM5. Normalised hippocampal volume (NHV) was defined as $NHV = HV \times TIV_m/TIV$, where TIV_m was the average of TIV of all the subjects. Normalised volumes of right (HRN) and left (HLN) hippocampi were used in this work.

Visual assessment of medial temporal atrophy

Six radiologists rated separately the 90 scans, identified with a number, in random order using a five-point rating scale (0–4) described by Scheltens et al. [9] based on width of choroid fissure, width of temporal horn of lateral ventricle and height of hippocampal formation (Fig. 1). Radiologists presented two levels of experience, two were experts (more than 10 years neuroradiological practice and 100 MRI examinations per year in AD patients) and four were non-experts (resident, beginning of the first training course in neuroradiology with no neuroradiological experience at time of reading or less than ten readings of MRI scans in AD patients). The experts were radiologists 1 and 2 and the non-experts were radiologists 3 to 6. Radiologists were blind to clinical data but provided information about number of subjects, age range of patients and controls, informed that there were three diagnostic categories (AD, MCI and CN) for differentiation and that categories were age-matched and equal in number. Radiologists had the Scheltens scale in front of them (description and illustration), but they were not given any sort of project training or practice datasets before being asked to apply the five-point visual Scheltens scale. Software used for visualisation was MRIcro [26]. Two readings were made by each radiologist. The interval between the two readings was 15 days. No duration limit was set for each reading. For each reading, the order of subjects

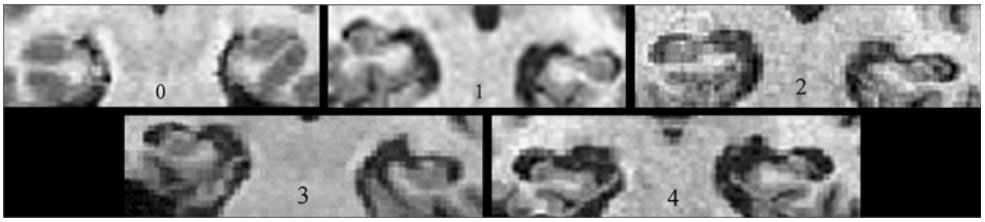

Fig. 1 Representative coronal MRI image of the 0–4 visual score of medial temporal atrophy

was randomised differently. Each hippocampus was evaluated separately. The three criteria of the rating scale and the scores were evaluated on each side and results are presented separately for each hemisphere.

Automatic classification

Values derived from automatic segmentation (more precisely NHV) and from visual assessment were used as features in an automatic classification method. The goal was to evaluate their performances compared with one another and to clinical diagnosis. We used the same methodology as the one employed in a previous study [25], with a nearest mean approach and bootstrap for training set selection to obtain robust estimates of classification rate, sensitivity, specificity and cut-off score. In this procedure, we drew without replacement approximately 75 % of each group to obtain a training set. On this training set, we estimated the mean value (normalised hippocampal volume or visual score) for each group. Each participant in the remaining 25 % was then assigned to the group which mean was closest to the value of this participant. Specifically, if $S1$ and $S2$ were two groups of participants with respective means, defined as $m1$ and $m2$, a new individual with value x was assigned to S1 if (x-$m1$) was less than (x-$m2$) and to S2 if otherwise. The procedure was repeated 5,000 times. We thus obtained the correct classification rates, sensitivity, specificity and cut-off scores for the 5,000 drawings. The cut-off score was only determined by automatic classification.

Statistical analysis

Statistical comparison of clinical characteristics of CN, MCI and AD subjects were performed with ANOVA for age, Chi-square for gender and Kruskal–Wallis for MMSE, hippocampal volume and visual assessment. Post-hoc analyses were performed with a Mann–Whitney U test for MMSE and multiple comparisons of mean ranks for hippocampal volume and visual assessment. Kappa coefficient was used to assess interobserver reliability of quality control of automatic segmentation and intraobserver reliability of visual assessment. To compare the assessment of medial temporal lobe atrophy made visually by radiologists and automatically by hippocampal volumetry, we correlated visual assessment and NHV using Spearman's rank coefficient. Automatic classifications with NHV or visual assessment as features were compared using the McNemar test. Statistical analyses were performed using MedCalc for Windows, version 12.2.1 (MedCalc Software, Mariakerke, Belgium).

