Contents lists available at ScienceDirect

Neurobiology of Aging

journal homepage: www.elsevier.com/locate/neuaging

High-dimensional morphometry

Empowering imaging biomarkers of Alzheimer's disease

Boris A. Gutman [a], Yalin Wang [b], Igor Yanovsky [c], Xue Hua [a], Arthur W. Toga [h], Clifford R. Jack Jr [d], Michael W. Weiner [e,f,g], Paul M. Thompson [a,h,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] USC Imaging Genetics Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[b] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA
[c] UCLA Joint Institute for Regional Earth System Science and Engineering, Los Angeles, CA, USA
[d] Mayo Clinic, Rochester, MN, USA
[e] Department of Radiology and Biomedical Imaging, UC San Francisco, San Francisco, CA, USA
[f] Department of Medicine, UC San Francisco, San Francisco, CA, USA
[g] Department of Psychiatry, UC San Francisco, San Francisco, CA, USA
[h] Institute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Received 6 May 2013
Received in revised form 22 May 2014
Accepted 23 May 2014
Available online 27 August 2014

Keywords:
Linear discriminant analysis
Shape analysis
Tensor-based morphometry
ADNI
Lateral ventricles
Alzheimer's disease
Mild cognitive impairment
Biomarker
Drug trial
Machine learning

ABSTRACT

In a previous report, we proposed a method for combining multiple markers of atrophy caused by Alzheimer's disease into a single atrophy score that is more powerful than any one feature. We applied the method to expansion rates of the lateral ventricles, achieving the most powerful ventricular atrophy measure to date. Here, we expand our method's application to tensor-based morphometry measures. We also combine the volumetric tensor-based morphometry measures with previously computed ventricular surface measures into a combined atrophy score. We show that our atrophy scores are longitudinally unbiased with the intercept bias estimated at 2 orders of magnitude below the mean atrophy of control subjects at 1 year. Both approaches yield the most powerful biomarker of atrophy not only for ventricular measures but also for all published unbiased imaging measures to date. A 2-year trial using our measures requires only 31 (22, 43) Alzheimer's disease subjects or 56 (44, 64) subjects with mild cognitive impairment to detect 25% slowing in atrophy with 80% power and 95% confidence.

1. Introduction

Imaging biomarkers of Alzheimer's disease (AD) must offer sufficient power to detect brain atrophy in subjects scanned repeatedly over time (Cummings, 2010; Ross et al., 2012; Wyman et al., 2012). The expected cost of a drug trial may be prohibitively high, unless we can reasonably expect disease-slowing effects to be detected quickly enough and with reasonably few subjects. Imaging measures from standard structural magnetic resonance imaging (MRI) show considerable promise. Their use stems from the premise that longitudinal changes may be more precisely and reproducibly measured with MRI than comparable changes in clinical, cerebrospinal fluid (CSF), or proteomic assessments; clearly, whether that is true depends on the measures used. The use of MRI in a drug trial has some caveats; most MR studies from published drug trials have detected no effect or even a small, and possibly irrelevant but significant, increase in atrophy in the treatment group. Brain measures that are helpful for diagnosis, such as positron emission tomography (PET) scanning, may not be stable for large multicenter (N = several hundred) longitudinal trials that aim to slow disease progression. Other markers, such as CSF measures of amyloid and tau proteins to assess brain amyloid, may suffer the opposite problem of showing too little change during the clinical AD period. As a result, there is interest in testing the reproducibility of biomarkers, as well as methods to optimally combine them (Yuan et al., 2012).

Recent studies have tested the reproducibility and accuracy of a variety of MRI-derived measures of brain change. Several of these are highly correlated with clinical assessments and can predict

[*] Corresponding author at: Laboratory of Neuro Imaging, Department of Neurology, UCLA School of Medicine, Imaging Genetics Center, Neuroscience Research Building 225E, 635 Charles Young Drive, Los Angeles, CA 90095-1769, USA. Tel.: +1 310 206 2101; fax: +1 310 206 5518.
E-mail address: pthomp@usc.edu (P.M. Thompson).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

0197-4580/$ – see front matter © 2015 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neurobiolaging.2014.05.038

future decline on their own or in combination with other relevant measures. Although not the only important consideration, some analyses have assessed which MRI-based measures show greatest effect sizes for measuring brain change over time, while avoiding issues of bias and asymmetry that can complicate longitudinal image analysis (Fox et al., 2011; Holland et al., 2011; Hua et al., 2013), and while avoiding removing scans from the analysis that may lead to unfairly optimistic sample size estimates (Hua et al., 2013; Wyman et al., 2012). Promising MRI-based measures include the brain boundary shift integral (Leung et al., 2012; Schott et al., 2010), the ventricular boundary shift integral (Schott et al., 2010), and measures derived from anatomic segmentation software such as Quarc or FreeSurfer, some of which have been recently modified to handle longitudinal data more accurately (Fischl and Dale, 2000; Holland and Dale, 2011; Reuter et al., 2012; Smith et al., 2002).

Although several power estimates are possible, the analysis advocated by the Alzheimer's Disease Neuroimaging Initiative (ADNI) Biostatistics Core (Beckett, 2000) is to estimate the minimal sample size required to detect, with 80% power, a 25% reduction in the mean annual change, using a 2-sided test and standard significance level $\alpha = 0.05$ for a hypothetical 2-arm study (treatment vs. placebo). The estimate for the minimum sample size is computed from the formula below. $\widehat{\beta}$ denotes the annual change (average across the group) and $\widehat{\sigma}_D^2$ refers to the variance of the annual rate of change.

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

Here, z_α is the value of the standard normal distribution for which $P[Z < z_\alpha] = \alpha$ the sample size required to achieve 80% power is commonly denoted by n80. Typical n80s for competitive methods are under 150 AD subjects and under 300 mild cognitive impairment (MCI) subjects; the larger numbers for MCI reflect the fact that brain changes tend to be slower in MCI than AD, and MCI is an etiologically more heterogeneous clinical category. For this reason, it is harder to detect a modification of changes that are inherently smaller, so greater sample sizes are needed to guarantee sufficient power to detect the slowing of disease.

Many algorithms can detect localized or diffuse changes in the brain, creating detailed 3D maps of changes (Avants et al., 2008; Leow et al., 2007; Shi et al., 2009), but the detail in the maps they produce is often disregarded when making sample size estimates according to Equation 1 as the formula expects a single univariate measure of change. In other words, it requires a single number or "numeric summary" to represent all the relevant changes occurring within the brain. To mitigate this problem, Hua et al. (2009) defined a "statistical ROI" based on a small sample of AD subjects by thresholding the *t*-statistic of each feature (voxel) and summing the relevant features over the ROI; this approach was initially advocated in the FDG-PET literature to home in on regions that show greatest effects (Chen et al., 2010). In spirit, the statistical ROI is a rudimentary supervised learning approach, as it finds regions that show detectable effects in a training sample and uses them to empower the analysis of future samples; the samples used are nonoverlapping and independent to avoid circularity. However, a simple threshold-based masking is known to potentially eliminate useful features as binarization loses a lot of the information present in continuous weights (Duda et al., 2001). Although many studies have used machine learning to predict the progression of neurodegenerative diseases and differentiate diagnostic groups such as AD, MCI, and controls (Kloppel et al., 2012; Kohannim et al., 2010; Vemuri et al., 2008), we found no attempts in the literature that used learning to directly optimize power to detect brain change.

To address this issue, we observed that minimizing Equation 1 is exactly analogous to one-class linear discriminant analysis (LDA). We applied the method to surface-based longitudinal expansion rates of the ventricular boundary (Gutman et al., 2013), achieving the lowest sample size estimates of any ventricle-based measure of AD to date, both in terms of absolute and control-adjusted atrophy. Here, we apply the LDA-based weighting to recently reported maps of whole brain volume change based on tensor-based morphometry (Hua et al., 2013). Further, we combine ventricular surface and tensor-based morphometry (TBM) volume measures into one combined atrophy score. Our results show a marked improvement over the stat-ROI approach, achieving substantively lower sample size estimates than any ADNI-based report to date.

2. Methods

2.1. Alzheimer's Disease Neuroimaging Initiative

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and nonprofit organizations as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. 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. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 adults, aged 55–90 years, 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.

Longitudinal brain MRI scans (1.5 Tesla) and associated study data (age, sex, diagnosis, genotype, and family history of AD) were downloaded from the ADNI public database (http://www.loni.ucla.edu/ADNI/Data/) on July 1, 2012. The first phase of ADNI, that is, ADNI-1, was a 5-year study launched in 2004 to develop longitudinal outcome measures of Alzheimer's progression using serial MRI, PET, biochemical changes in CSF, blood, and urine, and cognitive and neuropsychological assessments acquired at multiple sites similar to typical clinical trials.

