Comparing 3 T and 1.5 T MRI for Tracking Alzheimer's Disease Progression with Tensor-Based Morphometry

April J. Ho,[1] Xue Hua,[1] Suh Lee,[1] Alex D. Leow,[1,2] Igor Yanovsky,[3]
Boris Gutman,[1] Ivo D. Dinov,[1] Natasha Leporé,[1] Jason L. Stein,[1]
Arthur W. Toga,[1] Clifford R. Jack, Jr.,[4] Matt A. Bernstein,[4]
Eric M. Reiman,[5] Danielle J. Harvey,[6] John Kornak,[7] Norbert Schuff,[8,9]
Gene E. Alexander,[11] Michael W. Weiner,[8,9,10] Paul M. Thompson[1]*
and the Alzheimer's Disease Neuroimaging Initiative

[1]Laboratory of Neuro Imaging, Department of Neurology, UCLA School of Medicine,
Los Angeles, California
[2]Resnick Neuropsychiatric Hospital at UCLA, Los Angeles, California
[3]Department of Mathematics, University of California, Los Angeles, California
[4]Mayo Clinic, Rochester, Minnesota
[5]Banner Alzheimer's Institute, Department of Psychiatry, University of Arizona, Phoenix, Arizona
[6]Department of Public Health Sciences, UC Davis School of Medicine, Davis, California
[7]Department of Radiology and Biomedical Imaging and Department of Epidemiology and Biostatistics,
UC San Francisco, San Francisco, California
[8]Veteran Affairs Medical Center and Department of Radiology and Biomedical Imaging, UC San
Francisco, California
[9]Department of Radiology, UC San Francisco, California
[10]Department of Medicine and Psychiatry, UC San Francisco, San Francisco, California
[11]Department Psychology and Evelyn F. McKnight Brain Institute, University of Arizona,
Tucson, Arizona

◆   ◆

Abstract: A key question in designing MRI-based clinical trials is how the main magnetic field strength of the scanner affects the power to detect disease effects. In 110 subjects scanned longitudinally at both 3.0 and 1.5 T, including 24 patients with Alzheimer's Disease (AD) [74.8 ± 9.2 years, MMSE: 22.6 ± 2.0 at baseline], 51 individuals with mild cognitive impairment (MCI) [74.1 ± 8.0 years, MMSE: 26.6 ± 2.0], and 35 controls [75.9 ± 4.6 years, MMSE: 29.3 ± 0.8], we assessed whether higher-field MR imaging offers higher or lower power to detect longitudinal changes in the brain, using tensor-based morphometry (TBM) to reveal the location of progressive atrophy. As expected, at both field strengths,

progressive atrophy was widespread in AD and more spatially restricted in MCI. Power analysis revealed that, to detect a 25% slowing of atrophy (with 80% power), 37 AD and 108 MCI subjects would be needed at 1.5 T versus 49 AD and 166 MCI subjects at 3 T; however, the increased power at 1.5 T was not statistically significant ($\alpha = 0.05$) either for TBM, or for SIENA, a related method for computing volume loss rates. Analysis of cumulative distribution functions and false discovery rates showed that, at both field strengths, temporal lobe atrophy rates were correlated with interval decline in Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-cog), mini-mental status exam (MMSE), and Clinical Dementia Rating sum-of-boxes (CDR-SB) scores. Overall, 1.5 and 3 T scans did not significantly differ in their power to detect neurodegenerative changes over a year. *Hum Brain Mapp 31:499–514*, 2010. © 2009 **Wiley-Liss, Inc.**

Key words: Alzheimer's disease; tensor-based morphometry; MRI; field strength

◆ ═══════════════════════════════════════ ◆

INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia, affecting more than 26 million people worldwide [Wimo et al., 2006]. With the aging population living longer than ever before, AD is now a major public health concern with the number of affected patients expected to triple to reach 13.4 million, by the year 2050, in the United States alone [Mueller et al., 2005b]. Early signs of AD include loss of short-term memory functioning followed by a progressive decline in other cognitive domains including language, attention, orientation, visuospatial skills, and executive function, as well as emotional and behavioral disturbances. Several current therapeutic trials aim to delay disease progression by targeting patients with amnestic mild cognitive impairment (MCI), an intermediate risk state with a 5-fold increased annual conversion rate to AD compared to healthy population [Petersen, 2000; Petersen et al., 1994, 2001; Petersen and Negash, 2008].

Magnetic resonance imaging (MRI) is now widely used to detect changes in brain volume over time [Fox et al., 2000; Jack et al., 2003, 2008; Scheltens et al., 2002; Thompson et al., 2003]. As new treatments are developed to slow or delay disease progression, there is an urgent need to assess and compare the power of imaging methods for tracking and predicting disease progression, and discovering statistical effects of factors that may delay or accelerate disease onset (e.g., treatment, genotype, education, diet, and cardiovascular health). The Alzheimer's Disease Neuroimaging Initiative (ADNI), a collaborative project funded by the National Institute of Aging and the pharmaceutical industry, includes a major effort to optimize technical standards for image acquisition and analysis [Jack et al., 2008].

The U.S. Food and Drug Administration began approving 3 T brain MRI for clinical use in the late 1990s, presenting a new opportunity for imaging disease progression in the brain [Frayne et al., 2003]. Theoretically, increasing the magnetic field strength from standard 1.5 to 3 T roughly doubles the signal-to-noise ratio (SNR), and provides higher contrast to noise, per unit scan time, to better differentiate gray/white matter and other tissues. Even so, 3 T MR images often have an increased level of artifact compared to their 1.5 T counterparts [Bernstein et al., 2006]. For example, inhomogeneity in the RF transmit field can lead to an increased central brightening artifact at 3 T [Collins et al., 2005]. These artifacts can affect the accuracy of automated algorithms that classify tissue into gray and white matter components [Sled et al., 1998]. Also, as the field strength increases, the magnetic field inhomogeneity due to spatial variations in susceptibility increases [Schenck, 1996]. This can lead to local spatial distortion as well as artifactual local variations in image intensity. Consequently at 3 T, geometric distortions and signal drop-off can occur due to sharp changes in magnetic susceptibility at tissue/air interfaces, especially at the frontal and temporal poles [Frayne et al., 2003; Jack et al., 2008]; these effects are more problematic than at 1.5 T. Even so, higher-field imaging offers higher SNR for many other MRI-based acquisitions, such as blood-oxygenation level dependent (BOLD) contrast in functional MRI, diffusion tensor imaging, and MR spectroscopy. 3 T scanners now represent roughly 10% of the U.S. scanner market but they require the development of radio frequency antennas to accommodate the higher resonant frequency and other technical modifications that can handle increased chemical shift (as measured in Hz), a higher deposition of radio-frequency (RF) energy into the patient's tissue, increased acoustic noise, and greater need for safety precautions regarding implanted metallic devices [Bernstein et al., 2006; Frayne et al., 2003].

Few studies have directly compared 3 T and 1.5 T scanning for morphometric analyses, perhaps because this would require a relatively large cohort of subjects to be scanned at both field strengths. To help evaluate whether higher-field MRI is better for detecting structural brain changes in patients with AD, we conducted a study, as part of the ADNI, in which 25% of all subjects were

scanned at both 1.5 and 3 T at selected sites, using optimized MRI sequences at each respective field.

We analyzed 110 subjects scanned at both field strengths using tensor-based morphometry (TBM), a relatively new image analysis technique that identifies brain changes over time, based on the gradients of the deformation fields that align successive brain scans [Ashburner and Friston, 2003; Fox et al., 2001; Hua et al., 2008a,b; Leow et al., 2009; Studholme et al., 2001; Thompson et al., 2000]. We examined longitudinal brain changes, comparing maps of atrophic rates in groups of AD and MCI subjects relative to controls scanned at 1.5 and 3 T. To determine which field strength best detected progressive brain atrophy, we computed how many subjects would be needed to detect a 25% reduction in the mean annual rate of brain loss, a statistic that has been advocated as a measure of statistical power for clinical trials [Jack et al., 2008]. To boost power for sample size estimation, we used a technique recently advocated by Reiman and Chen [Hua et al., 2009; Reiman et al., 2008; Reiman and Langbaum, 2009], in which atrophic rates are summarized in a statistically predefined subregion of an anatomical ROI (such as the temporal lobe) showing the most active atrophy in an independent sample of AD subjects. Small sample sizes to detect active disease are a necessary but not sufficient condition for a valuable neuroimaging biomarker; it is also vital that the changes correlate with (or predict) cognitive decline, which we have previously found to be correlated with the atrophic rates in TBM [Leow et al., 2009]. We therefore also used cumulative distribution function (CDF) plots and false discovery rate (FDR) methods to compare the power of 1.5 versus 3 T scans of the same subjects to detect correlations between ongoing atrophy and cognitive decline. For this, we correlated temporal lobe rates of atrophy (at the voxel-wise level) with standard cognitive measures including the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog), mini-mental sate examination (MMSE), and clinical dementia rating (CDR), all standard tests that are widely used in studies of AD.

As 3 and 1.5 T scanning each have strengths and weaknesses, we assessed the hypothesis that estimated sample sizes for AD and MCI groups would differ at 1.5 versus 3 T, but we used a two-tailed hypothesis test, as there is an active debate regarding which is superior, depending on the application. We also tested whether declines in cognitive scores (ADAS-cog, MMSE, and CDR-SB scores) were strongly correlated with the detected rate of temporal lobe atrophy at 3 T, based on the notion that there may be greater signal drop-out and non-disease-related distortions at the temporal poles. Still, we expected this limitation to be partially mitigated by using a statistically predefined ROI that focused on areas where atrophy was detectable in an independent sample at each field strength, thereby explicitly avoiding voxels where power was diminished. Field strength effects were tested against the null hypothesis that the field strength made no difference; to test this, we used a permutation approach.

