Published in final edited form as:
Neuroimage. 2013 February 15; 66: 648–661. doi:10.1016/j.neuroimage.2012.10.086.

Unbiased tensor-based morphometry: Improved robustness and sample size estimates for Alzheimer's disease clinical trials

Xue Hua[a], **Derrek P. Hibar**[a], **Christopher R.K. Ching**[a], **Christina P. Boyle**[a], **Priya Rajagopalan**[a], **Boris A. Gutman**[a], **Alex D. Leow**[b,c], **Arthur W. Toga**[a], **Clifford R. Jack Jr.**[d], **Danielle Harvey**[e], **Michael W. Weiner**[f,g,h], **Paul M. Thompson**[a,i,*], and **the Alzheimer's Disease Neuroimaging Initiative**[1]

[a]Imaging Genetics Center, Laboratory of Neuro Imaging, Dept. of Neurology, UCLA School of Medicine, Los Angeles, CA, USA

[b]Dept. of Psychiatry, University of Illinois at Chicago, College of Medicine, Chicago, IL, USA

[c]Dept. of Bioengineering, University of Illinois at Chicago, Chicago, IL, USA

[d]Mayo Clinic, Rochester, MN, USA

[e]Dept. of Public Health Sciences, UC Davis School of Medicine, Davis, CA, USA

[f]Dept. of Radiology and Biomedical Imaging, UC San Francisco, San Francisco, CA, USA

[g]Dept. of Medicine, UC San Francisco, San Francisco, CA, USA

[h]Dept. of Psychiatry, UC San Francisco, San Francisco, CA, USA

[i]Dept. of Psychiatry, Semel Institute, UCLA School of Medicine, Los Angeles, CA, USA

Abstract

Various neuroimaging measures are being evaluated for tracking Alzheimer's disease (AD) progression in therapeutic trials, including measures of structural brain change based on repeated scanning of patients with magnetic resonance imaging (MRI). Methods to compute brain change must be robust to scan quality. Biases may arise if any scans are thrown out, as this can lead to the true changes being overestimated or underestimated. Here we analyzed the full MRI dataset from the first phase of Alzheimer's Disease Neuroimaging Initiative (ADNI-1) from the first phase of Alzheimer's Disease Neuroimaging Initiative (ADNI-1) and assessed several sources of bias that can arise when tracking brain changes with structural brain imaging methods, as part of a pipeline for tensor-based morphometry (TBM). In all healthy subjects who completed MRI scanning at screening, 6, 12, and 24 months, brain atrophy was essentially linear with no detectable bias in longitudinal measures. In power analyses for clinical trials based on these change measures, only 39 AD patients and 95 mild cognitive impairment (MCI) subjects were needed for a 24-month trial to detect a 25% reduction in the average rate of change using a two-sided test (α=0.05, *power*=80%). Further sample size reductions were achieved by stratifying the data into Apolipoprotein E (ApoE) ε4 carriers versus non-carriers. We show how selective data exclusion affects sample size estimates, motivating an objective comparison of different analysis techniques

*Corresponding author at: Imaging Genetics Center, Laboratory of Neuro Imaging, Dept. of Neurology, UCLA School of Medicine, Neuroscience Research Building 225E, 635 Charles Young Drive, Los Angeles, CA 90095-1769, USA. Fax: +1 310 206 5518. thompson@loni.ucla.edu. .

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but only some participated in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

based on statistical power and robustness. TBM is an unbiased, robust, high-throughput imaging surrogate marker for large, multi-site neuroimaging studies and clinical trials of AD and MCI.

Keywords

Alzheimer's disease; Mild cognitive impairment; Aging; ADNI; Tensor-based morphometry; Drug trial

Introduction

Alzheimer's disease (AD) affects 5.4 million people in the U.S. alone, and over 24 million people worldwide (Ferri et al., 2005). New treatments to slow or delay Alzheimer's disease progression must be rapidly and efficiently evaluated to alleviate a growing public health crisis. A related condition is mild cognitive impairment (MCI); people with MCI are at greatly increased risk of developing AD. As many as 10–25% of MCI subjects progress to probable AD per year (Petersen, 2000, 2003a, 2003b). Numerous therapeutic trials are underway to test novel compounds. Some of these trials use neuroimaging measures to assess treatment effects on brain measures, such as amyloid levels in the brain or rates of atrophy (Petersen, 2003a; Ross et al., 2012).

A wide variety of neuroimaging measures may be useful in tracking the progression of AD and MCI. The Alzheimer's Disease Neuroimaging Initiative (ADNI) was set up as one of several multi-center studies worldwide to develop and validate novel biomarkers to characterize, detect and track AD (Frisoni and Weiner, 2010; Mueller et al., 2005a, 2005b; Trojanowski et al., 2010; Weiner et al., 2010, 2012). In the first phase of ADNI (ADNI-1), 817 subjects received screening scans, including 188 early Alzheimer's patients, 400 subjects with MCI, and 229 healthy controls, who were studied at 6- or 12-month intervals for up to 36 months (Wyman et al., 2012). The entire dataset is publicly available (http://adni.loni.ucla.edu), offering a large test dataset to develop, validate, and compare biomarkers for disease classification and prognosis. A summary of approximately 200 published ADNI papers is provided in a recent review (Weiner et al., 2012).

High-resolution structural MRI is one of several imaging methods used to track AD, and numerous MRI-derived biomarkers have been thoroughly investigated, including but not limited to: (1) hippocampal volume (Jack et al., 1999, 2002; Morra et al., 2009a, 2009b; Schuff et al., 2009), (2) lateral ventricular volumes (Carmichael et al., 2006; Chou et al., 2008, 2009; Thompson et al., 2004), (3) gray matter volume or density, as measured using voxel-based morphometry (VBM) in the statistical parametric mapping (SPM) software package (Ashburner and Friston, 2000; Baron et al., 2001; Chetelat et al., 2005), (4) a measure of brain change over time known as the brain boundary shift integral (BBSI) (Fox et al., 2000; Freeborough and Fox, 1997), (5) automated methods for computing a variety of regional subvolumes, such as longitudinal FreeSurfer (Reuter et al., 2012), and the FMRIB Software Library (FSL) (Smith et al., 2007), (6) data-driven measures of temporal lobe atrophy using tensor-based morphometry (TBM) (Hua et al., 2009, 2010), and (7) measures of volume change in the entorhinal cortex, hippocampus, and whole brain using the commercial software known as quantitative anatomical regional change (Quarc) (Holland and Dale, 2011; Holland et al., 2009). Some of these methods also derive statistical *maps* of brain changes over time, as well as numeric summaries of atrophy from anatomically and statistically defined regions of interest. Earlier research applying pattern recognition and machine learning to medical image analysis has resulted in significant improvements in diagnostic accuracy and the specificity of AD imaging biomarkers (Davatzikos et al., 2008; Vemuri et al., 2008). In these studies, the goal was to create a tool to discriminate between

diagnostic groups, rather than to optimize the efficiency of a biomarker. In other words, applying these algorithms to explicitly minimize required sample sizes will require modifications. These modifications will likely lead to greatly reduced sample size requirements in a clinical trial. Some evidence for this can be found in a recent paper by (Gutman et al., 2012), where numeric summaries are computed from signals weighted using linear discriminant analysis, and others, like Hobbs et al. (2010), which use a linear support vector machine classifier.

With several imaging biomarkers currently being considered for therapeutic trials to track brain degeneration (Cummings, 2010), different approaches need to be compared. Ideally, biomarkers would show excellent effect sizes for detecting longitudinal changes, avoid sources of bias, and not fail on a substantial fraction of the data, as a real clinical trial would not allow the selective exclusion of data (Fox et al., 2011).

In the current paper, we had 3 goals: first, to report improved and highly competitive sample size estimates for TBM, showing that no bias is present. Second, to develop and test several new efforts to improve the robustness of TBM, making the results robust to outliers in the data and poor quality scans. Third, to test whether standard enrichment methods – preferential selection of subjects based on Apolipoprotein E (ApoE) ε4 genotype or family history – could further reduce the required sample sizes when used in conjunction with the proposed improvements. We also studied the effect of selective data exclusion on the sample size estimates, suggesting that sample size estimates may be unduly optimistic if any removal of outliers is allowed.

Materials and methods

Overall design

We employed TBM to analyze the full ADNI-1 dataset, including all available 1.5 Tesla MR images scanned at screening, with follow-up scans at 6, 12, 18, 24, and 36 months (N=3314), available for download at March 20, 2012. Numerical summaries were derived from a statistical region-of-interest (stat-ROI) inside the temporal lobes to quantify cumulative brain degeneration over time, and these were later used to compute sample size estimates for hypothetical clinical trials. We used a subgroup of healthy subjects with completed scan series at screening, 6, 12, and 24 months to assess whether our method was biased (in the sense of over- or under-estimating the true rate of change), and to confirm the biological plausibility of atrophy measures. We hypothesized that the healthy aging group would exhibit an essentially linear trend of minimal brain atrophy, with a zero intercept for the regression line fitted through all time points. We conducted power analyses to estimate sample size requirements for hypothetical clinical trials employing imaging outcome measures. We further tested the added effect of performing more standard drug trial enrichment strategies using ApoE status and family history of dementia. Finally, we conducted a simulation to demonstrate how sample size estimates were influenced by selective data removal, an effort that suggests reasonable recommendations for fair comparisons of methods in the future (cf. Wyman et al., 2012).

Alzheimer's Disease Neuroimaging Initiative

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging

(MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. MD, VA Medical Center and University of California–San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

MRI acquisition and image correction

All subjects were scanned with a standardized MRI protocol developed for ADNI (Jack et al., 2008). Briefly, high-resolution structural brain MRI scans were acquired at 59 ADNI sites using 1.5 Tesla MRI scanners (GE Healthcare, Philips Medical Systems, or Siemens). Using a sagittal 3D MP-RAGE scanning protocol, the typical acquisition parameters were repetition time (TR) of 2400 ms, minimum full echo time (TE) of 3 ms, inversion time (TI) of 1000 ms, flip angle of 8°, 24 cm field of view, 192×192×166 acquisition matrix in the x-, y-, and z- dimensions, yielding a voxel size of 1.25×1.25×1.2 mm^3, later reconstructed to 1 mm isotropic voxels. For every ADNI exam, the sagittal MP-RAGE sequence was acquired a second time, immediately after the first using an identical protocol. The MP-RAGE was run twice to maximize the chance that at least one scan would be usable for analysis.

The scan quality was evaluated by the ADNI MRI quality control (QC) center at the Mayo Clinic to exclude failed scans due to motion, technical problems, significant clinical abnormalities (e.g., hemispheric infarction), or changes in scanner vendor during the time-series (e.g., from GE to Philips). Image corrections were applied using a standard processing pipeline consisting of four steps: (1) correction of geometric distortion due to gradient non-linearity (Jovicich et al., 2006), i.e. "gradwarp" (2) "B1-correction" for adjustment of image intensity inhomogeneity due to B1 non-uniformity (Jack et al., 2008), (3) "N3" bias field correction for reducing residual intensity inhomogeneity (Sled et al., 1998), and (4) phantom-based geometrical scaling to remove scanner and session specific calibration errors (Gunter et al., 2006). The first three steps were applied to both the first and repeat MP-RAGE scans. The Mayo QC team then selected one of the preprocessed MP-RAGE scans with superior quality, and proceeded to step four for phantom-based geometrical scaling. The final corrected image was identified with the term "scaled" in the file name, to denote phantom-based scaling.

