J Neurol (1995) 242:557–560
© Springer-Verlag 1995

Philip Scheltens
Leonore J. Launer
Frederik Barkhof
Henri C. Weinstein
Willem A. van Gool

Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: interobserver reliability

Received: 5 August 1994
Received in revised form:
22 December 1994
Accepted: 12 January 1995

P. Scheltens (✉)
Department of Neurology,
Academisch Ziekenhuis VU, P.O. Box 7057,
1007 MB Amsterdam, The Netherlands

L. J. Launer
Department of Psychiatry,
Vrije Universiteit,
Amsterdam, The Netherlands

F. Barkhof
Department of Diagnostic Radiology,
Vrije Universiteit,
Amsterdam, The Netherlands

H. C. Weinstein
Department of Neurology,
St. Lucas Ziekenhuis,
Amsterdam, The Netherlands

W. A. van Gool
Department of Neurology,
Academic Medical Centre,
Amsterdam, The Netherlands

Abstract We conducted an interobserver study to assess agreement on visual rating of medial temporal lobe atrophy on coronal T1-weighted MRI. A total of 100 studies of elderly individuals, using two different MRI techniques (spin echo and inversion recovery sequences), were analysed by four raters (three neurologists and one neuroradiologist) using a five-point rating scale. Complete agreement was found in 37% of the total sample. Interobserver agreement as expressed by kappa values was 0.44 (95% CI=0.34–0.54) and 0.51 (95% CI=0.41–0.61) for the two techniques. After dichotomizing medial temporal lobe atrophy into present or absent, a post hoc analysis revealed higher complete agreeement (70%), with kappa values of 0.59 (95% CI=0.51–0.67) and 0.62 (95% CI=0.48–0.075), for the two techniques (all four raters). From this study we conclude that visual rating of medial temporal lobe atrophy on MRI in the coronal plane yields fair to good agreement among observers. We recommend this type of visual rating for use in clinical settings when a quick judgement on the presence of medial temporal lobe atrophy is needed.

Key words Magnetic resonance imaging · Interobserver agreement · Medial temporal lobe atrophy · Alzheimer's disease

Introduction

Quantitative studies have indicated that medial temporal lobe atrophy (MTA) on magnetic resonance imaging (MRI) appears significantly more severe in patients with Alzheimer's disease (AD) than in age-matched controls [5, 8, 9]. In 1992 we reported similar results using visual assessment of MTA on coronal MRI slices with a newly developed rating scale based on the assessment of the size of the hippocampal formation relative to the surrounding cerebrospinal fluid (CSF) spaces, and supplied guidelines for the use of the scale [12]. Recently we used the same method to identify AD patients among community-dwelling elderly satisfactorily [13]. We now report the results of the interobserver reliability of this scale based on 100 MRI studies analysed by four observers.

Materials and methods

In order to ensure inclusion of material showing a wide variety of MTA, we selected MRI scans of participants in the AMSTEL study (Amsterdam Study of the Elderly) [11], a population-based study of cognitive decline in the elderly, and consecutive images obtained for screening purposes in a memory clinic. The patient sample thus included a variety of elderly non-demented and demented subjects (with AD and other dementias). A total of 100 MRI scans were selected, blinded with respect to the age and the name of the subjects.

MRI was performed on a 0.6T unit employing two different techniques. Half of the images (50) were Tl-weighted spin-echo sequences (SE), with a TR of 300ms and a TE of 22ms, four excitations, scan time 6min, field of view 20cm, with six oblique slices and a slice thickness of 5mm (interslice gap 1mm, in-plane resolution 0.8 × 1.0mm) planned parallel to the brain stem axis and perpendicular to the hippocampal axis; this was the same method as described earlier [12]. The other 50 images were obtained using the inversion recovery (IR) technique (TR 2000, TI 500, TE 32ms, two excitations), with otherwise identical scan parameters and a scan time of 12min. The IR technique was implemented because it offers a better grey/white matter contrast sensitivity (Fig. 1).