MATERIALS AND METHODS

Subjects

Imaging data for this study was obtained from the Alzheimer's Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI) [Mueller et al., 2005a,b]. One of the largest studies of AD to date, ADNI is a 5-year collaborative project with support from the National Institute of Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), nonprofit organizations, and private pharmaceutical companies. The project began in 2003 and evaluates 800 adults, aged 55–90, including 200 elderly controls, 400 MCI subjects, and 200 patients with AD. The primary goal of ADNI is to determine whether serial MRI, positron emission tomography (PET; FDG and amyloid imaging), other biological markers, and clinical and neuropsychological assessments can be used as a reliable measure to track disease progression in patients with MCI and AD. Identifying specific markers sensitive to MCI and early AD is important for therapeutic development, and for monitoring treatment effectiveness in clinical trials when cost and time are considered. The Principal Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California, San Francisco.

1.5 and 3 T MRI scans were acquired at multiple sites and time points. Of all subjects scanned, 25% were scanned at 3 T at 31 of 59 participating sites. 3 and 1.5 T scans from the same subject are shown in Figure 1 for purposes of visual comparison. There are no striking visible differences, although the gray/white matter contrast appears slightly greater at 3 T, at least in this randomly selected subject. In this article, 110 subjects scanned at both 1.5 and 3 T were analyzed over a 1-year follow-up interval to assess structural brain change. Although the ADNI dataset contains many more 3 and 1.5 T scans than analyzed in this study, we restricted our attention to subjects with baseline and 12-month follow-up scans from both 1.5 and 3 T MRI scanners. This was done to avoid cohort effects; we were concerned that if we analyzed a different set of subjects at each field strength, it would be unclear whether any detected differences might be partly attributable to differences in the cohorts (e.g., age, sex, educational level, severity of AD, or other unidentified factors, etc.). While, in principle, this additional variability could also be corrected by including these attributes as covariates, such models are invariably imperfect and some bias due to imperfect matching would remain. Subjects were divided into three groups: 24 patients with AD (baseline age: 74.8 ± 9.2 years), 51 amnestic MCI subjects (baseline age: 74.1 ± 8.0 years), and 35 healthy elderly controls (baseline age: 75.9 ± 4.6 years; subject demographics are shown in Table I).

All subjects completed detailed clinical and cognitive assessments including the Alzheimer's Disease Assessment Scale (ADAS-Cog), Mini-Mental State Examination
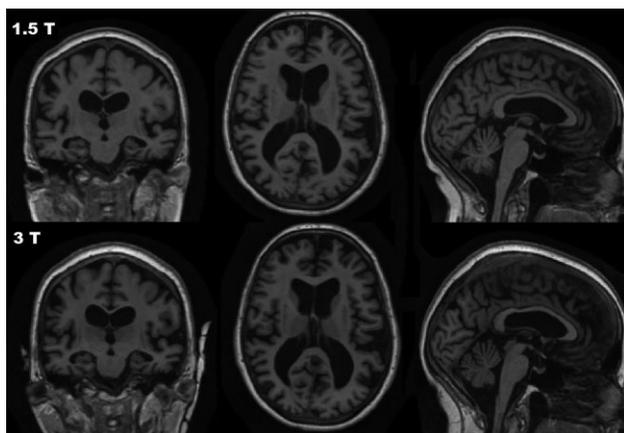
Figure 1.
1.5 T and 3 T MRI from the same subject. Note the marginally higher gray/white matter contrast in some regions (e.g., the internal capsule) and slightly higher spatial resolution at 3 T, but no other striking differences are apparent.

(MMSE), and the Clinical Dementia Rating Sum-of-Boxes score (CDR-SB) at the time of the baseline and follow-up scans. ADAS-Cog is based on a 70-point scale designed to measure the severity of cognitive impairment, and is currently the most widely used cognitive measure in AD trials [Rosen et al., 1984]. It consists of 11 tasks assessing learning and memory, language production and comprehension, constructional and ideational praxis, and orientation. The MMSE, with scores ranging from 0 to 30, provides a global measure of mental status based on five cognitive domains: orientation registration, attention and calculation, recall, and language [Cockrell and Folstein, 1988; Folstein et al., 1975]. Scores lower than 24 are typically associated with dementia. The sum-of-boxes clinical dementia rating (CDR-SB), ranging from 0 to 18, measures dementia severity by evaluating patients' performance in six domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care [Berg, 1988; Hughes et al., 1982; Morris, 1993]. All

patients with AD met NINCDS/ADRDA criteria for probable AD [McKhann et al., 1984]. On average, patients with AD in this study were considered to have mild to moderate, but not severe AD with baseline MMSE score 22.6 ± 1.96, CDR-SB score of 4.1 ± 5.1, and ADAS-Cog of 17.7 ± 5.66. Average MMSE, CDR-SB, and ADAS-Cog scores for each group are displayed in Table I. Detailed exclusion criteria may be found in the ADNI protocol [Mueller et al., 2005a,b].

MRI Acquisition, Image Calibration, and Correction

All subjects were scanned at multiple ADNI sites, with 31 of the total 59 sites acquiring both 1.5 and 3 T scans, according to a standardized protocol developed after a major effort to evaluate 3D T1-weighted sequences for morphometric analyses [Jack et al., 2008a; Leow et al., 2006]. High-resolution structural brain MRI scans were acquired using 1.5 and 3 T MRI scanners from General Electric Healthcare, Philips, and Siemens Medical Solutions (Table II shows the breakdown of the number of patients by scanner vendor).

In the 1.5 T scanning protocol, each subject underwent two 1.5 T T1-weighted MRI scans using a 3D sagittal volumetric magnetization prepared rapid gradient echo (MP-RAGE) sequence. As described in Jack et al. [2008], typical 1.5 T acquisition parameters are repetition time (TR) of 2,400 ms, minimum full TE, inversion time (TI) of 1,000 ms, flip angle 8°, 24 cm field of view, with a 256 × 256 × 170 acquisition matrix in the x-, y-, and z-dimensions yielding a voxel size of 1.25 × 1.25 × 1.2 mm³. In-plane, zero-filled reconstruction yielded a 256 × 256 matrix for a reconstructed voxel size of 0.9375 × 0.9375 × 1.2 mm³. For 3 T scans, acquisition parameters were repetition time (TR) of 2,300 ms, minimum full TE, inversion time (TI) of 900 ms, flip angle 8°, 26 cm field of view, with a 256 × 256 × 170 acquisition matrix in the x-, y-, and z-dimensions yielding a voxel size of 1.0 × 1.0 × 1.2 mm³. In plane, zero-filled reconstruction yielded a 256 × 256 matrix for a reconstructed voxel size of 1.0 × 1.0 × 1.2 mm³,

TABLE I. Patient demographics and cognitive scores

Measures	CTL	MCI	AD
Number of subjects	35 (13 M/22 F)	51 (30 M/21 F)	24 (10 M/14 F)
Age (years)	75.9 ± 4.6	74.1 ± 8.0 (NS)	74.8 ± 9.2 (NS)
Baseline CDR-SB	0.0 ± 0.0	1.66 ± 0.86**	4.1 ± 1.53**
Change in CDR-SB	0.057 ± 0.21	1.03 ± 1.52**	1.04 ± 1.19**
Baseline ADAS-Cog	5.05 ± 2.45	11.7 ± 4.39**	17.7 ± 5.66**
Change in ADAS-Cog	0.18 ± 2.85	2.43 ± 5.27*	3.29 ± 5.01**
Baseline MMSE	29.3 ± 0.80	26.6 ± 2.0**	22.6 ± 1.96**
Change in MMSE	−0.06 ± 1.31	−1.18 ± 2.90*	−0.96 ± 3.81 (NS)

Two-sample t-tests show which measures differed from those in controls: (NS) denotes not significant.
*$P < 0.05$; **$P < 0.01$. Two-tailed tests were used to test for age differences; one-tailed tests were used for cognitive scores and changes, on the premise that scores would be expected to be poorer and declines greater in AD and MCI than in controls.

Vendor	Siemens	Philips	GE Medical
1.5 T	20	13	77
3 T	65	29	16

Most 3 T scans were acquired on Siemens scanners, but most of the 1.5 T scans were collected on GE scanners.

although this reconstructed voxel size can be further decreased with sinc interpolation, if desired. The ADNI MR imaging protocol [Jack et al., 2008] compensated for the increased chemical shift and susceptibility artifacts observed at 3 T by doubling the receive bandwidth compared to the 1.5 T acquisition. This change costs a factor of $\sqrt{2}$ in the signal-to-noise ratio (SNR). SNR approximately doubles at 3 T compared to 1.5 T; the remaining factor of $\sqrt{2}$ was used to increase the spatial resolution of the 3 T protocol as described earlier. When necessary, the transmit bandwidth of the inversion RF pulse was also increased at 3 T to eliminate incomplete inversion artifacts [Bernstein et al., 2006]. On modern systems with phased array receive coils, the acquisition time at 1.5 T was approximately 7.7 min, compared to 9.3 min at 3 T. Because of differences in hardware, spin relaxation properties, chemical shift properties, and susceptibility artifacts at 1.5 and 3 T, the sequence parameters were not identical on the two scanners. Even so, sequences were optimized as much as possible to obtain similar tissue contrast at both field strengths.

Additional image corrections were also applied, using a processing pipeline at the Mayo Clinic, consisting of the following: (1) a procedure termed *GradWarp* for correction of geometric distortion due to gradient nonlinearity [Jovicich et al., 2006], (2) a "B1-correction," to adjust for image intensity inhomogeneity due to B1 nonuniformity using calibration scans [Jack et al., 2008], (3) "N3" bias field correction, for reducing residual intensity inhomogeneity [Sled et al., 1998], and (4) geometrical scaling, according to a phantom scan acquired for each subject [Jack et al., 2008], to adjust for scanner- and session-specific calibration errors. In addition to the original uncorrected image files, images with all of these corrections already applied (Grad-Warp, B1, phantom scaling, and N3) are available to the general scientific community (at www.loni.ucla.edu/ADNI).