The ADNI-1 dataset

The ADNI MRI Core has attempted to create a standard dataset to facilitate unbiased comparisons of quantitative methods (Wyman et al., 2012). The dataset aims to include all ADNI-1 1.5 Tesla scans that passed QC and went through all steps of image corrections. Successively, subsets of data were defined based on subjects with complete 1.5-Tesla MRI scan series for one or two years. As the process is ongoing, we based our paper on the full dataset available for download at March 20, 2012. In our analysis, we reported results for 100% of the data downloaded. In other words, no data exclusion was permitted.

In this manuscript, we reported results using the full dataset as well as for the subset of subjects with complete visits up to and including their 2-year visit.

The Full Dataset included serial brain MRI scans (*N*=3314; Table 1) from 188 probable AD patients (age at screening: 75.4±7.5 years, 99 Male (M) / 89 Female (F)), 400 individuals with amnestic MCI (age: 74.8±7.4 years, 257 M/143 F), and 229 healthy elderly controls (age: 76.0±5.0 years, 119 M/110 F). Subjects were scanned at screening and followed up at 6, 12, 18 (MCI only), 24, and 36 months (MCI and normal only).

The Complete 2-year Visit Subset (*N*=2079) included 98 AD (age: 75.2±7.4 years, 52 M/ 46 F), 207 MCI (age: 74.9±7.0 years, 139 M/68 F), and 163 healthy subjects (age: 76.0±4.9 years, 83 M/80 F) scanned at screening, 6, 12, 18 (MCI only) and 24 months.

All raw scans, images with different steps of corrections, and the standard ADNI-1 collections are available to the general scientific community at http://adni.loni.ucla.edu.

Image pre-processing

To adjust for linear drifts in head position and scale within the same subject, the follow-up scan (6-, 12-, 18-, 24-, or 36-month) was linearly registered to its matching screening scan using 9-parameter (9P) registration, driven by a mutual information (MI) cost function (Collins et al., 1994). 9P linear registration was chosen to correct for scanner voxel size variations in large longitudinal studies and any residual scaling errors after phantom-based image correction (Clarkson et al., 2009). Additionally, to account for global differences in brain scale across subjects, the mutually aligned time-series of scans was then linearly registered to the International Consortium for Brain Mapping template (ICBM-53) (Mazziotta et al., 2001), applying the same 9P transformation to both mutually aligned scans. Intermediate transformation matrices were concatenated into a single transformation file so that both screening and follow-up scans were resampled once during the linear registration (see (Yushkevich et al., 2010) on the need for equivalent resampling of both images to avoid one source of bias in analyzing longitudinal data). Globally aligned images were re-sampled in an isotropic space of 220 voxels along *x*-, *y*- and *z*-dimensions with a final voxel size of 1 mm^3.

Brain masks that excluded skull, other non-brain tissues, and the image background were generated automatically using a parameterless robust brain extraction tool (ROBEX) (Iglesias et al., 2011). Separate ROBEX masks were created for mutually aligned screening and follow-up scans in the ICBM space. A joint mask was then created using the union of two masks, followed by 2 iterations of morphological dilation using the mean dilation tool with a box kernel of size 3×3×3 in FSLMATHS (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Fslutils), to ensure that all brain tissues were included. Finally, we applied the dilated joint mask to uniformly “skull-strip” the screening and 9P registered follow-up scans, which were later used to compute the longitudinal change maps, also known as the “Jacobian” maps.

Average group template – minimal deformation target

A minimal deformation target (MDT) was created from the scans of 40 randomly selected normal subjects to serve as an unbiased average template image (Good et al., 2001) (Fig. 1). MDT construction has been detailed previously (Hua et al., 2008) and is described only briefly here. To construct an MDT, we first created an initial affine average template by taking a voxel-wise average of the 9P globally aligned scans after intensity normalization. Next, a non-linear average template was built after warping individual brain scans to the affine template (Yanovsky et al., 2008, 2009). The above steps were repeated until a full-resolution image registration was achieved. Lastly, the MDT was generated by applying the

inverse geometric centering of the displacement fields to the non-linear average (Kochunov et al., 2002, 2005).

Tensor-based morphometry and 3D longitudinal change maps

TBM is an image analysis technique that measures brain structural differences from the gradients of deformation fields that align one image to another (Ashburner and Friston, 2003; Chung et al., 2001; Freeborough and Fox, 1998; Riddle et al., 2004; Thompson et al., 2000; Toga, 1999). Individual Jacobian maps were created to estimate 3D patterns of structural brain change over time by warping the 9P-registered and 'skull-stripped' follow-up scan to match the corresponding screening scan. We used a non-linear inverse consistent elastic intensity-based registration algorithm (Leow et al., 2005), which optimizes a joint cost function based on mutual information (MI) and the elastic energy of the deformation. The deformation field was computed using a spectral method to implement the Cauchy–Navier elasticity operator (Marsden and Hughes, 1983; Thompson et al., 2000) using a Fast Fourier Transform (FFT) resolution of 64×64×64. This corresponds to an effective voxel size of 3.4 mm in the *x*, *y*, and *z* dimensions (220 mm / 64=3.4 mm). Color-coded maps of the Jacobian determinants were created to illustrate regions of ventricular/CSF expansion (i.e., with $det\ J(r)>1$), or brain tissue loss (i.e., with $det\ J(r)<1$) over time. These longitudinal maps of tissue change were also spatially normalized across subjects by nonlinearly aligning all individual Jacobian maps to a MDT, for regional comparisons and group statistical analyses.

Group average maps

To illustrate the average amount of atrophy at each follow-up time-point, relative to the screening visit, we computed the voxel-wise mean Jacobian map across subjects. These maps were color-coded to show the average percentage of regional brain tissue loss and ventricular/CSF expansion, relative to the screening scan (baseline).

Data-driven measures of temporal lobe atrophy

The use of a statistically-defined ROI based on an independent training sample was first proposed for positron emission tomography images (Chen et al., 2009, 2010; Reiman et al., 2008). We created a statistically-defined ROI (stat-ROI) based on voxels with significant atrophic rates over time ($p<0.00001$) within the temporal lobes, in a non-overlapping training set of 20 AD patients (age at baseline: 74.8±6.3 years; 7 men and 13 women) scanned at baseline and 12-months. The notion of using a statistical ROI has been described in prior work, and can be extended to use a variety of weighting methods (Gutman et al., 2012). In this study, we computed a numerical summary of the 3D Jacobian map to estimate the amount of cumulative atrophy, by taking an average within the data-driven, stat-ROI. For the 20 AD patients selected to create the stat-ROI, we used a leave-one-out strategy so that they could all be included in the final analysis (i.e., 19 AD patients were used for creating a stat-ROI, which was used to derive a numerical summary for the left-out subject, and this process was repeated by leaving out each of the other subjects).

Power analysis and sample size calculations

A power analysis was defined by the ADNI Biostatistics Core to estimate the sample size required to detect a 25% reduction in the mean annual rate of atrophy, using a two-sided test and standard significance level ($\alpha=0.05$) for a hypothetical two-arm study (treatment versus placebo). The estimated minimum sample size for each arm is computed from the formula below. Briefly, β denotes the estimated change and σ_D refers to the standard deviation of the rate of atrophy across subjects.

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

Here z_α is the value of the standard normal distribution for which $P[Z<z_\alpha]=\alpha$ and α is set to its conventional value of 0.05 (Rosner, 1990). The sample size required to achieve 80% power was computed in this study, referred to as *n80*. As the observation time ranged from 6 to 36 months, we computed the number of subjects required to detect a 25% reduction in the overall atrophy, for clinical trials with a duration of 6, 12, 18, 24, and 36 months respectively. Sample size estimates directly relate to the mean and standard deviation of the atrophy measures. The 95% confidence interval (*c*) for the *n80* statistic was computed based on bootstrap resampling with 10,000 samples, with a bias corrected and accelerated percentile method (Davison and Hinkley, 1997; Efron and Tibshirani, 1993).

Bias estimation

To estimate any potential additive bias from the TBM method, we assessed the linearity of the brain change over time in a subgroup of *N*=163 healthy subjects who had a complete scan series at screening, 6, 12, and 24 months. Bias was quantified by estimating the offset or intercept, with 95% confidence intervals, at time zero, by fitting a linear mixed effects model through measures of cumulative atrophy at 6, 12, and 24 months. The *lmer* and other statistical functions from the R statistical package (version 2.14.0: library (lme4)) were used to estimate the intercept and 95% confidence intervals.

Brain atrophy rates

When changes are small (e.g., over short intervals), the percentage of *cumulative atrophy* is a close approximation to the *rate of atrophy*. The relationship between the rate of atrophy and the eventual total cumulative volume loss is formulated below (*t* is the scan interval, in months):

$$\text{Cumulative atrophy (or percentage of tissue change)}=1-(1-\text{atrophyrate})^{\frac{t}{12}}$$

Or alternatively,

$$\text{Atrophyrate}=1-exp\left(\frac{log_e\left(1-\text{cumulative atrophy}\right)}{t}\times 12\right)$$

Excluding data and its impact on n80 estimates

We did not exclude any data in our TBM analysis; instead we used the entire ADNI1 dataset available at the time of download. To show how selective data exclusion affects the power analysis, we simulated a process of gradual data removal and demonstrated how it might affect *n80* estimates. We identified MCI subjects with positive numerical summaries of temporal lobe atrophy (meaning an apparent “gain” in tissue, which is not biologically plausible). We gradually removed them – i.e., a small proportion of the overall data – from the study population, and computed *n80* estimates and confidence intervals after each data point removed, at 12 and 24 months respectively. The MCI group consisted of elderly subjects with possible prodromal Alzheimer’s disease, so no tissue growth was expected.

Any positive numerical summaries may result from imaging noise, or unknown issues in image acquisition, pre and post image processing, or some combination of the above.

Results

TBM as an unbiased imaging biomarker

It is well established that healthy aging brains exhibit a fairly constant and very low rate of atrophy over time. We therefore used a subgroup of healthy subjects (*N*=163) who had complete visits at 6, 12, and 24 months to estimate a linear model for brain atrophy over successive visits. As an estimate of bias, the intercept estimate and its 95% confidence intervals were 0.06% [−0.07, 0.18] for TBM-derived numerical summaries of cumulative atrophy (Fig. 2). The actual scan acquisition intervals deviate from the nominal ones in certain subjects. We also estimated the linear model for brain atrophy using the actual scan interval in days (Inline Supplementary Fig. S1). The fitted intercept and confidence intervals were 0.009% [−0.12, 0.13]. We concluded that there was no evidence of methodological bias, as the 95% confidence interval covered zero, and the intercept estimate was extremely close to zero.

Inline Supplementary Fig. S1 can be found online at http://dx.doi.org/10.1016/j.neuroimage.2012.10.086.

To visualize the progression of measured brain atrophy within each subject, please see the scatter plot of cumulative atrophy (%) with lines connecting the points at 6, 12, and 24-months for each individual (Inline Supplementary Fig. S2).

Inline Supplementary Fig. S2 can be found online at http://dx.doi.org/10.1016/j.neuroimage.2012.10.086.