Results

Subjects

Demographic and neuropsychological data are given in Table 1. Among the 30 MCI patients, 11 converted to AD during the 18-month follow-up (MCI converters (MCIc)), and 19 remained stable (MCI non-converters (MCInc)). The four groups did not differ for age (ANOVA, $F_3^{86}=1.5$, $p=0.22$) and gender (Chi-square test, $\chi^2=0.85$, 3 df, $p=0.83$). There was a significant difference between groups for the MMSE (Kruskal–Wallis $H_3^{90}=63.6$, $p<0.0001$) and in all pair-wise comparisons using Mann–Whitney U test at $P<.0001$ except between MCInc and MCIc ($p=0.28$). The hippocampal volumes differed between the four groups for the right (Kruskal–Wallis, $H_3^{90}=29$, $p<0.0001$) and left hemispheres (Kruskal–Wallis, $H_3^{90}=30$, $p<0.0001$). HRN were significantly different between CN and MCIc ($p<0.0001$) and between CN and AD ($p<0.0001$). HLN were significantly different between CN and MCIc ($p<0.0001$), CN and AD ($p<0.0001$) and between MCIc and MCInc ($p<0.05$). The visual assessment were significantly different between the four groups for reading 1 for the right (Kruskal–Wallis, $H_3^{90}=30.42$, $p<0.0001$) and the left hemispheres (Kruskal–Wallis, $H_3^{90}=28.6$, $p<0.0001$) with a significantly difference between CN and AD ($p<0.0001$). Visual assessment was significantly different between MCInc and AD only for the left hemispheres ($p<0.05$). The visual assessment were significantly different between the four groups for reading 2 for the right (Kruskal–Wallis, $H_3^{90}=31.17$, $p<0.0001$) and the left hemispheres (Kruskal–Wallis, $H_3^{90}=30.53$, $p<0.0001$) with a significantly difference between CN and AD ($p<0.0001$) and MCInc and AD ($p<0.05$).

Table 1 Baseline characteristics of CN, MCI and AD subjects

	CN	MCI		AD
		MCInc	MCIc	
Sample size	30	19	11	30
Age (years)	74.2±4.3	73.9±7.5	75.6±4.1	74.9±6.8
Gender (F/M)	11/19	7/12	5/6	14/16
MMSE score	29.2±1.0	26.7±1.8	26±1.5	23.0±1.9
CDR (*n*)				
0	30	0	0	0
0.5	0	19	11	14
1	0	0	0	16
NHV (cm^3; right/left)	2.4 (0.3)/2.5 (0.5)	2.1 (0.4)/2.3 (0.5)	1.5 (0.5)/1.6 (0.4)	1.8 (0.7)/1.8 (0.6)
Score of visual assessment				
Reading 1 (right/left)	1.2 (1)/1.2 (1)	1.8 (1.1)/1.8 (1.1)	1.8 (1.3)/1.7 (1.2)	2.7 (1)/2.6 (1.1)
Reading 2 (right/left)	1.2 (1)/1.1 (0.9)	1.7 (1)/1.8 (1.1)	1.9 (1.2)/1.9 (1.2)	2.7 (1.1)/2.7 (1.1)

Data are given as mean±standard deviation for age and MMSE score. For NHV and visual score, data are given for right and left hemisphere (standard deviation). For CDR, the numbers of subjects per score are presented. Score of visual assessment are the average of values for all six raters

MMSE Mini-mental state examination, *CDR* Clinical Dementia Rating scale, *NHV* right and left normalised hippocampal volume, *CN* cognitively normal, *MCI* mild cognitive impairment, *MCInc* mild cognitive impairment non-converter, *MCIc* mild cognitive impairment converter, *AD* Alzheimer's disease