Image Preprocessing

To adjust for global differences in brain positioning and scale across individuals, all scans were linearly registered to the stereotactic space defined by the International Consortium for Brain Mapping (ICBM-53) [Mazziotta et al., 2001] with a 9-parameter (9P) transformation (3 translations, 3 rotations, 3 scales) using the Minctracc algorithm [Collins et al., 1994]. Follow-up scans were linearly registered to its matching baseline scan using a 9P registration. Both mutually aligned scans were then linearly registered to the ICBM-53. Globally aligned images were resampled in an isotropic space of 220 voxels along each axis (x, y, and z) with a final voxel size of 1 mm^3.

Minimal Deformation Target

For each field strength, a separate minimal deformation target (MDT), or group mean template, was constructed. This has been advocated in prior studies to reduce bias and improve statistical power [Hua et al., 2008a,b; Kochunov et al., 2001; Leporé et al., 2008]. The MDT was constructed using 40 normal controls' baseline scans as in our prior studies [Hua et al., 2008a,b]. A separate MDT template was created for the 1.5 T and for the 3 T scans (these average brain templates are shown in Fig. 2). To create the MDT, we first created an affine average template using an average of the globally-aligned scans after 9-parameter (9P) normalization. Next, a nonlinear average template was made by warping individual brain scans to the initial affine template. We used a nonlinear inverse consistent elastic intensity-based registration algorithm [Leow et al., 2005], which optimizes a joint cost function based on mutual information (MI) and the smoothness of the deformation fields. The deformation field was computed using a spectral method to implement the Cauchy–Navier elasticity operator [Marsden et al., 1983; Thompson et al., 2000] using a Fast Fourier Transform (FFT) resolution of 32 × 32 × 32. After the 40 scans were nonlinearly registered to the affine template, the average of these scans was used to create a nonlinear average intensity template. Then the MDT is created after applying inverse geometric centering of the displacement fields to the nonlinear average template (see
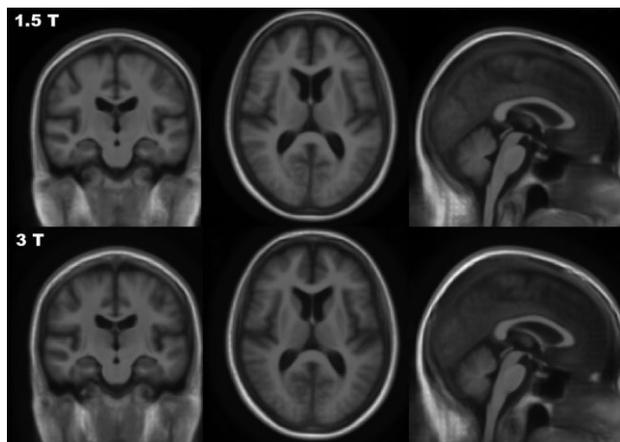
Figure 2.
Minimal Deformation Template (MDT) based on 1.5 T scans (top) and 3.0 T scans (bottom). There are no obvious differences in contrast or anatomical structure differentiation in these group mean templates.

Kochunov et al., 2002, 2005; Lepore et al., 2008, for related work and the rationale for this step).

To quantify 3D patterns of volumetric brain atrophy over time for each subject, an individual brain change map (Jacobian determinant map) was created using an unbiased symmetric Kullback-Leibler (sKL) method based on mutual information [Yanovsky et al., 2007, 2008]. 1.5 T baseline scans ($N = 110$) were first nonlinearly registered to the MDT specific for the 1.5 T normal group, and all 3 T baseline scans ($N = 110$) were nonlinearly registered to the MDT specific to the 3 T normal group [Hua et al., 2008a,b; Yanovsky et al., 2008]. After each scan was aligned to the MDT for its respective field strength, a Jacobian matrix field reflecting the gradients of the deformation field was derived for each subject. For 12-month follow-up scans, the follow-up scan for each subject was linearly and then nonlinearly registered to its corresponding baseline scan again using the same registration algorithm. Maps of change were shown on the baseline image warped to the MDT space.

Statistical Tests

Group comparisons

To illustrate systematic differences in atrophic rates between groups (AD or MCI vs. normal), we constructed voxel-wise statistical maps based on the Student's t-statistic. We corrected for the multiple comparisons implicit in making a statistical map, by using permutation tests [Bullmore et al., 1999; Chiang et al., 2007; Nichols and Holmes, 2002; Thompson et al., 2003]. In brief, a null distribution for the group differences in atrophic rates (Jacobian values) at each voxel was constructed using 5,000 random permutations. For each test, the subjects' diagnosis was randomly permuted and voxel-wise t-statistics were calculated. A ratio, describing the fraction of the time the t-statistic was more extreme in the randomized tests than the original test, was calculated to give a permutation-based P-value for the significance at each voxel. A "global P-value," describing the fraction of the time the suprathreshold volume ($P < 0.01$, uncorrected) was greater in the randomized maps than the real effect (the original labeling), was calculated to determine whether any significant changes could be detected across the brain. This procedure has been used in many prior reports [Braskie et al., 2008; Chiang et al., 2007; Chou et al., 2009]. The permutation testing therefore controlled for the number of vertices above $P < 0.01$ in the entire map (0.01 was chosen as the primary threshold at the voxel level, although other values could arguably be used). This is one of several standard ways to set up a permutation test and is sometimes called set-level inference. It deems a map significant when the total quantity of voxels with P-values lower than a fixed *a priori* threshold exceeds that obtained in 95% of random simulations.

Cumulative distribution function plots

Cumulative distribution function (CDF) plots were compiled based on the P-values generated in the two-sample t-tests. These were used to compare the effect sizes of effects of covariates of interest in all three groups [Lepore et al., 2008, Hua et al., 2008a,b; Morra et al., 2008]. The false discovery rate (FDR) method was used to assign overall significance values to each statistical map, based on the expected proportions of voxels with statistics exceeding any given threshold under the null hypothesis [Benjamini and Hochberg, 1995; Genovese et al., 2002; Storey, 2002].

Correlations of structural brain differences with cognitive measures

Correlations were computed at every voxel between rates of atrophy and cognitive scores using the Spearman's correlation. Interval changes (over 1 year) in scores from the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog), Mini-Mental State Examination (MMSE), and the Clinical Dementia Rating Sum-of-Boxes scales (CDR-SB) were correlated, at the voxel level, with structural brain changes over time after controlling for age and sex [Hua et al., 2008a,b; Leow et al., 2009; Morra et al., 2008]. All correlation maps were corrected for multiple comparisons as described earlier, using the FDR method.

Sample size

Using a statistically defined ROI based on voxels with significant atrophic rates ($P < 0.00001$) in a nonoverlapping training set of 22 patients with AD, a mean atrophic rate was computed for each subject [Hua et al., 2009; Reiman and Langbaum, 2009; Reiman et al., 2008]. A statistically-defined ROI was created for each group of scans. Prior studies have found that sample size estimates are relatively stable with respect to the statistical threshold used to define the statistically predefined ROI [Hua et al., 2009]. For each subject, the average annual change across all voxels within the predefined ROI was computed and used to estimate the sample size needed to detect a treatment effect of known magnitude in a hypothetical clinical trial. Using these numeric summaries, we computed the number of subjects needed to detect 25% reduction in the mean annual rate of brain change with 80 or 90% confidence and a false positive probability of $\alpha = 0.05$ [Rosner, 1990]. We estimated the sample size required to achieve 80% and 90% power (subsequently we will refer to these as $n80$ and $n90$). These power estimates were generated to evaluate the effects of field strength (1.5 T versus 3 T) on estimated minimal sample sizes. The estimated minimum sample size for each arm was computed from the formula:
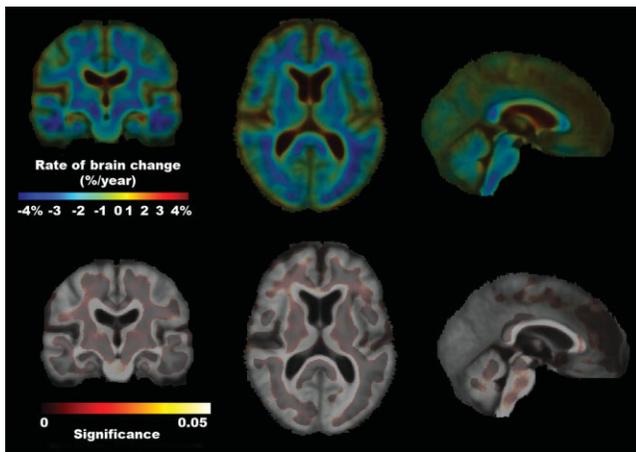
Figure 3.
AD ($n = 24$) versus controls ($n = 35$) scanned at 3 T. The top panels reflect the mean level of atrophy as a percentage reduction in volume over one year (%/year) beyond that found in controls (in other words, the mean control rate of atrophy has been subtracted out). Blue colors represent volume reduction while red colors represent volume expansion. Corresponding color-coded P maps show where atrophy in AD is significantly faster than in controls (bottom panels). The majority of the brain shows faster change in AD than controls. It is important to note that the regions of atrophy shown in this figure represent excess atrophy over and above age-related atrophy. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$n = \frac{2\hat{\sigma}_D^2 (z_{1-\alpha/2} + z_{\text{power}})^2}{(0.25\hat{\beta})^2}$$

Here z_α is the value of the standard normal distribution for which $P[Z < z_\alpha] = \alpha$ and in this case we set α to its conventional value of 0.05 [Rosner, 1990].

