Responsiveness of magnetic resonance imaging and neuropsychological assessment in memory clinic patients

Schmand, B.A.; Rienstra, A.; Tamminga, G.H.; Richard, E.; van Gool, W.A.; Caan, M.W.A.; Majoie, C.B.

Published in:
Journal of Alzheimer's Disease

DOI:
10.3233/JAD-131484

Journal of Alzheimer's Disease 40 (2014) 409–418
DOI 10.3233/JAD-131484
IOS Press

Responsiveness of Magnetic Resonance Imaging and Neuropsychological Assessment in Memory Clinic Patients

Ben Schmand[a,c,*], Anne Rienstra[c], Hyke Tamminga[c], Edo Richard[a], Willem A. van Gool[a], Matthan W.A. Caan[b] and Charles B. Majoie[b]
[a]*Department of Neurology, Academic Medical Center, Amsterdam, The Netherlands*
[b]*Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands*
[c]*Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands*

Handling Associate Editor: Andreas Monsch

Accepted 27 November 2013

Abstract.
Background: Scales of global cognition and behavior, often used as endpoints for intervention trials in Alzheimer's disease (AD) and mild cognitive impairment (MCI), are insufficiently responsive (i.e., relatively insensitive to change). Large patient samples are needed to detect beneficial drug effects. Therefore, magnetic resonance imaging (MRI) measures of cerebral atrophy have been proposed as surrogate endpoints.
Objective: To examine how neuropsychological assessment compares to MRI in this respect.
Methods: We measured hippocampal atrophy, cortical thickness, and performance on neuropsychological tests in memory clinic patients at baseline and after two years. Neurologists rated the patients as cognitively normal ($n = 28$; Clinical Dementia Rating, CDR = 0) or as impaired ($n = 34$; CDR > 0). We administered five tests of memory, executive functioning, and verbal fluency. A composite neuropsychological score was calculated by taking the mean of the demographically corrected standard scores. MRI was done on a 3 Tesla scanner. Volumetric measurements of the hippocampus and surrounding cortex were made automatically using FreeSurfer software.
Results: The composite neuropsychological score deteriorated 0.6 SD in the impaired group, and was virtually unchanged in the normal group. Annual hippocampal atrophy rates were 3.4% and 0.6% in the impaired and normal cognition groups, respectively. Estimates of required sample sizes to detect a 50% reduction in rate of change were larger using rate of hippocampal atrophy ($n = 131$) or cortical thickness ($n = 488$) as outcome compared to change scores on neuropsychological assessment ($n = 62$).
Conclusion: Neuropsychological assessment is more responsive than MRI measures of brain atrophy for detecting disease progression in memory clinic patients with MCI or AD.

Keywords: Alzheimer's disease, cognition, hippocampus, longitudinal design, magnetic resonance imaging, mild cognitive impairment, neuropsychological tests, responsiveness

*Correspondence to: Ben Schmand, AMC Neurology H2-262, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. Tel.: +31 20 566 8632/+31 20 525 6849; Fax: +31 20 566 9217; E-mail: b.schmand@amc.nl.

ISSN 1387-2877/14/$27.50

INTRODUCTION

Responsiveness is an important quality of evaluative measures that assess effects of experimental treatments in intervention trials. It refers to the capacity of detecting changes in health status. A responsive instrument detects changes over time when they actually occur, whereas it does not show any change when the patient remains stable [1, 2]. Thus, the more responsive an instrument, the better it uncovers real change. An important practical consequence is that the statistical power of a trial is larger if a more responsive outcome measure is used, and the sample size can be smaller, or the duration of the intervention can be shorter.

In the context of Alzheimer's disease (AD) and mild cognitive impairment (MCI), outcome measures used in clinical trials are cognitive and behavior rating scales, mostly the Alzheimer's Disease Assessment Scale (ADAS) [3] and the Clinical Dementia Rating (CDR) [4]. Until today, these scales are gold standards for pharmacological research on AD [5], although regulatory authorities are considering changing the requirements [6]. The ADAS and the CDR are often used as co-primary endpoints [7]. Other clinical scales frequently used are the Mini-Mental State Examination (MMSE) [8], and, in Europe, the cognitive section of the Cambridge Mental Disorders of the Elderly Examination (Cam-cog) [9].

Such scales are often considered unreliable [10, 11] and not sufficiently responsive [12–14] except in moderately severe AD [7, 15]. Several authors therefore proposed to replace these clinical scales as the gold standard with more reliable or more reproducible measures, in particular magnetic resonance imaging (MRI) measures of cerebral atrophy [16–19]. For example, if atrophy of the hippocampus would serve as the main outcome in trials of anti-AD compounds, the required sample sizes would be much smaller than those in current clinical trials. Calculations based on the dataset of the Alzheimer's Disease Neuroimaging Initiative (ADNI) showed that samples might be reduced by about a factor 5 in trials with AD patients, and even more in trials with MCI patients [17, 18]. Meta-analytic studies of the rate of hippocampal atrophy and of decline on global cognitive scales in untreated patients corroborate that effect sizes are larger for MRI than for most clinical scales [20–22]. Also the cortical thickness of temporal areas surrounding the hippocampus might be sensitive measures in this context [23, 24].

The status of the ADAS and similar scales as the gold standard for pharmacological trials can also be criticized on psychometric grounds [12, 25–28]. The question arises, therefore, how the comparison with MRI would be if neuropsychological tests with sound psychometric properties were used instead of global scales with intrinsic psychometric shortcomings. Primary candidates would be tests of memory, verbal fluency (i.e., speeded naming), and executive functioning, because these domains are affected most in MCI and early AD [7, 12, 29]. Moreover, these outcome measures directly assess the actual symptoms. This is in general preferable over indirect outcome measures assessing a biological process without direct link to clinical symptoms.