Subjects with complete 2-year visits

As different numbers of subjects are available at various follow-ups due to attrition or other reasons, it is difficult to compare patterns of brain degeneration over time using the full sample, as slightly different subjects would be represented at each time-point. We therefore used a subset of subjects with complete scan series at screening, 6, 12, 18 (MCI only), and 24 months to establish the trends of brain atrophy over a 2-year follow-up period, which include 98 AD (age at screening: 75.2±7.4 years, 52 M/46 F), 207 MCI (age: 74.9±7.0 years, 139 M/68 F), and 163 healthy subjects (age: 76.0±4.9 years, 83 M/80 F; same subjects as in Section 3.1) scanned at screening, 6, 12, 18 (MCI only) and 24 months.

Cumulative atrophy and n80 estimates

The color panels in Fig. 3 show average maps of cumulative tissue change at 6, 12, 18 (MCI only), and 24 months. In AD, MCI and normal groups respectively, a greater amount of brain degeneration, indicated by ventricle/CSF expansion (*red color*) and tissue loss in the temporal lobe areas (*blue color*), was observed over a longer follow-up period (24>18>12>6 months). As expected, the AD group showed the most severe regional brain degeneration: about 10–20% ventricular expansion and 5–10% tissue loss in the temporal lobes over 2 years. In contrast, the healthy elderly group showed only mild ventricular expansion with little to no temporal lobe tissue loss. The MCI group showed an intermediate level of atrophy. The average rates of ventricular expansion are 8.7%, 5.9%, and 3.5% per year, in AD, MCI, and controls, respectively, comparable to ventricular volume change measures in the other studies (Gutman et al., 2012; Nestor et al., 2008).

Next, numerical summaries were derived from a "stat-ROI" (statistical region of interest) to quantify cumulative temporal lobe atrophy, which were later used to compute sample size estimates (*n80*s) in hypothetical clinical trials. As summarized in Fig. 4a and Table 2, a greater amount of cumulative atrophy was observed for each subsequent follow-up period in every diagnostic group. The MCI group showed about twice as much, and AD showed 3–4 times, the rate of atrophy observed in normal aging.

In power analyses, fewer subjects (i.e., smaller *n80s*) were needed for a hypothetical clinical trial when the follow-up period was longer (Fig. 4b and Table 2). For hypothetical clinical trials intended to slow the rate of brain atrophy in AD, 106, 58, and 39 patients were necessary for hypothetical trials of duration 6, 12, and 24-months respectively. For clinical trials aimed at early or minimally symptomatic patients, i.e., the MCI group, 312, 124, 111, and 95 subjects were required for trials with duration of 6, 12, 18, and 24-months respectively. Finally, some trials aim to prevent disease onset and progression even in cognitively normal, healthy controls (Eastman, 2012). We therefore calculated sample size estimates for the healthy population, and the *n80*s were 785, 201, and 116 for trials with duration of 6, 12, and 24-month respectively (see Discussion for issues with computing power estimates for controls). Compared to AD and MCI, normally aging subjects showed a very slow and steady rate of degeneration, with lower variance overall, within the group.

Brain atrophy rates and evidence of deceleration

Brain atrophy rates were 2.7% in AD, 1.7% in MCI, and 0.8% in normal controls, estimated using the 12-month cumulative atrophy measures of subjects with complete 2-year visits (Table 2). The 12-month data was chosen rather than the 6-month data (first time point), as atrophy measurements over a shorter interval show a lower signal to noise ratio. We then used the formula described earlier to compute "expected cumulative atrophy" based on an assumption of constant atrophy rate, and compared the results with the "measured cumulative atrophy" (Table 2) to illustrate the trend of deceleration or acceleration (Fig. 5). There was a small but noticeable deceleration. At 24 months, the amount of deceleration is the greatest in AD, less in MCI, and minimal in controls, which accounts for 6.1%, 6.9%, and 6.0% of the total measured cumulative atrophy in AD, MCI, and controls, respectively.

Full ADNI-1 data set

We analyzed all available 1.5 Tesla MR scans that had passed QC and gone through the 4-step image correction pipeline, identified as "scaled" images in the ADNI database. Numerical summaries were derived from a stat-ROI inside the temporal lobes. From these, we computed sample size estimates in hypothetical clinical trials. Similar to results derived from subjects with complete 2-year visits (Table 2), the full data set showed a greater amount of cumulative atrophy with longer follow-up intervals (up to 3 years) in each diagnostic group (Table 3). Fewer subjects were available at later time points due to attrition. The average annual attrition rate for the entire ADNI-1 study was around 20%, (1st year: 19%, 2nd year 21%) with higher attrition in AD (1st year: 27%, 2nd year 24%) and MCI (1st year: 19%, 2nd year 25%) compared to normal individuals (1st year: 14%, 2nd year 12%). Common reasons for attrition included various health-related or personal limitations (~60%), death (5–10%), adverse events (4–8%), etc.

Drug trial enrichment

Many drug trials preferentially enroll participants who are more likely to decline, based on the premise that therapeutic effects may be easier to detect in subjects with greater brain changes. This approach is sometimes known as "enrichment". We tested two widely-used drug trial enrichment strategies, based on ApoE status and family history of dementia, using the 24-month data (*N*=521). Although these methods are used in current clinical trial design,

it is important to test whether they offer demonstrable advantages when used in conjunction with the MRI measures of this paper.

ApoE status

Risk for late-onset Alzheimer's disease is associated with a person's *ApoE* genotype on chromosome 19 (Pericak-Vance et al., 1991). The *ApoE* gene comes in three major forms or alleles: ε2, ε3, and ε4. The *ApoE* ε2 variant is rare and has been shown to be protective against Alzheimer's disease (Corder et al., 1994). *ApoE* ε3, the most common allele, may be considered as neutral (Saunders et al., 1993b; Schachter et al., 1994). *ApoE* ε4, present in ~40% of people with late-onset AD but only a sixth to a quarter of the normal elderly population, is linked to increased risk of developing AD, and the risk is dose-related (Corder et al., 1993; Saunders et al., 1993a).

We divided each diagnostic group into "ε4 carriers" (ε4/ε3 or ε4/ε4) and "non-carriers" (ε2/ε2, ε2/ε3 or ε3/ε3), and computed cumulative atrophy at 24 months, as well as *n80* estimates. Sixty-nine AD (46 ε4/ε3 and 23 ε4/ε4), 124 MCI (93 ε4/ε3 and 31 ε4/ε4), and 48 normal individuals (43 ε4/ε3 and 5 ε4/ε4) carried one or two copies of ε4, while 33 AD (3 ε2/ε3 and 30 ε3/ε3), 111 MCI (11 ε2/ε3 and 100 ε3/ε3) and 122 normal (2 ε2/ε2, 20 ε2/ε3 and 100 ε3/ε3) were non-carriers. We excluded a small number of subjects with a genotype of ε2/ε4, because it is not clear how to treat the aggregate effect of the two opposing alleles (3 AD, 9 MCI, and 2 normal), and this might complicate interpretation. As summarized in Table 4, a faster rate of brain atrophy was observed in ε4 carriers versus non-carriers, in all three diagnostic groups. For the MCI group, the sample size estimate of a 24-month trial in ε4 carriers (*n80*=73, *c*=[57,94]), was *half* of the estimate for non-carriers (*n80*=145, *c*=[111,206]). A trend-level enrichment effect of the ApoE status was observed for the AD group, but not for the normal aging group. The difference in statistical significance here could just be due to the smaller sample sizes for AD and controls, relative to the larger group of MCI subjects.

Family history of dementia

Information on each participant's family history of dementia is available via the Family History Questionnaire from the ADNI clinical data. We created two categories based on parental family history. We found no difference in the 24-month cumulative rate of atrophy, or in the *n80* estimates, comparing subjects with parental family history of dementia versus those without, in all three diagnostic groups (data not shown), using the TBM approach.

Selective data exclusion as a source of bias on sample size estimates

In our TBM analysis, we did not exclude data for any reason. Substantial bias can be introduced if selective data exclusion is allowed. Typical reasons for selective data exclusion include (1) software failure on certain scans, (2) outlier or extreme outcome measures identified by statistical analyses, and (3) biologically implausible measures, that might be noticed by a knowledgeable observer visually rating the scans.

We used the MCI sample at the 12- and 24-month time points to demonstrate how selective data removal impacts the sample size estimates. As "tissue expansions" (*i.e.* positive numerical summaries for temporal lobe atrophy) are unexpected in this elderly population with possible prodromal Alzheimer's disease, we identified the MCI subjects with positive tissue change and successively removed them, with a rank order to eliminate the greatest outliers first. We note that this would not be regarded as good statistical practice, but is in fact done in many studies where the analysis software does not give a reasonable answer. In many cases, outliers are removed because the program fails altogether, and there is no sensible value to use. We computed *n80* estimates and confidence intervals after each data

point removal, at 12 and 24 months respectively, to simulate the potential effect of data removal on sample size estimates (Fig. 6). At 12 months, 25 out of 326 MCIs had positive numerical summaries. The *n80* dropped from 135 [114,167] with no data removal to 93 [79,111] after removal of all 25 potential "outliers" – a 31% reduction of *n80* with 7.7% of data thrown out (Fig. 6a). Only 4 MCI subjects out of a total of 244 had positive numerical summaries at 24 months. The *n80* changed from 109 [92,131] to 100 [85,119] after removal of the 4 subjects or 1.6% of the data with positive measures (Fig. 6b).

Discussion

ADNI is one of the world's largest neuroimaging consortia studying Alzheimer's disease and aging. The entire dataset and a representative set of derived analytical results are shared among the academic community and available to the general public, allowing methodological scrutiny, replication, and direct comparison of different image analysis methods (Fox et al., 2011; Weiner et al., 2012). Registration-based imaging biomarkers are liable to three common but avoidable sources of bias, including asymmetric interpolation in global image registration (Yushkevich et al., 2010), failure to fully enforce inverse-consistency or transitivity (Christensen and Johnson, 2001; Hua et al., 2011; Thirion, 1998), and selective data exclusion. Any or all of these problems may lead to biased comparisons of methods or undue optimism about how the methods would perform in a real clinical trial (Fox et al., 2011).

The first source of bias in longitudinal processing can be addressed by using identical interpolation for both the baseline and follow-up scans, and treating scans in a way that does not depend on their order. Inverse consistent registration algorithms ensure that the correspondence does not depend on the order of the two images. In practice, inverse consistency can be quite intricate to achieve as many so-called inverse-consistent methods penalize deviations from inverse-consistency rather than completely removing it (Hua et al., 2011). Methods that enforce inverse consistency do not however explicitly enforce transitivity; in general, transitivity errors will remain even after inverse consistency is enforced. For example, it is possible to create a registration method that simply cannot produce maps that are *not* inverse consistent – this can be done by simultaneously computing the forward and backward maps as a pair, using methods such as those in Leow et al., 2005. By contrast, transitivity errors have to be computed and minimized for each dataset consisting of 3 or more time points, by registering all brains at once, so that the vectors aligning them obey the law of vector addition. For a time-series of images from an individual, all data can be given to the algorithm at once, and, if the scans' temporal order is known, and provided as an additional constraint on the allowable mappings, a set of deformation fields can be computed that obeys transitivity. If this is done, the results are "transitive by construction". Because this involves redistributing error vectors among a set of actual mappings, the mappings of earlier time-points may change if additional time-points are added. For the final source of bias – selective data exclusion – a standard dataset will help to ensure a more meaningful comparison of different analysis methods.

In an alternative indirect approach for bias correction, a linear regression of atrophy measures at multiple time points is used to estimate a non-zero intercept, which serves as a proxy for bias estimation and is subtracted from the observed change (Holland et al., 2012; Yushkevich et al., 2010). The non-zero intercept is an indirect measure of bias, which could be an aggregate of multiple sources of bias, so simply subtracting the intercept might not correct the bias. We advocate trying to identify and address any sources of bias at each step of the image processing pipeline, although this is more difficult.