Quality of fully automated segmentation of the hippocampus

For the 30 AD patients, the segmentation proved correct (≥3) for 17 (56.7 %) patients, acceptable (≥2) for 11 (36.6 %) patients and not satisfactory (<2) for 2 (6.7 %). For the 30 MCI patients, the segmentation proved correct (≥3) for 21 (70 %) patients, acceptable (≥2) for 9 (30 %) patients and not satisfactory (<2) for 0. For the 30 CN patients, the segmentation proved correct (≥3) for 24 (80 %) patients, acceptable (≥2) for 5 (16.7 %) patients and not satisfactory (<2) for 1 (3.3 %). There was good to very good interobserver agreement, with kappa coefficients of 0.84, 0.76 and 0.83 for right, left, and mean quality-control scores.

Visual assessment

The time needed for rating was approximately 1 to 2 min by hemispheres. There was moderate to good intraobserver agreement, with kappa coefficients between 0.58±0.06 and 0.63±0.05 for expert readers and between 0.45±0.08 and 0.74±0.04 for non-expert readers.

Correlation between visual assessment and NHV

Visual assessment of medial temporal lobe atrophy correlated significantly with normalised hippocampal volumes for all radiologists and all readings (Fig. 2) except for the right hemispheres on the first reading of one non-expert reader (Table 2). Correlation coefficients varied between −0.17 and −0.53 for the first reading and between −0.26 and −0.49 for the second reading. p values were similar between readers and readings ($p<0.0001$) except for the first readings of two non-expert readers which were not ($p=0.09$) or less ($p=0.01$) correlated with normalised hippocampal volumes and for the right hemispheres of one non-expert reader on the second reading ($p=0.01$).

Comparison of classification accuracy of automatic classification using visual assessment and NHV as discriminant features

Accuracy, sensitivity, specificity and cut-off values for the classification of AD versus CN and MCIc versus CN using NHV obtained by automated segmentation or visual assessment are given in Tables 3 and 4. Mean classification accuracies over the two readings of expert readers were 82 % for AD versus CN, 72 % for MCIc versus CN and 71.5 % for MCInc versus CN. Mean classification accuracies over the two readings of non-expert readers were 75 % for AD versus CN, 58 % for MCIc versus CN and 56 % for MCInc versus CN. Classification accuracies with automatic volumetry were 76 % for AD versus CN, 89.5 % for MCIc versus CN and 64 % for MCInc versus CN. For AD versus CN classification, the difference between the classification accuracy of automatic classification with automated segmentation and visual assessment was statistically significant only on the first reading of one non-expert reader for the left