Cross-sectional studies in prevalent AD patients and healthy elderly show large effect sizes for memory tests, especially delayed recall, much larger than for hippocampal atrophy [30, 31]. This is not surprising, because amnesia is a defining characteristic of AD. Also longitudinal studies in healthy elderly and MCI patients show larger differences for memory tests than for hippocampal atrophy between subjects who converted to dementia within a few years and those who did not [29, 32].

In view of these promising research findings on neuropsychological tests, and given the psychometric weaknesses of global scales, we conducted a longitudinal study to directly compare the responsiveness of MRI measures of brain atrophy and tests of memory, fluency, and executive functioning in an unselected sample of memory clinic patients. We expected that the cognitive tests would be at least as responsive as the rate of atrophy of the hippocampus and surrounding areas in the subset of patients who showed clinical signs of cognitive decline. We also calculated required sample sizes for a hypothetical trial with neuropsychological assessment, hippocampal atrophy, or cortical thickness as surrogate endpoints.

MATERIALS AND METHODS

This study is part of the project "Improving the early Diagnosis of Alzheimer's Disease and Other dementias" (IDADO) of the department of Psychology, University of Amsterdam, and the departments of Neurology and Radiology of the Academic Medical Centre (AMC), Amsterdam.

Participants

Included were consecutive patients of the AMC memory clinic who were between 50 and 85 years of age, and who had subjective cognitive complaints that might signify MCI, an early stage of AD (but with

CDR < 1), or another type of dementia. The inclusion lasted from February 2007 until November 2009. To support external validity, we attempted to create a naturalistic sample representative of memory clinic patients with possible pre-dementia and early dementia. Therefore, we used as few exclusion criteria as possible. Exclusion criteria were a diagnosis of AD or other dementia established by a dementia specialist (i.e., CDR $\geq$ 1), other brain or systemic disease sufficient to cause the mental complaints, current substance abuse or addiction, contra-indications for MRI scanning, insufficient command of Dutch language, pre-existent mental retardation, and serious somatic disease or handicaps that prevented neuropsychological evaluation. Psychiatric disorders were not excluded, except severe syndromes that could mimic dementia, such as chronic psychosis or severe mood disorder with psychotic features. An additional exclusion criterion was non-credible test performance, i.e., patients who do not invest a reasonable amount of effort, and thus do not perform to the best of their abilities. These patients were excluded because they obscure brain-behavior relations [33].

Patients gave written informed consent at study entry. The ethical review board of the AMC approved the project.

Procedures

At baseline and at follow-up after two years all patients had a neurological consultation at the memory clinic. The neurologist (WAvG, ER, GW) performed a neurological examination including history taking, interview with an informant (if available), and administration of MMSE and CDR.

After this initial consultation, a provisional (research) diagnosis was given, without considering the results of MRI and neuropsychological tests. Subsequently, a neuropsychological evaluation was administered and a structural MRI scan was made. If necessary, this was done during a second visit, but within a month after the first. The neuropsychological testing was done in a quiet room by a neuropsychologist or a master student supervised by the neuropsychologist (AR, HT). Sixteen tests were administered, of which part is reported here (see next paragraph). We selected tests of memory, verbal fluency, and executive functioning. Testing took 2.5 hours (45 minutes for the tests reported here).

The neurologist received a written report of neuropsychological test results and was given access to the MRI. If necessary, the provisional diagnosis was revised. In a third visit, the neurologist discussed the results, possible diagnoses, and treatment options with the individual patients and their relatives.

Materials

Two instruments were used during the neurological consultation, viz. the MMSE [8] and the CDR [4]. The neuropsychological evaluation contained tests of verbal fluency (Controlled Oral Word Association Task), memory (subtest Prose Recall from the Rivermead Behavioural Memory Test and Rey's Auditory Verbal Learning Test), and executive function (Trail Making Test and Stroop Color-Word Test). Parallel test versions were used at follow-up, except for the Stroop test. The Test of Memory Malingering and Word Memory Test were used to detect non-credible responding. For references of the tests, see Lezak et al. [34]. Raw test scores were transformed into T-scores corrected for gender, age, and level of education using recently published Dutch normative data [35]. Level of education was scored on the UNESCO ISCED scale [36].

MRI acquisition and processing

Imaging was performed at baseline and at 2-year follow-up on a 3.0 Tesla MRI system (Philips Intera, Best, The Netherlands) with a 6-channel SENSE head coil. A gradient echo 3D FFE, T1-weighted, sagittal sequence was used with the following pulse sequence parameters: echo time [TE] = 3.5 ms, repetition time [TR] = 9 ms, field of view [FOV] = 256, 232, 170 mm (FH, AP, RL), scanning matrix = 256 $\times$ 231, flip angle = 8°, voxel size = 1 $\times$ 1 $\times$ 1 mm. Volumetric measurements of the hippocampus and measurement of cortical thickness of the entorhinal, parahippocampal, and middle temporal cortices were made automatically using the FreeSurfer image analysis suite, release version 5.2.0 (http://surfer.nmr.mgh.harvard.edu) [37, 38]. MR images were processed using the longitudinal processing stream [39] (https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferMethodsCitation), with the exception that only two structural scans were available for every patient and that surface statistics were not taken into account. Details of the procedures for subcortical segmentation are described by Fischl et al. [40]. Thickness of the entorhinal cortex was determined with the help of an *ex vivo* atlas [41]. Validation studies showed reasonable overlap in absolute hippocampal volumes between FreeSurfer and manual volumetry [42, 43]. FreeSurfer is considered a reasonable substitute for manual tracing,

with only a small size increase compared to manual tracing [44]. The left and right hippocampi were selected as MRI measures in this study, because these regions are considered to best differentiate between controls and AD subjects [43]. Images were visually inspected for gross structural abnormalities, presence of artifacts and accurate hippocampal segmentations. The hippocampal volumes were described in cubic millimeters, and the total volume of the hippocampus was also expressed as percentage of the total intracranial volume (% ICV) to control for variation in head size. Statistical calculations were done with the % ICV measure. Cortical thickness was not corrected and was expressed in millimeters.