All subjects underwent thorough clinical and cognitive assessment at the time of scan acquisition. All AD patients met NINCDS/ADRDA criteria for probable AD (McKhann et al., 1984). The ADNI protocol lists more detailed inclusion and exclusion criteria (Mueller et al., 2005a, 2005b), available online (http://www.alzheimers.org/clinicaltrials/fullrec.asp?PrimaryKey=208). The study was conducted according to the Good Clinical Practice guidelines, the Declaration of Helsinki and the United States, 21 CFR Part 50-Protection of Human Subjects and Part 56-Institutional Review Boards. Written informed consent was obtained from all participants before performing experimental procedures, including cognitive testing.

2.2. 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). Additional data were collected at 3-T but is not used here as it was only collected on a subsample that is too small for making comparative assessments of power. Using a sagittal 3D MP-RAGE scanning protocol, the typical acquisition parameters were repetition time of 2400 ms, minimum full echo time, inversion time 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 improve the chance that at least 1 scan would be usable for analysis and for signal averaging if desired.

The scan quality was evaluated by the ADNI MRI quality control center at the Mayo Clinic to exclude failed scans because of 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 4 steps: (1) correction of geometric distortion because of gradient nonlinearity (Jovicich et al., 2006), that is, “gradwarp”; (2) “B1-correction” for adjustment of image intensity inhomogeneity because of B1 nonuniformity (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).

2.3. The ADNI-1 data set

For our experiments, we analyzed data from 683 ADNI subjects with baseline and 1 year scans, and 542 subjects with baseline, 1 year, and 2 years scans. The former group consisted of 144 AD subjects (age at screening: 75.5 ± 7.4 years, 67 females [F], and 77 males [M]), 337 subjects with MCI (74.9 ± 7.2 years, 122 F and 215 M), and 202 age-matched healthy controls (NC) (76.0 ± 5.1 years, 95 F and 107 M). The 2-year group (i.e., people with scans at baseline and after a 1-year and 2-year interval) had 111 AD (75.7 ± 7.3, 52 F and 59 M), 253 MCI (74.9 ± 7.1, 87 F and 166 M), and 178 NC (76.2 ± 5.2, 85 F and 93 M) subjects. All raw scans, images with different steps of corrections, and the standard ADNI-1 collections are available to the general scientific community at http://www.loni.ucla.edu/ADNI/Data/. We used exactly all ADNI subjects available to us (on February 1, 2012) who had both baseline and 12 months scans, and all subjects with 24 months scans (available July 1, 2012) (Table 1). The use of all subjects without data exclusion has been advocated by Wyman et al. (2012) and Hua et al. (2013), because any scan exclusion can lead to power estimates that are unfairly optimistic, and many drug trials prohibit the exclusion of any scans at all.

2.4. Surface extraction and analysis

Our surfaces were extracted from 9-parameter affine-registered, fully processed, T1-weighted anatomic scans. We used a modified version of Chou registration-based segmentation (Chou et al., 2008), using inverse-consistent fluid registration with a mutual information fidelity term (Leow et al., 2007). To avoid issues of bias and nontransitivity, we segmented each of our subjects' 2 or 3 scans separately. In this approach, a set of hand-labeled “templates” are aligned to each scan, with multiple atlases being used to greatly reduce error. There were 2 templates from each of the 3 diagnostic groups, with 1 male and 1 female subject in each. The template surfaces were registered as a group following a medial-spherical registration method (Gutman et al., 2012). To improve on the standard multi-atlas segmentation, which generally involves a direct or a weighted average of the warped binary masks, we selected an individual template that best fits the new boundary at each boundary point. A naïve formulation of this synthesis can be written as:

$$S(p) = \sum_i W^i(p)T_i(p), \quad W^i(p) = \begin{cases} 1 \; if \; s(I,I_i)[p] > s(I,I_j)[p] \, \forall j \neq i \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{2}$$

Here, I, S are the new image and boundary surface, $\{I_i, Ti\}i$ are template surfaces and images warped to the new image, and $s(I, \mathrm{I}_{i\mathrm{l}})$ $[p]$ is some local normalized similarity measure at point p. Normalized mutual information around a neighborhood of each point was used to measure similarity. This approach allows for more flexible segmentation, in particular for outlier cases. Even a weighted average, with a single weight applied to each individual template, often distorts geometric aspects of the boundary that are captured in only a few templates, perhaps only in one. However, to enforce smoothness of the resulting surface, care must be taken around the boundaries of the surface masks W^i. An effective approach is to smooth the masks with a spherical heat kernel so that our final weights are $W^i_\sigma(q) = \int_{\mathbb{S}^2} K_\sigma(p,q)W^i(p)\mathrm{d}p$. This approach is similar to Yushkevich et al. (2010b), differing mainly in the fact that it is a surface-based rather than a voxel-based approach.

Local surface-based maps of atrophy were then generated using the algorithm described in (Gutman et al., 2012, 2013). Briefly, the algorithm deforms a curve to minimize the medial energy associated with the shape, which may be written as:

$$R(c,c',\mathscr{M}) = \int_0^1 \int_{p \in \mathscr{M}} w(c(t),c'(t),p,\mathscr{M})|c(t) - \; p|^2 d\mathscr{M}\mathrm{d}t \tag{3}$$

The term $w(c(t),c'(t),p,\mathscr{M})$ represents the medial weight for each pair of curve and surface points, which is described in detail in Gutman et al. (2012). Two surface-based feature functions are generated based on the curve representing shape geometry: thickness and the global orientation function (Gutman et al., 2012). We nonlinearly register shapes, first longitudinally and then to a mean template by parametrically minimizing sum of square differences between corresponding feature functions. Our mean template is generated by averaging the hand-traced templates in a groupwise fashion as described in Gutman et al. (2012). The thickness change maps represent change in the distance to the medial axis from any given point on the ventricular boundary or intuitively change in thickness of the shape.

2.5. Tensor-based morphometry

TBM is an image analysis technique that measures brain structural differences from the gradients of deformation fields that align 1 image to another (Ashburner and Friston, 2003; Freeborough and Fox, 1998; Leow et al., 2007). 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 nonlinear inverse consistent elastic intensity-based registration algorithm (Leow et al., 2007), which optimizes a joint cost function based on mutual information 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 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 and/or 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 minimal deformation template, for regional comparisons and group statistical analyses. See Hua et al. (2013) for more details.

2.6. LDA for empowering biomarkers

In designing an imaging biomarker, one generally seeks to balance the intuitiveness of the measure and its power to track disease progression. In this study, we choose to use, alternatively, radial expansion of the lateral ventricles, local tissue loss as measured by Jacobian determinants of nonlinear longitudinal warps, or the combination of the two. Having made this choice, we would now like to find an optimal linear weighting for each surface vertex and image voxel to maximize the effect size of our combined global measure of change. A linear model may not have the intuitive clarity of a binary weighting (i.e., specifying or masking a restricted region to measure), but its meaning is still sufficiently clear and can be easily visualized. Thus, we would like to minimize our sample size estimate as a function of the weights, w:

$$n(w) = C\frac{\frac{1}{N-1}\sum\left(x_i^{\mathrm{T}}w - m^{\mathrm{T}}w\right)^2}{\left(m^{\mathrm{T}}w\right)^2} = \frac{1}{N-1}C\frac{w^{\mathrm{T}}S_W w}{w^{\mathrm{T}}S_B w} \tag{4}$$

Here $C = 32\left(z_{1-\alpha/2} + z_{power}\right)^2$, x_i is the thickness change for the ith subject, m is the mean vector, the covariance matrix $S_W = \sum_{i=1}^{N}(x_i - m)(x_i - m)^T$ and $S_B = mm^{\mathrm{T}}$. Minimizing Equation 4 is equivalent to maximizing.

$$J(w) = \frac{w^{\mathrm{T}}S_B w}{w^{\mathrm{T}}S_W w} \tag{5}$$

which is a special case of the LDA cost function, with a maximum given by

$$w = S_W^{-1}m \tag{6}$$

For our purposes, m represents the mean of the diseased group. We denote this by $m = m_{AD,\,MCI}$, where $m_{AD,\,MCI}$ stands for the mean expansion vector in the combined MCI and AD group. We make no distinction between these 2 groups during LDA training. Maximizing Equation 5 directly is generally not stable when S_W has a high condition number. Further, when the feature space is large enough, as in the case of Jacobian fields with roughly 2 million features, storing the dense 2 M × 2 M covariance matrix directly simply becomes impossible. We resolve this issue by applying principal components analysis (PCA) to our training sample, storing the first k principal components in the rows of a matrix, P and computing the corresponding k eigenvalues λ_j. This is a standard approach when applying LDA to actual 2-class problems, as it makes the mixed covariance matrix nearly diagonal. In our case, the covariance in PCA space is exactly diagonal, which reduces Equation 6 to a direct computation:

$$w = P^{\mathrm{T}}\omega, \quad \text{where} \quad \omega_j = [Pm]_j \big/ \lambda_j \tag{7}$$

This approach is very fast. One can compute the first k eigenvectors and eigenvalues of S_W without explicitly computing S_W itself. Although alternative, possibly more flexible basis function sets are possible, we choose PCA for its simplicity.