RESULTS

3D Maps of Brain Atrophy

Mean brain structural change maps were derived from averaging individual rate-of-atrophy (Jacobian) maps within each group (AD, MCI, and normal control), reflecting mean percent tissue loss over 1 year. Statistical maps were derived comparing AD with controls (Figs. 3 and 4) and MCI with controls (Figs. 5 and 6). Maps comparing patients with AD to normal controls show a widespread atrophic pattern, with faster ongoing atrophy in AD especially in the temporal lobe, and faster expansion of ventricular and CSF spaces in AD versus controls. Maps comparing MCI to normal controls reflect a much more restricted region with faster atrophic rates in MCI. Intriguingly, 3D maps comparing patients with AD and normal

controls showed a much more widespread pattern of significant atrophy when scanned at 3 T (Fig. 3) versus 1.5 T (Fig. 4), in the sense that the number of voxels passing the weak $P = 0.05$ voxel-level statistical threshold was greater. This may be due to the marginally higher spatial resolution and contrast of the 3 T scans. Both 3 and 1.5 T scans showed significant temporal lobe atrophy as expected. Permutation tests were conducted to determine the overall significance of the maps in Figures 3–6 (bottom panel), corrected for multiple comparisons. The estimated rates of atrophy were higher in the white matter than in the cortex (see Discussion).

Estimates of Minimal Sample Sizes

To determine whether 3 or 1.5 T MRI had greater power in detecting effects on temporal lobe volume loss over 1 year, we computed the sample size per arm needed to measure a 25% slowing of the atrophic rate with 80 and 90% power ($\alpha = 0.05$) (Fig. 7). For both the AD and MCI groups, 1.5 T MRI ($n80 = 37$ for AD, 108 for MCI) did not show a statistically different sample size estimate to detect temporal lobe atrophy when compared to 3 T ($n80 = 49$ for AD, 166 for MCI). To determine whether these sample size estimates were statistically different, we ran 10,000
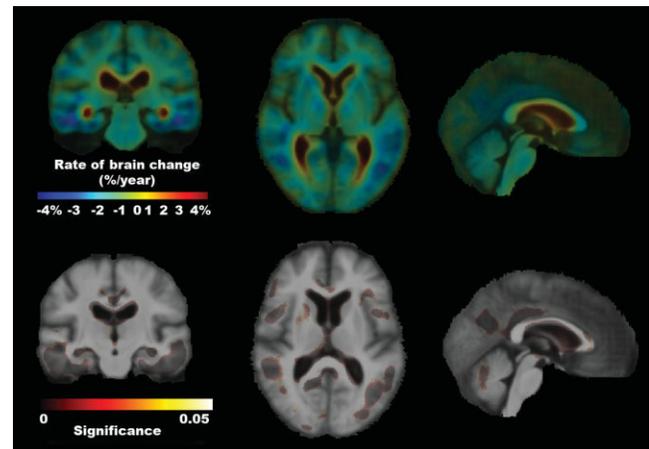

Figure 4.
AD ($n = 24$) versus controls ($n = 35$) scanned at 1.5 T. The top panels reflect the mean level of atrophy as a percentage reduction in volume over one year (%/year) by comparing AD versus controls. Again, the mean rate of atrophy in controls has been subtracted out. Blue colors represent progressive volume reductions; red colors represent volume expansions (e.g., in the lateral ventricles, which expand in AD). Corresponding color-coded P maps show the local significance of these differences indicated in the bottom panels. It is important to note that the regions of atrophy shown in this figure represent excess atrophy over and above age-related atrophy. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]
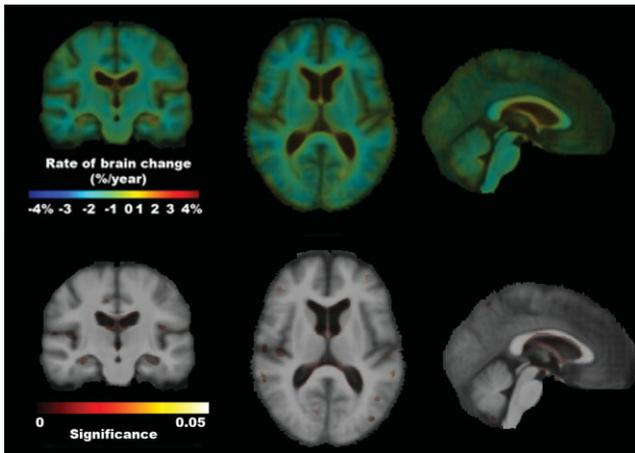
Figure 5.

MCI (*n* = 51) versus controls (*n* = 35) scanned at 3 T. The top panels reflect the mean level of atrophy as a percentage reduction in volume over one year (%/year) by comparing MCI versus controls. Again, the mean rate of atrophy in controls has been subtracted out. Blue colors represent progressive volume reductions; red colors represent volume expansions. Corresponding color-coded P maps show the local significance of these differences indicated in the bottom panels. It is important to note that the regions of atrophy shown in this figure represent excess atrophy over and above age-related atrophy. [Color figure can be viewed in the online issue, which is available at www.interscience. wiley.com.]
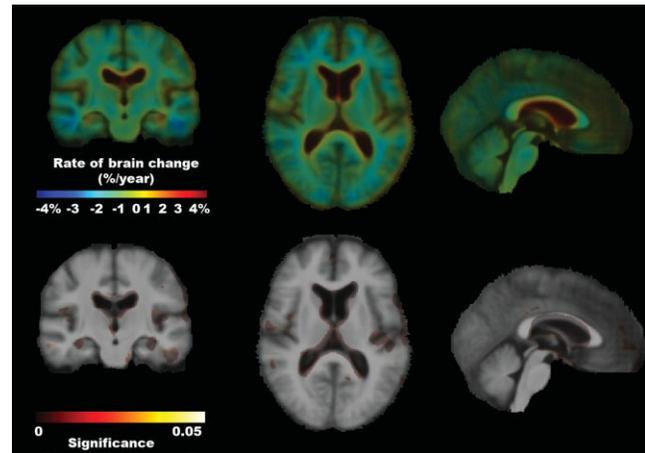

Figure 6.

MCI (*n* = 51) versus controls (*n* = 35) scanned at 1.5 T. The top panels reflect the mean level of atrophy as a percentage reduction in volume over one year (%/year) by comparing MCI versus controls. Blue colors represent volume reduction while red colors represent volume expansion. Corresponding color-coded P maps show the local significance of these differences indicated in the bottom panels. It is important to note that the regions of atrophy shown in this figure represent excess atrophy over and above age-related atrophy. [Color figure can be viewed in the online issue, which is available at www.interscience. wiley.com.]

permutations of a mixed sample of 50% 1.5 T scans and 50% 3 T from each diagnostic group (AD and MCI) to obtain a null distribution of sample size estimates based on the null hypothesis that the scanner type makes no difference (see histogram in Fig. 8 [top row]). Next, we ranked the 1.5 T power estimates for both AD and MCI and found that neither was in the outer 5% (i.e., $P < 0.05$) of the null distribution, showing that the 1.5 T power estimates were not significantly better.

To explore further whether the results regarding differences between scanners were dependent on the method used (TBM), we also used an independent method to compute a measure of the overall percentage brain volume change, and thus a second set of sample size estimates (Fig. 8, bottom row). We used Structural Image Evaluation, using Normalisation, of Atrophy (SIENA), an FSL program that estimates a two time-point percentage brain volume change [Smith et al., 2002, 2004]. SIENA estimates the percentage brain volume change (PBVC) between two input images from the same subject, by calling a series of FSL programs to strip the non-brain tissue from the two images, register the two brains (using the scalp as a constraint to hold the scaling constant during the registration) and estimates the brain change between the two time points. The estimated sample sizes were greater for the SIENA analysis (AD: $n80 = 116$ for 1.5 T and 92 for 3 T;
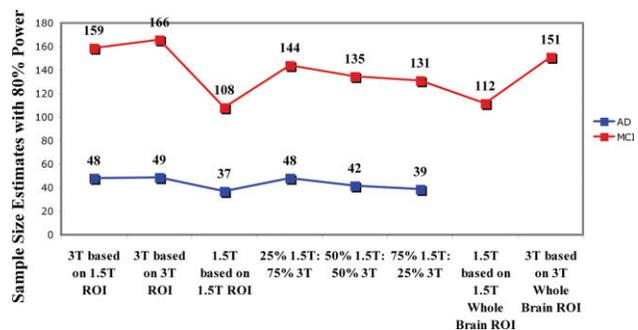

Figure 7.