Statistical analyses

At follow-up, the neurologists divided the patients into two groups based on the CDR scores. One group had normal cognition (CDR = 0), whereas the other group showed clinical signs of cognitive impairment (CDR $\geq$ 0.5). For each patient, the mean T-score of the neuropsychological tests and the hippocampal volumes at baseline and at follow-up were calculated. At baseline, two patients in the impaired group had one missing test value. At follow-up, one of the impaired patients refused to do the neuropsychological tests (diagnosed as AD by the treating neurologist). One patient was unable to do the tests (diagnosed as mixed AD/vascular). Two patients had one or two missing values. For patients with a missing value, mean T-scores were calculated using their remaining test results; the patient who was untestable was given the lowest mean T-score of the group; the patient who refused was given her baseline T-score minus the mean decline. Responsiveness of neuropsychological and MRI measures was expressed as the standardized mean change, i.e., the ratio of the mean change between baseline and follow-up relative to the standard error of the difference (SED) or the standard deviation of the change scores: (M0–M1)/SED.

Required sample sizes for a hypothetical two-arm intervention trial of treatment versus placebo were calculated using the following formula:

$$n = 2\sigma^2 (Z_{1-\alpha/2} + Z_{power})^2/(0.25\Delta)^2 \text{ [17, 45].}$$

In this formula Δ denotes the mean change in the intervention group minus the mean change in the placebo group, σ represents the standard deviation of the change scores, and n is the required number per study arm. A level of significance of $\alpha = 0.05$ was used. Z_{power} denotes z-value of the power with which a hypothetical treatment effect of 25% reduction of the rate of decline (0.25 in this formula) can be detected in the calculated sample size n. Power was set at 80%. In the present analyses, we substituted Δ by the mean change in the impaired group minus the mean change in the cognitively normal group, and σ by the SD of the change scores in the impaired group.

Statistical analyses were done with SPSS version 19.0 (SPSS/IBM 2010).

RESULTS

Seventy-one patients had full baseline and follow-up assessments and MRI (see Fig. 1). Nine patients were excluded, five because of non-credible responding (i.e., insufficient effort), three because of other than neurodegenerative diseases that might have caused their cognitive decline (history of malignant hypertension and diabetes mellitus with severe white matter abnormalities; Gaucher disease; stroke during the follow-up interval), and one because of insufficient scan quality.

Sixty-two patients satisfied the inclusion and exclusion criteria. After two-year follow-up, 28 patients had normal cognition (CDR = 0), and 34 showed clinical signs of cognitive impairment (CDR $\geq$ 0.5) according to the neurologist's provisional diagnosis.

The groups were comparable in level of education and gender distribution (Table 1), but the normal cognition group was younger than the impaired group. Note that the neurologist made this group division without considering MRI and neuropsychological evaluations. Two patients, who the neurologist considered to be normal at follow-up, appeared to satisfy MCI criteria after MRI and neuropsychological evaluation. Three patients, who the neurologist considered to

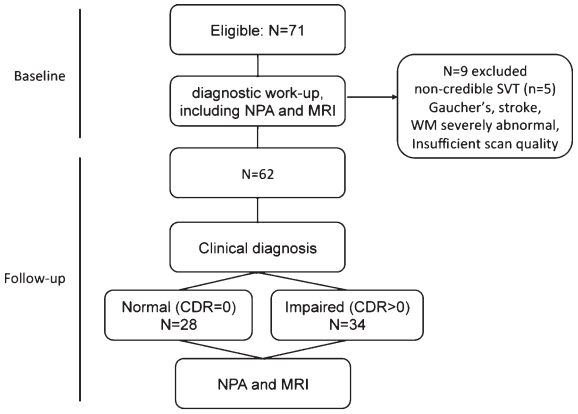

Fig. 1. Flow chart of the study.

Table 1
Demographic and clinical characteristics. Patients who at follow-up after two years were considered to have normal cognition or to show cognitive impairment after neurological consultation. Diagnoses at baseline and at follow-up were the revised diagnoses after taking MRI and neuropsychological results into consideration

	Normal cognition at follow-up (CDR = 0; $n = 28$)	Cognitive impairment at follow-up (CDR > 0; $n = 34$)	p-value
Male/female	14/14	18/16	0.82[a]
Age	62.0 (8.1)	70.3 (8.9)	<0.001[b]
Education level (0–6 ISCED)	4.1 (1.4)	3.8 (1.2)	0.27[c]
MMSE at baseline	28.0 (1.8)	26.7 (2.3)	0.011[c]
MMSE at follow-up	28.2 (1.3)	23.0 (4.5)	<0.001[c]
Revised diagnosis at baseline			
Dementia	0	8	
MCI	1	21	
psychiatric disorder	8	1	
worried well	19	4	
Revised diagnosis at follow-up			
dementia	0	14	
MCI	2	15	
psychiatric disorder	5	2	
worried well	21	3	

[a]chi-square; [b] t-test; [c] Mann-Whitney test.

