Test-Retest Reliability of Memory Task Functional Magnetic Resonance Imaging in Alzheimer Disease Clinical Trials

Alireza Atri, MD, PhD; Jacqueline L. O'Brien, BA; Aishwarya Sreenivasan, MS, BA; Sarah Rastegar, BA; Sibyl Salisbury, RN; Amy N. DeLuca, BS; Kelly M. O'Keefe, BA; Peter S. LaViolette, MS; Dorene M. Rentz, PsyD; Joseph J. Locascio, PhD; Reisa A. Sperling, MD, MSc

Objective: To examine the feasibility and test-retest reliability of encoding-task functional magnetic resonance imaging (fMRI) in mild Alzheimer disease (AD).

Design: Randomized, double-blind, placebo-controlled study.

Setting: Memory clinical trials unit.

Participants: We studied 12 patients with mild AD (mean [SEM] Mini-Mental State Examination score, 24.0 [0.7]; mean Clinical Dementia Rating score, 1.0) who had been taking donepezil hydrochloride for more than 6 months from the placebo arm of a larger 24-week study (n = 24, 4 scans on weeks 0, 6, 12, and 24, respectively).

Interventions: Placebo and 3 face-name, paired-associate encoding, block-design blood oxygenation level–dependent fMRI scans in 12 weeks.

Main Outcome Measures: We performed whole-brain t maps (P < .001, 5 contiguous voxels) and hippocampal regions-of-interest analyses of extent (percentage of active voxels) and magnitude (percentage of signal change) for novel-greater-than-repeated face-name contrasts. We also calculated intraclass correlation coefficients and power estimates for hippocampal regions of interest.

Results: Task tolerability and data yield were high (95 of 96 scans yielded favorable-quality data). Whole-brain maps were stable. Right and left hippocampal regions-of-interest intraclass correlation coefficients were 0.39 to 0.87 and 0.67 to 0.74, respectively. To detect 25.0% to 50.0% changes in week-0 to week-12 hippocampal activity using left-right extent or right magnitude with 80.0% power (2-sided α = .05) requires 14 to 51 patients. Using left magnitude requires 125 patients because of relatively small signal to variance ratios.

Conclusions: Encoding-task fMRI was successfully implemented in a single-site, 24-week, AD randomized controlled trial. Week 0 to 12 whole-brain t maps were stable, and test-retest reliability of hippocampal fMRI measures ranged from moderate to substantial. Right hippocampal magnitude may be the most promising of these candidate measures in a leveraged context. These initial estimates of test-retest reliability and power justify evaluation of encoding-task fMRI as a potential biomarker for signal of effect in exploratory and proof-of-concept trials in mild AD. Validation of these results with larger sample sizes and assessment in multisite studies is warranted.

Arch Neurol. 2011;68(5):599-606

©2011 American Medical Association. All rights reserved.
ability of fMRI in 12 patients with mild AD randomized to the placebo arm of the study.

METHODS

STUDY PARTICIPANTS

Twelve patients with mild AD (Mini-Mental State Examination [MMSE] scores, 16–26) were randomized to the 12-week placebo arm of a larger (n = 24 patients) and longer (24 weeks) AD pharmacologic fMRI study. Inclusion criteria were National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer’s Disease and Related Disorders Association criteria for probable AD, fluency in English, lack of focal lesions on neuroimaging scans, taking a stable dosage of donepezil hydrochloride for longer than 6 months, and having a study partner (eg, spouse or relative) to monitor adherence. Exclusion criteria were unstable or severe medical or psychiatric illness, contraindication to fMRI, use of another investigational agent within 2 months, use of a cholinesterase inhibitor other than donepezil hydrochloride or an antipsychotic within 6 months, and ever having taken memantine hydrochloride. Patients and partners provided consent in accordance with the Human Research Committee guidelines. Study participants were remunerated $50 after each fMRI.