Sample size estimates with 80% power when mixing 1.5 T and 3 T scans. The lowest sample size number (*n*80), indicating greatest power, is found in the 1.5 T group when using the statistically-defined ROI based on 1.5 T scans. The worst performance arises when analyzing the 3 T scans using the 1.5 T-specific statistical ROI. Power can be increased in the 3 T group if a statistically-defined ROI based on 3 T scans is used instead, perhaps because some voxels with artifacts are avoided. Mixing 1.5 and 3 T scans does not affect the *n*80 adversely, and sample sizes for mixed scanner designs lie in between those using 1.5 or 3 T exclusively. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]
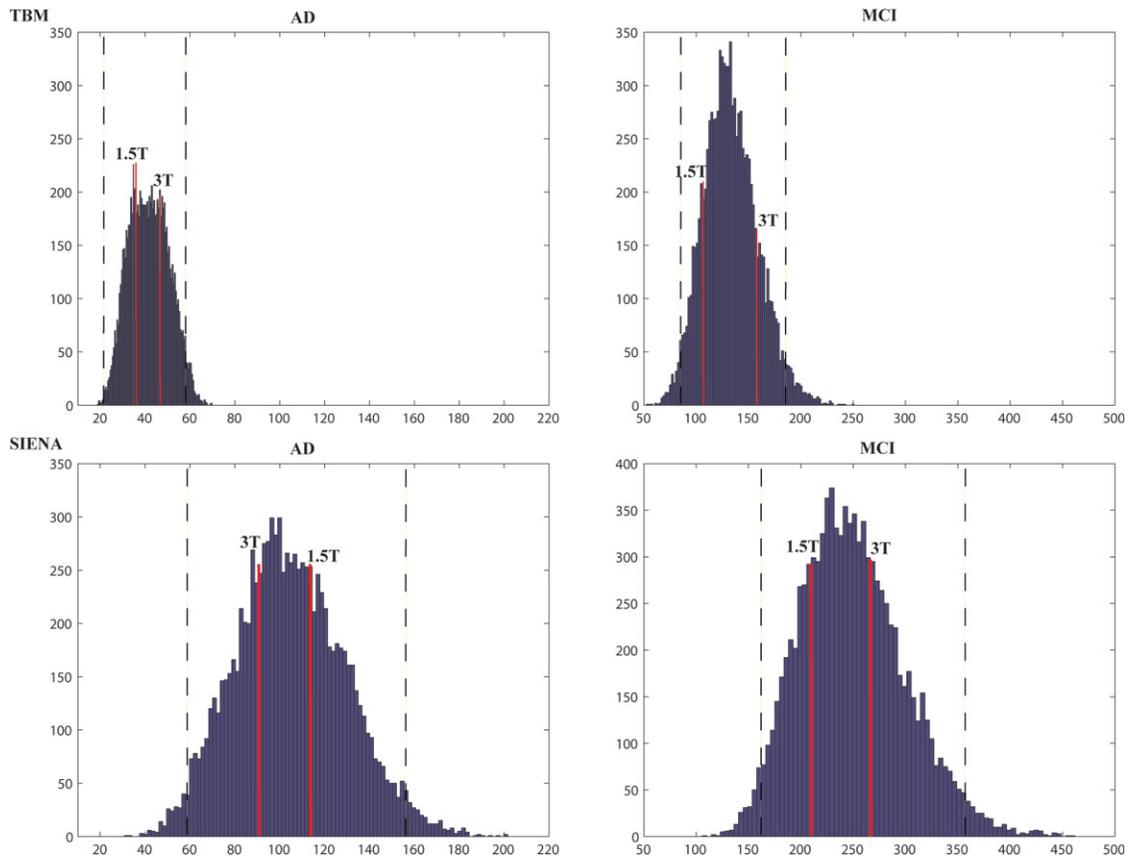
Figure 8.

Distribution of power estimates based on the null hypothesis that the scanner type makes no difference. Here we computed minimum sample size estimates based on 10,000 random permutations mixing 1.5 and 3 T scans for TBM analysis (top row) and for SIENA analysis (bottom row). For TBM, the "power estimates" on the x-axis represent the estimated sample sizes (AD: n80 = 37 for 1.5 T and 49 for 3 T; MCI: n80=107 for 1.5 T and 166 for 3 T) required to detect a 25% reduction in atrophy (with 80% power) for AD (left) and MCI (right) groups. For SIENA, the estimated sample sizes are numerically greater (AD: n80 = 116 for 1.5 T and 92 for 3 T; MCI: n80 = 207 for 1.5 T and 265 for 3 T). The red lines indicate the estimated sample sizes for designs that exclusively used 1.5 T scans or exclusively used 3 T; dashed lines indicate the boundary that excludes the outer 5% of the null distribution of sample sizes that occur from randomly mixing scanners. None of the red lines falls in the outer 5% of the null distribution, so we cannot reject the null hypothesis that the scanner field strength makes no difference. In other words, there is no detectable difference in power; power was excellent at both field strengths, and, as expected, was somewhat higher for AD than for MCI. For TBM analysis, the 3 T scanner gave a power estimate, for the MCI group, that was within the outer 11% of the power estimates obtained by randomly mixing scanners. The 3 T scanner gave a power estimate, for the AD group, that was within the outer 25% of the power estimates obtained by randomly mixing scanners. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

MCI: n80 = 207 for 1.5 T and 265 for 3 T) when compared to the power numbers computed using TBM (Table III). Even so, the pattern of results was entirely consistent between the two methods: in general, there was no evidence that one field strength gave better power than the other, for either analysis method. Power was somewhat higher for TBM than for SIENA, and it was higher for analyses of AD than for MCI; even so, power was relatively good for both methods.

In addition, we computed estimates of the minimal sample size for various study designs that allowed mixing of images from both 3 and 1.5 T scanners (Fig. 7). This corresponds to the practical situation of running a multisite clinical trial where not all sites can scan at the same field strength. For each combined group of scans (25% 3 T, 75% 1.5 T; 50% 3 T, 50% 1.5 T; 75% 3 T, 25% 1.5 T), 1.5 and 3 T scans were selected at random, while ensuring that the number of subjects from each diagnostic group (AD, MCI, and controls) remained consistent. The n80 numbers in Figure 7 reflect average values after repeated random permutations for each combination of scans; we bootstrapped these estimates to avoid any dependency on the particular

TABLE III. Sample size estimates (with 80% power) for TBM-derived measures versus SIENA-derived measures

TBM			SIENA		
1.5 T MRI n80	AD 37	MCI 107	1.5 T MRI n80	AD 116	MCI 207
3 T MRI n80	AD 48	MCI 159	3 T MRI n80	AD 92	MCI 265

Power was excellent at both field strengths and somewhat higher for AD than for MCI, as expected.

individuals assigned to each field strength. Regardless of these estimates, for practical reasons, a multisite study may be easier to design if scanners of two different field strengths can be accommodated, as some sites have only one scanner. Naturally, however, any given subject should be scanned exclusively at single field strength during the course of any longitudinal study.

One might hypothesize that mixing data from different field strengths would incur a severe loss of power relative to using only one field strength, but that was not the case. In Figure 7, the second and third columns show that the minimal sample sizes are numerically slightly larger at 3 T for MCI (n80: 166 for 3 T versus 107 for 1.5 T), but they are very similar for AD (49 at 3 T and 37 for 1.5 T). However, field strength had no detectable effect on these power estimates, in either the MCI or AD groups, because the n80s for 1.5 and 3 T scans only, did not fall in the outer 5% of the null distribution (see histogram in Fig. 8). Before any judgment is made as to whether these differences in estimated sample sizes are practically significant or not, it is worth noting that they are around six times lower (i.e., better) than the estimated sample sizes for the best clinical measures, CDR-SB for detecting change between AD and controls as well as MCI subjects and controls (highlighted in Table IV).

Such a sample size difference of around 58 MCI subjects for 3 versus 1.5 T might be regarded as somewhat trivial when past studies using ADAS-Cog or MMSE would require over a thousand subjects to detect the same percent slowing of disease progression in MCI. Second, the power for 3 T slightly worsened when summarizing atrophic rates using the statistical ROI derived at 1.5 T, relative to using the statistical ROI derived at 3 T. This is to be expected, as the main reason to develop a predefined ROI in an independent sample is to rule out voxels that are showing lower effect sizes. In a 3 T study where some temporal lobe distortions are expected, the 1.5 T ROI is slightly larger than the 3 T ROI, so by definition it is including voxels with lower effect sizes than would have been the case if the 3 T ROI were used. The next three data points in Figure 7 (columns 4–6) show power estimates for scans in various ratios, including 75% 3 T scans and 25% 1.5 T scans, equal numbers of scans at each field strength, and a 75:25 mix with 1.5 T scans outnumbering 3 T scans. Interestingly, power estimates were not substan-

tially worse—when using mixes of scanners—than they were when using one field strength exclusively; sample size requirements were intermediate between those achievable when using each field strength exclusively. There is no mathematical reason why mixing field strengths would be advantageous; even so, mixing scanners, which may be more practically feasible, does not result in a drastic depletion of power. Finally, the power does not increase when using a whole brain ROI, which may capture regions with ongoing atrophy that do not fall in the temporal lobe ROI. In other words, it is helpful to restrict the ROI based on both anatomic criteria (temporal lobe only) and statistical training (voxels with high effect sizes in independent training data). The last two columns in Figure 7 (columns 7 and 8) reflect power numbers that are no better than the ROI specific to the temporal lobe, with the 3 T group again having a worse power estimate than the 1.5 T group.

Correlations of Temporal Lobe Atrophy With Cognitive Decline

We assessed how these brain changes relate to measures of cognitive decline over the 1-year period, by correlating changes in cognitive scores (ADAS-cog, MMSE, and CDR-SB) with longitudinal rates of temporal lobe atrophy within the statistically-defined ROI at each field strength, after controlling for age and sex. (Note that in a real clinical trial, it may make more sense to use only one single ROI, but here, because we wanted to study field strength effects specifically, being fair to each field strength, we made separate ROIs for 3 and 1.5 T here to avoid biasing the results in favor of one field strength). Here, we used CDF plots to display the relative effect sizes for the associations between rates of temporal lobe atrophy and changes in ADAS-Cog, MMSE, and CDR-SB scores (Fig. 9). The clinical score that correlated the most strongly with higher rates of temporal lobe atrophy was 1-year a decrease in

TABLE IV. Comparison of sample size estimates (with 80% power) between TBM-derived measures and clinical measures

	AD	MCI
Loss rate %/year (1.5 T)	37	108
Loss rate %/year (3 T)	49	166
Change in CDR-SB	327	549
Change in ADAS-Cog	583	1183
Change in MMSE	3962	1531

These numbers are calculated based on the same sample of subjects. Both 1.5 and 3 T scans (using a TBM-derived loss rate in %/year) have a much lower n80 compared to the cognitive measures. These estimates may differ slightly from those in prior reports, as they depend to some degree on the specific cohort studied. Also, perhaps paradoxically, the change in MMSE has greater effect size in MCI than AD, so for that score, the sample size estimates are smaller for MCI than for AD.
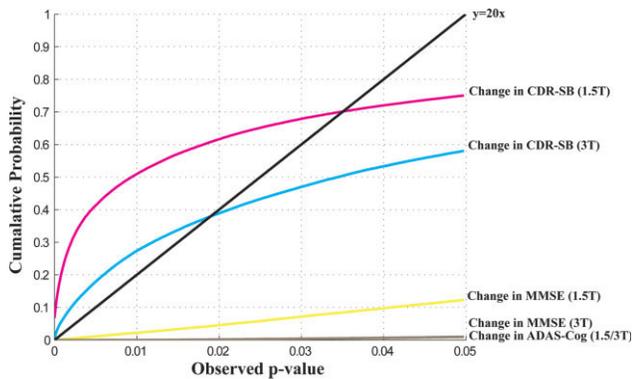
Figure 9.