Table 2a
Demographically corrected T-scores on neuropsychological tests of patients who at follow-up were considered by neurological consultation to have normal cognition or to show cognitive impairment. Mean (standard deviation)

	Baseline test scores		Follow-up test scores	
	Normal cognition at follow-up (CDR = 0; $n = 28$)	Cognitive impairment at follow-up (CDR > 0; $n = 34$)	Normal cognition at follow-up (CDR = 0; $n = 28$)	Cognitive impairment at follow-up (CDR > 0; $n = 34$)
Letter fluency COWAT	48.6 (9.5)	40.5 (9.5)	51.6 (10.8)	39.4 (11.6)
RAVLT immediate recall	51.9 (8.1)	36.4 (10.5)	51.6 (10.3)	28.5 (9.4)
RAVLT delayed recall	54.1 (5.8)	36.5 (12.3)	50.5 (11.6)	27.9 (10.4)
RBMT prose immediate recall	48.6 (10.7)	39.4 (11.3)	52.6 (11.0)	36.3 (11.0)
RBMT prose delayed recall	46.8 (10.6)	37.2 (10.9)	51.7 (11.9)	35.1 (10.8)
Stroop CWT interference	49.0 (8.8)	39.8 (10.7)	48.2 (9.7)	36.5 (10.3)
Trail Making Test part B	51.5 (9.8)	34.2 (19.5)	49.3 (10.7)	30.2 (19.0)
Mean T-score	49.8 (5.9)	37.7 (7.8)	49.9 (7.6)	31.4 (9.8)
Mean T change			*0.112 (4.626)*	*–6.286 (6.339)*

be impaired, appeared to have normal cognition at neuropsychological testing; they were reclassified as worried well. The latter label was used for people who were cognitively normal, did not have cognitive test scores in the impaired range, and who did not meet criteria for a psychiatric disorder. Although they had cognitive complaints and were unsure about their mental status, these worries appeared to be unjustified. Two patients, who had MCI at baseline, had improved clinically and were considered normal on follow-up. Psychiatric comorbidity mostly concerned mood disorders (dysthymia and depression). Note that the presence of a psychiatric syndrome does not conflict with a diagnosis of normal cognition. All dementia cases were diagnosed as probable AD except two who had mixed AD-vascular dementia, and one patient who had frontotemporal dementia.

These revised diagnoses are given to describe the sample. However, all statistical analyses were done with the CDR classification to prevent incorporation bias, i.e., circularity of diagnostic reasoning.

The normal cognition group performed around the population mean of T = 50 (Table 2a), and did not decline during follow-up. The impaired group performed on most tests about one SD below the population mean at baseline, and declined 0.6 SD (mean of 6.3 T-score points) during the follow-up interval. The difference in decline between both groups was 6.4 T-score points (95% CI 3.5–9.3; $p = 0.00002$ two-tailed t-test).

At baseline, the magnitudes of group differences in cortical thickness measures were around 0.3–0.4 SD (Table 2b). In the impaired group, the mean thickness of the entorhinal, middle temporal, and parahip-

Table 2b
Cortical thickness (in mm) of entorhinal, middle temporal and parahippocampal cortices, and hippocampal volumes (in mm^3, as percentage of intracranial volume, and hippocampal atrophy rate) of patients who at follow-up were considered by neurological consultation to have normal cognition or to show cognitive impairment. Mean (standard deviation)

	Baseline cortical thickness and hippocampal volume		Follow-up cortical thickness and hippocampal volume	
	Normal cognition at follow-up (CDR = 0; $n = 28$)	Cognitive impairment at follow-up (CDR > 0; $n = 34$)	Normal cognition at follow-up (CDR = 0; $n = 28$)	Cognitive impairment at follow-up (CDR > 0; $n = 34$)
Entorhinal cortex left (mm)*	3.4 (0.5)	3.2 (0.5)	3.4 (0.5)	3.1 (0.6)
Entorhinal cortex right (mm)*	3.7 (0.4)	3.4 (0.5)	3.7 (0.5)	3.4 (0.5)
Entorhinal cortex mean (mm)*	3.6 (0.4)	3.3 (0.5)	3.6 (0.4)	3.2 (0.6)
Middle temporal cortex left (mm)	2.9 (0.3)	2.8 (0.2)	2.9 (0.2)	2.8 (0.3)
Middle temporal cortex right (mm)	3.0 (0.3)	2.9 (0.3)	3.0 (0.2)	2.9 (0.3)
Middle temporal cortex mean (mm)	3.0 (0.3)	2.9 (0.2)	2.9 (0.2)	2.8 (0.2)
Parahippocampal cortex left (mm)	2.9 (0.5)	2.7 (0.4)	2.9 (0.4)	2.6 (0.4)
Parahippocampal cortex right (mm)	2.9 (0.5)	2.7 (0.4)	2.8 (0.4)	2.6 (0.4)
Parahippocampal cortex mean (mm)	2.9 (0.5)	2.7 (0.3)	2.8 (0.4)	2.6 (0.4)
Mean thickness (ER, MT, PH) (mm)	3.1 (0.3)	3.0 (0.3)	3.1 (0.3)	2.9 (0.3)
Mean change in thickness (mm)			*–0.026 (0.155)*	*–0.064 (0.106)*
Hippocampal volume left (mm^3)	3520 (436)	2936 (471)	3440 (394)	2717 (515)
Hippocampal volume right (mm^3)	3571 (427)	3005 (539)	3553 (400)	2835 (585)
Hippocampal volume total (mm^3)	7091 (829)	5941 (966)	6993 (759)	5553 (1051)
Hippocampal volume %ICV	0.46 (0.07)	0.39 (0.09)	0.45 (0.07)	0.37 (0.09)
Mean change %ICV			*–0.007 (0.023)*	*–0.025 (0.026)*
Atrophy rate %/year			0.6 (2.5)	3.4 (3.3)

*determined with *ex vivo* atlas (Fischl et al. [41]).