STUDY DESIGN AND PROCEDURES

The overall study spanned 9 visits in 24 weeks and used an RCT (50.0% memantine and donepezil hydrochloride, 50.0% placebo and donepezil hydrochloride) parallel-group design for 12 weeks, followed by a 12-week, single-blind period when all patients received drug therapy (100% memantine and donepezil hydrochloride). The reliability study data reported were obtained from the fMRI scan results at weeks 0 (baseline or T1), 6 (T2), and 12 (T3) in the placebo group only. Neuropsychological and clinical assessments included the MMSE, the AD Assessment Scale–Cognitive, and the Clinical Dementia Rating (CDR) scale.

fMRI PARADIGM

The details of the fMRI paradigm, sequencing, and preprocessing activities are described in previously published studies1,2,4,6,7,27 and online (eAppendix; http://www.archneur:com). The paradigm is composed of 3 conditions presented in successive blocks: novel face-name pairs, repeated face-name pairs, and fixation cross. Eighty-four novel pairs and 42 repeated pairs were displayed for 5 seconds each across 6 runs. Study participants were instructed to try to remember the name paired with each face. Immediately after scanning, 2 postscan behavioral/memory tests were administered: a face recognition (yes/no reply) and free recall of name (for yes responses) task and a 2-alternative forced-choice name-recognition-of-face task.

IMAGE ACQUISITION, PREPROCESSING, AND QUALITY ASSURANCE

Data were acquired on a 3-T GE scanner (GE Healthcare, Chalfont St Giles, England). Each functional run was 4 minutes, 15 seconds (102 time points; first 4 discarded for T1 stabilization). Preprocessing was performed with SPM2 statistical software (http://www.fil.ion.ucl.ac.uk/spm/software/spm2/), using 3 × 3 × 3-mm resectioning in Montreal Neurological Institute space, 8-mm full width at half maximum gaussian smoothing, and a 260-second high-pass filter.

Because of scanner repair, 1 patient underwent scans at 6 and 12 weeks, respectively, via a 3-T Siemens scanner (Siemens Medical, Munich, Germany). Data quality assurance included manual inspection of all images sequentially for scanner spiking and excessive motion and automated artifact detection algorithms that repaired any time point with a mean signal greater than 3 SDs of the mean global signal of each patient using an interpolation from surrounding scans (9 patients were affected). One week 6 scan (T2) was irreparable because of excessive intrascan movement and was imputed using T1 and T3 averages from the patient.

STATISTICAL ANALYSIS: fMRI AND CLINICAL MEASURES

Prespecified analyses focused on 2 methods to assess changes in magnitude, calculated as the percentage of signal change, and extent, calculated as the percentage of active voxels, of activations for novel-greater-than-repeated (N > R) stimuli between week 0 (baseline or T1) and week 12 fMRIs (ie, T1 to T3 change in N > R contrast). Prespecified primary analyses were whole-brain t test analyses and statistical parametric maps with significance thresholds at P < .001 and extent threshold of 5 contiguous voxels and hippocampal region-of-interest (ROI) analyses with small volume corrections for multiple comparisons within a priori, anatomically defined hippocampal ROIs.28 For comparison, secondary analyses were performed on non-hippocampal, a priori, anatomically defined ROIs, including bilateral precuneus and posterior cingulate cortices obtained in template space using the MarsBaR application (http://marsbar.sourceforge.net/), which have previously shown robust and selective task-related and spatiotemporally correlated activity in this fMRI paradigm.29 Repeated-measures analyses of covariance (ANCOVAs) assessed changes in clinical measures in the 12 weeks between T1 and T3 fMRIs.

TEST-RETEST RELIABILITY OF fMRI ACTIVATIONS IN RELIABILITY ROIs

Test-retest reliability for the extent and magnitude of N > R activity from baseline (T1) to week-12 (T3) fMRIs was assessed using 2 complementary approaches: intraclass correlation coefficients (ICCs), using a variation of ICC assessing agreement of score values (not merely correlation) for random-effects models, referred to by Shroot and Fleiss30 as ICC (2,1) for reliability at a single point in time and ICC (2,k) for that of an average score across a number of time points of assessment (eAppendix) and power analysis and sample size determination.31 Because no standard or widely accepted definition for general adjectives exists that describes reliability measures, ICC values, and ranges, we chose to adopt conservative terms by using the definition proposed by Shroot32 when qualitatively referring to ICC values and ranges: virtually none (0.00, 0.10), slight (0.11, 0.40), fair (0.41, 0.60), moderate (0.61, 0.80), or substantial (0.81, 1.0).33 We did not use the widely quoted but much more liberal terms of Landis and Koch34 when describing reliability values (slight [0.0, 0.20], fair [0.21, 0.40], moderate [0.41, 0.60], substantial [0.61, 0.80], or almost perfect [0.81, 1.00]) or several other proposed descriptors.33-34 Power analyses estimated required sample sizes sensitive to 25%, 50%, and 75% changes (up or down) from baseline in extent and magnitude at power levels of 70%, 80%, and 90% with 2-sided α < .05.30 To determine whether effects of demographic (eg, age and educational level), clinical (eg, CDR, CDR sum of boxes [CDR-SB], and MMSE scores), and postscan memory and behavioral measures contributed to hippocampal ROI test-retest variability, thereby requiring adjustment for them in the ICC calcula-
changes were found between baseline and week 12.