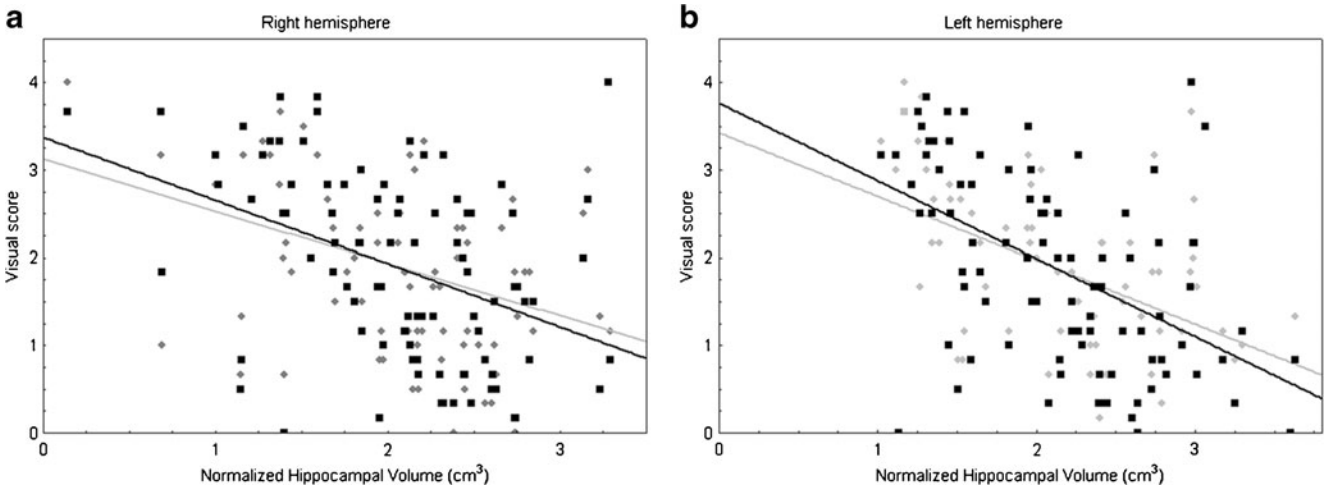

Fig. 2 Correlation between normalised hippocampal volume and visual scores of all raters for *right* (**a**) and *left* (**b**) hemisphere. Reading 1 is displayed in *gray*, reading 2 in *black*. For the right hemisphere, $r=-0.36$ ($p<0.001$) on the reading 1 and -0.41 ($p<0.001$) on the reading 2. For the left hemisphere, $r=-0.44$ ($p<0.001$) on the reading 1 and -0.50 ($p<0.001$) on the reading 2

side ($p=0.03$). For MCIc versus CN, the classification accuracy with automatic segmentation was significantly better than those obtained with visual assessment for all non-expert readers except for the left hemisphere on the second reading of one non-expert ($p=0.09$), and better than one obtained with visual assessment for one expert reader on the first reading for the right hemisphere ($p=0.04$) (Fig. 3). For MCInc versus CN classification, the difference between the classification accuracy of automatic classification with automated segmentation and visual assessment was not statistically significant in all readers ($p>0.05$).

The cut-off scores for the visual assessment varied between readers from 1.6 to 2.2 for the classification of AD versus CN for the first reading and from 1.5 to 2.2 for the second reading. For the classification of MCIc versus CN, cut-off scores varied from 1.2 to 2 for the first reading and from 1.2 to 1.8 for the second reading. For the classification of MCInc versus CN, cut-off scores varied from 1.2 to 1.9 for the first reading and from 1.3 to 1.5 for the second reading.

Discussion

Correlation between visual assessment and NHV

Visual assessment was efficient to evaluate hippocampal atrophy as compared with hippocampal segmentation. This result is in line with previous studies [27–30], which used the same visual scale [9]. Measurements performed in these previous studies included the width of temporal horn [27] or manual segmentation [27–30] for obtaining hippocampal volumes. Here, we extend these results by assessing the effects of expertise and practice. To our knowledge, this is the first time that the Scheltens visual rating scale with expert and non-expert readers has been compared with a computerised method. As expected, practice improved performances in non-expert readers as visual assessment of medial temporal lobe atrophy of two non-expert readers correlated with the automatic volumes better for the second than the first reading. In clinical practice, expert neuroradiologists work primarily in specialised centres whereas MRI

Table 2 Spearman rank correlations (r) between visual rating scores and hippocampal volume for each of the two readings

Rater No.	First reading		Second reading	
	Right	Left	Right	Left
Expert radiologists				
1	−0.40 (<0.0001)	−0.53 (<0.0001)	−0.42 (<0.0001)	−0.47 (<0.0001)
2	−0.41 (<0.0001)	−0.43 (<0.0001)	−0.41 (<0.0001)	−0.49 (<0.0001)
Non-expert radiologists				
3	−0.39 (0.0002)	−0.46 (0.0002)	−0.39 (<0.0001)	−0.49 (<0.0001)
4	*−0.17 (0.09)*	−0.27 (0.01)	−0.39 (0.0002)	−0.45 (<0.0001)
5	−0.36 (0.0004)	−0.37 (0.0003)	−0.26 (0.01)	−0.41 (<0.0001)
6	−0.25 (0.02)	−0.24 (0.02)	−0.41 (<0.0001)	−0.48 (<0.0001)