Compared to Hua et al. (2011), we implemented two major changes in this analysis, in addition to analyzing the full ADNI-1 dataset. First, a uniform brain mask was generated based on each longitudinal scan pair including a screening and follow-up scan. Brain masks were used to exclude non-brain tissues and the image background, prior to nonlinear registration. This implementation made TBM more robust, by minimizing the influence of non-uniform background intensities in some scans. Second, the current manuscript used the non-linear inverse consistent elastic intensity-based registration algorithm, known as "3DMI" (Leow et al., 2005), which replaced the inverse consistent nonlinear registration algorithm using a regularization term of the symmetrized Kullback–Leibler distance, known as "ic-sKL-MI" (Hua et al., 2011). The ic-sKL-MI was initially designed to improve image registration resolution and accuracy. There was a very small residual intercept of 0.28%, when the level of atrophy over successive time-points was modeled using linear regression, so we discontinued the use of ic-sKL-MI until further improvements can be made. It is worth recalling that the two kinds of registration program produce warping fields with different formal properties – the deformation maps over time obey mathematical laws that differ depending on the formulation. Registration methods based on elasticity tend to minimize an elastic energy, which penalizes severe linear and volumetric compressions and expansions. The sKL-based methods are somewhat different as they increase the overall uniformity of the Jacobian field – implicitly, they also penalize severe volumetric compressions and expansions. Further study of these functionals is needed, but since 3DMI (the elastic method) tended to eliminate small offsets in longitudinal time-series, it seems preferable at this time.

Defining a standard ADNI dataset to ensure an objective methodological comparison

As noted in Wyman et al. (2012), the ADNI MRI Core recently made an effort to define a standard dataset to help compare different analysis methods side-by-side. Most ADNI publications, to date, have used a different subset of data; much of this was unintentional and unavoidable as a series of publications came out as more data progressively became available. In addition, problems with a small number of scans were only noticed well into the study (e.g., scan series in which the scanner vendor was changed during the time-series); (Wyman et al., 2012). Recognizing the effort and time necessary to finalize a standard dataset, we analyzed all available scans at the time of download. Data analysis was concluded on June 1, 2012. We provide a full list of subjects analyzed in the paper (Inline Supplementary Table S1). For scans processed in this paper, as well as scans added after the completion of data analysis, we will upload the full results to the ADNI website (https://ida.loni.ucla.edu). The findings of the paper are unlikely to change when a few scans are changed. Given the rate at which errors are discovered, it might be an unachievable goal to finalize a standard dataset that is free from all possible sources of error. Even so, it provides a reasonable practice to report all available data, while acknowledging that subjects may still be added or removed from the standard ADNI-1 dataset.

Inline Supplementary Table can be found online at http://dx.doi.org/10.1016/j.neuroimage.2012.10.086.

Robustness of image analysis techniques or failure rate

The robustness of an image analysis technique may be defined as percentage of analyzable scans relative to all available data. As clinical trials typically must report outcomes on all subjects assessed, this presents serious problems for methods where an algorithm fails or has low robustness. If a method is based on segmentation, for example, some algorithms may not give reliable segmentations on some fraction of the data. If this data is then excluded from the results, the performance – and even the applicability – of the method in a real clinical trial may be unclear. For example, if method A fails on 10% of the data but those

subjects are subsequently excluded, then as we have shown, sample size estimates based on that data may be more than 30% too low. In our TBM analysis, we did not exclude any scans, and analyzed the full ADNI-1 dataset that passed basic image quality QC. As other analysis methods do report successful results after removing outliers (e.g., around 10% of the data were excluded in (Holland et al., 2009, 2012), a simulation was run to show how selective data exclusion impacts sample size estimates (see Section 3.5).

Relative versus absolute change

A critical consideration when estimating sample sizes for treatment response is whether to include effects seen in normal aging as potentially treatable effects. Some researchers argue that the relative change – or the rate of change corrected for normal aging – should be defined as the only treatable effect (Holland et al., 2012). The power analysis was conducted by calculating sample size estimates using the variance parameters from the patient cohort, with the treatment effect defined as the difference between the mean rates of change in the patients and healthy controls. The advantage of this approach is that it can partially cancel out any systematic methodological bias, reducing the overoptimism of power calculations. This suggestion, however, creates several challenges. First, many current MCI or AD trials do not enroll healthy subjects as controls. Second, a growing number of prevention trials now enroll healthy subjects and treat them (Eastman, 2012; Ross et al., 2012), in which case they would be considered as a “treatment group”. A large body of work shows that some pathological processes – such as vascular degeneration – undoubtedly occur in normal controls, and treatments may resist their progression to some extent, whether or not a person is considered ill or shows clinical signs sufficient for an AD or MCI diagnosis.

The ADNI Biostatistics Core advocates using the absolute change, as the potentially treatable rate of atrophy, as atrophy and cognitive decline with normal aging cannot be considered impervious to treatments developed to resist AD, as many of the contributing biological processes are the same. In real clinical trials, a treatment group is typically compared against a placebo group to assess drug effects. The subtraction of the placebo group mean could serve a dual purpose of isolating treatment effects and reducing some sources of bias in power analyses.

In this manuscript, we computed sample sizes needed to detect a mean reduction in the absolute rate of change in each diagnostic group, while providing the full data necessary to compute effects corrected for normal aging (Tables 2 and 3).

Confounding factors in methods comparisons

Several research laboratories have computed longitudinal brain imaging measures, and made their results publicly available. There is a great interest in determining which of the many reported techniques is most sensitive for measuring brain changes and factors that influence them, while remaining robust enough to give plausible measures for all available scans that pass basic scan quality QC, and while avoiding known sources of measurement bias (Wyman et al., 2012).

In a recent publication comparing longitudinal brain structural measures in ADNI, a commercial method known as quantitative anatomical regional change (Quarc) was compared to several methodologies including FreeSurfer Longitudinal v.4.4, FreeSurfer Cross-sectional v.4.3, Boundary Shift Integral (BSI), and TBM (Holland et al., 2012). The authors concluded that Quarc provided the most powerful change biomarker among all methods, in the sense of requiring low sample sizes to detect a given degree of slowing in the rate of atrophy. As a critical difference and possibly a major source of improvement, Quarc was run on “averaged” MRI scans, by averaging the MP-RAGE scans acquired back-

to-back as part of ADNI, while the other processing sequences used only one of the scans selected by the Mayo QC pipeline. This seemingly minor point may have given Quarc an apparent 40% SNR advantage over the others that would not be achievable in the more common situation where only one scan is collected. While the developers may advocate that all available data should be used, it is not clear that future studies will collect a back-to-back pair of MP-RAGE scans; for objective comparisons, future studies may need to target the same scan data. Moreover, Quarc used additional QC steps (QCPASS=1) to exclude 17% of data that either failed during Quarc processing or generated visually unsatisfactory results. This is another significant source of bias in the comparison, and may make the method unusable for many practical situations, including clinical trials.

To achieve a meaningful comparison and accelerate the development of imaging biomarkers compatible with rigorous clinical trial regulations, we suggest the following practices:

1. Analyze a standard dataset when available
2. When a standard dataset is unavailable, report the date of image download, number and IDs of the downloaded subjects, and number and IDs of subjects included in the final report, to facilitate an assessment of robustness of the image analysis techniques
3. Disclose full details of all pertinent image processing steps
4. Test for and address sources of bias at each of the image processing steps
5. Describe additional QC steps required for specific algorithms, and specify reasons for drop-out or exclusion of unanalyzable endpoints, with any data throw-out avoided or justified.

Evidence of deceleration in brain atrophy

It is important to note that the acceleration or deceleration of atrophy can only be reliably evaluated using the subjects with a common and complete set of visits, as results shown in Table 2. The full sample, which contained a gradually reduced sample size at each time-point (Table 3), was liable to attrition bias, so it is important not to simply regress all available scan data against time, in estimating group trends. For example, in the full sample, people who remained in the study tended to be more healthy than people who dropped out, so the later time points (e.g., 24, 36 months) might accumulate a group of subjects who were healthier or less impaired at baseline than the people represented at the earlier time points.

We observed a small but noticeable deceleration in brain atrophy, comparing the “cumulative atrophy expected assuming a constant rate of decline” versus the “measured cumulative atrophy” (Fig. 5). The deceleration accounts for about 6–7% of the overall measured change at 24-months. This could be attributed to a biological deceleration, transitivity errors, and/or regularization effect. Transitivity error refers to a difference between (1) the total atrophy estimated from the direct mapping of 0 to 24 month scans, relative to (2) the composition of mappings from 0 to 12, and 12 to 24 months (Hua et al., 2010). This is a common source of error in nonlinear registration, which has been modeled and extensively discussed in a prior publication (Hua et al., 2010). Specifically, registration problems involve the tuning of deformation field parameters, and the state vector of the parameters is not guaranteed to follow the same path through the 12-month time point if that data is not used as a constraint. Such a transitivity error could be “hidden” in a fully-4D registration method that includes all the time points and performs registration on time-series of scans all at once, while adjusting the mappings to reduce the transitivity errors. As noted in Hua et al. (2011), methods to do this include transform reconciliation (Woods et al., 1998), and group-wise registration (Leporé et al., 2008). These methods can compute a set

of mappings between all *N* brains in a study, and they use the internal consistency among mappings (or triplets of brains) as a means to reduce errors of various kinds, or simply to redistribute the mean error among all the mappings. Another possibility for an apparent slight deceleration of atrophy is due to the regularization effect, where all follow-up scans are longitudinally aligned to the corresponding screening scans, using the same Cauchy–Navier elasticity operator, irrespective of the time interval or diagnostic group. As a result, the atrophy progression might appear to artificially decelerate, with some dependency on the overall amount of atrophy. We expect a minimal impact as the maximum change in the study is around 5% (AD-24Mo), a relatively small change compared to prior cross-sectional studies using the same algorithm to detect large scale changes (10–30%). However we cannot entirely rule out the possibility that the slight trend for apparent deceleration in atrophy rates is due to the regularization in spatial warping.

Bias-variance trade-offs

It could be argued that if we computed the atrophy over each successive interval instead of always to baseline, much or all of the deceleration would disappear, but at the price of additional variability due to noise, and the difficulty of computing robust mappings when changes in the images are extremely small. Given the importance of the accuracy of a biomarker for clinical trials, one could argue that a less biased measure might be preferable unless it was catastrophically more variable. In support of this line of argument, in Hua et al., 2011, we modeled and discussed the transitivity error extensively. We tested the hypothesis of a systematic transitivity error – the direct mapping from 0 to 24 months (which is used to estimate atrophy) showed slightly less change than the composition of mappings from 0 to 12, and 12 to 24 months. The transitivity error was small in all areas of the brain, around 20 times smaller than the estimate of the true change. As this error was weakly correlated with the true biological change, subtracting it may even reduce the discriminative power of the measures. Clearly, when evaluating a registration method on a time-series, one could consider a step-wise mapping of all the scans, but this could even cause an over- or under-estimate of the true change, if there were a lot of noise in the scans. Ultimately, fully-4D registration methods may be desirable that can enforce a trajectory through all the scan data. Some of these methods have the undesirable effect that new scans might change the estimate of atrophy for earlier ones. This is reasonable from a Bayesian point of view, but perhaps somewhat disconcerting to have to re-run the full analysis, and change prior results, when new scans come in.