CDF plots for voxel-wise correlations of temporal lobe atrophic rates across all subjects (n = 110, including AD, MCI, and controls) with 1-year interval changes in cognitive scores ADASCog, MMSE, and CDR-SB. Overall, scores were correlated with slightly higher effect sizes at 1.5 T compared to 3 T. CDR-SB was the clinical score that correlated most strongly with temporal lobe atrophy, at both 1.5 and 3 T. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

CDR-SB scores for both field strengths. As expected, CDR-SB decline correlated with marginally higher effect sizes in the 1.5 T MRI (critical value = 70%) when compared to 3 T MRI (critical value = 38%). In false discovery rate theory, the critical value is the highest fraction of the image that can be shown as significant while keeping the expected false discovery rate below 5%. Interval decline in MMSE and ADAS-Cog scores did not show significant associations with brain changes in this sample, when corrected for multiple comparisons with FDR, at either field strength. As we noted before, these correlations are undoubtedly detectable in a larger sample, but here our sample size was limited to 110 subjects as we wanted to include only subjects scanned on two different scanners. As in our prior study of 100 subjects scanned 1 year apart at 1.5 T [Leow et al., 2009], we also analyzed correlations between atrophic rates and CSF-derived measures of A-beta and Tau proteins, but these were not significant at either field strength. This is probably because only 60 of the 110 subjects in this had available data on CSF-derived measures of A-beta and Tau proteins; in our prior study we included more subjects with pathology measures as we did not require that subjects were also scanned on two different scanners. CSF measures of pathology may also represent trait rather than state markers and may not change much with disease progression. If that is the case, then there would not be a strong expectation that the rate of atrophy would correlate with CSF-derived measures of A-beta and Tau proteins.

DISCUSSION

In this article, we found that sample size estimates derived from TBM measures in both 3 and 1.5 T groups

were substantially better than all those based on cognitive or clinical measures MMSE, CDR-SB, and ADAS-Cog. The best functional measure for detecting MCI, in terms of requiring the smallest samples, was the CDR-SB, but this was still five times worse than TBM (549 versus 108 for TBM at 1.5 T; Table IV). In a sense, the CDR-SB is a functional measure rather than a cognitive score, (i.e. it is an informant-based assessment). Even so, the overall message would be that structural MRI imaging at any field strength can provide dramatically reduced sample sizes than even the best cognitive scores. Sample size estimates for detecting a 25% slowing of MCI were not statistically worse at 3 T versus 1.5 T ($n80$ = 166 at 3 T versus 108 at 1.5 T). Even so, the slightly higher sample size numbers to detect changes in the 3 T MCI group, also found by another group studying the same population [Alexander et al., personal communication], may be due to minor geometric distortions, residual intensity inhomogeneities, magnetic susceptibility effects, increased patient motion due to longer scan times and acoustic noise, and other artifacts that are generally harder to control at higher field strengths. Perhaps surprisingly, mixing scanners with different field strengths does not result in a drastic loss of power relative to using images collected at only one field strength, although power was marginally worse than using 1.5 T scanners only.

Several papers have investigated brain change on MRI over one year in the ADNI dataset [Hua et al., 2009; Leow et al., 2009; Misra et al., 2008; Morra et al., 2008; Nestor et al., 2008; Schuff et al., 2009]; to our knowledge however, our article is the only study to compare longitudinal data at 1.5 and 3 T MRI. Other groups have investigated field strength effects on the detection of signal abnormalities [Di Perri et al., 2009], reliability of imaging measures [Jovicich et al., 2006], measurement of image-derived parameters [Lu et al., 2005], and diagnostic benefits [Frayne et al., 2003], primarily focusing on 1.5 versus 3 T scanning.

This study further confirms past independent reports that neuroimaging measures require a drastically lower sample size than cognitive measures to detect neurodegenerative changes [Fox et al., 2000; Jack et al., 2004; Schuff et al., 2009]. Volumes of the hippocampus and entorhinal cortex are effective neuroimaging markers compared to cognitive scores, with sample size estimates about 10 times lower [Jack et al., 2004]. TBM measures based on the temporal lobe [Hua et al., 2008a,b] derived here from an empirically-defined statistical ROI (see Hua et al., 2009 for details) have similar advantages over cognitive scores. In this article, the smallest sample sizes were required for the 1.5 T scans ($n80$: 37 for AD, 108 for MCI) using a 1.5 T-specific statistically-defined ROI. The 3 T power estimates ($n80$: 49 for AD, 166 for MCI), based on a 3 T-specific statistically-defined ROI, were slightly poorer, but not statistically different.

Even though the sample sizes needed to detect a fixed percent reduction in the rate of progression are lower for TBM than for clinical scores, we must bear in mind that a

given effect size on a clinical scale may have very different consequences for the patient than an effect of the same magnitude on an MRI scale. In other words, power comparisons between imaging and clinical measures should be performed cautiously, as a certain percent reduction in the rate of progression may have very different meanings for clinical scores versus MRI. MRI measures, in particular, may include some regional changes that do not have a direct bearing on the cognition or well-being of the patient. A change with a certain fixed effect size on a clinical scale may be of more importance than a comparable reduction in the atrophic rate.

Our sample size estimates are based on assuming a 25% slowing of the rate of atrophy. In reality, treatments may slow atrophy to different degrees. Even so, the sample size estimates required to detect a k% slowing of atrophy can be easily derived by multiplying the numbers in this paper by $(25/k)^2$. To see this, we note that the estimated minimum sample size for each arm is computed from the formula:

$$n = \frac{2\hat{\sigma}_D^2(z_{1-\alpha/2} + z_{power})^2}{(0.25\hat{\beta})^2}$$

where z_a is the value of the standard normal distribution for which $P[Z < z_a] = \alpha$, and in this case we set α to its conventional value of 0.05 [Rosner, 1990]. The number 0.25 appears in this formula as a multiplier on the effect size, beta, and represents an assumption of a 25% slowing of atrophy. Assuming, more generally, that there is a k% slowing of the atrophic rate, the required sample size to detect it, n, is proportional to $1/k^2$. This inverse-square law means that a 10% slowing of atrophy would need four times as many subjects to detect as a 20% slowing of atrophy, and a 5% slowing of atrophy would need 16 times as many subjects to detect as a 20% slowing of atrophy. This quadratic dependency is illustrated in Figure 10. The effect on the histograms of assuming a k% slowing of atrophy, rather than a 25% slowing of atrophy, would be to stretch the histograms horizontally by a factor of $(25/k)^2$.

The effect of assuming any other fixed percentage slowing of atrophy can therefore be computed by multiplying all the numbers in this article by a fixed number. Consequently, it would make no difference to the findings reported here, if we assumed a treatment could slow atrophy by a different proportion. The significance of all the statistical tests would be unaffected, as multiplying all the variables by a fixed constant does not alter any effect sizes in the statistical tests.

Here we based our sample size calculations on a statistical test that would have known power (80%) to detect a certain percent slowing (25% slowing) of the rate of atrophy. This definition has been adopted in other studies—one study used 25% and 50% slowing of the average rate of change with 80% and 90% power [Jack et al., 2004], and another study used a 25% slowing with 90% power [Schuff et al., 2009]. One could also consider an alternative sample size definition based on how many subjects would be
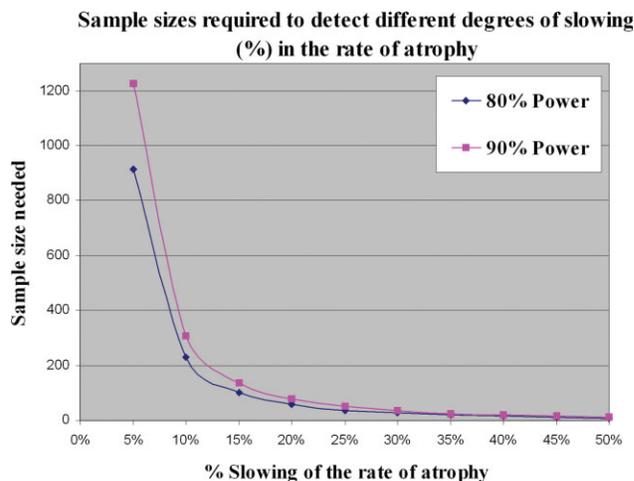

Figure 10.
Sample sizes required to detect different degrees of slowing in the rate of atrophy are shown to have a quadratic dependency. Our sample size estimates are based on assuming a 25% slowing of the rate of atrophy, whereas in reality, treatments may slow atrophy to different degrees. Even so, the sample size estimates required to detect a k% slowing of atrophy can be easily derived by multiplying the numbers in this article by $(25/k)^2$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

needed to detect a 25% reduction in brain volume over an interval, with a specific level of power (e.g., 80% or 90%). One issue with aiming to detect a certain % reduction in brain volume is that the loss of volume is not uniform across the brain, so an analysis method focusing on a small number of voxels with high effect sizes would appear to have a very high power, even if the treatment effects on other regions of the brain were also of interest. A more common question for treatment trials asks how rapid the atrophic rate truly is in disease, and then considers the situation where treatment slows the atrophy by some fixed percentage.

Even so, defining power based on % slowing of atrophy has some acknowledged limitations. First, it does not take into account the rate of atrophy, or its variance, in a comparison group of healthy normal subjects. This is because most placebo-controlled treatment trials do not evaluate normal subjects, but only assess people with the disease or those at increased risk (e.g., MCI subjects) who are randomized to different treatments. Second, if some proportion of the atrophy in disease also occurs in normals, then it may be unrealistic to expect treatments to reverse that part of the atrophy, although that is implicit in basing power computations on the atrophic rates in one group only. Even so, one advantage of the definition used here is that it can be readily applied to any longitudinal assessments that give numeric summaries, and can then be used to compare analysis methods head-to-head.