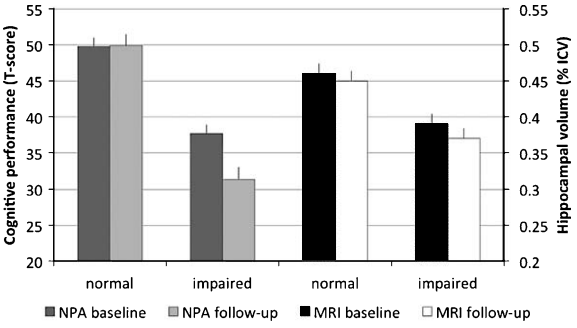

Fig. 2. Cognitive performance and hippocampal volume of patients who were cognitively normal or impaired. Cognitive performance (left y-axis) is expressed as the mean T-score of five neuropsychological tests; scores are corrected for age, gender, and educational level. Hippocampal volume (right y-axis) is expressed as percentage of intracranial volume. Patients were rated by neurologists at follow-up as cognitively normal ($n = 28$) or as cognitively impaired ($n = 34$). Follow-up interval was 2 years. NPA, neuropsychological assessment. Bars indicate standard error of the mean.

pocampal cortices decreased with 0.06 mm during the follow-up interval, which amounts to a decrease of about 0.2 SD (Table 2b). The normal cognition group showed less cortical thinning, except in the middle temporal cortex, where it was a small fraction larger than in the impaired group. The difference in mean cortical thickness between both groups was not significant ($p = 0.27$; two-tailed t-test).

The baseline group difference in (uncorrected) total hippocampal volume was about one SD, and the volume decreased 0.4 SD in the impaired group (i.e., 5941–5553 = 388/966; Table 2b). This corresponds to a hippocampal atrophy rate of 3.4% per year. The normal cognition group showed an atrophy rate in the order of 0.1 SD or 0.6% per year. The difference between both groups was significant (95% CI 1.3–4.3; $p = 0.0005$ two-tailed t-test). The longitudinal FreeSurfer processing stream could not analyze three scan pairs; they were segmented separately. Cognitive decline and hippocampal atrophy are illustrated in Fig. 2.

Substituting the mean change in the impaired group minus the mean change in the normal group for Δ in the above formula, and the SD of the mean change in the impaired group for σ, the estimated required sample sizes per study arm are $n = 1952$ for mean cortical thickness, $n = 523$ for hippocampal atrophy, and $n = 246$ for neuropsychological assessment (Table 3). These sample sizes would have a power of 80% to detect an effect size of 25% reduction in atrophy rate or cognitive decline, respectively, at 5% significance. Samples of $n = 488$, $n = 131$, or $n = 62$ per study arm would be required to detect 50% reduction with the same power and significance. Thus, a theoretical study with hippocampal atrophy as the endpoint would need to include about twice the number of participants than a study with NPA as the endpoint. If cortical thickness

Table 3
Minimum *n* per arm for each variable if it would be an isolated outcome measure in a theoretical trial to detect 25% or 50% reduction in rate of change at 80% study power

	25% reduction	50% reduction
Letter fluency COWAT	489	122
RAVLT immediate recall	393	98
RAVLT delayed recall	1381	345
RBMT prose immediate recall	444	111
RBMT prose delayed recall	464	116
Stroop CWT interference	401	100
Trail Making Test part B	1768	442
Mean T-score	*246*	*62*
Entorhinal cortex thickness	853	213
Middle temporal cortex thickness	n.a.	n.a.
Parahippocampal cortex thickness	3429	857
Mean cortical thickness (ER, MT, PH)	*1952*	*488*
Left hippocampus volume (% ICV)	561	140
Right hippocampus volume (%ICV)	646	162
Both hippocampi volume (%ICV)	*523*	*131*

n.a. not applicable because of slightly faster thinning in normal than in impaired group.

would be the outcome measure, the required sample sizes would be about eight times larger. These analyses were repeated without the three patients whose scans had to be analyzed separately and without the patient with frontotemporal dementia. This did not change the results.

Application of analysis of variance covarying for age only marginally affected the estimated means and standard deviations, and left the resulting numbers needed per arm virtually unchanged.

The neuropsychological measure that performed best in isolation was Rey's AVLT total learning score. The best performing isolated MRI measure was volume of the left hippocampus (Table 3).

DISCUSSION

Our analysis suggests that neuropsychological assessment is more responsive than hippocampal atrophy or cortical thickness of temporal lobe areas in patients with MCI and early dementia, using the neurologist's clinical diagnosis of progressive disease as gold standard. Our estimates of required sample sizes for intervention trials in MCI and AD are about 50% smaller if a brief neuropsychological test battery would be the outcome measure instead of hippocampal atrophy. Cortical thickness of temporal areas as outcome measure would require much larger samples.