Using the composition of mappings to measure change, in theory, could reduce the transitivity error by performing repeated compositions. For example, if we break down the direct mapping of scans from 0 to 36 months into compositions of mappings from 0–6+6–12+12– 18+18–24+24–36 (5 registration steps instead of 1), the transitivity error may be reducible by around 4-fold; even so, other types of errors associated with image registration (e.g., inverse consistency errors) could be amplified by 4 times and remain concealed.

In Hua et al., 2011, we presented a detailed discussion on ways to further reducing transitivity errors, but as noted before, most of the proposed solutions do not remove the error but rather redistribute the errors.

Linear versus exponential model

The annual rates of atrophy are low, 2.7% in AD, 1.7% in MCI, and 0.8% in normal elderly, and the observation time is short, 0.5 to 3 years, the difference between a linear and an exponential model is very hard to detect. The R^2 (reflecting goodness of fit) for a regression model fitting the Jacobian values against time is 0.99904 using the linear model, and 0.9992 using the exponential model, using the MCI data at 6, 12, 18, and 24 months. We chose the

linear over the exponential model for simplicity. However we also consider that in reality there may be several sources of nonlinearity, both biological and technical.

Mathematically, an exponential decline in volume (Jacobians) may be expected under the assumption that the rate of volume loss depends on the amount of tissue present (section 2.11). When changes are small (e.g. over short intervals), the percentage of *cumulative atrophy* is a close approximation to the *rate of atrophy*. In other words, if the overall atrophy that has accumulated at the end of the study is divided into equal time intervals, those changes may be reasonable approximations of how much change actually occurred over those intervals. This assumes a linear volumetric loss of tissue, or a time interval so short that departures from linearity are not detectable. One could also posit an "*exponential decay*" model where the volume of tissue lost over each successive time interval is a fixed proportion of the tissue remaining at the start of the interval. Although we have studied both of these models, the overall duration of ADNI is still relatively short, and it is difficult to make a strong case for one model in favor of the other.

Biologically implausible measures

In MCI, it is unlikely that "tissue growth" occurs, or that any biological process is leading to positive numerical summaries of temporal lobe atrophy. While neuroplasticity is possible in principle, by far the most likely effect is that noise or error during acquisition or pre-processing does lead some measures to show a positive change, even if their true mean is zero (this is evident by processing short-interval scan pairs from the same subject). We identified 25 MCI subjects at 12-months and 4 MCI subjects at 24-months, who had positive numerical summaries, i.e., biologically implausible measures. They accounted for 7.7% and 1.6% of the total available number of subjects at 12- and 24-months respectively. Possible sources of biologically implausible measures include imperfect global image registration, slightly different noise levels for the scans at different time-points, subject motion, or different intensity profiles in the baseline and follow-up images, and scaling errors due to scanner calibration and/or image corrections.

Random and non-random missing data

In modeling the effect of selective data removal, we note that removing subjects with greatest tissue "gain" is "non-random" – it focuses on results that are biologically implausible. Even so, some might advocate that other kinds of data exclusion – especially data removed "at random" might be defensible or even desirable. In some studies, a subject may be excluded because they have an unusual anatomy (e.g., very large ventricles) at all timepoints, which may have a lesser effect on the group average rate of change. As described in the methods, the ADNI MRI quality control center at the Mayo Clinic excluded failed scans due to motion, technical problems, significant clinical abnormalities (e.g., hemispheric infarction), or changes in scanner vendor during the time-series (e.g., from GE to Philips). Subjects with unusual anatomy were removed at this step. The remaining subjects were uploaded to the ADNI database and they have no major scan quality issues and should be analyzed by all sites. Several image analysis approaches have additional built-in quality control processes in the image processing pipelines to exclude failed scans, e.g. some methods use terminology of the following kinds to flag excluded scans: OVERALLQC="Pass" or "Partial"; Boundary Shift Integral, VENTACCEPT=1, REGRATING<=3, KMNREGRATING<=3; Quarc, QCPASS=1. Scans that failed QC will have missing values, which are more likely to be "non-random missing data" as they are perhaps not so dependent on the actual rate of atrophy in the subject.

Sample size estimates

Sample size estimates directly relate to the mean and variance (standard deviation) of the atrophy measures. Both the mean level of cumulative atrophy and its variance increased monotonically with longer inter-scan intervals (Table 2 and 3). As the mean rose faster than the variance with longer intervals, incrementally greater effect sizes and smaller sample size estimates were observed. In theory, the random measurement errors arising from small variations in scanner calibration and RF bias fields remain stable while systematic atrophies accumulate over longer intervals, leading to greater signal to noise ratios. However, longer intervals do not necessarily translate to better sample size estimates due to the trade-off between observation time and attrition (Hua et al., 2010).

ApoE genotyping for drug trial enrichment

Strategies for drug trial enrichment using ApoE status were highly effective for the MCI cohort, marginally effective for the AD cohort and not obviously effective at all in the healthy controls. The sample sizes of AD and controls are substantially smaller than that of the MCI group, which might have affected the power of statistical analysis, thus we should not rule out the possibility of using ApoE genotyping for trial enrichment in AD and controls. Fewer controls had ε4 compared to MCI and AD, which might further undermine the statistical power in detecting an ε4 effect in the control group. For the 24-month data, there were 20 (12%), 100 (59%), 43 (25%), and 5 (3%) controls who had the genetic profiles of ε2/ε3, ε3/ε3, ε4/ε3, and ε4/ε4 respectively, compared to 11 (5%), 100 (42%), 93 (40%), and 31 (13%) in MCI, and 3 (3%), 30 (29%), 46 (45%), 23 (23%) in AD. As shown in Fig. 3, both AD and MCI had prominent atrophy localized in the temporal lobe areas, while the healthy controls had a very low rate of atrophy spread somewhat diffusely throughout the brain. The stat-ROI was trained on 20 AD subjects to identify the brain regions most likely to show significant atrophy in AD (a focal effect), but they were not optimized for picking up ApoE's effect on normal aging (a diffuse effect).

Family history of dementia

We found no difference in the 24-month cumulative rate of atrophy, or in the *n80* estimates, comparing subjects with versus those without parental family history of dementia. This does not mean that it is pointless to use family history as a basis for enrichment, only that we did not detect any benefit of using it for the TBM measures used here in the full ADNI sample.

Localization of changes with TBM

The volumetric change in AD appears to be most severe in temporal lobe white matter (WM) rather than gray matter (GM) (Fig. 3), which might seem contradictory as AD is widely accepted to be a predominantly hippocampal and cortical gray matter pathology. It had been difficult to quantify WM change in conventional MRI due to the lack of visible anatomical boundaries that would be required to parcellate WM, until the recent development of voxel-based approaches. Several studies using diffusion tensor imaging (DTI), relaxometry, and functional connectivity studies have provided substantial evidence for diffuse WM abnormalities in AD (Buckner et al., 2009; Wozniak and Lim, 2006). Myelin breakdown and Wallerian degeneration both lead to WM atrophy, perhaps secondary to the effect of cortical neuronal loss in AD (Bartzokis, 2011; Bartzokis et al., 2006, 2007; Spires-Jones et al., 2009). As both GM and WM changes are occurring in AD, a key question is which of the MRI-derived measures is the most reliable for detecting dynamic changes over time with greatest effect sizes and accuracy. As a percentage, more cortical and hippocampal GM may be lost over time than WM. Even so, the effect sizes for the changes in GM may be lower than expected, as these structures are convoluted and difficult to measure accurately. The cortex is thin and the hippocampus is narrow, accounting for a

small proportion of the total voxels in whole brain TBM analysis. As a result, the changes in the cortex and hippocampus may be greatest, as a percentage of their volume, but when pooling data across subjects voxel-by-voxel, the interiors of large white matter structures tend to be better registered than the cortical and hippocampal boundaries once all the data are aligned. Therefore, coherent patterns of WM atrophy are more likely to be reinforced across all members of a group than at boundary voxels where loss patterns may be less well registered, even after nonlinear registration.

Geometrical scaling in ADNI (scaled versus scaled_2)

ADNI-1 employed phantom-based geometrical scaling of MR images to improve spatial calibration of scans and longitudinal stability across all acquisition sites. For a subset of scans, both “scaled” and “scaled_2” images are provided when errors in phantom based scaling were identified and reversed to no scaling. Note that in scaled_2 images, the scaling errors were not corrected; the scaling factors were changed to 1 for all axes that were identified as faulty. The difference in scaling was in the range of 10^{-4}, so the choice of scaled versus scaled_2 scans is not likely to affect conclusions substantially. Even so, we have tested this formally, and a recent publication from our group found a high degree of correlation and no detectable difference between all available scaled and scaled_2 images, using the TBM approach (Ching et al., 2012).

Average brain template

The average brain template or MDT was created from 40 randomly selected normal subjects. Another design is to create the MDT based on randomly selected subjects from the entire study, including subjects from different diagnostic groups. The later design has the advantage of equally representing the entire study population but it might add complications in interpreting the results of several sets of studies including different groups, especially for cross-sectional studies where the MDT serves as the reference. For longitudinal studies such as the current manuscript, the choice of MDT has a negligible impact, as it does not affect the estimates of brain change rates in each person. The brain change rates are estimated by nonlinearly registering a follow-up scan to its screening scan, with the screening scan serving as the reference, not the MDT. Spatial normalizations among different brains enable regional comparisons and group analyses to be performed.

Limitations

Some limitations of this study must be mentioned. The use of *n80* as the sole guide for estimating sample sizes for real clinical trials has been questioned from multiple points of view, and its limitations should be understood by those using it as a guide. First, a real treatment effect may slow cognitive decline but not atrophy, or a treatment may slow atrophy with no detectable clinical benefit. For that reason, imaging measures would not be used as the only outcome measure in any clinical trial. Second, the use of a statistical region of interest will isolate regions most likely to show deterioration, but a real treatment may exert its effects on brain regions other than these, or even on processes not observable with anatomical MRI or any imaging modality. As a result, multiple imaging measures and modalities, as well as scanning methods not yet developed, are likely to be advantageous to avoid missing potentially beneficial effects. Finally, any direct comparison of sample size estimates for different imaging metrics may overlook the differential value of slowing cognitive decline versus imaging decline. Clearly, the benefit to the patient of a 25% reduction in the rate of amyloid accumulation or brain atrophy may be very different from the value of a 25% reduction in the rate of cognitive decline, even if the latter requires far more subjects to detect, or requires a more expensive clinical trial. For the same reason, it may not make sense to compare imaging measures head-to-head that do not assess the same

part of the brain or do not assess the same signal. Even so, any trial using MRI may be able to benefit from the diverse range of measures now available as surrogate markers of AD progression.

Although the leave-one-out approach used to implement the stat-ROI should give unbiased numerical summaries for the left out subjects, these summaries may not be totally independent of each other. The loss of the independent and identically distributed (i.i.d.) variance may cause a slight bias in statistical tests requiring an i.i.d. noise assumption. The MCI group contains a much larger proportion of men (139 M:68 F) than both the AD (52:46) and healthy control (83:80) groups. Sample size estimates based on the full ADNI cohort might be partially influenced by sex and other demographic factors not controlled for in ADNI. Finally, 9-parameter linear registration was used to align scans longitudinally, starting with 3 translations, and followed by 3 rotations and 3 scales. This step is not entirely inverse consistent, i.e., scale followed by rotation gives a different family of transformations than rotation followed by scale. The effects are likely to be relatively small as rotations and scales are very small as ADNI implemented stringent calibration procedures to ensure consistency in global scaling and voxel size calibration over time and across different sites (Ching et al., 2012; Jack et al., 2008).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514. Algorithm development for this study was also funded by the NIA, NIBIB, the National Library of Medicine, and the National Center for Research Resources (AG016570, EB01651, LM05639, RR019771 to PT). Author contributions were as follows: XH, DH, CC, CB, PR, BG, AT, and PT performed the image analyses; AL, BG, and PT developed algorithms used in the analyses; CJ, DH, and MW contributed substantially to the image and data acquisition, study design, quality control, calibration and pre-processing, databasing and image analysis.