The order of subjects in each diagnostic group is randomly changed to eliminate the confound because of different scanning protocols at different ADNI acquisition sites. This step is needed mainly to ensure a roughly equal distribution of sites in each fold, as ADNI subjects are ordered by site by default. Where the subjects are scanned is known to correlate with reliability in many morphometric measures, and we have found that our LDA measures are affected by the site distribution as well. This is only done once before LDA training, with the same order and same subdivision of diagnostic groups used for each method.

To validate our data-driven weighting approaches, we create 2 groups of equal size, with an equal number of MCI and AD subjects in each. Each of these folds is then used to optimize the number of principal components k. This is done by subdividing the training fold further into 2 subfolds of equal size, computing principal components separately on each subfold, and training a different LDA model using all PC's up to k, with k varying from 1 to the total number of subjects in the subfold. A sample size for each subfold's model is computed by applying the linear weights to the other subfold. The optimal k is chosen so that the mean of the 2 subfolds' sample size estimates is minimized. Further, to avoid circularity, we do not use the 6 hand-traced subjects used in generating the ventricular surface template for model training or testing. For TBM, such circularity is avoided entirely, as the minimum distance template (Hua et al., 2008) is based on 40 control subjects, which are not used during the training or testing stages. This approach is an adaptation of the standard nested cross-validation technique in machine learning.

Because Jacobian determinants have a skewed distribution due to the nature of the measurement, we perform LDA training on the logarithm of the Jacobian maps, which in the first approximation is equivalent to actual atrophy rates over a given time interval. This step ensures that the Gaussian assumption in LDA is more closely satisfied.

3. Results

In the following, we compare the performance of our LDA-based vertex weighting of ventricular expansion (Medial Vent LDA), the LDA-based voxel weighting of TBM maps (TBM-LDA), the combination of the 2 LDA measures into 1 score and the LDA Stat-ROI method previously reported in Hua et al. (2013). Although in general, absolute ventricular expansion may not be specific to AD pathology; its finely resolved surface-based signature is used here as a surrogate

Table 1
Available scans for ADNI-1 on February 1, 2012 for 12 months and July 1, 2012 for 24 months. Total number of scans used: N = 2065

	Screening	12 mo	24 mo
AD	200	144	111
MCI	408	337	253
Normal	232	202	178
Total	840	683	542

Key: AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; MCI, mild cognitive impairment.

Table 2
Sample size estimates for clinical trials using anatomic biomarkers of change over 12 months as an outcome measure

	MCI	AD	Mean MCI	Mean AD
Vent-LDA	111/96 (85–150)/(75–127)	65/86 (46–92)/(64–128)	104 (94–139)	75 (64–102)
TBM-LDA Whole	85/99 (67–110)/(77–131)	48/50 (34–70)/(35–85)	92 (77–111)	49 (38–66)
TBM-LDA GM	110/93 (85–145)/(73–122)	48/49 (33–74)/(35–76)	101 (84–122)	49 (37–64)
Vent + TBM	83/72 (66–112)/(56–92)	41/46 (28–65)/(32–68)	**78 (63–90)**	**43 (33–58)**
TBM stat-ROI	—	—	135 (114–167)	64 (51–86)

Depending on how we weigh the features on the ventricular surfaces, the sample size estimates can be reduced, and the power of the study increased. “Whole” stands for whole-brain TBM of Fig. 1 and “GM” means the TBM model restricted to gray matter from Fig. 5. Mean sample size estimates are computed as the average of the 2 folds' estimates. The values in parentheses represent 95% confidence intervals.
The lowest sample size estimates for each group are in bold.
Key: AD, Alzheimer's disease; GM, gray matter; MCI, mild cognitive impairment; ROI, region of interest; TBM, tensor-based morphometry.

measure of AD-related atrophy in addition to what can be learned from TBM. In testing each of these weighting methods, we used nested 2-fold cross-validation. Only AD and MCI subjects were used in the training stage. Further, we restricted our training sample to include only 1-year changes. Twenty-four month data was only used for testing, applying 1-year models to the nonoverlapping subgroups of the 24-month data. Tables 2 and 3 summarize sample size estimates for 1-year and 2-year clinical trials for each of the 4 biomarkers. The linear weight maps are visualized in Figs. 1 and 2. To visualize the difference between a multivariate approach and a mass-univariate type of weighting as done in the stat-ROI approach, we also display maps of t-statistics in Figs. 3 and 4. The t-maps were computed to test the null hypothesis that no change takes place among the AD and MCI subjects at each spatial location over 1 year. In another test, we restricted the PCA feature space to the gray matter voxels, segmented by BrainSuite (Shattuck et al., 2001) and computed the resulting power estimates. The weight maps are visualized in Fig. 5. To assess the reproducibility of our sample sizes, we also computed bootstrapped 95% confidence intervals for our sample size estimates (DiCicio and Efron, 1996).

For ventricular surface measures, the optimal number of principal components was found to be 28 and 47, for folds 1 and 2, respectively. For Jacobian maps, the smallest sample size was achieved at $k = 115$ and 103 for whole-brain LDA and at $k = 98$ and 95 for LDA restricted to gray matter.

We compared the sample size estimates of the stat-ROI approach with TBM-LDA in Table 4. The LDA measures significantly outperformed the stat-ROI measure for MCI subjects and trended better for AD subjects.

To assess whether there is any evidence of longitudinal bias of our weighted measures, we applied our 1-year models to healthy controls at 12 and 24 months. Using a method similar to Hua et al. (2011), we used the y-intercept of the linear regression as a measure of bias (bearing in mind the caveats noted that there may be some biological acceleration or deceleration that could appear to be a bias). We again used bootstrapping to estimate the intercept and linear fit confidence intervals (DiCicio and Efron, 1996), with the exception of TBM stat-ROI, which we reprint from Hua et al. (2013). We note that using standardized linear fit model, CI's leads to intervals that are more than twice as wide for the LDA models, implying that our CI's are quite conservative. Fig. 6 shows the regression plots for all LDA models over the 2 follow-up time points. Confidence intervals for the linear fits are shown in dotted green lines. The bias test results are summarized in Table 5. We note that the intercept shows virtually zero bias for all the LDA models, as it is 2 orders of magnitude lower than change in controls at 1 year.

4. Discussion

Here, we continued the effort started in Gutman et al. (2013) to increase the efficiency of clinical trials in AD and MCI, based on multiple neuroimaging features. We applied a 1-class linear discriminant analysis to a set of TBM features as well as a combination of TBM and ventricular surface features. Based on a nonparametric comparison, the resulting sample size estimates are significantly better than the stat-ROI approach, which has been the standard feature weighting method to date. The linear feature weighting also produces an intuitive, univariate measure of change—a single number summary that can be correlated to other relevant variables and outcome measures. The linear weights can be easily visualized, adding insight into the pattern and 3D profile of disease progression.

4.1. Machine learning in AD

Machine learning has been applied to classify AD and MCI subjects based on brain images in many studies. Fan et al. (2008) applied Support Vector Machine (SVM) to RAVENS maps, an approach similar to modified Voxel-Based Morphometry (Good et al., 2002), incorporating partial tissue classification and a high-dimensional nonlinear volume registration. Vemuri et al. (2008) used a similar method with tissue probability maps. Kloppel et al. (2008) further showed that this linear model is stable across different data sets. In general, classification algorithms can achieve AD-NC cross-validation accuracy in the mid 90s (approximately 95%) within the same data set, although performance inevitably degrades when applied to new data sets because of differences in demographics and scanning protocols.