crease on postscan memory test 2 scores, no significant
reported elsewhere in an analysis of potential antide-
(n=12), which is not included in this report and will be
from the drug arm (n=12) did not differ widely from those of the larger group
and the placebo arm (n=12) showed similar regional ac-
brain N
fMRI scans yielded acceptable quality data. Baseline whole-
24-week study with 4 fMRI scans. A total of 95 of the 96
All patients enrolled in the larger study completed the

culations (with and without adjustment) and their rationale.

RESULTS

DEMOGRAPHICS AND CLINICAL
CHARACTERISTICS

Baseline characteristics (Table 1) of the placebo arm
(n=12) did not differ widely from those of the larger group
(n=24; MMSE score range, 18-26) or from the drug arm
(n=12), which is not included in this report and will be
reported elsewhere in an analysis of potential antide-
mentia drug effects on fMRI signals. Except for a
decrease on postscan memory test 2 scores, no significant
changes were found between baseline and week 12.

FEASIBILITY, TOLERABILITY,
AND DATA QUALITY

All patients enrolled in the larger study completed the
24-week study with 4 fMRI scans. A total of 95 of the 96
fMRI scans yielded acceptable quality data. Baseline whole-
brain N > R activation maps for the group of 24 patients
and the placebo arm (n=12) showed similar regional ac-
tivity (Figure 1), and difference maps between them were
null (ie, had no significant clusters; data not shown).

STABILITY OF IMRI WHOLE-BRAIN MAPS
AND HIPPOCAMPAL ROI ACTIVITY
ACROSS 12 WEEKS

Reliability analyses were performed in the placebo sub-
group (n=12). Regional activity patterns for N > R con-
trasts were consistent with those of past studies using the
same paradigm.1,2 At each scan, areas of significant N > R
activity were found in the bilateral hippocampi, right in-
ferior frontal cortex, right cingulate, and right prefrontal
cortex (Figure 1B-C and Table 2). Also, whole-
brain N > R activation maps for all permutations of
difference maps among time points T1, T2, and T3 (eg,
T1-T2 and T1-T3) were stable and showed no clusters
of significant activity differing between sessions.

The mean extent and magnitude for right and left hip-
ppocampal ROIs did not significantly vary across ses-
sions (weeks 0, 6, and 12) (Figure 2) or from the larger
group of 24 patients at baseline (eFigure). Repeated-
measure ANCOVAs revealed no significant changes for
hippocampal ROI signals with or without covariance ad-
justments of baseline characteristics. Sensitivity analy-
B Table 1. Demographics and Clinical and Memory Measures at Baseline (T1) and Weeks 6 (T2) and 12 (T3) for the Larger Study
and the Placebo Arm

<table>
<thead>
<tr>
<th>Variable</th>
<th>All Patients (n=24)</th>
<th>Placebo Arm (n=12)</th>
<th>T2 Placebo Arm (n=12)</th>
<th>T3 Placebo Arm (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women, No. (%)</td>
<td>15 (62.5)</td>
<td>8 (66.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, y</td>
<td>71.6 (1.7)</td>
<td>71.8 (2.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational attainment, y</td>
<td>16.0 (0.6)</td>
<td>16.2 (0.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMSE, correct</td>
<td>24.0 (0.7)</td>
<td>24.3 (0.8)</td>
<td>23.6 (1.0)</td>
<td>23.6 (0.7)</td>
</tr>
<tr>
<td>ADAS-Cog, errors b</td>
<td>26.2 (1.9)</td>
<td>25.1 (2.6)</td>
<td>25.6 (3.3)</td>
<td>25.5 (2.7)</td>
</tr>
<tr>
<td>CDR-SB score b</td>
<td>4.7 (0.5)</td>
<td>4.0 (0.6)</td>
<td>4.0 (0.6)</td>
<td>4.3 (0.7)</td>
</tr>
<tr>
<td>CDR score b</td>
<td>1.0 (0.1)</td>
<td>0.9 (0.1)</td>
<td>1.0 (0.1)</td>
<td>1.0 (0.1)</td>
</tr>
<tr>
<td>Postscan memory tests, No. (%) correct</td>
<td>67.0 (1.9)</td>
<td>65.7 (2.5)</td>
<td>58.3 (2.8)</td>
<td>63.0 (4.1)</td>
</tr>
<tr>
<td>2-Alternative forced choice name recognition</td>
<td>68.8 (3.3)</td>
<td>73.2 (4.7)</td>
<td>59.0 (4.7)</td>
<td>60.7 (3.5)</td>
</tr>
</tbody>
</table>