Data are given as rho value (p value). Non-significant p value is set in italic

Table 3 Sensitivity, specificity, percent correct classification and cut-off scores or volume of AD versus CN with automatic classification using automated hippocampal volumetry and visual assessment of radiologists

Rater No.	Right			Left		
	Sensitivity/ specificity classification (%)	Cut-off	*p* value	Sensitivity/ specificity classification (%)	Cut-off	*p* value
Volumetry (cm^3)	70/82/76	2.1		73/79/76	2.2	
Expert radiologists						
1						
R1	86/77/81	1.8	0.64	86/80/83	1.7	0.34
R2	85/77/81	1.8	0.61	85/77/81	1.8	0.58
2						
R1	90/80/85	1.7	0.30	86/83/85	1.6	0.27
R2	77/83/80	1.7	0.79	73/90/82	1.5	0.51
Non-expert radiologists						
3						
R1	75/85/80	2.1	0.58	77/92/84	2.1	0.30
R2	70/86/78	2.1	1	64/85/75	2.1	1
4						
R1	79/64/71	1.9	0.66	75/73/74	2	1
R2	79/70/75	1.9	1	72/83/78	2.1	0.79
5						
R1	75/67/71	1.9	0.81	82/67/74	1.8	1
R2	85/70/78	1.8	1	83/70/77	1.8	1
6						
R1	63/63/63	2.2	0.11	63/59/61	2.2	*0.03*
R2	77/83/80	2.2	0.77	78/83/80	2.1	0.61

p values were obtained using a MacNemar test. Significant *p* value is set in italic

R1 first reading, *R2* second reading, *AD* Alzheimer disease, *CN* cognitively normal

examinations of MCI and AD patients are often performed by radiologists with less neuroradiological experience.

Comparison of classification accuracy of automatic classification using visual assessment and NHV as discriminant features

In the present study, classification accuracy of automatic classification was 76 % for AD versus CN, 89.5 % for MCIc versus CN and 64 % for MCInc versus CN using automated hippocampal segmentation, in line with previous results using a larger database and automatic hippocampus segmentation [25] and also numerous other hippocampal volumetry studies [13, 31–33]. Classification accuracy for AD versus CN was less good using hippocampal segmentation than whole brain atrophy [14] but similar for MCIc versus CN. Lower classification accuracy in AD subjects was probably due to the lower performances of hippocampus segmentation in these patients compared with MCI or control subjects. Automated segmentation may be easily performed in clinical practice, with fully automated segmentation of the hippocampus requiring only a T1-weighted volume, and no manual intervention from the radiologist. Using visual assessment, the overall classification accuracy for AD versus CN, MCIc versus CN and MCInc versus CN was better for scores of expert than non-expert radiologists.

For AD versus CN and MCInc versus CN, the classification accuracy of automatic classification was similar for hippocampal volumes and visual assessment. Some studies reported that visual assessment were more discriminant than volumetry to distinguish AD from control [8, 28] but others did not confirm an advantage of visual assessment [27, 30, 34]. For MCIc versus CN in non-expert readers, the classification accuracy was better using hippocampal volumes than visual assessment and the second reading was not better than the first reading. In MCI patients, visual assessment of medial temporal lobe atrophy was more difficult than in AD as atrophy was less severe. This explains why expert readers and hippocampal segmentation were more accurate than non-expert readers.