To calculate power estimates and compute the CDF plots, we used an empirically derived predefined statistical ROI, a method recently advocated by Reiman and Chen for PET analysis [Reiman et al., 2008; Chen et al., 2009]; we adapted this for MRI analysis in [Hua et al., 2009]. The statistically-defined ROI is based on an independent training sample and improves power by concentrating on changes typically observed in patients with AD. The statistical ROI is also adaptive to the data. Using it assumes that a potential treatment for AD would slow rates in the same regions as those where atrophy has the highest effect size, which is plausible, but is not automatically the case (this may depend on the treatment). Additionally, one could argue that the statistical ROI is not easy to specify as an outcome measure independent of the dataset. In clinical trials regulated by the FDA, outcomes must be specified before the trial begins. Our prior study of 515 subjects at 1.5 T [Hua et al., 2009] found that the statistical threshold used to define the ROI does not greatly affect sample sizes estimates, thresholds of p = 0.001, 0.0001, and 0.00001 gave sample size estimates of 48, 50, and 52 subjects for AD, and 88, 91, and 95 subjects, respectively, for MCI. This relative insensitivity to the threshold means that the lower values at 3 T are unlikely to be due to suboptimal selection of the threshold, or due to inherent biases in the way the ROIs are generated. Paradoxically, a more sensitive method might pick up additional voxels in the ROI, that when averaged into the ROI could artificially reduce the SNR. Future work will focus on improving the way in which the statistical ROI is applied to the data (e.g., weighting data from different voxels according to their effect sizes, or using a machine learning principle such as adaptive boosting; see, e.g., Morra et al., 2009].

One notable aspect of the topography of brain matter loss in Figures 3–6 is that the greatest proportion of brain matter loss appears to lie in the white matter rather than the cortical surface. This is mainly because (1) the registration fields in TBM are spatially smooth and partial volume averaging effects diminish the signal somewhat at tissue boundaries, such as the cortex/CSF interface, and (2) the registration accuracy of TBM is poorer at the cortical surface, at least relative to some approaches that explicitly model the cortical surface. As noted in prior work [Hua et al., 2008a,b; Leow et al., 2009], to better sensitize the TBM approach for detecting cortical gray matter loss, several approaches have been considered: (1) using voxel-based morphometry (VBM; Ashburner and Friston, 2000] or a related approach termed RAVENS [Davatzikos et al., 2001], (2) adaptively smoothing deformation-based compression signals at each point based on the amount of gray matter lying under the filter kernel [Studholme et al., 2003], or (3) running deformation maps at a very high-spatial resolution and with less spatial regularization or with a regularization term that enforces continuity but not smoothness [Leow et al., 2009].

Although we found no statistical difference in power between the 1.5 and 3 T groups, there are several issues associated with higher field strengths to consider. Our analysis concentrated on changes observed in AD and MCI within the temporal lobe. In this region, susceptibility-induced geometric distortion and signal losses may increase noise for derived parameter estimates. These effects are less easy to control at higher field strength. In addition, other minor disadvantages associated with higher field strength images include chemical shift artifacts, adjustments of pulse sequence parameters to account for changes in relaxation and susceptibility, and the cost of installation, which may be higher at 3 T [Frayne et al., 2003]. At higher field, there are also safety issues due to the higher radio-frequency specific absorption rate (SAR), especially for RF-intensive sequences, but 3D T1-weighted sequences such as MP-RAGE have relatively low power deposition and are not limited by SAR considerations at 3 T. When the ADNI MRI protocol was designed, some of the increased SNR at 3 T was traded off for reductions in chemical shift and susceptibility artifacts by increasing the read-out bandwidth at 3 T versus 1.5 T. Conversely, 3 T MRI offers many benefits (i.e., increased SNR) for functional imaging, diffusion studies, and white matter lesion detection [Di Perri et al., 2009].

Although this study examined morphometric features measurable at 1.5 and 3 T, very high field strength studies may reveal still finer-scale features not observable at lower field, including hippocampal subfields that may be relevant to tracking AD or MCI (see Augustinack et al., 2005, for detection of entorhinal layer II with 7 T MRI). Van Leemput [2009] used 3 T scans with a 0.38 mm in-plane resolution to segment hippocampal subfields, and Mueller and Weiner [2009] used 4 T scanning to assess effects of age and genotype on hippocampal subfields. The increased contrast at higher field is likely to assist future morphometric studies, especially when scans are collected with more RF receiver channels and parallel imaging to reduce scan time, which in turn reduces potential motion artifact. Although the 3 T acquisition in the ADNI protocol developed in 2004 was over a minute longer than at 1.5 T, with today's technology the 3 T acquisitions are typically 2–3 min shorter.

One of the more surprising outcomes of this study was that mixing data from different field strength scanners did not cause a drastic loss of power compared to acquiring data at a single field strength. This implies that field strength induces relatively little bias and/or variation compared to other sources such as variations between subjects and between MRI sites. It needs to be seen, however, if this still holds for MRI scans with more than two serial observations per subject. Whether scanners can be mixed depends on the quality control procedures (including phantom-based calibration scans), the tendency for each participating site to allow drifts in spatial calibration over time, and the adequacy of subsequent image corrections. Sample size estimates may also be lower for a study conducted at a single site. In a recent study evaluating the impact of image acquisition variables, combining data

across platforms (i.e., vendors) and across field-strength caused small volume difference biases, depending on the brain structure and MRI vendor/field strength combination [Jovicich et al., 2006]. In a multisite study with different field strengths and vendors, such as ADNI, these confounds are important to evaluate. Even so, in this multi-site study (which performed 3 T scanning at 31 different locations), mixing 3 and 1.5 T scans did not greatly reduce power.

In summary, both 1.5 and 3 T MRI required a dramatically smaller sample size to detect changes in AD and MCI groups when compared to the sample sizes needed for the standard functional measures, ADAS-Cog, MMSE, or CDR-SB. Different MRI field strengths did not affect the power to detect 25% slowing of atrophy (with 80% power) and mixing 1.5 and 3 T scans did not greatly reduce power and is likely to be acceptable for future clinical studies. Currently, most MRI studies are conducted at 1.5 T; however, with more studies using higher field strength scanners, the next generation of 3 T scanners may become the gold standard for research and clinical studies.

REFERENCES

Ashburner J, Friston KJ (2003): Morphometry. Human Brain Function. San Diego, CA: Academic Press.

Augustinack JC, van der Kouwe AJ, Blackwell ML, Salat DH, Wiggins CJ, Frosch MP, Wiggins GC, Potthast A, Wald LL, Fischl BR (2005): Detection of entorhinal layer II using 7 Tesla [corrected] magnetic resonance imaging. Ann Neurol 57:489–494. Erratum in: Ann Neurol 2005;58:172.

Benjamini Y, Hochberg Y (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Statist Soc B 57:289–300.

Berg L (1988): Clinical dementia rating (CDR). Psychopharmacol Bull 24:637–639.

Braskie MN, Klunder AD, Hayashi KM, Protas H, Kepe V, Miller KJ, Huang SC, Barrio JR, Ercoli LM, Siddarth P, Satyamurthy N, Liu J, Toga AW, Bookheimer SY, Small GW, Thompson PM (2008): Plaque and tangle imaging and cognition in normal aging and Alzheimer's disease. Neurobiol Aging, November 10 [Epub ahead of print].

Bernstein MA, Huston J III, Ward HA (2006): Imaging artifacts at 3.0T. J Magn Reson Imaging 24:735–746.

Bullmore ET Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer MJ (1999): Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans Med Imaging 18:32–42.

Chen K, Reschke C, Lee W, Napatkamon A, Liu X, Bandy D, Langbaum J, Alexander GE, Foster NL, Koeppe RA, Jagust WJ, Weiner MW, Reiman EM (2009): Cross-sectional and longitudinal analyses of fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative. In ADNI Data Presentations Meeting, Seattle, WA.

Chiang MC, Dutton RA, Hayashi KM, Lopez OL, Aizenstein HJ, Toga AW, Becker JT, Thompson PM (2007): 3D pattern of brain atrophy in HIV/AIDS visualized using tensor-based morphometry. Neuroimage 34:44–60.

Chou YY, Lepore N, de Zubicaray GI, Carmichael OT, Becker JT, Toga AW, Thompson PM (2008): Automated ventricular mapping with multi-atlas fluid image alignment reveals genetic effects in Alzheimer's disease. Neuroimage 40:615–630.

Chou YY, Lepore N, Chiang MC, Avedissian C, Barysheva M, McMahon KL, de Zubicaray GI, Meredith M, Wright MJ, Toga AW, Thompson PM (2009): Mapping genetic influences on ventricular structure in twins. Neuroimage 44:1312–1323.

Cockrell JR, Folstein MF (1988): Mini-Mental State Examination (MMSE). Psychopharmacol Bull 24:689–692.

Collins CM, Liu W, Schreiber W, Yang QX, Smith MB (2005): Central brightening due to constructive interference with, without, and despite dielectric resonance. J Magn Reson Imaging 21:192–196.

Collins DL, Neelin P, Peters TM, Evans AC (1994): Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J Comput Assist Tomogr 18(2):192–205.

Davatzikos C, Genc A, Xu D, Resnick SM (2001): Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy. Neuroimage 14:1361–1369.

Di Perri C, Dwyer MG, Wack DS, Cox JL, Hashmi K, Saluste E, Hussein S, Schirda C, Stosic M, Durfee J, Poloni GU, Nayyar N, Bergamaschi R, Zivadinov R (2009): Signal abnormalities on 1.5 and 3 Tesla brain MRI in multiple sclerosis patients and healthy controls. A morphological and spatial quantitative comparison study. Neuroimage 47:1352–1362.

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.

Fox NC, Cousens S, Scahill R, Harvey RJ, Rossor MN (2000): Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: Power calculations and estimates of sample size to detect treatment effects. Arch Neurol 57:339–344.