Abbreviations: ADAS-Cog, Alzheimer’s Disease Assessment Scale–cognitive subscale; CDR, Clinical Dementia Rating; CDR-SB, Clinical Dementia Rating scale

Table 3 lists the hippocampal ICCs, with and without adjustment for potential baseline CDR-SB score ×
Change maps with extent (percentage of voxels active [PVA]) of novel-greater-than-repeated contrast blood oxygenation level (BOLD)–dependent functional magnetic resonance imaging (fMRI) signal in the left and right hippocampal regions of interest across the 3 scans: T1 (baseline, week 0), T2 (week 6), and T3 (week 12). Hippocampal regions of interest demonstrated similar extent and magnitude of activation at each fMRI session. Error bars represent SEM.

For the right hippocampus, higher baseline CDR-SB scores were associated with larger rates of decrease in hippocampal activity. Adjusted ICCs, which preremoved variance due to CDR-SB score × time interactions, increased ICC estimates for the right hippocampus only. For right hippocampus extent, a raw individual ICC of 0.33 yielded an adjusted individual ICC of 0.59, but a raw mean ICC of 0.50 yielded an adjusted mean ICC of 0.75. For the right hippocampus magnitude, a raw individual ICC of 0.67 yielded an adjusted individual ICC of 0.87, but a raw mean ICC of 0.80 yielded an adjusted mean ICC of 0.93. For comparison, ICCs for the precuneus and posterior cingulate, important hubs in the default intrinsic connectivity network, were lower (range, 0.33–0.60) and unaffected by adjustments (eTable 1).

For 80.0% power and a group-level change of 50.0% from baseline in extent to be detected in the left hippocampus, 15 patients would be required. For similar power and 50.0% change in magnitude, 125 patients would be needed. For similar 50.0% changes to be detected in the right hippocampus for extent or magnitude, 14 patients would be required. At every power level (70.0%, 80.0%, 90.0%), left hippocampal magnitude was predicted to require sample sizes of approximately 1 order of magnitude greater than the other measures (left-side or right-side extent, right-side magnitude) (eTable 2).

This study demonstrates the feasibility of implementing task-related fMRI within the typical format of an AD RCT. Test-retest reliability of encoding-related fMRI was assessed using patients from the placebo arm who underwent fMRIs 12 weeks apart. Changes in fMRI activity were assessed globally via whole-brain map-level t tests and regionally via ICCs for magnitude and extent of N > R activity in a priori structurally defined hippocampal ROIs. Test-retest reliability occurred mostly in the moderate-

Figure 1. Stability of whole-brain statistical parametric maps.
Novel-greater-than-repeated contrast maps (P < .001, 5-voxel extent) for the same template coordinate (−24, −24, −9) for the placebo arm (n = 12) at baseline (T1) (A), week 6 (T2) (n = 11) (B), and week 12 (T3) (n = 12) (C). Mapwise activity patterns are stable and consistent with patterns found in previous studies.1,2 Difference maps with P < .01, 5-voxel extent threshold (not shown) show no significant clusters (ie, no significant differences between baseline, week 6, and week 12 scans).

Figure 2. Mean (SEM) magnitude (percentage of signal change [PSC]) and extent (percentage of voxels active [PVA]) of novel-greater-than-repeated contrast blood oxygenation level (BOLD)–dependent functional magnetic resonance imaging (fMRI) signal in the left and right hippocampal regions of interest across the 3 scans: T1 (baseline, week 0), T2 (week 6), and T3 (week 12). Hippocampal regions of interest demonstrated similar extent and magnitude of activation at each fMRI session. Error bars represent SEM.