Mean cut-off scores from experts ranged from 1.2 to 1.5 to distinguish MCIc from CN and from 1.5 to 1.8 to distinguish AD from CN, in line with previously reported cut-off

Table 4 Sensitivity, specificity, percent correct classification and cut-off scores or volume of MCIc versus CN with automatic classification using automated hippocampal volumetry and visual assessment of radiologists

Rater No.	Right			Left		
	Sensitivity/ specificity classification (%)	Cut-off	*p* value	Sensitivity/ specificity classification (%)	Cut-off	*p* value
Volumetry (cm^3)	82/95/91	2		90/87/88	2.08	
Expert radiologists						
1						
R1	44/76/67	1.3	*0.04*	46/77/69	1.2	0.12
R2	73/77/76	1.5	0.23	63/76/63	1.4	0.51
2						
R1	64/80/75	1.4	0.23	64/83/78	1.3	0.23
R2	45/83/73	1.4	0.11	45/90/78	1.3	0.45
Non-expert radiologists						
3						
R1	62/53/55	1.6	*0.003*	64/57/59	1.7	*0.01*
R2	72/63/66	1.7	*0.02*	72/63/65	1.6	0.09
4						
R1	30/52/46	1.4	*0.0007*	30/48/43	1.5	*0.0007*
R2	64/64/64	1.6	*0.007*	63/60/61	1.6	*0.01*
5						
R1	51/61/58	1.4	*0.0042*	53/66/62	1.3	*0.02*
R2	31/61/53	1.2	*0.0042*	44/68/61	1.3	*0.035*
6						
R1	55/55/55	2	*0.001*	40/48/45	1.9	*0.001*
R2	64/66/65	1.8	*0.007*	64/67/66	1.7	*0.04*

p values were obtained using a MacNemar test. Significant *p* values are set in italics

R1 first reading, *R2* second reading, *MCIc* mild cognitive impairment converter, *CN* cognitively normal

score of 1.33 using a modified version of Scheltens scale, the Visual Rating System [35], whereas dichotomised scores as 2 or less (for CN) or more than 2 (for AD) using Scheltens scale were reported [9, 36].

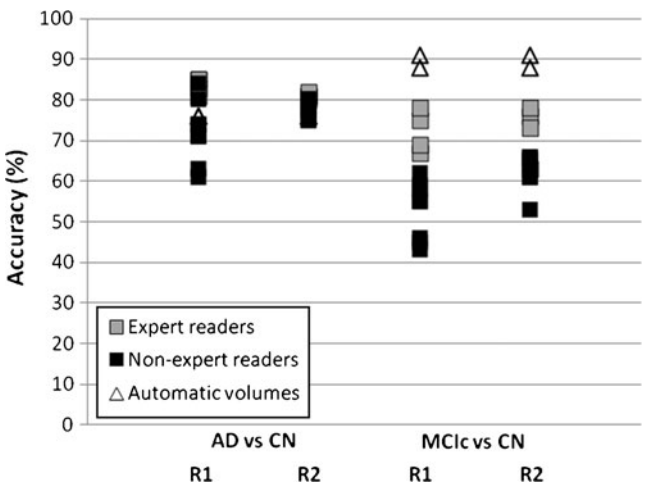

Fig. 3 Classification accuracy for AD, MCIc and CN for readings 1 (*R1*) and 2 (*R2*) compared with the results derived from the automatic volumes

Limitations

Best quality of segmentation was obtained with CN, then with MCI and finally with AD, in accordance with a previous study, which had shown that the errors of segmentation did not influence the performances of classification on larger samples [25].

This study was conducted on a research cohort with optimised MRI acquisition procedure therefore the scans were likely of better quality than scans available in clinical settings. Another limitation of the study was that medial temporal lobe atrophy is not specific of AD, as it has been reported in numerous other dementias including fronto-temporal lobar dementia, dementia with Lewy body or vascular dementia as well as in patients with temporal epilepsy due to hippocampal sclerosis [37].

Conclusions

Concordance between visual assessment and automatic hippocampal segmentation was high, showing that radiologists accurately evaluate hippocampal atrophy. Automated segmentation,

requiring no manual intervention from the radiologist, gives the same results whatever the person who uses it. Radiological assessment took less time and was as accurate as computer-based volumetry for the classification of AD and MCInc versus CN but less accurate for the classification of MCIc versus CN.