Abbreviations

AD	Alzheimer's disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
ApoE	Apolipoprotein E
BBSI	Brain boundary shift integral
FSL	FMRIB Software Library
MRI	Magnetic resonance imaging

MCI	Mild cognitive impairment
MDT	Minimal deformation target
MI	Mutual information
Quarc	Quantitative anatomical regional change
SPM	Statistical parametric mapping
stat-ROI	Statistical region-of-interest
TBM	Tensor-based morphometry
VBM	Voxel-based morphometry
TR	Repetition time
TE	Echo time
TI	Inversion time
QC	Quality control

References

Ashburner J, Friston KJ. Voxel-based morphometry–the methods. NeuroImage. 2000; 11:805–821. [PubMed: 10860804]

Ashburner, J.; Friston, KJ. Human Brain Function. Academic Press; 2003. Morphometry.

Baron JC, Chetelat G, Desgranges B, Perchey G, Landeau B, de la Sayette V, Eustache F. In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. NeuroImage. 2001; 14:298–309. [PubMed: 11467904]

Bartzokis G. Alzheimer's disease as homeostatic responses to age-related myelin breakdown. Neurobiol. Aging. 2011; 32:1341–1371. [PubMed: 19775776]

Bartzokis G, Lu PH, Geschwind DH, Edwards N, Mintz J, Cummings JL. Apolipoprotein E genotype and age-related myelin breakdown in healthy individuals: implications for cognitive decline and dementia. Arch. Gen. Psychiatry. 2006; 63:63–72. [PubMed: 16389198]

Bartzokis G, Lu PH, Geschwind DH, Tingus K, Huang D, Mendez MF, Edwards N, Mintz J. Apolipoprotein E affects both myelin breakdown and cognition: implications for age-related trajectories of decline into dementia. Biol. Psychiatry. 2007; 62:1380–1387. [PubMed: 17659264]

Buckner RL, Sepulcre J, Talukdar T, Krienen FM, Liu H, Hedden T, Andrews-Hanna JR, Sperling RA, Johnson KA. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. J. Neurosci. 2009; 29:1860–1873. [PubMed: 19211893]

Carmichael OT, Thompson PM, Dutton RA, Lu A, Lee SE, Lee JY, Kuller LH, Lopez OL, Aizenstein HJ, Meltzer CC, Liu Y, Toga AW, Becker JT. Mapping ventricular changes related to dementia and mild cognitive impairment in a large community-based cohort. IEEE ISBI. 2006:315–318.

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. Cross-sectional and longitudinal analyses of fluorodeoxyglucose positron emission tomography images from the Alzheimer's disease neuroimaging initiative. ADNI Data Presentations Meeting; Seattle, WA: 2009.

Chen K, Langbaum JB, Fleisher AS, Ayutyanont N, Reschke C, Lee W, Liu X, Bandy D, Alexander GE, Thompson PM, Foster NL, Harvey DJ, de Leon MJ, Koeppe RA, Jagust WJ, Weiner MW, Reiman EM. Twelve-month metabolic declines in probable Alzheimer's disease and amnestic mild cognitive impairment assessed using an empirically pre-defined statistical region-of-interest: findings from the Alzheimer's Disease Neuroimaging Initiative. NeuroImage. 2010; 51:654–664. [PubMed: 20202480]

Chetelat G, Landeau B, Eustache F, Mezenge F, Viader F, de la Sayette V, Desgranges B, Baron JC. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. NeuroImage. 2005; 27:934–946. [PubMed: 15979341]

Ching CRK, Hua X, Ward C, Gunter J, Bernstein M, Jack CR Jr. Weiner MW, Thompson PM. Phantom-based MRI corrections and power to track brain change. IEEE International Symposium on Biomedical Imaging. 2012:1172–1175.

Chou YY, Lepore N, de Zubicaray GI, Carmichael OT, Becker JT, Toga AW, Thompson PM. Automated ventricular mapping with multi-atlas fluid image alignment reveals genetic effects in Alzheimer's disease. NeuroImage. 2008; 40:615–630. [PubMed: 18222096]

Chou YY, Lepore N, Avedissian C, Madsen SK, Parikshak N, Hua X, Shaw LM, Trojanowski JQ, Weiner MW, Toga AW, Thompson PM. Mapping correlations between ventricular expansion and CSF amyloid and tau biomarkers in 240 subjects with Alzheimer's disease, mild cognitive impairment and elderly controls. NeuroImage. 2009; 46:394–410. [PubMed: 19236926]

Christensen GE, Johnson HJ. Consistent image registration. IEEE Trans. Med. Imaging. 2001; 20:568–582. [PubMed: 11465464]

Chung MK, Worsley KJ, Paus T, Cherif C, Collins DL, Giedd JN, Rapoport JL, Evans AC. A unified statistical approach to deformation-based morphometry. NeuroImage. 2001; 14:595–606. [PubMed: 11506533]

Clarkson MJ, Ourselin S, Nielsen C, Leung KK, Barnes J, Whitwell JL, Gunter JL, Hill DL, Weiner MW, Jack CR Jr. Fox NC. Comparison of phantom and registration scaling corrections using the ADNI cohort. NeuroImage. 2009; 47:1506–1513. [PubMed: 19477282]

Collins DL, Neelin P, Peters TM, Evans AC. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. 1994; 18:192–205. [PubMed: 8126267]

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993; 261:921–923. [PubMed: 8346443]

Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC Jr. Rimmler JB, Locke PA, Conneally PM, Schmader KE, et al. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat. Genet. 1994; 7:180–184. [PubMed: 7920638]

Cummings JL. Integrating ADNI results into Alzheimer's disease drug development programs. Neurobiol. Aging. 2010; 31:1481–1492. [PubMed: 20447734]

Davatzikos C, Resnick SM, Wu X, Parmpi P, Clark CM. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. NeuroImage. 2008; 41:1220–1227. [PubMed: 18474436]

Davison, AC.; Hinkley, DV. Bootstrap methods and their application. Cambridge University Press; 1997.

Eastman P. Plans under way for Alzheimer's prevention trial. Neurology Today. 2012:1–14.

Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman & Hall; New York: 1993.

Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, Hall K, Hasegawa K, Hendrie H, Huang Y, Jorm A, Mathers C, Menezes PR, Rimmer E, Scazufca M. Global prevalence of dementia: a Delphi consensus study. Lancet. 2005; 366:2112–2117. [PubMed: 16360788]

Fox NC, Cousens S, Scahill R, Harvey RJ, Rossor MN. 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. 2000; 57:339–344. [PubMed: 10714659]

Fox NC, Ridgway GR, Schott JM. Algorithms, atrophy and Alzheimer's disease: cautionary tales for clinical trials. NeuroImage. 2011; 57:15–18. [PubMed: 21296168]

Freeborough PA, Fox NC. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Trans. Med. Imaging. 1997; 16:623–629. [PubMed: 9368118]

Freeborough PA, Fox NC. Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images. J. Comput. Assist. Tomogr. 1998; 22:838–843. [PubMed: 9754126]

Frisoni GB, Weiner MW. Alzheimer's Disease Neuroimaging Initiative special issue. Neurobiol. Aging. 2010; 31:1259–1262. [PubMed: 20570400]

Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage. 2001; 14:21–36. [PubMed: 11525331]

Gunter, J.; Bernstein, M.; Borowski, B.; Felmlee, J.; Blezek, D.; Mallozzi, R.; Levy, J.; Schuff, N.; Jack, CR, Jr.. Validation testing of the MRI calibration phantom for the Alzheimer's Disease Neuroimaging Initiative Study. ISMRM 14th Scientific Meeting and Exhibition; 2006.

Gutman, B.; Rajagopalan, P.; Hua, X.; Thompson, PM. Maximizing Power to Track Alzheimer's Disease Progression by LDA-Based Weighting of Longitudinal Ventricular Surface Features. The 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Novel Neuroimaging Biomarkers for Alzheimer's Disease and Related Disorders Workshop; Nice, France. 2012.

Hobbs NZ, Henley SM, Ridgway GR, Wild EJ, Barker RA, Scahill RI, Barnes J, Fox NC, Tabrizi SJ. The progression of regional atrophy in premanifest and early Huntington's disease: a longitudinal voxel-based morphometry study. J. Neurol. Neurosurg. Psychiatry. 2010; 81:756–763. [PubMed: 19955112]

Holland D, Dale AM. Nonlinear registration of longitudinal images and measurement of change in regions of interest. Medical Image Analysis. 2011; 15:489–497. [PubMed: 21388857]

Holland D, Brewer JB, Hagler DJ, Fennema-Notestine C, Dale AM. Subregional neuroanatomical change as a biomarker for Alzheimer's disease. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:20954–20959. [PubMed: 19996185]

Holland D, McEvoy LK, Dale AM. Unbiased comparison of sample size estimates from longitudinal structural measures in ADNI. Hum. Brain Mapp. 2012; 33:2586–2602. [PubMed: 21830259]

Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, Jack CR Jr. Weiner MW, Thompson PM. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. NeuroImage. 2008; 43:458–469. [PubMed: 18691658]

Hua X, Lee S, Yanovsky I, Leow AD, Chou YY, Ho AJ, Gutman B, Toga AW, Jack CR Jr. Bernstein MA, Reiman EM, Harvey DJ, Kornak J, Schuff N, Alexander GE, Weiner MW, Thompson PM. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: an ADNI study of 515 subjects. NeuroImage. 2009; 48:668–681. [PubMed: 19615450]

Hua X, Lee S, Hibar DP, Yanovsky I, Leow AD, Toga AW, Jack CR Jr. Bernstein MA, Reiman EM, Harvey DJ, Kornak J, Schuff N, Alexander GE, Weiner MW, Thompson PM. Mapping Alzheimer's disease progression in 1309 MRI scans: power estimates for different inter-scan intervals. NeuroImage. 2010; 51:63–75. [PubMed: 20139010]

Hua X, Gutman B, Boyle CP, Rajagopalan P, Leow AD, Yanovsky I, Kumar AR, Toga AW, Jack CR Jr. Schuff N, Alexander GE, Chen K, Reiman EM, Weiner MW, Thompson PM. Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. NeuroImage. 2011; 57:5–14. [PubMed: 21320612]

Iglesias JE, Liu CY, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging. 2011; 30:1617–1634. [PubMed: 21880566]

Jack CR Jr. Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, Boeve BF, Waring SC, Tangalos EG, Kokmen E. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology. 1999; 52:1397–1403. [PubMed: 10227624]

Jack CR Jr. Dickson DW, Parisi JE, Xu YC, Cha RH, O'Brien PC, Edland SD, Smith GE, Boeve BF, Tangalos EG, Kokmen E, Petersen RC. Antemortem MRI findings correlate with hippocampal neuropathology in typical aging and dementia. Neurology. 2002; 58:750–757. [PubMed: 11889239]

Jack CR Jr. Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell JL, 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. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging. 2008; 27:685–691. [PubMed: 18302232]

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. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. NeuroImage. 2006; 30:436–443. [PubMed: 16300968]

Kochunov P, Lancaster J, Thompson P, Toga AW, Brewer P, Hardies J, Fox P. An optimized individual target brain in the Talairach coordinate system. NeuroImage. 2002; 17:922–927. [PubMed: 12377166]

Kochunov P, Lancaster J, Hardies J, Thompson PM, Woods RP, Cody JD, Hale DE, Laird A, Fox PT. Mapping structural differences of the corpus callosum in individuals with 18q deletions using targetless regional spatial normalization. Hum. Brain Mapp. 2005; 24:325–331. [PubMed: 15704090]

Leow, A.; Huang, SC.; Geng, A.; Becker, JT.; Davis, S.; Toga, AW.; Thompson, PM. Information Processing in Medical Imaging. Glenwood Springs, Colorado, USA: 2005. Inverse Consistent Mapping in 3D Deformable Image Registration: Its Construction and Statistical Properties; p. 493-503.