Table 3
Sample size estimates for clinical trials, using anatomical biomarkers of change over 24 months as an outcome measure

	MCI	AD	Mean MCI	Mean AD
Vent-LDA	80/62 (65, 108)/(44, 86)	67/47 (47, 122)/(31, 67)	71 (65, 98)	57 (45, 89)
TBM-LDA Whole	61/64 (47, 81)/(50, 81)	28/33 (19, 44)/(21, 56)	63 (52, 75)	**31 (22, 43)**
TBM-LDA GM	73/66 (58, 92)/(51, 88)	38/31 (25, 60)/(19, 51)	69 (57, 81)	34 (25, 47)
Vent + TBM	53/58 (40, 72)/(46, 73)	28/34 (19, 43)/(22, 62)	**56 (44, 64)**	32 (22, 44)
TBM stat-ROI	—	—	109 (92, 131)	41 (33, 55)

The values in parentheses represent 95% confidence intervals.
The lowest sample size estimates for each group are in bold.
Key: AD, Alzheimer's disease; GM, gray matter; MCI, mild cognitive impairment; ROI, region of interest; TBM, tensor-based morphometry.

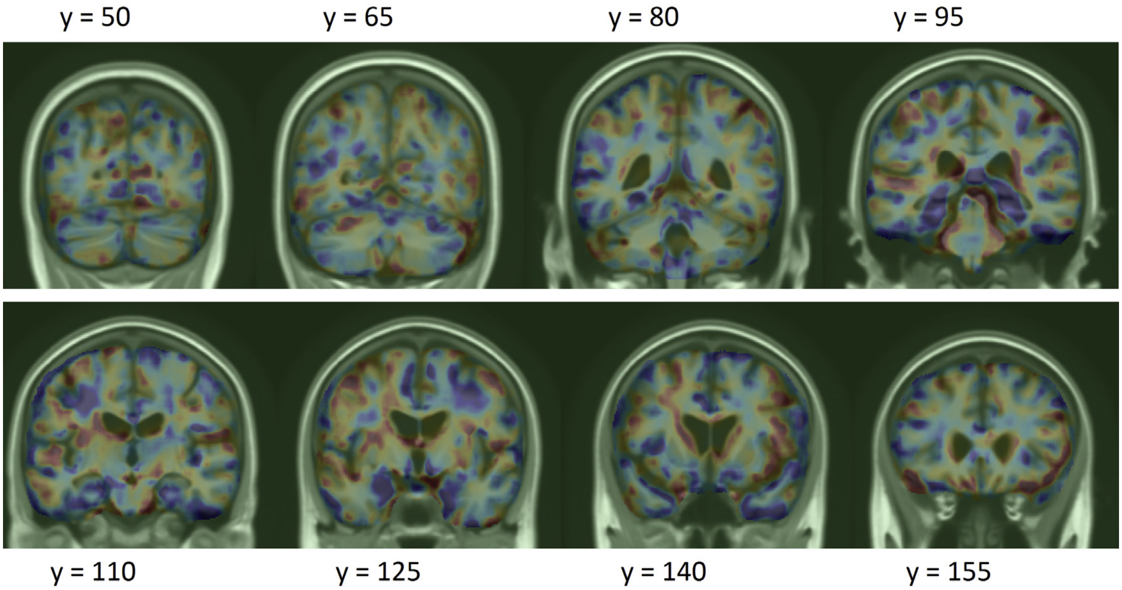

Fig. 1. Log-Jacobian (TBM) LDA weighting, scaled by standard deviation of the weights. Red regions expect expansion, and blue regions expect atrophy. Abbreviation: TBM, tensor-based morphometry. (For interpretation of the references to color in this Figure, the reader is referred to the web version of this article.)

Cuingnet et al. (2010) developed a Laplacian-regularized SVM approach for classifying AD and NC subjects, which bears similarity to our Tikhonov-regularized LDA (Gutman et al., 2013). The Laplacian regularizer is shown to improve classification rates for AD versus NC subjects. SVM has also been used, in our prior work, to separate AD and NC subjects based on hippocampal shape invariants and spherical harmonics (Gutman et al., 2009). Cho et al. (2012) smoothed surface atlas-registered cortical thickness data with a low-pass filter of the Laplace-Beltrami operator. Following this procedure, PCA was performed on the smoothed data, and LDA was applied on a subset of the PCA coefficients to train a linear classifier. The resulting classification accuracy is very competitive. Another surface-based classifier (Gerardin et al., 2009) uses the SPHARM-PDM approach to classify AD and NC subjects based on hippocampal shape. SPHARM-PDM (Styner et al., 2005) computes SPHARM coefficients based on an area-preserving spherical parameterization and defines correspondence via the first-order ellipsoid. This leads to a basic surface registration and a spectral shape decomposition. Gerardin et al. (2009) reported competitive classification rates compared with whole-brain approaches. Shen et al. (2010) used a Bayesian feature selection approach and classification on cortical thickness data, showing competitive AD-NC and MCI-NC classification accuracy with SVM. Zhang et al. (2011) developed a multiple kernel SVM classifier to further improve diagnostic multimodality AD and MCI classification.

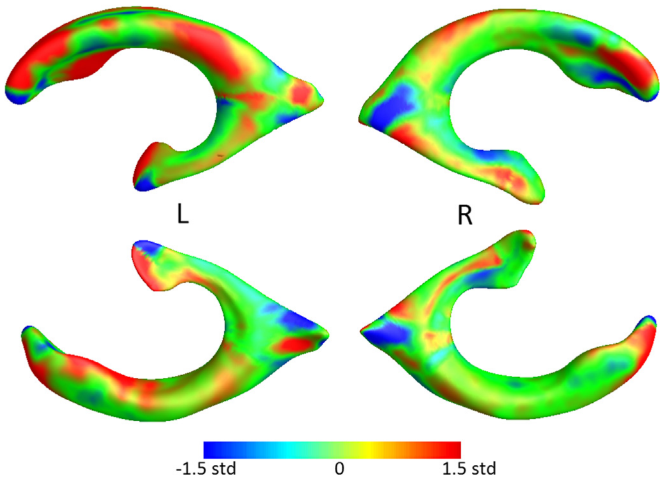

Fig. 2. Ventricular LDA weighting, scaled by standard deviation of the weights.

4.2. *Classifiers and biomarkers*

It is important to stress that although many studies have used machine learning to derive a single measure of “AD-like” morphometry for discriminating AD and MCI subjects from the healthy group; no study, we are aware of, has used machine learning to maximize the power of absolute atrophy rates in AD. We have attempted this by using a straightforward application of LDA. The fundamental difference between classification accuracy and biomarker reliability lies in the difference of the underlying goals. Regardless of the regularization, the goal in classification is to separate 2 classes of subjects in a generalizable way. As a result, subjects which are most difficult to classify will play a disproportionately large role in defining an atrophy measure. For example, we see that this is true of the 2 most popular classification algorithms: AdaBoost and Support Vector Machines. SVM considers only the “support vectors,” and AdaBoost greedily up-weights the difficult cases.

However, in the context of a drug trial, the main concern is not prediction of disease but the identification of a measurable effect on brain degeneration in the whole population because of a new drug. This difference exists regardless of the fine details of statistical analysis and machine learning algorithms, such as whether the test applied to detect drug effects should make Gaussian assumptions, or whether for example one uses a hard margin or a soft-margin SVM approach. Ultimately, the best classifier may ignore or downplay the very substrate of the diseased population that is most helped by a drug in favor of correctly discriminating the nearly normal-appearing subjects who do not experience the beneficial effect. Good classification accuracy and high biomarker power are, in principle, different goals precisely because a good biomarker