Acknowledgements The authors thank Christie Fock-Yee, Flore Viry and Patricia Ziggiatti Cavalheiro (AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Département de Neuroradiologie, Paris, France) for their data interpretation. Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

Conflict of interest We declare that we have no conflict of interest.

References

1. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34(7):939–944
2. Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P (2007) Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol 6(8):734–746
3. Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH (2011) Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 7(3):257–262
4. Bottino CM, Castro CC, Gomes RL, Buchpiguel CA, Marchetti RL, Neto MR (2002) Volumetric MRI measurements can differentiate Alzheimer's disease, mild cognitive impairment, and normal aging. Int Psychogeriatr 14(1):59–72
5. Jack CR, Petersen RC, O'Brien PC, Tangalos EG (1992) MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. Neurology 42(1):183–188
6. Lehéricy S, Baulac M, Chiras J, Piérot L, Martin N, Pillon B, Deweer B, Dubois B, Marsault C (1994) Amygdalohippocampal MR volume measurements in the early stages of Alzheimer disease. AJNR Am J Neuroradiol 15(5):929–937
7. Colliot O, Chételat G, Chupin M, Desgranges B, Magnin B, Benali H, Dubois B, Garnero L, Eustache F, Lehéricy S (2008) Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248(1):194–201
8. Wahlund LO, Julin P, Johansson SE, Scheltens P (2000) Visual rating and volumetry of the medial temporal lobe on magnetic resonance imaging in dementia: a comparative study. J Neurol Neurosurg Psychiatry 69(5):630–635
9. Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, Kuiper M, Steinling M, Wolters EC, Valk J (1992) Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatry 55(10):967–972
10. Scheltens P, Launer LJ, Barkhof F, Weinstein HC, van Gool WA (1995) Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: interobserver reliability. J Neurol 242(9):557–560
11. Teipel SJ, Born C, Ewers M, Bokde AL, Reiser MF, Möller HJ, Hampel H (2007) Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. NeuroImage 38(1):13–24
12. Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, Boeve BF, Petersen RC, Jack CR (2008) Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39(3):1186–1197
13. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O, Initiative TAsDN (2010) Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage 56:766–781
14. Matsuda H, Mizumura S, Nemoto K, Yamashita F, Imabayashi E, Sato N, Asada T (2012) Automatic voxel-based morphometry of structural MRI by SPM8 plus diffeomorphic anatomic registration through exponentiated lie algebra improves the diagnosis of probable Alzheimer disease. AJNR Am J Neuroradiol (in press)
15. Driscoll I, Davatzikos C, An Y, Wu X, Shen D, Kraut M, Resnick SM (2009) Longitudinal pattern of regional brain volume change differentiates normal aging from MCI. Neurology 72(22):1906–1913
16. Karas GB, Burton EJ, Rombouts SA, van Schijndel RA, O'Brien JT, Scheltens P, McKeith IG, Williams D, Ballard C, Barkhof F (2003) A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. NeuroImage 18(4):895–907
17. Klöppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, Mader I, Mitchell LA, Patel AC, Roberts CC, Fox NC, Jack CR, Ashburner J, Frackowiak RS (2008) Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. Brain 131(Pt 11):2969–2974
18. Folstein MF, Folstein SE, McHugh PR (1975) Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12(3):189–198
19. Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL (1982) A new clinical scale for the staging of dementia. Br J Psychiatry 140:566–572
20. Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell LJ, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DL, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW (2008) The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 27(4):685–691
21. Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, Kennedy D, Schmitt F, Brown G, Macfall J, Fischl B,