Table 2. Peak Voxels With Significant Differences in N > R Contrast at Baseline Scan for 12 Patients

<table>
<thead>
<tr>
<th>Region</th>
<th>MNI b</th>
<th>Talairach b</th>
<th>P Value</th>
<th>Corrected c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right hippocampus</td>
<td>27, −27, −9</td>
<td>−27, −27, −6</td>
<td>.02</td>
<td>.01</td>
</tr>
<tr>
<td>Left hippocampus</td>
<td>−24, −27, −9</td>
<td>24, −27, −6</td>
<td>.01</td>
<td>.007</td>
</tr>
<tr>
<td>Right inferior</td>
<td>42, 30, −9</td>
<td>42, 29, −9</td>
<td>.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Frontal cortex</td>
<td>9, 30, 33</td>
<td>9, 31, 29</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>Right prefrontal cortex</td>
<td>30, 45, 0</td>
<td>30, 44, −2</td>
<td>.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Abbreviations: MNI, Montreal Neurological Institute; N > R, novel-greater-than-repeated face-name pair.

a Whole-brain statistical parametric map analysis for significance threshold of P < .001 with extent threshold of 5 contiguous voxels.

b Coordinates are shown for the peak voxel in a cluster.

c Represents P value for peak voxel in a cluster with small volume correction accounting for region-of-interest size (number of voxels in region of interest)–dependent multiple-comparison adjustment (ie, multiple-comparison adjustment of P value based only on voxels in the region of interest, not the whole brain).

COMMENT

This study demonstrates the feasibility of implementing task-related fMRI within the typical format of an AD RCT. Test-retest reliability of encoding-related fMRI was assessed using patients from the placebo arm who underwent fMRIs 12 weeks apart. Changes in fMRI activity were assessed globally via whole-brain map-level t tests and regionally via ICCs for magnitude and extent of N > R activity in a priori structurally defined hippocampal ROIs. Test-retest reliability occurred mostly in the moderate-
to–substantial range; whole-brain contrast maps showed stability, and hippocampal ICs, adjusted for baseline disease severity by time-related decline (which only affected right-side magnitude), ranged from 0.6 to 0.9. If a priori focus is directed at the right hippocampus or changes in extent (ie, percentage of active voxels), power estimates predict that for this paradigm, relatively modest sample sizes may detect group-level, 12-week fMRI changes in the 25.0% to 50.0% range.

We have demonstrated the feasibility of implementing multiple fMRI sessions in a longitudinal AD RCT format. Study participants tolerated an intensive imaging protocol with high yield of favorable-quality data (95 of 96 scans yielded acceptable data). Our results support the feasibility of successfully implementing task-related fMRI paradigms in mild AD across multiple scans and weeks.

The other objective of the study, to assess whole-brain, map-level fMRI and hippocampal test-retest reliability, was assessed in individuals randomized to the placebo arm. This allowed power calculations to predict sample sizes needed to accurately detect significant changes in hippocampal activity. These estimates may inform design and interpretation of future exploratory and proof-of-concept trials that use fMRI as a potential AD biomarker.

The strengths of the study include its rigorous RCT design; the inclusion of well-characterized patients undergoing stable, long-term cholinergic therapy; high compliance and follow-up; the use of a robust and well-characterized associative memory battery; its block-design encoding paradigm; and the use of standard fMRI software, tools, and processing streams that increase generalizability. Also, reliability was assessed for convergence using several approaches, sensitivity analysis showed robustness of extent and magnitude values to perturbations in statistical and extent of contiguous-voxels thresholds, and power projections were obtained to guide sample sizes for future early-phase fMRI RCTs, particularly those at single sites involving patients with mild AD.

The patterns of regional fMRI activity are consistent with those in previous studies and support the validity of focusing on changes in a priori–defined hippocampal and related ROIs in which drug-related effects on episodic memory encoding are observed, particularly in this encoding paradigm. These studies suggest specificity for hippocampal activity and inversely related activity between the hippocampus and precuneus for subsequent memory success or failure and face-name, encoding-related activity. Reassuringly, hippocampal ROIs showed the highest ICs compared with several other preselected regions in a distributed memory network.