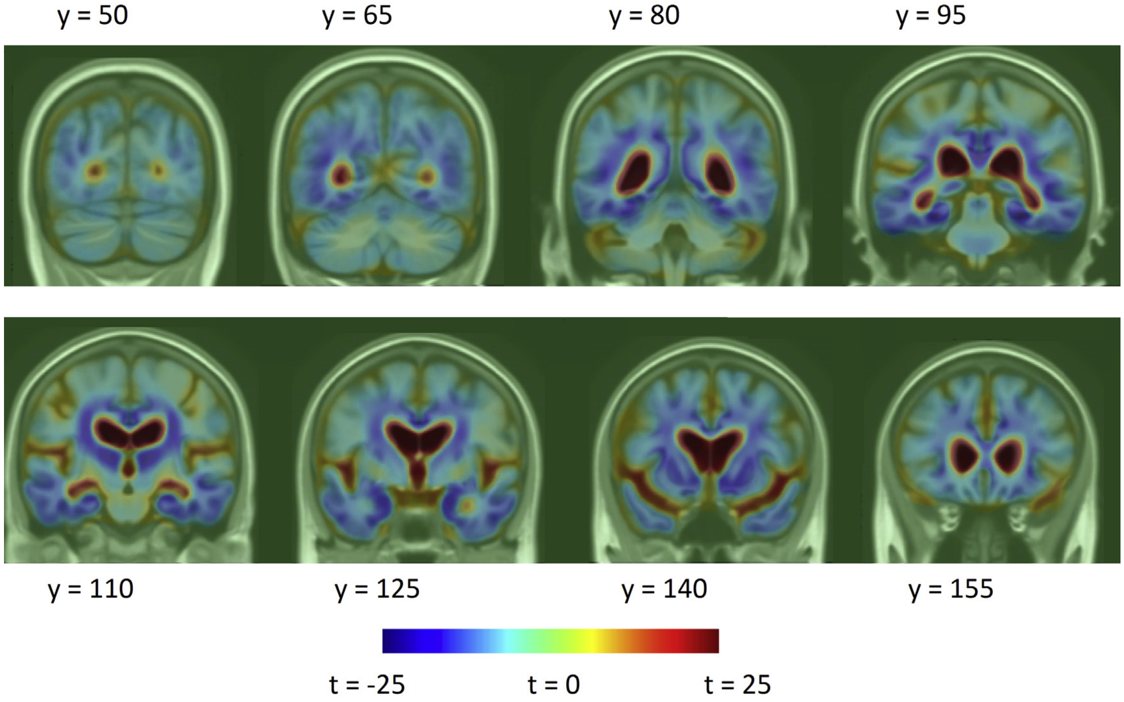

Fig. 3. Log-Jacobian (TBM) t-maps, based on the null hypothesis that there is no change over 1 year in AD and MCI subjects in each voxel. The difference between these maps and Fig. 1 shows the difference between a multivariate and a mass-univariate approach in weighting Jacobian maps. Abbreviations: AD, Alzheimer's disease; MCI, mild cognitive impairment.

must treat all subjects equally. This is why the best classifier will not, in general, be the best biomarker. The requirement for equal treatment of all subjects also implies greater computational burdens when optimizing an imaging biomarker compared with a classifier.

A related question examines whether a Gaussian assumption made in the power estimate is appropriate. Although several arguments can be made on the subject, it must be noted that the assumption is not made by this work, or any other work concerned with biomarker power in ADNI, but by hypothetical trial design itself. Because the trial is based on a test with Gaussian assumptions (Beckett, 2000), the only appropriate power estimate must make the same assumptions as well. In fact, the power estimate used here assesses in part how much a measure's deviation from Gaussianity will affect its sensitivity in the hypothetical test.

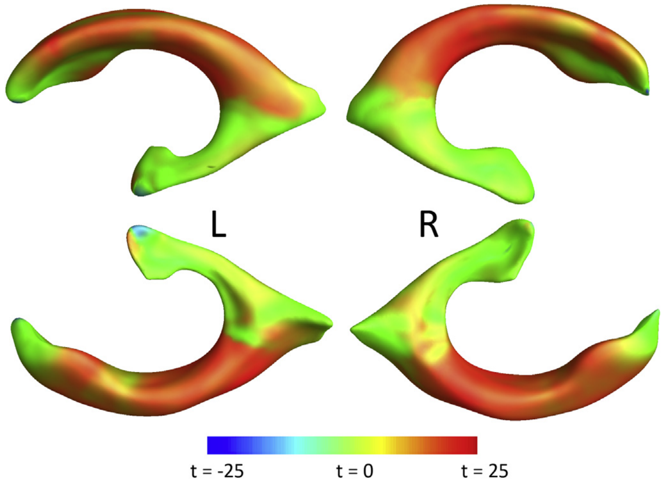

Fig. 4. Ventricular thickness t-maps, based on the null hypothesis that there is no change over 1 year in AD and MCI subjects at each mesh vertex. The difference between these maps and Fig. 2 shows the difference between a multivariate and a mass-univariate approach in weighting Jacobian maps. Abbreviations: AD, Alzheimer's disease; MCI, mild cognitive impairment.

Outside of Alzheimer's literature, we found one approach for explicitly minimizing sample size estimates (Qazi et al., 2010) and another that uses SVM for classification of Huntington disease patients versus controls, with reduced sample sizes as a by-product (Hobbs et al., 2010). The first article is methodologically closest in spirit to this work: a fidelity term is explicitly defined to be the control-adjusted sample size estimate. A number of nonlinear constraints are then added: the total variation norm (TV1-norm), sparsity, and nonnegativity. Although the first 2 have analogs that can be linearly optimized as we do here (TV2 and L^2 norm), the third constraint forces the authors to use nonlinear conjugate gradient, which leads to far slower convergence. More importantly, because of the differences in the nature of their data—knee cartilage CT images—and ours, the sparsity and nonnegativity constraints are perhaps not appropriate for brain imaging. We expect the effect over soft tissue to be diffuse without many discontinuities, and nonnegativity is generally not appropriate in brain MR either. This is because of the fact that we expect some brain regions to grow and others to shrink over time. Further, conjugate gradient optimization would be impossibly slow to apply to brain MR images with millions of features, although it may still make sense to do for the far sparser knee CT images. The second article (Hobbs et al., 2010) uses leave-one-out linear SVM weighting of fluid registration-based TBM maps to derive an atrophy measure. No spatial regularization or sample size-specific modification to the learning approach is used. In both of these cases, the measure used is based on the difference between the mean of controls and the diseased group, which is not the main goal of the present work. Our main contribution, absent

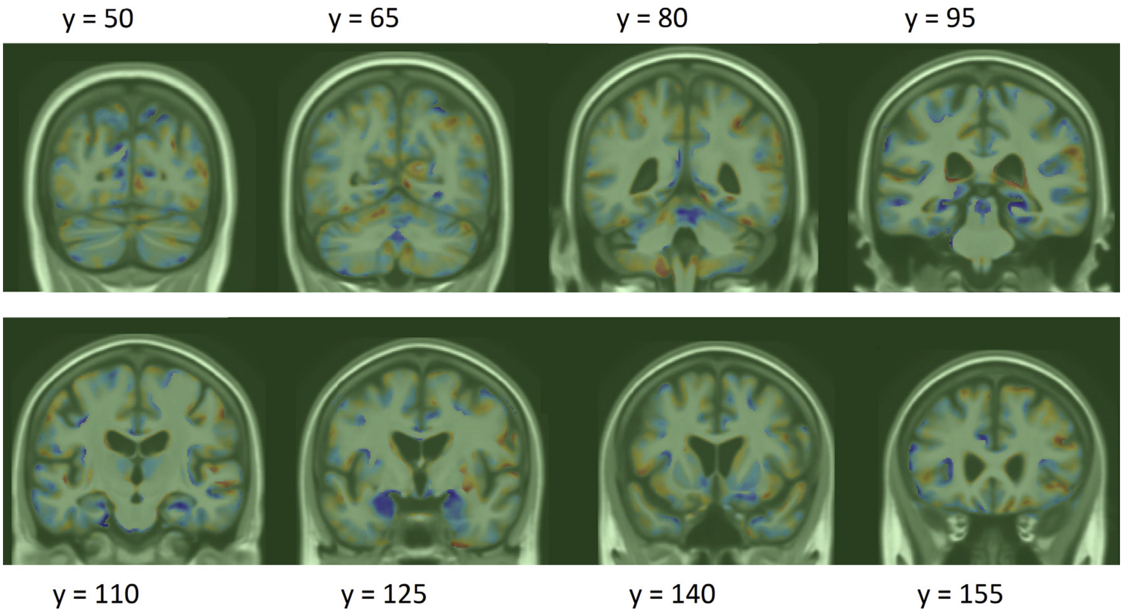

Fig. 5. Log-Jacobian (TBM) LDA weighting restricted to gray matter regions, scaled by standard deviation of the weights. Red regions expect expansion, and blue regions expect atrophy. Abbreviation: TBM, tensor-based morphometry. (For interpretation of the references to color in this Figure, the reader is referred to the web version of this article.)

in the previously mentioned works, is to optimize a univariate measure of brain degeneration over time.