Dale A (2006) Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. NeuroImage 30(2):436–443
22. Narayana PA, Brey WW, Kulkarni MV, Sievenpiper CL (1988) Compensation for surface coil sensitivity variation in magnetic resonance imaging. Magn Reson Imaging 6(3):271–274
23. Chupin M, Mukuna-Bantumbakulu AR, Hasboun D, Bardinet E, Baillet S, Kinkingnéhun S, Lemieux L, Dubois B, Garnero L (2007) Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: method and validation on controls and patients with Alzheimer's disease. NeuroImage 34(3):996–1019
24. Chupin M, Hammers A, Liu RS, Colliot O, Burdett J, Bardinet E, Duncan JS, Garnero L, Lemieux L (2009) Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. NeuroImage 46(3):749–761
25. Chupin M, Gérardin E, Cuingnet R, Boutet C, Lemieux L, Lehéricy S, Benali H, Garnero L, Colliot O, AsDN I (2009) Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19(6):579–587
26. Rorden C, Brett M (2000) Stereotaxic display of brain lesions. Behav Neurol 12(4):191–200
27. Bresciani L, Rossi R, Testa C, Geroldi C, Galluzzi S, Laakso MP, Beltramello A, Soininen H, Frisoni GB (2005) Visual assessment of medial temporal atrophy on MR films in Alzheimer's disease: comparison with volumetry. Aging Clin Exp Res 17(1):8–13
28. Wahlund LO, Julin P, Lindqvist J, Scheltens P (1999) Visual assessment of medical temporal lobe atrophy in demented and healthy control subjects: correlation with volumetry. Psychiatry Res 90(3):193–199
29. Ridha BH, Barnes J, van de Pol LA, Schott JM, Boyes RG, Siddique MM, Rossor MN, Scheltens P, Fox NC (2007) Application of automated medial temporal lobe atrophy scale to Alzheimer disease. Arch Neurol 64(6):849–854
30. Visser PJ, Verhey FR, Hofman PA, Scheltens P, Jolles J (2002) Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. J Neurol Neurosurg Psychiatry 72(4):491–497
31. de Leon MJ, Convit A, George AE, Golomb J, de Santi S, Tarshish C, Rusinek H, Bobinski M, Ince C, Miller D, Wisniewski H (1996) In vivo structural studies of the hippocampus in normal aging and in incipient Alzheimer's disease. Ann N Y Acad Sci 777:1–13
32. Frisoni GB, Laakso MP, Beltramello A, Geroldi C, Bianchetti A, Soininen H, Trabucchi M (1999) Hippocampal and entorhinal cortex atrophy in frontotemporal dementia and Alzheimer's disease. Neurology 52(1):91–100
33. Pennanen C, Kivipelto M, Tuomainen S, Hartikainen P, Hänninen T, Laakso MP, Hallikainen M, Vanhanen M, Nissinen A, Helkala EL, Vainio P, Vanninen R, Partanen K, Soininen H (2004) Hippocampus and entorhinal cortex in mild cognitive impairment and early AD. Neurobiol Aging 25(3):303–310
34. Westman E, Cavallin L, Muehlboeck JS, Zhang Y, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Soininen H, Spenger C, Lovestone S, Simmons A, Wahlund LO, consortium A (2011) Sensitivity and specificity of medial temporal lobe visual ratings and multivariate regional MRI classification in Alzheimer's disease. PLoS One 6(7): e22506
35. Duara R, Loewenstein DA, Potter E, Appel J, Greig MT, Urs R, Shen Q, Raj A, Small B, Barker W, Schofield E, Wu Y, Potter H (2008) Medial temporal lobe atrophy on MRI scans and the diagnosis of Alzheimer disease. Neurology 71(24):1986–1992
36. DeCarli C, Frisoni GB, Clark CM, Harvey D, Grundman M, Petersen RC, Thal LJ, Jin S, Jack CR, Scheltens P, Group AsDCS (2007) Qualitative estimates of medial temporal atrophy as a predictor of progression from mild cognitive impairment to dementia. Arch Neurol 64(1):108–115
37. Knoops AJ, van der Graaf Y, Appelman AP, Gerritsen L, Mali WP, Geerlings MI (2009) Visual rating of the hippocampus in non-demented elders: does it measure hippocampal atrophy or other indices of brain atrophy? The SMART-MR study. Hippocampus 19 (11):1115–1122