At baseline, the cognitive test scores in the declining group were between 1.0 and 1.6 SD below the demographically corrected mean, and the scores of the normal group were around the population mean. Thus, the latter group indeed had normal cognitive functions and did not decline during follow-up, unlike the patients who were rated as impaired. These patients declined 0.6 SD (6 T-score points). The hippocampal volumes we reported might have been systematically overestimated by FreeSurfer [44], but the annual atrophy rates of the hippocampi of 0.6% and 3.4% in the normal and declining groups respectively, are in line with the literature. Two meta-analyses found mean rates of hippocampal atrophy of 1.4% per year in healthy elderly and 4.7% in AD patients [46], with MCI patients in an intermediate position [20]. Our patients were relatively young compared to the subjects in these meta-analyses, and our declining group was a mix of MCI and dementia cases. Both these factors likely resulted in somewhat lower atrophy rates. Nevertheless, the resulting sample size estimates for hippocampal atrophy were in the same order of magnitude, i.e., several hundred per arm, as those found in ADNI [17]. As an aside, the absolute values of these estimates are only indicative. They strongly fluctuate with relatively small changes in deltas and SDs of the n-per-arm formula. At any rate, we conclude that our cognitive and neuroimaging results are plausible, and compatible with previous findings.

Our sample was relatively small. Furthermore, we did not exclude psychiatric comorbidity, unlike many other similar studies. This may be unusual, but it ensures that our analyses reflect daily clinical practice. We treated the patients with normal cognition as if they were healthy controls. Although one could question if patients with subjective cognitive complaints should be considered 'normal', our data show that this was acceptable, since their cognitive scores were normal and did not decline, and the age-related rate of atrophy of their hippocampi was within normal limits. Thus, the fact that some patients had a mood disorder or another psychiatric problem was apparently irrelevant. A similar reasoning may justify the composition of our impaired group. This group was diluted somewhat with psychiatric cases and even with a few worried well, who were misclassified by the neurologists. Nevertheless, at group level their cognitive scores and hippocampal volumes, and the changes over time of these measures, including the greater change in the left than in the right hippocampus, corresponded to what is reported in the literature, as we saw above. Eight patients in this group appeared to be demented at baseline when their diagnosis was revised (Table 1), and we did not treat them differently from the majority, i.e., the MCI patients. Note however, that the revised demen-

tia diagnosis could be made only after MRI scanning and neuropsychological assessment. Had we excluded the dementia cases, we would have introduced incorporation bias or circularity of diagnostic reasoning. Moreover, the line between MCI and dementia is hard to draw in practice, and there is a trend toward investigating these categories jointly [47]. Our study design conformed to the STARD and QUADAS-2 criteria [48, 49].

The most likely explanation of our results is perhaps that neurologists base their decision whether or not a patient's cognition is normal mainly on the patient history, the informant report, and on their own observations during the consultation, supported by results of a dementia-screening instrument such as the MMSE. Thus, their decision is based on behavioral information, and rightly so because the dementias are defined by behavioral symptoms. Neuropsychological evaluation assesses these behavioral symptoms. It is therefore to be expected that clinical judgment correspond closer to neuropsychological assessment than to volumetric measures.

Neuropsychological assessment largely is a quantitative technique, and it is more precise and more reliable than the clinician's naked eye, which serves a more qualitative judgment. This precision of neuropsychological assessment and its focus on the defining symptoms of dementia probably explain its superior responsiveness. Another important characteristic of neuropsychological tests is that they allow to correct the scores for age, educational level, and if necessary for gender and ethnicity in a statistically sound way. This removes a major portion of the variance in final scores, and allows more clearly identifying the cognitive deficits due to the disease.

There are caveats to our study. First, we simply took the mean standard score of the cognitive tests. However, more powerful statistical methods are available to detect cognitive changes [50]. Second, the FreeSurfer software could not analyze the MRI scans of three patients longitudinally. This was not a random phenomenon; it concerned declining patients. However, it did not distort our results. Finally, we did not consider other biomarkers than hippocampal atrophy and cortical thickness, nor did we take genetic variation, such as APOE genotype, into account. It seems unlikely that cerebrospinal fluid biomarkers, such as amyloid and tau concentrations, will reflect cognitive and behavioral deterioration better than neuropsychological assessment [23, 51, 52]. However, if one would combine neuropsychological measures as endpoints with sample enrichment strategies by neuroimaging, genetic, and neurochemical biomarkers, the statistical power of intervention studies might be increased even further [5, 13, 18].

Our results have important implications for pharmaceutical research into MCI and AD. Hundreds of compounds for symptomatic or disease-modifying treatment of AD have been tested by now, and they all failed, apart from a handful of modestly effective drugs [53]. We have shown that neuropsychological tests can provide the evaluative measures that are so badly needed to speed up the discovery of disease modifying or preventive drugs [7]. However, in clinical practice the effects of drugs should in the first place be visible to the patients and their relatives.

ACKNOWLEDGMENTS

The Psychology Research Institute of the University of Amsterdam, the Departments of Neurology and Radiology of the Academic Medical Center Amsterdam, and the Graduate School of Neurosciences Amsterdam financed this study. We thank Dr. Thelma Schilt and Dr. Gerard Walstra for critical reading of the manuscript. Dr. Walstra also greatly contributed to the data collection.

Authors' disclosures available online (http://www.j-alz.com/disclosures/view.php?id=2042).

REFERENCES

[1] Guyatt G, Walter S, Norman G (1987) Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chronic Dis* **40**, 171-178.

[2] Revicki D, Hays RD, Cella D, Sloan J (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* **61**, 102-109.

[3] Rosen WG, Mohs RC, Davis KL (1984) A new rating scale for Alzheimer's disease. *Am J Psychiatry* **141**, 1356-1364.

[4] Morris JC (1993) The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43**, 2412-2414.

[5] Cummings J, Gould H, Zhong K (2012) Advances in designs for Alzheimer's disease clinical trials. *Am J Neurodegener Dis* **1**, 205-216.