Table 1 Visual rating of medial temporal lobe atrophy [12] (↑ = increase, ↓ = decrease, *N* = normal)

Score	Width of choroide fissure	Width of temporal horn	Height of hippocampal formation
0	N	N	N
1	↑	N	N
2	↑ ↑	↑	↓
3	↑ ↑ ↑	↑ ↑	↓ ↓
4	↑ ↑ ↑	↑ ↑ ↑	↓ ↓ ↓

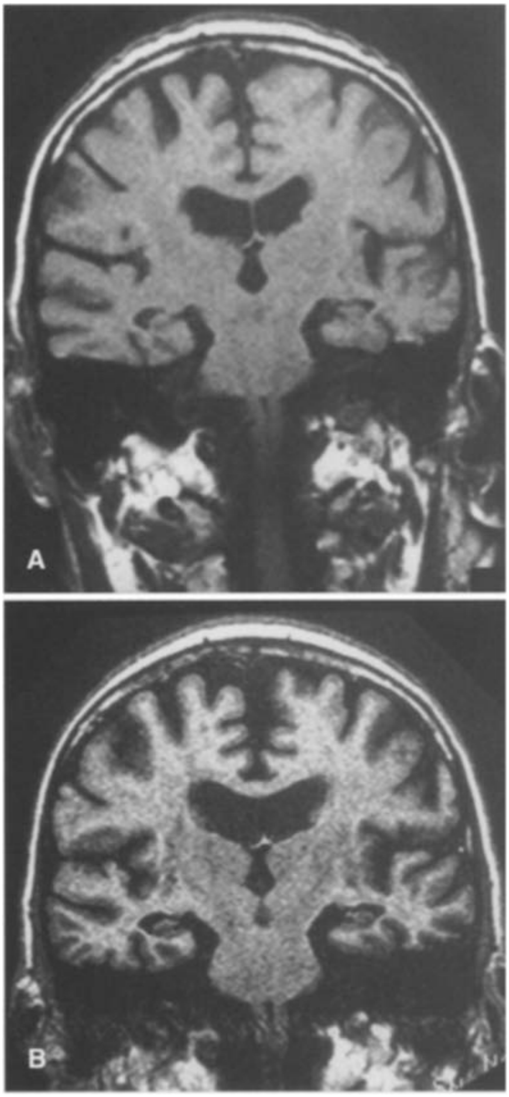


Fig. 1 Two coronal MRI slices of the same patient, using spin-echo (**A**) and inversion recovery technique (**B**) (See text for scan parameters). Scan B was made 1 year after scan A. The panel unanimously gave A a score of 1 and B a score of 2

The five-point rating scale (0–4) is based on the width of the surrounding CSF spaces and the height of the hippocampal formation, which includes the hippocampus proper, the subiculum, and the parahippocampal and dentate gyri [12]. A score of 0 is assigned when no CSF is seen surrounding the hippocampus. A score of 4 is given when there is severe atrophy of the medial temporal lobe and the normal anatomy of the hippocampus is no longer visible, with enlargement of the temporal horn and the choroid fissure (Table 1). Both sides are included in the assessment. In cases of severe asymmetry the score of the more affected side is given. In all studies, the raters were asked to judge the degree of MTA on the slice that best depicted the hippocampal formation and surrounding structures.

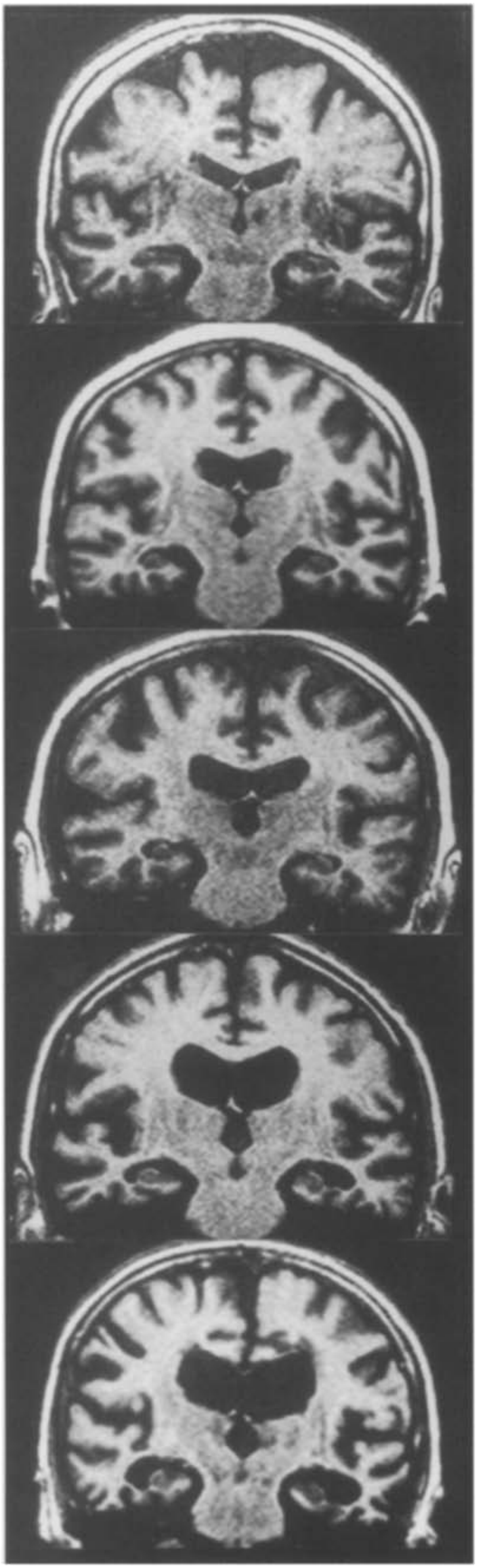
The raters were three neurologists and one radiologist. Two of them (raters 1 and 2) were involved in the development of the scale and could be considered to have more experience with rating MTA on MRI than the other two. The scans were identified by a number only, to make sure all raters were blinded to the identity, age, sex and diagnosis of the subjects. After a short training session, the raters judged all scans in random order, which was different for each rater.