Leporé, N.; Brun, CC.; Chou, YY.; Lee, AD.; Barysheva, M.; de Zubicaray, GI.; Meredith, M.; McMahon, K.; Wright, MJ.; Toga, AW.; Thompson, PM. Multi-atlas Tensorbased Morphometry and its Application to a Genetic Study of 92 Twins. Med Image Comput Comput Assist Interv Workshop on Mathematical Foundations of Computational Anatomy (MFCA); 2008.

Marsden, J.; Hughes, T. Mathematical foundations of elasticity. Prentice-Hall; 1983.

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. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos. Trans. R. Soc. Lond. B Biol. Sci. 2001; 356:1293–1322. [PubMed: 11545704]

Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, Parikshak N, Hua X, Toga AW, Jack CR Jr. Schuff N, Weiner MW, Thompson PM. Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Hum. Brain Mapp. 2009a; 30:2766–2788. [PubMed: 19172649]

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. 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. 2009b; 45:S3–S15. [PubMed: 19041724]

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 2005a; 15:869–877. xi–xii. [PubMed: 16443497]

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement. 2005b; 1:55–66. [PubMed: 17476317]

Nestor SM, Rupsingh R, Borrie M, Smith M, Accomazzi V, Wells JL, Fogarty J, Bartha R. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. Brain. 2008; 131:2443–2454. [PubMed: 18669512]

Pericak-Vance MA, Bebout JL, Gaskell PC Jr. Yamaoka LH, Hung WY, Alberts MJ, Walker AP, Bartlett RJ, Haynes CA, Welsh KA, et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am. J. Hum. Genet. 1991; 48:1034–1050. [PubMed: 2035524]

Petersen RC. Aging, mild cognitive impairment, and Alzheimer's disease. Neurol. Clin. 2000; 18:789–806. [PubMed: 11072261]

Petersen RC. Mild cognitive impairment clinical trials. Nat. Rev. Drug Discov. 2003a; 2:646–653. [PubMed: 12904814]

Petersen, RC. Mild Cognitive Impairment: Aging to Alzheimer's Disease. Oxford University Press; New York: 2003b.

Reiman EM, Chen K, Ayutyanont N, Lee W, Bandy D, Reschke C, Alexander GE, Weiner MW, Koeppe RA, Foster NL, Jagust WJ. Twelve-month cerebral metabolic declines in probable Alzheimer's disease and amnestic mild cognitive impairment: Preliminary findings from the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimers Dement. 2008; 4:T110–T111.

Reuter M, Schmansky NJ, Rosas HD, Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage. 2012; 61:1402–1418. [PubMed: 22430496]

Riddle WR, Li R, Fitzpatrick JM, DonLevy SC, Dawant BM, Price RR. Characterizing changes in MR images with color-coded Jacobians. Magn. Reson. Imaging. 2004; 22:769–777. [PubMed: 15234445]

Rosner, B. Fundamentals of Biostatistics. PWS-Kent Publishing Company; Boston: 1990.

Ross, J.; Thompson, PM.; Tariot, P.; Reiman, EM.; Schneider, L.; Frigerio, E.; Fiorentini, F.; Giardino, L.; Calzà, L.; Norris, D.; Cicirello, H.; Casula, D.; Imbimbo, BP. Primary and Secondary Prevention Trials in Subjects at Risk of Developing Alzheimer's Disease: the GEPARD-AD (Genetically Enriched Population At Risk of Developing Alzheimer's Disease) Studies. CTAD conference; Monte Carlo, Monaco. 2012.

Saunders AM, Schmader K, Breitner JC, Benson MD, Brown WT, Goldfarb L, Goldgaber D, Manwaring MG, Szymanski MH, McCown N, et al. Apolipoprotein E epsilon 4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases. Lancet. 1993a; 342:710–711. [PubMed: 8103823]

Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Alberts MJ, et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology. 1993b; 43:1467–1472. [PubMed: 8350998]

Schachter F, Faure-Delanef L, Guenot F, Rouger H, Froguel P, Lesueur-Ginot L, Cohen D. Genetic associations with human longevity at the APOE and ACE loci. Nat. Genet. 1994; 6:29–32. [PubMed: 8136829]

Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, Thompson PM, Jack CR Jr. Weiner MW. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain. 2009; 132(4):1067–1077. [PubMed: 19251758]

Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging. 1998; 17:87–97. [PubMed: 9617910]

Smith SM, Rao A, De Stefano N, Jenkinson M, Schott JM, Matthews PM, Fox NC. Longitudinal and cross-sectional analysis of atrophy in Alzheimer's disease: cross-validation of BSI, SIENA and SIENAX. NeuroImage. 2007; 36:1200–1206. [PubMed: 17537648]

Spires-Jones TL, Stoothoff WH, de Calignon A, Jones PB, Hyman BT. Tau pathophysiology in neurodegeneration: a tangled issue. Trends Neurosci. 2009; 32:150–159. [PubMed: 19162340]

Thirion JP. Image matching as a diffusion process: an analogy with Maxwell's demons. Medical Image Analysis. 1998; 2:243–260. [PubMed: 9873902]

Thompson PM, Giedd JN, Woods RP, MacDonald D, Evans AC, Toga AW. Growth patterns in the developing brain detected by using continuum mechanical tensor maps. Nature. 2000; 404:190–193. [PubMed: 10724172]

Thompson PM, Hayashi KM, De Zubicaray GI, Janke AL, Rose SE, Semple J, Hong MS, Herman DH, Gravano D, Doddrell DM, Toga AW. Mapping hippocampal and ventricular change in Alzheimer disease. NeuroImage. 2004; 22:1754–1766. [PubMed: 15275931]

Toga, AW. Brain Warping. 1 ed. Academic Press; San Diego: 1999.

Trojanowski JQ, Vandeerstichele H, Korecka M, Clark CM, Aisen PS, Petersen RC, Blennow K, Soares H, Simon A, Lewczuk P, Dean R, Siemers E, Potter WZ, Weiner MW, Jack CR Jr. Jagust W, Toga AW, Lee VM, Shaw LM. Update on the biomarker core of the Alzheimer's Disease Neuroimaging Initiative subjects. Alzheimer's & Dementia : the Journal of the Alzheimer's Association. 2010; 6:230–238.

Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, Boeve BF, Petersen RC, Jack CR Jr. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage. 2008; 39:1186–1197. [PubMed: 18054253]

Weiner MW, Aisen PS, Jack CR Jr. Jagust WJ, Trojanowski JQ, Shaw L, Saykin AJ, Morris JC, Cairns N, Beckett LA, Toga A, Green R, Walter S, Soares H, Snyder P, Siemers E, Potter W, Cole PE, Schmidt M. The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimer's & Dementia : the Journal of the Alzheimer's Association. 2010; 6:202–211. e207.

Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ. The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. Alzheimer's & Dementia : the Journal of the Alzheimer's Association. 2012; 8:S1–68.

Woods RP, Grafton ST, Watson JD, Sicotte NL, Mazziotta JC. Automated image registration: II. Intersubject validation of linear and nonlinear models. J. Comput. Assist. Tomogr. 1998; 22:153–165. [PubMed: 9448780]

Wozniak JR, Lim KO. Advances in white matter imaging: a review of in vivo magnetic resonance methodologies and their applicability to the study of development and aging. Neurosci. Biobehav. Rev. 2006; 30:762–774. [PubMed: 16890990]

Wyman, BT.; Harvey, DJ.; Crawford, F.; Bernstein, MA.; Cole, PE.; DeCarli, C.; Fox, NC.; Gunter, JL.; Hill, D.; Killiany, RJ.; Pachai, C.; Shchwarz, AJ.; Schuff, N.; Suhy, J.; Thompson, PM.; Weiner, M.; Jack, CR, Jr.. Standardization of Analysis Sets for Reporting Results from ADNI MRI data. 2012.

Yanovsky I, Thompson P, Osher S, Leow AD. Asymmetric and symmetric unbiased image registration: statistical assessment of performance. IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis. 2008; 1:1–8.

Yanovsky I, Leow AD, Lee S, Osher SJ, Thompson PM. Comparing registration methods for mapping brain change using tensor-based morphometry. Medical Image Analysis. 2009; 13:679–700. [PubMed: 19631572]

Yushkevich PA, Avants BB, Das SR, Pluta J, Altinay M, Craige C. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3 T MRI data. NeuroImage. 2010; 50:434–445. [PubMed: 20005963]

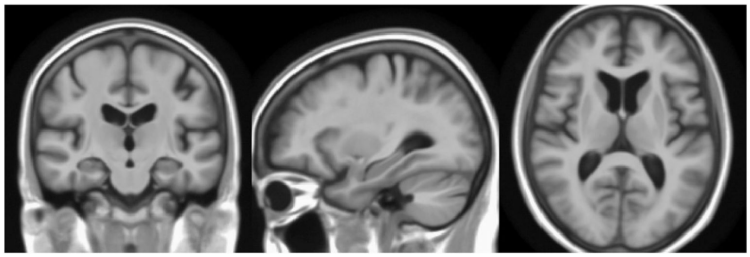

Fig. 1.

High-resolution average group template – the minimal deformation target (MDT). The MDT is shown here using the radiological convention (with slices at x=140, y=110, z=110, in a coordinate system whose image centroid is at (110,110,110)).

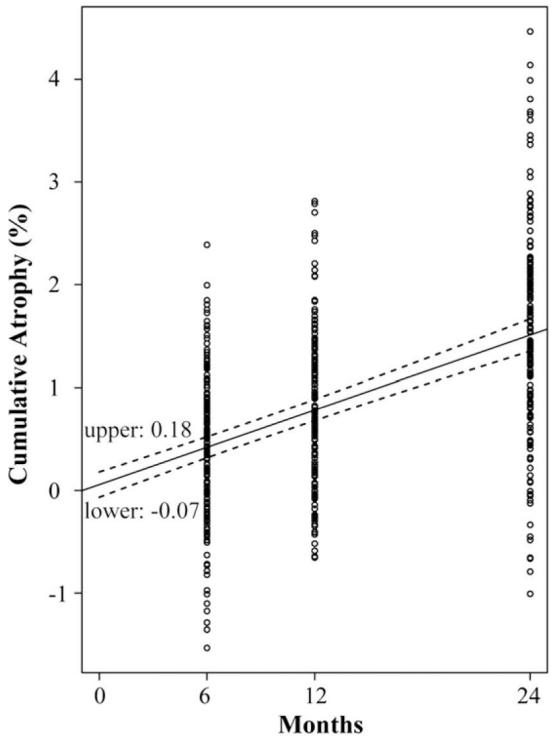

Fig. 2.