4.3. *Power estimates of other measures in AD*

Our change measures outperformed all other published unbiased measures as an AD biomarker with respect to the sample size requirements, assuming of course that the reference data are comparable. In the following section we compare each method's best measure as reported in Holland et al. (2011) and 2 other methods against our TBM-LDA and TBM + Vent LDA measures. FreeSurfer ventricular measures give a 2-year estimates of 90 (68, 128) for AD and 153 (126, 194) for MCI. An FSL tool, known as SIENA (Cover et al., 2011; Smith et al., 2002), achieved a 1-year point estimate for sample size of 132 for AD and 278 for MCI. Quarc entorhinal achieved a 2-year whole brain estimates of 44 (33, 63) for AD and 134 (110, 171) for MCI. KN-BSI, a whole brain gray matter atrophy measure (Schott et al., 2010) required 1-year samples of 81 (64, 109) for AD and 149 (122, 188) for MCI. For a 2-year trial, Holland et al. (2011) estimate KN-BSI power at 75 (58, 104) for AD and 142 (115, 182) for MCI. Hua et al. (2013) used improved TBM with the stat-ROI voxel weighting to achieve 2-year sample sizes of 41 (33, 55) for AD and 109 (92, 131) for MCI. Wolz et al. (2010) measured hippocampal volume change based a longitudinal adaptation of the LEAP algorithm, achieving 24-month power estimates of 46 for AD and 121 for MCI and 12-month estimates of 67 and 206. Some confusion has resulted because of the use of the term "two-arm" to describe a study of treatment versus placebo groups in (Wolz et al., 2010). The power estimates are, in fact, computed identically and are directly comparable with the others', as can be seen by comparing Equation (1) mentioned previously and Equation (4) in Wolz et al. (2010). The estimates "per arm" in other previously mentioned studies have the same meaning as the estimates "for both arms" [sic] in Wolz et al. (2010), without need to adjust them by a factor of 2. This can also be confirmed by applying Equation (1) to their reported means and standard deviations. We note that both the 24-month LEAP and the SIENA estimates are based on a much smaller sample of subjects—(83, 165) and (85, 195)—than the other methods mentioned previously, and any comparisons must be made with the appropriate reservations. These comparisons are summarized in Fig. 7.

Table 4
Bootstrapped *p*-values, stat-ROI versus TBM-LDA measures

	12 mo		24 mo	
	GM-LDA versus stat-ROI	Whole LDA versus stat-ROI	GM-LDA versus stat-ROI	Whole LDA versus stat-ROI
AD	0.0683	0.0795	0.162	0.0631
MCI	**0.014**	**0.0019**	**0.0001**	**<0.0001**

Nonparametric test assessing the probability that the stat-ROI measure leads to lower or equal required sample size compared with the given LDA measure. Significant results at the $p = 0.05$ level are in bold.
Key: AD, Alzheimer's disease; GM, gray matter; MCI, mild cognitive impairment; ROI, region of interest; TBM, tensor-based morphometry.

A likely reason for such a favorable comparison with existing atrophy scores is because of the multivariate nature of our raw atrophy measures. Unlike the other methods used in ADNI, most of which are ROI volume measures or their combinations, our measure is based on a spatially distributed map. This presents a challenge and an opportunity to optimally combine thousands or even millions of features into a useful biomarker. The simplest approach, linear weighting, outperforms other methods in terms of power estimates. However, we do not wish for this simplicity to be misleading; the linear model uses the fine-grained spatial analysis from TBM and surface features, which is not available in other popular ADNI measures. Although one could use the same approach to optimize power by, for example, combining all FreeSurfer regional volumes optimally that approach would still not offer the voxelwise accuracy of TBM and local surface-based measures.

4.4. *Algorithmic bias*

We showed that our measures are longitudinally unbiased according to the intercept CI test (Yushkevich et al., 2010a). The test

Table 5
Longitudinal bias analysis of AD imaging biomarkers

Vent-LDA	TBM-LDA	Vent + TBM	TBM stat-ROI	TBM-LDA GM only
0.0064 (−0.0218 to 0.06)	-1.48×10^{-5} (-5.1×10^{-4}, 4.9×10^{-4})	0.077 (−0.48, 0.67)	0.06 (−0.07, 0.18)	-1.02×10^{-4} (-5.6×10^{-4}, 3.9×10^{-4})

Change in healthy controls is linearly regressed over 2 time points. The intercept is very close to zero with the confidence interval clearly containing zero for each method. The LDA-based measures do not show any algorithmic bias according to the CI test.
Key: AD, Alzheimer's disease; CI, confidence interval; GM, gray matter; TBM, tensor-based morphometry.

addresses an issue raised by Thompson and Holland (2011) about overly optimistic power estimates caused by additive algorithmic bias. The fact that the baseline and follow-up scans were processed identically, and independently, avoids several sources of subtle bias in longitudinal image processing that can arise from not handling the images in a uniform way (Thompson and Holland, 2011). Some issues have been raised regarding the validity of the intercept CI test as a test for bias in estimating rates of change. The CI test assumes that the true morphometric change from baseline increases in magnitude linearly over time in healthy controls. Relying on this assumption, the test examines whether the intercept of the linear model, fitted through measures of change at successive time intervals in controls, is zero. If this is not the case, the measure of change is said to have additive bias. We address the common criticisms of this test in our previous report (Gutman et al., 2013) and conclude that the test remains appropriate so long as it is only applied to control subjects.

4.5. Total and relative atrophy

There has been some recent debate regarding the need to subtract the mean of the healthy controls when estimating sample sizes for a drug trial. Some ADNI collaborators seem to have rejected this idea (Gutman et al., 2013; Hua et al., 2013), in part because real drug trials do not tend to enroll controls, and even if they did, many controls already harbor incipient Alzheimer pathology or some degree of vascular pathology that may also be resisted by treatment. However, the idea is not completely without merit, because all meaningful trials must compare a treatment against another (placebo or established) treatment group. Further, any additive algorithmic bias could be excluded by subtracting the mean rate of controls. We addressed this issue in our previous report on ventricular LDA biomarkers (Gutman et al., 2013) by computing an additional linear ventricular expansion model specific to AD and MCI progression. We did this by directly applying a 2 class, as opposed to 1 class, LDA with the covariance defined strictly by the diseased group, as required by the current practice of NC-adjusted sample size estimates. The resulting power estimates for NC-adjusted atrophy outperformed all previous ventricular measures.

4.6. Future work

Future work will include utilization of additional biomarkers, including other imaging biomarkers, such as measures based on diffusion imaging or even nonimaging biomarkers (such as CSF or proteomic measures) into the framework. We would like to extend the use of supervised learning to further reduce our sample size estimates. For example, in the PCA experiment, we simply used all principal components up to a cutoff value. Although the power estimates were impressive, the spatial patterns of the weights contained high-frequency components without clear anatomic meaning. A greedy boosting-type search over the principal components as in Lu et al. (2003) may lead to better performance, with the goal of making the pattern more generalizable and more congruent across the folds. As our linear weighting is likely to contain a combination of disease effect and systematic registration artifact, a boosting approach over the principle components could potentially isolate and discount any principal components containing the artifactual portion of the variance. Alternatively, a more comprehensive set of basis functions could be utilized to describe the TBM atrophy patterns, yet enable whole sample learning on conventional computers. Additional improvements in sample size estimates could potentially be achieved by controlling for

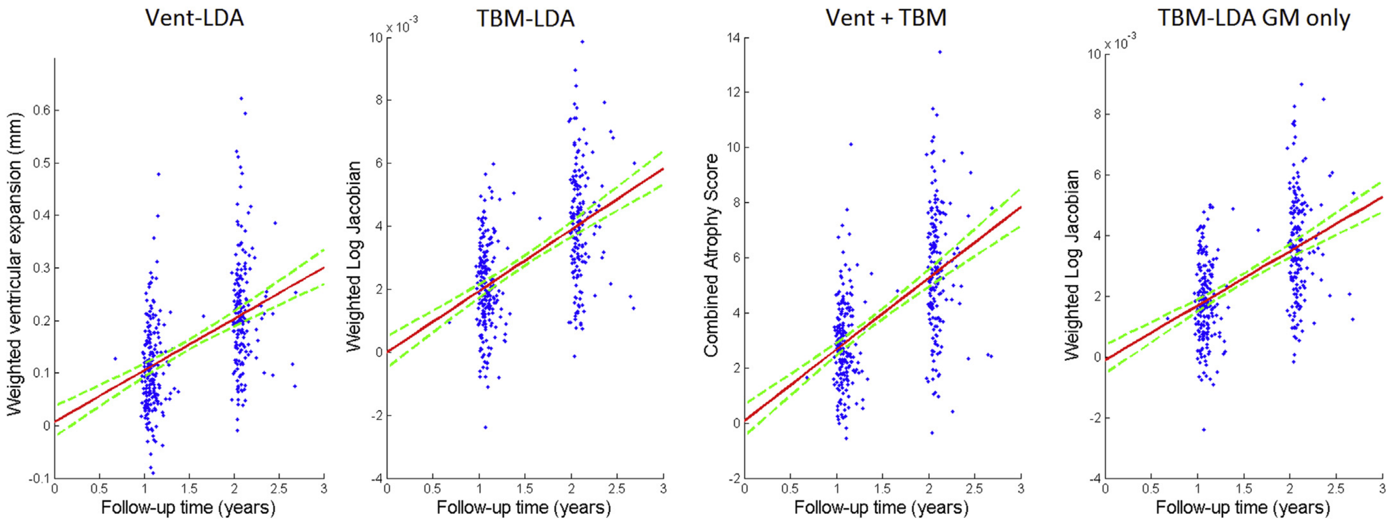

Fig. 6. Regression plots for LDA-based atrophy measures in controls. Green dotted lines show 95% confidence belts for the regression models. All LDA models are longitudinally unbiased, because the zero intercept is contained in the 95% confidence interval on the intercept, for each of the methods. (For interpretation of the references to color in this Figure, the reader is referred to the web version of this article.)