The degree of agreement among observers was expressed by means of kappa statistics [6]. According to Fleiss [6] a kappa value below 0.40 denotes poor agreement, between 0.40 and 0.75 means fair to good agreement, and a kappa value above 0.75 indicates excellent agreement. The overall kappa [95% confidence interval (CI)] was calculated for all four raters for both techniques separately. In addition, kappa values were calculated for the agreement between each pair of raters for both techniques.

Results

Inspection of the raw data (not shown) disclosed that all raters agreed completely on 37 of 100 scans. The largest difference between raters was 2 points (scores 0 and 2 for the same scan). On average, rater 2 (P.S.) gave lower scores than the other three raters (1.62 vs 1.79, 1.90 and 1.91). A score of 4 was only assigned to 2 scans, a score on which all raters agreed. A score of 0 was given 9 times by rater 2, but never by rater 3. In Fig. 2 examples of the scores are given.

The overall kappa value for the SE technique was 0.44 (95% CI=0.34–0.54) and for the IR technique 0.51 (95% CI=0.41–0.61), indicating poor to fair agreement for both techniques. The kappa values for agreement be-



◀ **Fig. 2** Five coronal MRI slices made with the inversion recovery technique. All raters agreed completely on the scores of 2, 3 and 4 for the lower three scans. However, there was disagreement on the score for the first scan (0 by two raters and 1 by two raters) and on the score for the second scan (2 by two raters and 1 by two raters)

Table 2 Kappa values for the interobserver agreement on assessing medial temporal lobe atrophy on MRI among four raters using spin-echo (SE)/inversion recovery (IR) techniques

	Rater 2	Rater 3	Rater 4
Rater 1	0.38/0.57	0.48/0.58	0.63/0.51
Rater 2		0.26/0.46	0.36/0.34
Rater 3			0.58/0.55

Table 3 Kappa values for the interobserver agreement on assessing medial temporal lobe atrophy dichotomized into absent (scores 0, 1) versus present (scores 2–4), among all raters with SE/IR techniques

	Rater 2	Rater 3	Rater 4
Rater 1	0.68/0.70	0.63/0.61	0.67/0.45
Rater 2		0.51/0.59	0.54/0.45
Rater 3			0.77/0.58

tween the different pairs of raters are given in Table 2. As can be seen, agreement varies between 0.26 and 0.63 for the SE technique and between 0.34 and 0.57 for the IR technique.

In accordance with a previous study [13] the scores were dichotomized into MTA absent (scores 0 and 1) and MTA present (scores 2–4). A post hoc analysis revealed that the overall kappa value for the SE technique was 0.59 (95% CI=0.51–0.67) and for the IR technique 0.62 (95% CI=0.48–0.75), indicating good agreement for both techniques. The kappa values for agreement between the different pairs of raters are given in Table 3.

Discussion

We found that the assessment of MTA on MRI using visual inspection showed fair to good interrater agreement, regardless of the experience of the rater. Although the IR technique yields a higher contrast sensitivity, the interobserver agreement was on average only slightly higher when using this technique.

Post hoc analysis of our data revealed that agreement improved when only the presence or absence of MTA was considered. The same was found by De Leon et al. [4], who reported 100% agreement using a dichotomized scale. This method of scoring may be more relevant in clinical settings [13], and it would thus be worthwhile to consider condensing the rating system into a 1–3 or 0/1 scale. A similar conclusion was drawn from our visual-volumetric correlation study [14].