[6] Isaac M, Vamvakas S, Abadie E, Jonsson B, Gispen C, Pani L (2011) Qualification opinion of novel methodologies in the predementia stage of Alzheimer's disease: Cerebrospinal-fluid related biomarkers for drugs affecting amyloid burden – Regulatory considerations by European Medicines Agency focusing in improving benefit/risk in regulatory trials. *Eur Neuropsychopharmacol* **21**, 781-788.

[7] Vellas B, Andrieu S, Sampaio C, Coley N, Wilcock G, European Task Force, Group (2008) Endpoints for trials in Alzheimer's disease: A European task force consensus. *Lancet Neurol* **7**, 436-450.

[8] 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**, 189-198.

[9] Roth M, Tym E, Mountjoy CQ, Huppert FA, Hendrie H, Verma S, Goddard R (1986) CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *Br J Psychiatry* **149**, 698-709.

[10] Zamrini E, De Santi S, Tolar M (2004) Imaging is superior to cognitive testing for early diagnosis of Alzheimer's disease. *Neurobiol Aging* **25**, 685-691.

[11] de Leon MJ, Mosconi L, Blennow K, DeSanti S, Zinkowski R, Mehta PD, Pratico D, Tsui W, Saint Louis LA, Sobanska L, Brys M, Li Y, Rich K, Rinne J, Rusinek H (2007) Imaging and CSF studies in the preclinical diagnosis of Alzheimer's disease. *Ann N Y Acad Sci* **1097**, 114-145.

[12] Harrison J, Minassian SL, Jenkins L, Black RS, Koller M, Grundman M (2007) A neuropsychological test battery for use in Alzheimer disease clinical trials. *Arch Neurol* **64**, 1323-1329.

[13] Monsell SE, Liu D, Weintraub S, Kukull WA (2012) Comparing measures of decline to dementia in amnestic MCI subjects in the National Alzheimer's Coordinating Center (NACC) Uniform Data Set. *Int Psychogeriatr* **24**, 1553-1560.

[14] Schneider LS, Sano M (2009) Current Alzheimer's disease clinical trials: Methods and placebo outcomes. *Alzheimers Dement* **5**, 388-397.

[15] Coley N, Andrieu S, Jaros M, Weiner M, Cedarbaum J, Vellas B (2011) Suitability of the Clinical Dementia Rating-Sum of Boxes as a single primary endpoint for Alzheimer's disease trials. *Alzheimers Dement* **7**, 602-610.e2.

[16] Jack CR Jr, Slomkowski M, Gracon S, Hoover TM, Felmlee JP, Stewart K, Xu Y, Shiung M, O'Brien PC, Cha R, Knopman D, Petersen RC (2003) MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. *Neurology* **60**, 253-260.

[17] Ard MC, Edland SD (2011) Power calculations for clinical trials in Alzheimer's disease. *J Alzheimers Dis* **26**(Suppl 3), 369-377.

[18] Grill JD, Di L, Lu PH, Lee C, Ringman J, Apostolova LG, Chow N, Kohannim O, Cummings JL, Thompson PM, Elashoff D, Alzheimer's Disease Neuroimaging Initiative (2013) Estimating sample sizes for predementia Alzheimer's trials based on the Alzheimer's Disease Neuroimaging Initiative. *Neurobiol Aging* **34**, 62-72.

[19] Hampel H, Wilcock G, Andrieu S, Aisen P, Blennow K, Broich K, Carrillo M, Fox NC, Frisoni GB, Isaac M, Lovestone S, Nordberg A, Prvulovic D, Sampaio C, Scheltens P, Weiner M, Winblad B, Coley N, Vellas B, Oxford Task Force Grp (2011) Biomarkers for Alzheimer's disease therapeutic trials. *Prog Neurobiol* **95**, 579-593.

[20] Shi F, Liu B, Zhou Y, Yu C, Jiang T (2009) Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. *Hippocampus* **19**, 1055-1064.

[21] Han L, Cole M, Bellavance F, McCusker J, Primeau F (2000) Tracking cognitive decline in Alzheimer's disease using the mini-mental state examination: A meta-analysis. *Int Psychogeriatr* **12**, 231-247.

[22] Ito K, Ahadieh S, Corrigan B, French J, Fullerton T, Tensfeldt T, Alzheimer's Disease Working Group (2010) Disease progression meta-analysis model in Alzheimer's disease. *Alzheimers Dement* **6**, 39-53.

[23] Gomar JJ, Bobes-Bascaran MT, Conejero-Goldberg C, Davies P, Goldberg TE, Alzheimer's Disease Neuroimaging Initiative (2011) Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative. *Arch Gen Psychiatry* **68**, 961-969.

[24] Velayudhan L, Proitsi P, Westman E, Muehlboeck JS, Mecocci P, Vellas B, Tsolaki M, Kloszewska I, Soininen H, Spenger C, Hodges A, Powell J, Lovestone S, Simmons A, dNeuroMed Consortium (2013) Entorhinal cortex thickness predicts cognitive decline in Alzheimer's disease. *J Alzheimers Dis* **33**, 755-766.

[25] Wouters H, Appels B, van der Flier WM, van Campen J, Klein M, Zwinderman AH, Schmand B, van Gool WA, Scheltens P, Lindeboom R (2012) Improving the accuracy and precision of cognitive testing in mild dementia. *J Int Neuropsychol Soc* **18**, 314-322.

[26] Wouters H, van Gool WA, Schmand B, Lindeboom R (2008) Revising the ADAS-cog for a more accurate assessment of cognitive impairment. *Alzheimer Dis Assoc Disord* **22**, 236-244.

[27] Riepe MW, Janetzky W, Lemming OM (2011) Measuring therapeutic efficacy in patients with Alzheimer's disease: role of instruments. *Dement Geriatr Cogn Disord* **31**, 233-238.