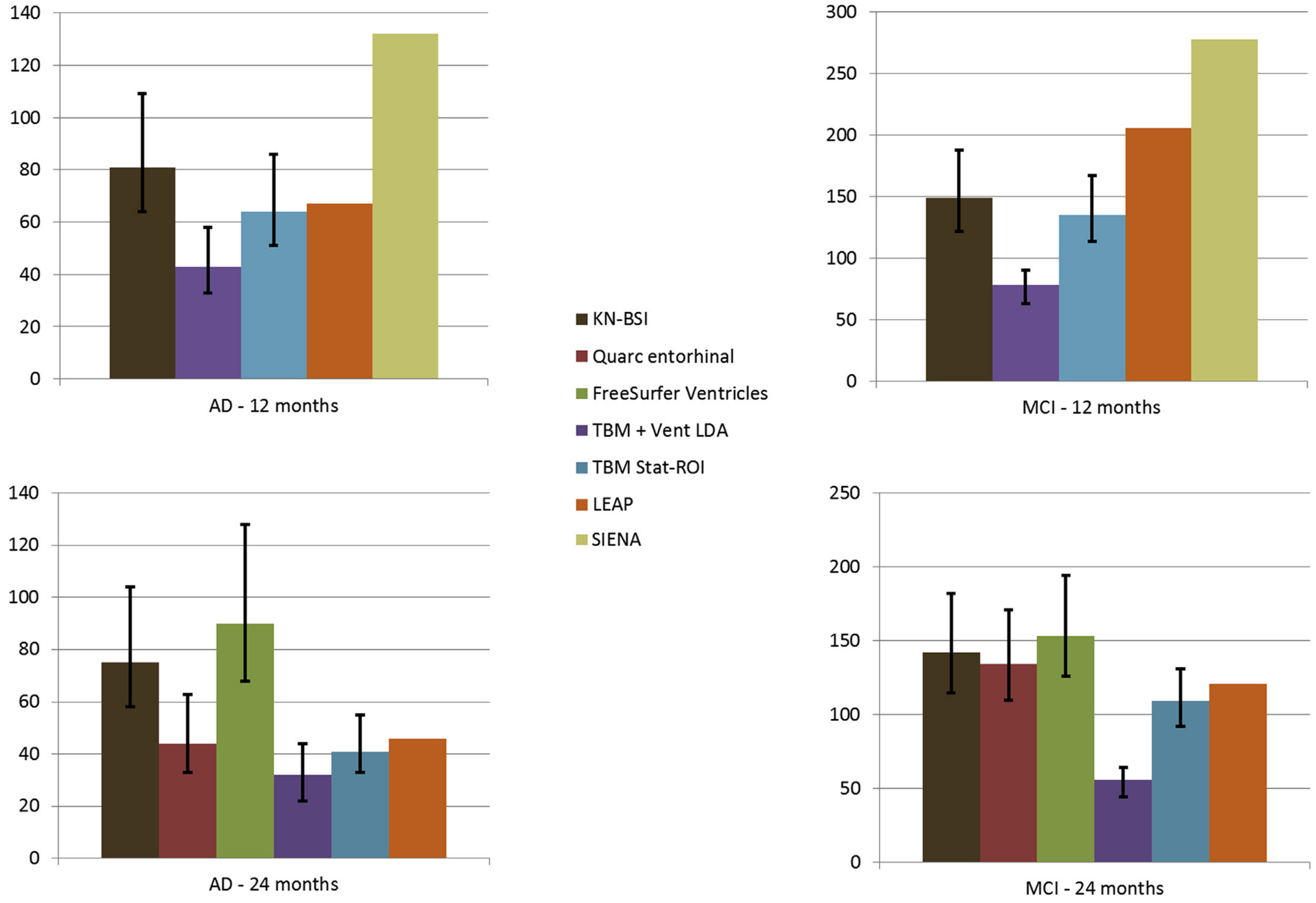

Fig. 7. Sample size estimates for different biomarkers for 1- and 2-year trials with 2 scans per subject. Black bars indicate 95% confidence intervals, where available.

confounding factors such as age and sex, as is in Schott et al. (2010), and by enrichment techniques accounting for ApoE genotype or family history of AD.

A potential limitation of a data-driven method such as what we have presented here pertains to its reliance on the specifics of the data. In particular, image quality and inclusion criteria of a hypothetical trial are assumed to be the same as in ADNI. Simpler univariate methods like LEAP and BSI do not suffer from this limitation to the same extent, as they do not make such strict assumptions about image quality and assume nothing about the subjects included in the trial. Nonetheless, as our measure outperforms other competitive measures by quite a few subjects, it is quite possible that a new trial with significantly different parameters may still be better served by the proposed method. In this case, some data may need to be set aside to train a new model specific to the trial. Whether this additional training set justifies the reduced number of test subjects required will be the subject of future work. In this article, we have simply assumed that the hypothetical trial will follow the design of ADNI, which justifies our direct head-to-head N80 comparisons. In this case, the new trial would simply use our existing weight maps to compute the aggregate atrophy measure without requiring any additional training subjects.

It is important to interpret biomarker power in its proper context. Basing a measure of brain change on a certain region or parameter of the brain may overlook valuable disease-modifying effects that affect other regions or measures. Perhaps even more importantly, the slowing of a change measure by 25% may have different value to the patient, depending on whether the measure is volumetric loss, amyloid clearance, or decline in cognition. We must therefore treat the n80 as a guide to biomarker utility weighing it against other relevant criteria, in much the same way as we advocated the weighting of multiple features within an image here, rather than relying on any one marker of disease progression.

Disclosure statement

The authors have no conflicts of interest to disclose.

Acknowledgements

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 Fund; Alzheimer's Association; Alzheimer Drug Discovery Foundation; Amorfix Life Sciences Ltd; AstraZeneca; Bayer Healthcare; BioClinica, Inc; Biogen; Bristol-Myers Squibb Foundation; Eisai; Elan; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech Foundation; 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; Servier; Synarc Inc; and Takeda Pharmaceuticals North America. 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 Rev March 26, 2012 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 National Institutes of Health grants P30 AG010129 and K01 AG030514. Algorithm development for this study was also funded by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, the U.S. National Library of Medicine, and the National Center for Research Resources (AG016570, EB01651, LM05639, and RR019771 to Paul M. Thompson). Author contributions were as follows: Boris A. Gutman wrote the manuscript. Boris A. Gutman, Xue Hua, Arthur W. Toga, and Paul M. Thompson performed image analyses. Boris A. Gutman, Yalin Wang, Igor Yanovsky, and Paul M. Thompson developed algorithms used in the analyses. Paul M. Thompson and Igor Yanovsky made substantial comments on the manuscript. Clifford R. Jack and Michael W. Weiner contributed substantially to the image and data acquisition, study design, quality control, calibration, and preprocessing, databasing, and image analysis.

References

Ashburner, J., Friston, K.J., 2003. Morphometry. In: Richard S.J. Frackowiak, et al., Human Brain Function, second ed. Elsevier Academic Press, Amsterdam, Boston.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41.

Beckett, L.A., 2000. Community-based studies of Alzheimer's disease: statistical challenges in design and analysis. Stat. Med. 19, 1469–1480.

Chen, K., Langbaum, J.B., Fleisher, A.S., Ayutyanont, N., Reschke, C., Lee, W., Liu, X., Bandy, D., Alexander, G.E., Thompson, P.M., Foster, N.L., Harvey, D.J., de Leon, M.J., Koeppe, R.A., Jagust, W.J., Weiner, M.W., Reiman, E.M., 2010. 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 51, 654–664.

Cho, Y., Seong, J.K., Jeong, Y., Shin, S.Y., 2012. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. Neuroimage 59, 2217–2230.

Chou, Y.Y., Lepore, N., de Zubicaray, G.I., Carmichael, O.T., Becker, J.T., Toga, A.W., Thompson, P.M., 2008. Automated ventricular mapping with multi-atlas fluid image alignment reveals genetic effects in Alzheimer's disease. Neuroimage 40, 615–630.