Cumulative temporal lobe atrophy demonstrates a linear trend, with an intercept that is essentially zero, in the healthy subjects (*N*=163) with brain scans at screening, 6, 12, and 24 months. Numerical summaries, representing the amount of cumulative temporal lobe atrophy, were fitted against time using a linear mixed effects model. The solid line shows the best linear fit. The dotted lines show the 95% confidence intervals. The plot was generated with *R*, using the *lme4* package. Fitted intercept and confidence intervals were 0.06% [−0.07, 0.18].

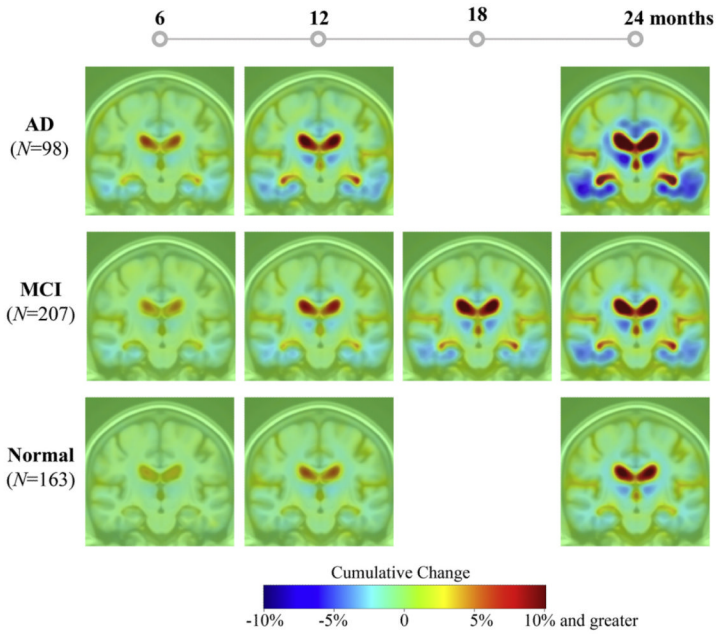

Fig. 3.

Cumulative brain change in the subjects with complete 2-year visits at screening, 6, 12, 18 (MCI only), and 24 months. Warmer (red) colors indicate ventricle/CSF expansion and cooler (blue) colors signify tissue loss.

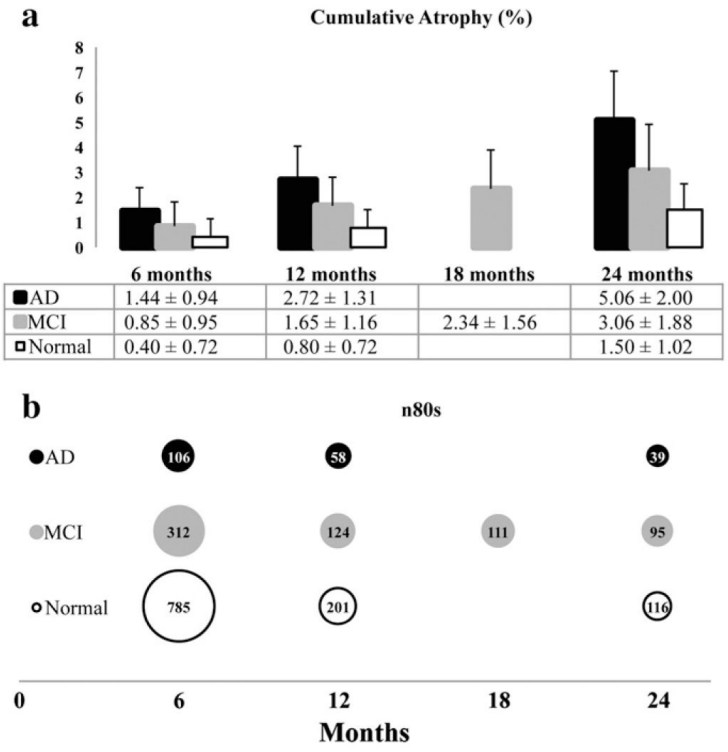

Fig. 4.

Cumulative atrophy and *n80* estimates. (a) The average amount of cumulative temporal lobe atrophy and its standard deviation were computed for AD, MCI, and normal groups, at 6, 12, 18 (MCI only), and 24 months. (b) *n80* estimates show the number of subjects necessary to detect a 25% reduction in average change in a hypothetical clinical trial with a 6, 12, 18 or 24-month duration (α=0.05, *power*=80%; see Discussion for assumptions).

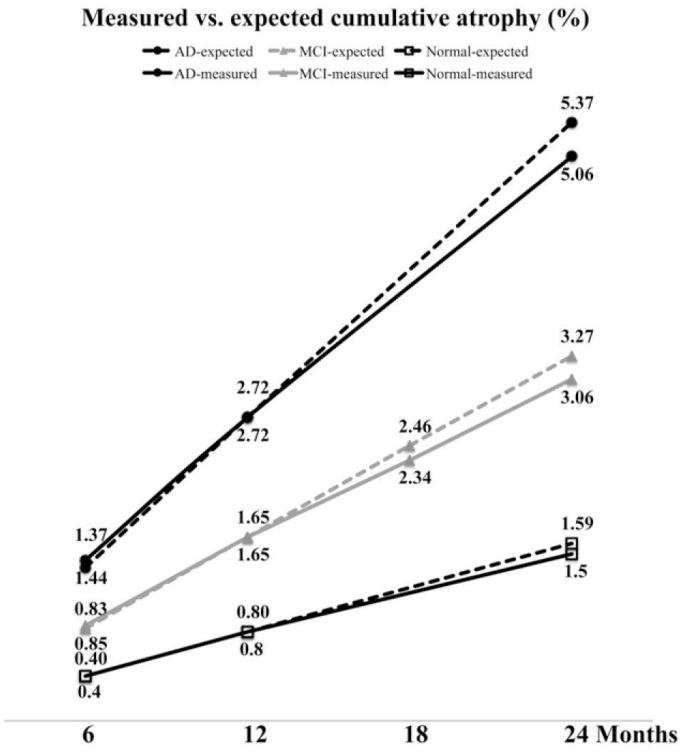

Fig. 5.

"Measured cumulative atrophy" (solid line) compared to "expected cumulative atrophy" (dotted line) based on an assumption of constant rate. Brain atrophy rates were estimated using the 12-month cumulative atrophy measures of subjects with complete 2-year visits.

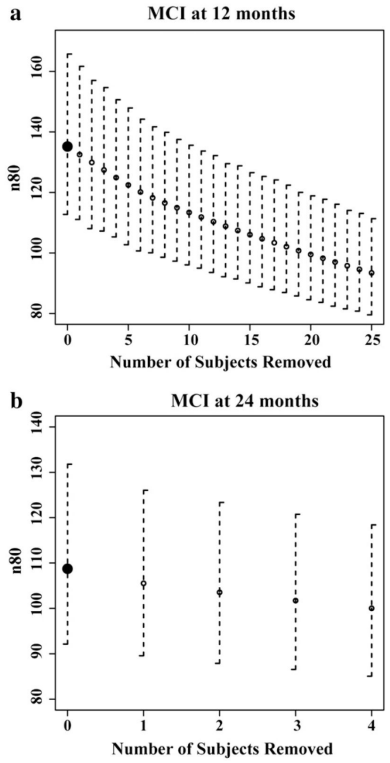

Fig. 6.

Effect of removing subjects with positive numerical summaries on apparent sample size requirements (*n80*), with the upper and lower bounds for the confidence intervals, for trials that last 12 (a) or 24 (b) months. Computed sample size requirements are 31% lower when only 7.7% of the scans are removed (25 out of 326 MCIs). Clearly, this vetting would lead to an overly optimistic sample size estimate.

Table 1

Available scans for ADNI-1 (*N*=3314) at March 20, 2012.

	Screening	6Mo	12Mo	18Mo	24Mo	36Mo
AD	188	159	138	*n/a*	105	*n/a*
MCI	400	346	326	286	244	170
Normal	229	208	196	*n/a*	172	147
Total	817	713	660	286	521	317

Table 2

Numerical summaries from subjects with a complete set of visits, up to and including 2 years. Average level of cumulative temporal lobe atrophy (*mean*), standard deviation (*std dev*), sample size estimates (*n80*) and 95% confidence intervals of the *n80* estimates (*c*) are summarized for the group of ADNI1 subjects with complete 2-year visits. Only MCI subjects were scanned at 18-months, so the other two groups have no data at that time-point.

		6Mo	12Mo	18Mo	24Mo
AD (*N*=98)	*mean*	1.44	2.72		5.06
	std dev	0.94	1.31		2.00
	n80	106	58		39
	c	[75,187]	[45,81]		[32,52]
MCI (*N*=207)	*mean*	0.85	1.65	2.34	3.06
	std dev	0.95	1.16	1.56	1.88
	n80	312	124	111	95
	c	[221,556]	[98,160]	[91,140]	[80,120]
Normal (*N*=163)	*mean*	0.40	0.80		1.50
	std dev	0.72	0.72		1.02
	n80	785	201		116
	c	[463,1672]	[151,284]		[90,160]

Table 3

Numerical summaries from the full ADNI-1 dataset. The number of subjects (*N*), average level of cumulative temporal lobe atrophy (*mean*), its standard deviation (*std dev*), sample size estimates (*n80*) and95% confidence intervals of the *n80* estimates (*c*) are summarized for the full ADNI-1 dataset.

		6Mo	12Mo	18Mo	24Mo	36Mo
AD	*N*	159	138	*n/a*	105	n/a
	Mean	1.46	2.70		5.03	
	std dev	0.99	1.37		2.04	
	n80	114	64		41	
	C	[87,166]	[51,86]		[33,55]	
MCI	*N*	346	326	286	244	170
	Mean	0.94	1.68	2.31	2.94	4.15
	std dev	0.93	1.23	1.62	1.94	2.45
	n80	248	135	124	109	87
	C	[194,358]	[114,167]	[105,153]	[92,131]	[74,108]
Normal	*N*	208	196	*n/a*	172	147
	Mean	0.40	0.82		1.48	2.02
	std dev	0.72	0.73		1.01	1.24
	n80	799	198		118	94
	C	[500,1528]	[154,275]		[92,161]	[77,125]

Table 4

Drug trial enrichment based on ApoE genotype. Cumulative atrophy, as a percent of baseline, and *n80* estimates at 24-month follow-up, for ApoE ε4 carriers versus non-carriers. *N*: number of subjects in the category, *atrophy*: cumulative atrophy at the 24-month follow-up (% change±standard deviation), *n80*: sample size estimate for a 24-month trial to detect a 25% reduction in average change with 80% power (two-sided test α=0.05), and *c*: confidence intervals, for *n80* estimates.

	ε4 carriers (ε4/ε3 or ε4/ε4)				non-carriers (ε2/ε2, ε2/ε3 or ε3/ε3)			
	N	***atrophy***	***n80***	***c***	***N***	***atrophy***	***n80***	***c***
AD	69	5.45±1.91	31	[23,41]	33	4.36±2.09	57	[38,93]
MCI	124	3.62±1.95	73	[57,94]	111	2.19±1.67	145	[111,206]
Normal	48	1.67±1.16	122	[80,229]	122	1.38±0.94	116	[87,171]