A robust fMRI biomarker of encoding and retrieval processes would ideally include 1 or more measures of shifting patterns of activity (signatures) in core network hubs that include, depending on cognitive load and task specificity, hippocampal and related medial temporal lobe areas; precuneus, posterior cingulate, and related medial and lateral parietal regions; and medial inferior and dorsal-lateral frontal cortices. Although this study primarily focuses on longitudinal fMRI feasibility and reliability in the hippocampus, a central node in memory acquisition and integration, future studies will leverage cognitive networks by integrating activity patterns in hubs, including medial and lateral parietal and medial and inferior frontal regions; assess reliability and power analysis for fMRI network signals; and explore potential drug-related effects.

Overall, we opted for a conservative bias and greater focus on generalizability. We used individual (single) ICs, not group (mean) ICs (arithmetic average for a group of scans) that would have provided higher values (Table 3). Calculated ICs also assumed random scans (ie, model 2 ICs), as opposed to fixed ones (ie, model 3 ICs), thereby increasing the generalizability of results. Hippocampal ICs, with or without adjustment for baseline CDR-SB scores, are generally higher than those recently reported in healthy elderly controls and patients with MCI in verbal episodic memory encoding and retrieval fMRI tasks 6 weeks apart. Also, power predictions for estimated sample sizes to detect changes in fMRI measures 12 weeks apart assume modest (25.0%–75.0%) and bidirectional changes (2-sided α values) in hippocampal activity. In similar paradigms, ROI effect sizes were larger or unidirectional, including in the hippocampi of young patients administered scopolamine (percentage change vs placebo, −53% for extent and −57%
for magnitude) and lorazepam (percentage change vs placebo, −52% for extent and −57% for magnitude).27 and in fusiform regions of AD patients administered rivastigmine (percentage change vs no rivastigmine, +95% for magnitude in left and +660% in right fusiform regions).8 With the use of exploratory analyses, unidirectional a priori hypotheses (eg, IMRI activity will increase with drug or intervention) or exclusion of left hippocampal magnitude as a primary signal measure may allow modest sample sizes to detect changes in the 50.0% range. For this paradigm, population (ie, mild AD), and interval of several days to weeks, the right hippocampal signal, especially the magnitude measure (ie, percentage change in right hippocampal BOLD signal),30 is likely to be most sensitive to physiologic, pathologic, and pharmacologic stressors; has substantial measurement reliability (during these short intervals unless corrected for trait instability); and potentially may be useful as an exploratory biomarker of trait, state, rate, or signal of effect. Later in the disease state, this may not be so because the neural correlates that affect the BOLD signal changes may be muted. Pairing an IMRI scan with a clinical visit at week 12 provides a parsimonious design consistent with proof-of-concept AD trials. For experimental drugs with potential subacute symptomatic effects, this provides a sufficient interval to detect signals of clinical efficacy beyond 4-week to 8-week windows when placebo effects may mingle with drug-related effects in such a way that the elements cannot be distinguished or separated.37,38 Finally, the tools and methods used for functional data analyses were simple, standard, and widely available (eg, statistical parametric map, Montreal Neurological Institute template space, MarsBaR).

It is important to recognize that the interaction of baseline CDR-SB score with a decrease in right hippocampal fMRI signal over time is not due to fMRI measurement inaccuracy or unreliability but is an estimable component of putative real variation that can be accounted for independently and removed from an adjusted ICC (through regression and residualizing methods), as was done in our study. Otherwise, it might confound as measurement unreliability and bias ICCs downward. Hence, we opted for the more conservative approach of removing this confounding source of variance from the denominator of the ICC formula but without adding it to the numerator (eAppendix). The finding that patients with greater impairment, ie, those with higher baseline CDR-SB scores, exhibited a greater decrease in right hippocampal activity 12 weeks later is consistent with studies that show a decreased hippocampal signal via fMRI in AD relative to cognitively intact older controls and patients with MCI.1,5,22 It is also consistent with the hypothesis that once AD patients meet criteria for mild dementia, task-related hippocampal activity may rapidly decline with advancing illness.3,39 Similarly, AD patients with smaller hippocampal volumes subsequently show a greater rate of decrease in hippocampal volumes during 1 year.39 Baseline levels of cognition and function and their interactions with time in study are also important determinants of clinical trajectory of decline.30,42 Finally, improvements compared with raw ICCs were specific to the right hippocampus; raw and adjusted ICCs were not substantially different in the left hippocampus and comparison ROIs (Table 3 and eTable 1). This is not surprising because the face-name paradigm provides greater novelty and cognitive demands in the visual domain, and previous studies1,3-7,27 have shown task-related, age-related, and disease-related sensitivity for this paradigm in the right hippocampus.