Fox NC, Crum WR, Scahill RI, Stevens JM, Janssen JC, Rossor MN (2001): Imaging of onset and progression of Alzheimer's disease with voxel-compression mapping of serial magnetic resonance images. Lancet 358:201–205.

Frayne R, Goodyear BG, Dickhoff P, Lauzon ML, Sevick RJ (2003): Magnetic resonance imaging at 3.0 Tesla: Challenges and advantages in clinical neurological imaging. Invest Radiol 38:385–402.

Genovese CR, Lazar NA, Nichols T (2002): Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15:870–878.

Hua X, Leow AD, Lee S, Klunder AD, Toga AW, Lepore N, Chou YY, Brun C, Chiang MC, Barysheva M, Jack CR Jr, Bernstein MA, Britson PJ, Ward CP, Whitwell JL, Borowski B, Fleisher AS, Fox NC, Boyes RG, Barnes J, Harvey D, Kornak J, Schuff N, Boreta L, Alexander GE, Weiner MW, Thompson PM; and the Alzheimer's Disease Neuroimaging Initiative (2008a): 3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry. Neuroimage 41:19–34.

Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, Jack CR Jr, Weiner MW, Thompson PM (2008b): Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: An MRI study of 676 AD, MCI, and normal subjects. Neuroimage 43:458–469.

Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL (1982): A new clinical scale for the staging of dementia. Br J Psychiatry 140:566–572.

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.

Jack CR Jr, Shiung MM, Gunter JL, O'Brien PC, Weigand SD, Knopman DS, Boeve BF, Ivnik RJ, Smith GE, Cha RH, Tangalos EG, Petersen RC (2004): Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology 62:591–600.

Jack CR Jr, Bernstein MA, Fox NC, Thompson PM, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell J, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DL, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW (2008a): The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 27:685–691.

Jack CR Jr, Petersen RC, Grundman M, Jin S, Gamst A, Ward CP, Sencakova D, Doody RS, Thal LJ (2008b): Longitudinal MRI findings from the vitamin E and donepezil treatment study for MCI. Neurobiol Aging 29:1285–1295.

Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, Kennedy D, Schmitt F, Brown G, Macfall J, Fischl B, Dale A (2006): Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. Neuroimage 30:436–443.

Kochunov P, Lancaster JL, Thompson P, Woods R, Mazziotta J, Hardies J, Fox P (2001): Regional spatial normalization: Toward an optimal target. J Comput Assist Tomogr 25:805–816.

Kochunov P, Lancaster J, Thompson P, Toga AW, Brewer P, Hardies J, Fox P (2002): An optimized individual target brain in the Talairach coordinate system. Neuroimage 17:922–927.

Kochunov P, Lancaster J, Hardies J, Thompson PM, Woods RP, Cody JD, Hale DE, Laird A, Fox PT (2005): Mapping structural differences of the corpus callosum in individuals with 18q deletions using targetless regional spatial normalization. Hum Brain Mapp 24:325–331.

Leow A, Huang SC, Geng A, Becker JT, Davis S, Toga AW, Thompson PM (2005): Inverse consistent mapping in 3D deformable image registration: Its construction and statistical properties. In Information Processing in Medical Imaging, Glenwood Springs, Colorado, USA, pp 493–503.

Leow AD, Klunder AD, Jack CR Jr, Toga AW, Dale AM, Bernstein MA, Britson PJ, Gunter JL, Ward CP, Whitwell JL, Borowski BJ, Fleisher AS, Fox NC, Harvey D, Kornak J, Schuff N, Studholme C, Alexander GE, Weiner MW, Thompson PM (2006): Longitudinal stability of MRI for mapping brain change using tensor-based morphometry. Neuroimage 31:627–640.

Leow AD, Yanovsky I, Parikshak N, Hua X, Lee S, Toga AW, Jack CR Jr, Bernstein MA, Britson PJ, Gunter JL, Ward CP, Borowski B, Shaw LM, Trojanowski JQ, Fleisher AS, Harvey D, Kornak J, Schuff N, Alexander GE, Weiner MW, Thompson PM (2009): Alzheimer's disease neuroimaging initiative: A one-year follow up study using tensor-based morphometry correlating degenerative rates, biomarkers and cognition. Neuroimage 45:645–655.

Lepore, N., Brun C, Chou YY, Chiang MC, Dutton RA, Hayashi KM, Luders E, Lopez OL, Aizenstein HJ, Toga AW, Becker JT, Thompson PM (2008): Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors. IEEE Trans Med Imaging 27:129–141.

Lu H, Nagae-Poetscher LM, Golay X, Lin D, Pomper M, van Zijl PC (2005): Routine clinical brain MRI sequences for use at 3.0 Tesla. J Magn Reson Imaging 22(1):13–22.

Marsden J, Hughes T (1983). Mathematical Foundations of Elasticity: Prentice-Hall.

Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L, Thompson P, MacDonald D, Iacoboni M, Schormann T, Amunts K, Palomero-Gallagher N, Geyer S, Parsons L, Narr K, Kabani N, Le Goualher G, Boomsma D, Cannon T, Kawashima R, Mazoyer B (2001): A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos Trans R Soc Lond B Biol Sci 356:1293–1322.

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984): Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34:939–944.

Misra C, Fan Y, Davatzikos C (2009): Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. Neuroimage 44:1415–1422.

Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, Parikshak N, Hua X, Toga AW, Jack CR Jr, Weiner MW, Thompson PM (2008): Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. Neuroimage 43:59–68.

Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, Parikshak N, Toga AW, Jack CR Jr, Schuff N, Weiner MW, Thompson PM (2009): Automated mapping of hippocampal atrophy in 1-year repeat MRI data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Neuroimage 45:S3–S15.

Morris JC (1993): The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology 43:2412–2414.

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005a): The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am 15:869–877,xi–xii.

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005b): Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement 1:55–66.

Mueller SG, Weiner MW (2009): Selective effect of age, Apo e4, and Alzheimer's disease on hippocampal subfields. Hippocampus 19:558–564.

Nestor SM, Rupsingh R, Borrie M, Smith M, Accomazzi V, Wells JL, Fogarty J, Bartha R (2008): Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. Brain 131(Pt 9):2443–2454.

Nichols TE, Holmes AP (2002): Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Petersen RC (2000): Aging, mild cognitive impairment, and Alzheimer's disease. Neurol Clin 18:789–806.

Petersen RC, Negash S (2008): Mild cognitive impairment: An overview. CNS Spectr 13:45–53.

Petersen RC, Smith GE, Ivnik RJ, Kokmen E, Tangalos EG (1994): Memory function in very early Alzheimer's disease. Neurology 44:867–872.

Petersen RC, Doody R, Kurz A, Mohs RC, Morris JC, Rabins PV, Ritchie K, Rossor M, Thal L, Winblad B (2001): Current concepts in mild cognitive impairment. Arch Neurol 58:1985–1992.

Reiman EM, Langbaum JBS (2009): Brain imaging in the evaluation of putative Alzheimer's disease slowing, risk-reducing and prevention therapies. In: Jagust W, D'Esposito M, editors. Imaging and the Aging Brain. New York: Oxford University Press. pp 1–16

Reiman EM, Chen K, Ayutyanont N, Bandy D, Lee W, Reschke C, Alexander GE, Weiner MW, Koeppe RA, Jagust WJ, (2008): Twelve-month cerebral metabolic declines in probable Alzheimer's disease and amnestic mild cognitive impairment: Preliminary findings from the Alzheimer's disease neuroimaging initiative. Chicago, Illinois: ICAD.

Rosen WG, Mohs RC, Davis KL (1984): A new rating scale for Alzheimer's disease. Am J Psychiatry 141:1356–1364.

Rosner B (1990): Fundamentals of Biostatistics. Boston: PWS-Kent Publishing Company.

Scheltens P, Fox N, Barkhof F, De Carli C (2002): Structural magnetic resonance imaging in the practical assessment of dementia: Beyond exclusion. Lancet Neurol 1:13–21.

Schenck JF (1996): The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds. Med Phys 23:815–850.

Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, Thompson PM, Jack CR Jr, Weiner MW (2009): MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 132:1067–1077.

Sled JG, Zijdenbos AP, Evans AC (1998): A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17:87–97.

Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, De Stefano N (2002): Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. Neuroimage 17:479–489.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (Suppl 1):S208–S219.

Storey JD (2002): A direct approach to false discovery rates. J R Stat Soc B 64 (Part 3):479–498.

Studholme C, Cardenas V, Schuff N, Rosen H, Miller B, Weiner M (2001): Detecting spatially consistent structural differences in Alzheimer's and fronto temporal dementia using deformation morphometry. MICCAI 2208:41–48.

Studholme C, Cardenas V, Maudsley A, Weiner M (2003): An intensity consistent filtering approach to the analysis of deformation tensor derived maps of brain shape. NeuroImage 19:1638–1649.

Thompson PM, Giedd JN, Woods RP, MacDonald D, Evans AC, Toga AW (2000): Growth patterns in the developing brain detected by using continuum mechanical tensor maps. Nature 404:190–193.

Thompson PM, Hayashi KM, de Zubicaray G, Janke AL, Rose SE, Semple J, Herman D, Hong MS, Dittmer SS, Doddrell DM, Toga AW (2003): Dynamics of gray matter loss in Alzheimer's disease. J Neurosci 23:994–1005.

Van Leemput K, Bakkour A, Benner T, Wiggins G, Wald LL, Augustinack J, Dickerson BC, Golland P, Fischl B (2009): Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus 19:549–557.

Wimo A, Jonsson L, Winblad B (2006): An estimate of the worldwide prevalence and direct costs of dementia in 2003: Dement Geriatr Cogn Disord 21:175–181.

Yanovsky I, Thompson P, Osher S, Leow A (2007): Topology releasing log-unbiased nonlinear image registration: theory and implementation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Yanovsky I, Thompson P, Osher S, Leow A (2008): Asymmetric and symmetric unbiased image registration: statistical assessment of performance. IEEE Computer Society. Workshop on Mathematical Methods in Biomedical Image Analysis.