Cover, K.S., van Schijndel, R.A., van Dijk, B.W., Redolfi, A., Knol, D.L., Frisoni, G.B., Barkhof, F., Vrenken, H., 2011. Assessing the reproducibility of the SienaX and Siena brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. Psychiatry Res. 193, 182–190.

Cuingnet, R., Chupin, M., Benali, H., Colliot, O., 2010. Spatial Prior in SVM-based Classification of Brain Images. Proc SPIE 7624, Medical Imaging 2010: Computer-Aided Diagnosis 7624.

Cummings, J.L., 2010. Integrating ADNI results into Alzheimer's disease drug development programs. Neurobiol. Aging 31, 1481–1492.

DiCicio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Stat. Sci. 11, 10.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. Wiley, New York.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39, 1731–1743.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. U.S.A 97, 11050–11055.

Fox, N.C., Ridgway, G.R., Schott, J.M., 2011. Algorithms, atrophy and Alzheimer's disease: cautionary tales for clinical trials. Neuroimage 57, 15–18.

Freeborough, P.A., Fox, N.C., 1998. Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images. J. Comput. Assist.Tomogr. 22, 838–843.

Gerardin, E., Chetelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehericy, S., Garnero, L., Eustache, F., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47, 1476–1486.

Good, C.D., Scahill, R.I., Fox, N.C., Ashburner, J., Friston, K.J., Chan, D., Crum, W.R., Rossor, M.N., Frackowiak, R.S., 2002. Automatic differentiation of anatomical patterns in the human brain: validation with studies of degenerative dementias. Neuroimage 17, 29–46.

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

Gutman, B.A., Hua, X., Rajagopalan, P., Chou, Y.-Y., Wang, Y., Yanovsky, I., Toga, A.W., Jack Jr., C.R., Weiner, M.W., Thompson, P.M., 2013. Maximizing power to track Alzheimer's disease and MCI progression by LDA-based weighting of longitudinal ventricular surface features. Neuroimage 70, 386–401.

Gutman, B., Wang, Y., Morra, J., Toga, A.W., Thompson, P.M., 2009. Disease classification with hippocampal shape invariants. Hippocampus 19, 572–578.

Gutman, B.A., Yalin, W., Rajagopalan, P., Toga, A.W., Thompson, P.M., 2012. Shape Matching with Medial Curves and 1-D Groupwise Registration. in: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on, pp 716–719.

Hobbs, N.Z., Henley, S.M.D., Ridgway, G.R., Wild, E.J., Barker, R.A., Scahill, R.I., Barnes, J., Fox, N.C., Tabrizi, S.J., 2010. The progression of regional atrophy in premanifest and early Huntington's disease: a longitudinal voxel-based morphometry study. J. Neurol. Neurosurg. Psychiatry 81, 756–763.

Holland, D., Dale, A.M., 2011. Nonlinear registration of longitudinal images and measurement of change in regions of interest. Med. image Anal. 15, 489–497.

Holland, D., McEvoy, L.K., Dale, A.M., 2011. Unbiased comparison of sample size estimates from longitudinal structural measures in ADNI. Hum. Brain Mapp. 33, 2586–2602.

Hua, X., Gutman, B., Boyle, C.P., Rajagopalan, P., Leow, A.D., Yanovsky, I., Kumar, A.R., Toga, A.W., Jack Jr., C.R., Schuff, N., Alexander, G.E., Chen, K., Reiman, E.M., Weiner, M.W., Thompson, P.M., 2011. Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. Neuroimage 57, 5–14.

Hua, X., Hibar, D.P., Ching, C.R., Boyle, C.P., Rajagopalan, P., Gutman, B.A., Leow, A.D., Toga, A.W., Jack Jr., C.R., Harvey, D., Weiner, M.W., Thompson, P.M., 2013. Unbiased tensor-based morphometry: improved robustness and sample size estimates for Alzheimer's disease clinical trials. Neuroimage 66, 648–661.

Hua, X., Lee, S., Yanovsky, I., Leow, A.D., Chou, Y.Y., Ho, A.J., Gutman, B., Toga, A.W., Jack Jr., C.R., Bernstein, M.A., Reiman, E.M., Harvey, D.J., Kornak, J., Schuff, N., Alexander, G.E., Weiner, M.W., Thompson, P.M., 2009. 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 48, 668–681.

Hua, X., Leow, A.D., Parikshak, N., Lee, S., Chiang, M.C., Toga, A.W., Jack Jr., C.R., Weiner, M.W., Thompson, P.M., 2008. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage 43, 458–469.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27, 685–691.

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

Kloppel, S., Abdulkadir, A., Jack Jr., C.R., Koutsouleris, N., Mourao-Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. Neuroimage 61, 457–463.

Kloppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S., 2008. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. Brain 131, 2969–2974.

Kohannim, O., Hua, X., Hibar, D.P., Lee, S., Chou, Y.Y., Toga, A.W., Jack Jr., C.R., Weiner, M.W., Thompson, P.M., 2010. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol. Aging 31, 1429–1442.

Leow, A.D., Yanovsky, I., Chiang, M.C., Lee, A.D., Klunder, A.D., Lu, A., Becker, J.T., Davis, S.W., Toga, A.W., Thompson, P.M., 2007. Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration. IEEE Trans. Med. Imaging 26, 822–832.

Leung, K.K., Ridgway, G.R., Ourselin, S., Fox, N.C., 2012. Neuroimaging AsD Consistent multi-time-point brain atrophy estimation from the boundary shift integral. Neuroimage 59, 3995–4005.

Lu, J., Plataniotis K.N., Venetsanopoulos, A.N., 2003. Boosting Linear Discriminant Analysis for Face Recognition. ICIP 657–660.

Marsden, J., Hughes, T., 1983. Mathematical Foundations of Elasticity. Prentice-Hall; Englewood Cliffs, N.J.

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

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005a. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15, 869–877.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005b. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement. 1, 55–66.

Qazi, A.A., Jorgensen, D.R., Lillholm, M., Loog, M., Nielsen, M., Dam, E.B., 2010. A framework for optimizing measurement weight maps to minimize the required sample size. Med. Image Anal. 14, 255–264.

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61, 1402–1418.

Ross, J., Thompson, P.M., Tariot, P., Reiman, E.M., Schneider, L., Frigerio, E., Fiorentini, F., Giardino, L., Calzà, L., Norris, D., Cicirello, H., Casula, D., Imbimbo, B.P., 2012. 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.

Schott, J.M., Bartlett, J.W., Barnes, J., Leung, K.K., Ourselin, S., Fox, N.C., 2010. Reduced sample sizes for atrophy outcomes in Alzheimer's disease trials: baseline adjustment. Neurobiol. Aging 31, 1452–1462, 1462 e1–2.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. Neuroimage 13, 856–876.

Shen, L., Qi, Y., Kim, S., Nho, K., Wan, J., Risacher, S.L., Saykin, A.J., 2010. Sparse bayesian learning for identifying imaging biomarkers in AD prediction. Med. Image Comput. Comput. Assist. Interv. 13, 611–618.

Shi, Y., Morra, J.H., Thompson, P.M., Toga, A.W., 2009. Inverse-consistent surface mapping with Laplace-Beltrami eigen-features. Inf. Process. Med. Imaging. 21, 467–478.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.

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

Styner, M., Lieberman, J.A., McClure, R.K., Weinberger, D.R., Jones, D.W., Gerig, G., 2005. Morphometric analysis of lateral ventricles in schizophrenia and healthy controls regarding genetic and disease-specific factors. Proc. Natl. Acad. Sci. U.S.A 102, 4872–4877.

Thompson, P.M., Giedd, J.N., Woods, R.P., MacDonald, D., Evans, A.C., Toga, A.W., 2000. Growth patterns in the developing brain detected by using continuum mechanical tensor maps. Nature 404, 190–193.

Thompson, W.K., Holland, D., 2011. Bias in tensor based morphometry Stat-ROI measures may result in unrealistic power estimates. Neuroimage 57, 1–4.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 39, 1186–1197.

Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., 2010. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. Neuroimage 52, 109–118.

Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., Decarli, C., Fox, N.C., Gunter, J.L., Hill, D., Killiany, R.J., Pachai, C., Schwarz, A.J., Schuff, N., Senjem, M.L., Suhy, J., Thompson, P.M., Weiner, M., Jack Jr., C.R., 2012. Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimers Dement. 9, 332–337.

Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J., 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. Neuroimage 61, 622–632.

Yushkevich, P.A., Avants, B.B., Das, S.R., Pluta, J., Altinay, M., Craige, C., 2010a. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3 T MRI data. Neuroimage 50, 434–445.

Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S., 2010b. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. Neuroimage 53, 1208–1224.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867.