[28] Balsis S, Unger AA, Benge JF, Geraci L, Doody RS (2012) Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: A comparison of item response theory-based scores and total scores. *Alzheimers Dement* **8**, 288-294.

[29] Backman L, Jones S, Berger AK, Laukka EJ, Small BJ (2005) Cognitive impairment in preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology* **19**, 520-531.

[30] Zakzanis KK (1998) Quantitative evidence for neuroanatomic and neuropsychological markers in dementia of the Alzheimer's type. *J Clin Exp Neuropsychol* **20**, 259-269.

[31] Clerx L, Visser PJ, Verhey F, Aalten P (2012) New MRI markers for Alzheimer's disease: A meta-analysis of diffusion tensor imaging and a comparison with medial temporal lobe measurements. *J Alzheimers Dis* **29**, 405-429.

[32] Schmand B, Huizenga HM, van Gool WA (2010) Meta-analysis of CSF and MRI biomarkers for detecting preclinical Alzheimer's disease. *Psychol Med* **40**, 135-145.

[33] Rienstra A, Groot PF, Spaan PE, Majoie CB, Nederveen AJ, Walstra GJ, de Jonghe JF, van Gool WA, Olabarriaga SD, Korkhov VV, Schmand B (2013) Symptom validity testing in memory clinics: Hippocampal-memory associations and relevance for diagnosing mild cognitive impairment. *J Clin Exp Neuropsychol* **35**, 59-70.

[34] Lezak MD (2012) *Neuropsychological Assessment*. Oxford University Press, Oxford; New York.

[35] Schmand B, Houx P, de Koning I (2012) *Normen van psychologische tests voor gebruik in de klinische neuropsychologie*, Nederlands Instituut van Psychologen, Utrecht. www.psynip.nl.

[36] UNESCO (1997) *International Standard Classification of Education*. ISBN 92-9189-035-9.

[37] Fischl B, Sereno MI, Dale AM (1999) Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9**, 195-207.

[38] Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179-194.

[39] Reuter M, Schmansky NJ, Rosas HD, Fischl B (2012) Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage* **61**, 1402-1418.

[40] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM (2002) Whole

brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341-355.

[41] Fischl B, Stevens AA, Rajendran N, Yeo BT, Greve DN, Van Leemput K, Polimeni JR, Kakunoori S, Buckner RL, Pacheco J, Salat DH, Melcher J, Frosch MP, Hyman BT, Grant PE, Rosen BR, van der Kouwe AJ, Wiggins GC, Wald LL, Augustinack JC (2009) Predicting the location of entorhinal cortex from MRI. *Neuroimage* **47**, 8-17.

[42] Tae WS, Kim SS, Lee KU, Nam EC, Kim KW (2008) Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. *Neuroradiology* **50**, 569-581.

[43] Lehmann M, Douiri A, Kim LG, Modat M, Chan D, Ourselin S, Barnes J, Fox NC (2010) Atrophy patterns in Alzheimer's disease and semantic dementia: A comparison of FreeSurfer and manual volumetric measurements. *Neuroimage* **49**, 2264-2274.

[44] Morey RA, Petty CM, Xu Y, Hayes JP, Wagner HR, 2nd, Lewis DV, LaBar KS, Styner M, McCarthy G (2009) A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *Neuroimage* **45**, 855-866.

[45] Hua X, Gutman B, Boyle CP, Rajagopalan P, Leow AD, Yanovsky I, Kumar AR, Toga AW, Jack CR Jr, Schuff N, Alexander GE, Chen K, Reiman EM, Weiner MW, Thompson PM, Alzheimer's Disease Neuroimaging Initiative (2011) Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. *Neuroimage* **57**, 5-14.

[46] Barnes J, Bartlett JW, van de Pol LA, Loy CT, Scahill RI, Frost C, Thompson P, Fox NC (2009) A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. *Neurobiol Aging* **30**, 1711-1723.

[47] Aisen PS, Andrieu S, Sampaio C, Carrillo M, Khachaturian ZS, Dubois B, Feldman HH, Petersen RC, Siemers E, Doody RS, Hendrix SB, Grundman M, Schneider LS, Schindler RJ, Salmon E, Potter WZ, Thomas RG, Salmon D, Donohue M, Bednar MM, Touchon J, Vellas B (2011) Report of the task force on designing clinical trials in early (predementia) AD. *Neurology* **76**, 280-286.

[48] Bossuyt PM, Reitsma JB, Standards for reporting of diagnostic, accuracy (2003) The STARD initiative. *Lancet* **361**, 71.

[49] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, QUADAS-2 Group (2011) QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* **155**, 529-536.

[50] Huizenga HM, Smeding H, Grasman RP, Schmand B (2007) Multivariate normative comparisons. *Neuropsychologia* **45**, 2534-2542.

[51] Zhou B, Teramukai S, Yoshimura K, Fukushima M (2009) Validity of cerebrospinal fluid biomarkers as endpoints in early-phase clinical trials for Alzheimer's disease. *J Alzheimers Dis* **18**, 89-102.

[52] Schmand B, Eikelenboom P, van Gool WA, Alzheimer's Disease Neuroimaging Initiative (2012) Value of diagnostic tests to predict conversion to Alzheimer's disease in young and old patients with amnestic mild cognitive impairment. *J Alzheimers Dis* **29**, 641-648.

[53] Pater C (2011) Mild cognitive impairment (MCI) – the novel trend of targeting Alzheimer's disease in its early stages – methodological considerations. *Curr Alzheimer Res* **8**, 798-807.