Other studies addressing the same issue seemed to yield better results, although true comparison is difficult, since some of the studies were performed using CT [3, 4, 10] or used another method of assessing MTA [7] or interobserver agreement [2–4, 7, 10]. De Leon et al. [4] reported a high correlation coefficient (0.92) between two observers (four-point scale) on a sample of 25 CT studies obtained with their negative angle technique (transverse slices). Kido et al. [10] also reported a high correlation (0.90) between two observers in rating temporal lobe atrophy on a three-point scale using CT scans of 24 AD patients and 18 controls. The consortium to establish a registry for AD (CERAD) used MRI and reported acceptable intraclass correlation coefficients (> 0.79) when 14 radiologists subjectively scored (four-point scale) the size of the lateral ventricles and the temporal horns, and an intraclass coefficient of 0.70 for Sylvian fissure enlargement and global atrophy of the brain [2]. Interpretation of these results should be done with caution, since correlation coefficients are not appropriate for measuring agreement [1].

We conclude that our scale for visual rating of MTA on mid-field MRI in the coronal plane yields fair to good agreement among observers. The method can be used on all MRI systems. The use of higher-field systems will increase the spatial resolution and probably also the interobserver agreement. We recommend that this type of visual rating of MTA be used in clinical settings when a quick judgement on the presence or absence of MTA is needed.

Acknowledgements The authors are grateful to C. de Lezenne Coulander for help with the statistical analysis. Dr. Scheltens is the recipient of a grant from the Universitair Stimulerings Fonds (USF).

References

1. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet I:307–310
2. Davis PC, Gray L, Albert M, Wilkinson W, Hughes J, Heyman A, Gado M, Kumar AJ, Destian S, Lee C, Duvall E, Kido D, Nelson MJ, Bello J, Weathers S, Jolesz F, Kikinis R, Brooks M (1992) The consortium to establish a registry for Alzheimer's disease (CERAD). III. Reliability of a standardized MRI evaluation of Alzheimer's disease. Neurology 42: 1676–1680
3. De Leon MJ, George AE, Stylopoulos LA, Smith G, Miller DC (1989) Early marker for Alzheimer's disease: the atrophic hippocampus. Lancet II:672–673
4. De Leon MJ, Golomb J, George AE, Convit A, Rusinek H, Morys J, Bobinski M, De Santi S, Tarshish C, Narkiewicz O, Wisniewski HM (1993) Hippocampal formation atrophy: prognostic significance for Alzheimer's disease. In: Corain B, Iqbal K, Nicolini M, Winblad B, Wisniewski HM, Zatta PF (eds) Alzheimer's disease: advances in clinical and basic research. Wiley, Chicester, pp 35–46
5. Erkinjuntti T, Lee DH, Gao F, Steenhuis R, Eliasziw M, Fry R, Merskey H, Hachinsky VC (1993) Temporal lobe atrophy on magnetic resonance imaging in the diagnosis of early Alzheimer's disease. Arch Neurol 50: 305–310
6. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382
7. Golomb J, Leon MJ De, Kluger A, George AE, Tarshish C, Ferris S (1993) Hippocampal atrophy in normal aging: an association with recent memory. Arch Neurol 50:967–973
8. Jack CR Jr, Petersen RC, O'Brien PC, Tangalos EG (1992) MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. Neurology 42:183–188
9. Kesslak JP, Nalcioglu O, Cotman CW (1991) Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer's disease. Neurology 41:51–54
10. Kido DK, Caine ED, Lemay M, Ekholm S, Booth H, Panzer R (1989) Temporal lobe atrophy in patients with Alzheimer disease: a CT study. AJNR 10:551–555
11. Launer LJ, Dinkgreve M, Jonker C, Hooyer C, Lindeboom J (1993) Are age and education independent correlates of the Mini-Mental state exam performance of community dwelling elderly? J Gerontol 48:271-277
12. Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, Kuiper MA, Steinling M, Wolters EC, Valk J (1992) Atrophy of the medial temporal lobes on MRI in ''probable'' Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatry 55:967–972
13. Scheltens P, Launer LJ, Weinstein HC, Barkhof F, Jonker C (1994) The value of SPECT and MRI in early diagnosis of Alzheimer's disease. Neurology 44 [Suppl 2]:A179
14. Vermersch P, Scheltens P, Leys D, Barkhof F (1994) Visual rating of hippocampal atrophy: correlation with volumetry. J Neurol Neurosurg Psychiatry 57:1015