These data and interpretations also have limitations and caveats. Although this study provides favorable internal validity and successful implementation at a single designed site, results could vary considerably across multiple sites, scanners, platforms, and AD populations. These results require validation in single-site studies and assessment of whether findings for whole-brain and hippocampal signal reliability will accurately reflect scaling in multisite studies. Our patients were experiencing the mild clinical stages of AD (CDR of 1; mean [SD] MMSE score, 24.0 [0.7]), were highly educated, and were receiving stable, long-term donepezil hydrochloride therapy. Although generalizable to most candidates with AD eligible for currently enrolling in experimental drug RCTs, extrapolation to those who are drug naive, use other antidementia medicines, or have low educational levels requires caution. Also, on the basis of our previous experience, it is likely that most patients with moderate-stage (CDR, 2) AD would have difficulty completing this fMRI paradigm and performing above chance levels. High internal validity and patient homogeneity in our study may have resulted in underestimation of ICCs due to low between-patient variance. Although we do not measure or adjust for individual or native hippocampal volumes or possible changes, given the low annual rates of hippocampal atrophy in AD, it is unlikely that atrophy during a 12-week period would significantly affect the accuracy of ROI boundaries and fMRI signals.39,45,46 Our results suggest extent and right hippocampal measures (extent and magnitude) may be more robust and efficient for power projections in visual-verbal, paired-associate paradigms. The left magnitude measure had only moderate ICCs (0.67), resulting in the need for many more patients to detect 25.0% to 50.0% effects. Except for the left hippocampus, magnitude ICCs were approximately 0.1 to 0.3 higher than extent ICCs (Table 2 and eTable 1). However, power analysis did not show an advantage of using magnitude measures, especially on the left side. This finding underscores that ICCs and sample size estimates provide somewhat complementary information for pragmatic design and interpretation of biomarkers in AD RCTs.

Importantly, our short-term study does not address the usefulness of fMRI in detecting disease-modifying effects in longer-term studies in AD populations. It is possible that a subacute fMRI effect will be predictive of longitudinal change, but as with positron emission tomography and structural MRI,45,46 the ultimate validation of fMRI as a potential biomarker of efficacy will require incorporation into an AD therapeutic trial demonstrating positive clinical benefit. Caution should be exercised in general pertaining to the nature of the BOLD fMRI signal as a surrogate for neural activity. Changes in the BOLD signal may reflect other neurophysiologic processes, including microvascular coupling, and not necessarily changes in dendritic synaptic local field potentials. Future studies will assess test-retest reliabili-
ity by defining ROIs in native space, leveraging network dynamics, and using modeling to quantify functional connectivity.

In conclusion, our study demonstrated moderate-to-substantial test-retest reliability for a face-name, paired-associate encoding, block-design fMRI paradigm performed by patients with mild AD at a single site. These highly focused findings suggest that significant BOLD fMRI changes in hippocampal signals occur acutely or subacutely within 12 weeks due to a potential intervention or disease progression, the signal, noise, and measurement variability characteristics of longitudinal fMRI measures using similar encoding paradigms may allow their detection with reasonable accuracy. Power analyses suggest that detection of changes from baseline hippocampal activity in the 50.0% range may require dozens, not hundreds, of study participants, especially if a priori or exploratory focus is on right hippocampal or extent measures. Meanwhile, small group-level changes in the 25.0% range may be detectable with sample sizes currently used in small phase 2 AD trials. These results support the feasibility of using fMRI as a potential biomarker in early-phase proof-of-concept RC Ts to detect whether a drug is acutely or subacutely reaching or affecting the brain or having a specific targeted or biological effect (as measured via BOLD fMRI) on a brain region or network. This study provides evidence that task-related fMRI is feasible to implement longitudinally in mild AD at a single site and may have sufficient test-retest reliability to be incorporated in early-phase clinical trials. In combination with other experimental measures, task fMRI may potentially help detect a signal of effect and guide early-development programs for novel AD therapeutics.

Accepted for Publication: October 14, 2010.
Correspondence: Alireza Atri, MD, PhD, Memory Disorders Unit, Massachusetts General Hospital, 15 Parkman St, Wang Ambulatory Care Center 715, Boston, MA 02114 (atri@nmr.mgh.harvard.edu).

Author Contributions: Drs Atri and Sperling had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Atri, Rentz, and Sperling. Acquisition of data: Atri, O’Brien, Rastegar, Salisbury, DeLuca, O’Keefe, and Rentz. Analysis and interpretation of data: Atri, Sreenivasan, LaViolette, Rentz, Locascio, and Sperling. Drafting of the manuscript: Atri, Sreenivasan, DeLuca, Rentz, and Sperling. Critical revision of the manuscript for important intellectual content: Atri, O’Brien, Rastegar, Salisbury, O’Keefe, LaViolette, Locascio, and Sperling. Statistical analysis: Atri, LaViolette, and Locascio. Obtained funding: Atri and Sperling. Administrative, technical, and material support: Atri, Sreenivasan, Rastegar, Salisbury, DeLuca, O’Keefe, and LaViolette. Study supervision: Atri, Rentz, and Sperling.

Financial Disclosure: Dr Atri has served as a consultant or on the scientific advisory board or has received lecture honoraria from Eisai Pharmaceuticals, Forest Pharmaceuticals Inc, H. Lundbeck A/S, Merck & Co Inc, Merz Pharmaceuticals, and Novartis AG. Dr Sperling has served as a consultant or on the scientific advisory board or has received lecture honoraria from Eisai Pharmaceuticals, Elan Corporation plc, Eli Lilly and Company, Forest Pharmaceuticals Inc, Merck & Co Inc, Pfizer Inc, and Wyeth Pharmaceuticals. The principal authors (Drs Atri and Sperling) retain full control of the data and publication rights.

Funding/Support: The study was supported by National Institute on Aging grantsIK23 AG027171(Dr Atri) and R01 AG027435 (Dr Sperling); the Harvard–Massachusetts Institute of Technology Health Sciences and Technology Pfizer-Merk Clinical Investigator Training Program (Dr Atri); the National Institutes of Health loan repayment program (Dr Atri); Investigator-Initiated Research Grants from Forest Pharmaceuticals Inc and the Harvard Center for Neurodegeneration and Repair; the Clinical, Neuroimaging, and Statistics Cores of the Massachusetts Alzheimer’s Disease Research Center (National Institute on Aging grant 5 P50AG05134 to Dr Growdon and Bradley T. Hyman, MD, PhD); and the Geriatric Research Education and Clinical Center at the Edith Nourse Rogers Memorial Veterans Administration Bedford Medical Center. Less than 30% of this research was supported by an Investigator-Initiated Research Grant from Forest Pharmaceuticals Inc.

Online-Only Materials: The eAppendix, eFigure, eTable 1, and eTable 2 are available at http://www.archneurol.com.

Additional Contributions: Kim Celone, PhD, Kristina Depeau, MPH, Eli Diamond, MD, Saul Miller, MS, Matja Pihlamajaki, MD, and Meghan Searl, PhD, provided assistance with data collection and preliminary data processing. Lynn Shaughnessy, MA, provided assistance with manuscript preparation. John Growdon, MD (Massachusetts General Hospital Memory Disorders Unit and Massachusetts Alzheimer’s Disease Research Center), provided significant assistance with recruitment of participants, obtaining space and resources, and guidance, and Bruce Rosen, MD, PhD (Martinos Center for Biomedical Imaging), provided guidance, space, and resources for this research. Finally, and most important, we express our deep gratitude for the commitment of the patients, family members, and caregivers without whose generous contribution and dedication this research would not be possible.

REFERENCES


8. Rombouts SARB, Barkhof F, Van Meel CS, Scheltens P. Alterations in brain ac-
11. Johnson SC, Schmitz TW, Mortiz CH, et al. Activation of brain regions vulnera-
campal activation—a pharmacological fMRI study. Brain. 2006;129(pt 1):141-
157.
14. Greicius MD, Srivastava G, Reiss AL, Menon V. Default-mode network activity distin-
14504-14509.
16. Petrella JR, Prince SE, Krishnan S, Husn H, Kelley L, Doraiswamy PM. Effects of donepezil on cortical activation in mild cognitive impairment: a pilot double-
17. Goekoop R, Rombouts SAR, Jonker C, et al. Challenging the cholinergic sys-
sponse among normal, MCI, and Alzheimer’s patients. Neurology. 2003;60(2):
253-260.
21. Remy F, Mirashed F, Campbell B, Richter W. Verbal episodic memory impair-
22. Rombouts SAR, Barkhof F, Veitman DJ, et al. Functional MR imaging in Alz-