Statistical Analysis of Longitudinal Neuroimage Data with Linear Mixed Effects Models

Jorge L. Bernal-Rusiel, Douglas N. Greve, Martin Reuter, Bruce Fischl, Mert R. Sabuncu

PII: S1053-8119(12)01068-3
DOI: doi: 10.1016/j.neuroimage.2012.10.065
Reference: YNIMG 9908

To appear in: *NeuroImage*

Accepted date: 22 October 2012

Please cite this article as: Bernal-Rusiel, Jorge L., Greve, Douglas N., Reuter, Martin, Fischl, Bruce, Sabuncu, Mert R., Statistical Analysis of Longitudinal Neuroimage Data with Linear Mixed Effects Models, *NeuroImage* (2012), doi: 10.1016/j.neuroimage.2012.10.065

Statistical Analysis of Longitudinal Neuroimage Data with Linear Mixed Effects Models

Jorge L. Bernal-Rusiel[1], Douglas N. Greve[1], Martin Reuter[1,2] , Bruce Fischl[1,3], and Mert R. Sabuncu[1,3];

for the Alzheimer's Disease Neuroimaging Initiative*

[1] Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School/Massachusetts General Hospital, Charlestown, MA

[2]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA

[3] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA

Corresponding Author:

Mert R. Sabuncu

Athinoula A. Martinos Center for Biomedical Imaging

Massachusetts General Hospital

Building 149, 13th Street, Room 2301

Charlestown, Massachusetts, USA 02129

Phone: 617 643-7460, Fax: 617 726-7422

Email: msabuncu@nmr.mgh.harvard.edu

* Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators is available at http://tinyurl.com/ADNI-main.

ABSTRACT

Longitudinal neuroimaging (LNI) studies are rapidly becoming more prevalent and growing in size. Today, no standardized computational tools exist for the analysis of LNI data and widely used methods are sub-optimal for the types of data encountered in real-life studies. Linear Mixed Effects (LME) modeling, a mature approach well known in the statistics community, offers a powerful and versatile framework for analyzing real-life LNI data. This article presents the theory behind LME models, contrasts it with other popular approaches in the context of LNI, and is accompanied with an array of computational tools that will be made freely available through FreeSurfer – a popular Magnetic Resonance Image (MRI) analysis software package.

Our core contribution is to provide a quantitative *empirical* evaluation of the performance of LME and competing alternatives popularly used in prior longitudinal structural MRI studies, namely repeated measures ANOVA and the analysis of annualized longitudinal change measures (e.g. atrophy rate). In our experiments, we analyzed MRI-derived longitudinal hippocampal volume and entorhinal cortex thickness measurements from a public dataset consisting of Alzheimer's patients, subjects with mild cognitive impairment and healthy controls. Our results suggest that the LME approach offers superior statistical power in detecting longitudinal group differences.

KEYWORDS

Longitudinal Studies, Linear Mixed Effects Models, Statistical Analysis

1 INTRODUCTION

Longitudinal neuroimaging (LNI) studies have become increasingly widespread over the last decade, e.g. (Asami et al., 2011; Blockx et al., 2011; Chetelat et al., 2005; Davatzikos and Resnick, 2002; Desikan et al., 2011; Driscoll et al., 2011; Fjell et al., 2009; Fotenos et al., 2005; Fouquet et al., 2009; Frings et al., 2011; Giedd et al., 1999; Hedman et al., 2011; Ho et al., 2003; Holland et al., 2009; Holland et al., 2011; Hua et al., 2010; Hua et al., 2009; Jack Jr et al., 2009; Jack Jr et al., 2008; Josephs et al., 2008; Kaladjian et al., 2009; Kalkers et al., 2002; Ment et al., 2009; Pantelis et al., 2003; Paviour et al., 2006; Resnick et al., 2010; Sabuncu et al., 2011; Schumann et al., 2010; Sidtis et al., 2010; Sluimer et al., 2008; Sluimer et al., 2009; Sullivan et al., 2011; Thambisetty et al., 2011; Thambisetty et al., 2010; Tosun et al., 2010; Whitwell et al., 2011; Whitwell et al., 2007). Compared to the cross-sectional approach, the longitudinal design can provide increased statistical power by reducing the confounding effect of between-subject variability (Thompson et al., 2011). Moreover, a serial assessment can be the only way to unambiguously characterize the effect of interest in a randomized experiment, such as a drug trial (Davis et al., 2005; Dickerson and Sperling, 2005; Ge et al., 2000). Finally, longitudinal studies provide unique insights into the temporal dynamics of the underlying biological process (Jack Jr et al., 2012; Sabuncu et al., 2011).

LNI studies have yielded novel discoveries, yet a careful scrutiny of the literature reveals that the statistical methods commonly lack maturity and sophistication. We believe that the underutilization of appropriate methodology in LNI studies is mainly due to two related reasons. Firstly, the relevant statistical tools are not readily available in user-friendly neuroimage analysis software environments (such as SPM (Friston, 2007; SPM), FSL (Smith et al., 2004), or FreeSurfer (Fischl, 2012)). Secondly, the technical intricacies of modeling longitudinal data are not well understood and/or appreciated.

In this article, we advocate the use of Linear Mixed Effects (LME) modeling, which provides a flexible and powerful statistical framework for the analysis of longitudinal data (Fitzmaurice et al., 2011; Verbeke and Molenberghs, 2000). We discuss the theoretical underpinnings of the LME framework and contrast it with other methods popular in LNI.

There are two alternative approaches most commonly applied to the analysis of prior LNI data. These are (1) *repeated measures analysis of variance* (or within-subject ANOVA) (Girden, 1992), e.g., (Asami et al., 2011; Blockx et al., 2011; Bonne et al., 2001; Fouquet et al., 2009; Giedd et al., 1999; Ho et al., 2003; Kaladjian et al., 2009; Mathalon et al., 2001; Pantelis et al., 2003; Resnick et al., 2010; Sidtis et al., 2010; Sluimer et al., 2009), ; and (2) *cross-sectional (General Linear Model –GLM- based) analysis of summary measurements*, such as percent annualized difference, e.g., (Desikan et al., 2011; Fjell et al., 2009; Fotenos et al., 2005; Fouquet et al., 2009; Frings et al., 2011; Hedman et al., 2011; Holland et al., 2009; Hua et al., 2010; Hua et al., 2009; Jack Jr et al., 2009; Josephs et al., 2008; Kalkers et al., 2002; Kasai et al., 2003; Paviour et al., 2006; Sabuncu et al., 2011; Sluimer et al., 2008; Whitwell et al., 2007). However, these methods are known to be sub-optimal for general real-life longitudinal data since they do not model the covariance structure of serial measurements appropriately and cannot handle imperfect timing and/or subject dropout (i.e., unbalanced data), in particular those cases with only a single time-point (Fitzmaurice et al., 2011).

Another related approach is the two-stage strategy for solving the hierarchical models adopted in functional neuroimaging (Friston, 2007). Yet these tools typically rely on assumptions that are unrealistic for the LNI design we consider here[1]. For example, in LNI studies one usually has only a handful of scans per subject and not hundreds of time-points. Furthermore, a pre-whitening step is unlikely to be suitable since LNI data are not sampled at uniform time intervals and do not obey a stationary autoregressive structure.

The contributions of this article are multi-fold. First, we present a thorough overview of the LME approach in the context of longitudinal studies. Computational tools implementing this approach will accompany this article as a part of *FreeSurfer*, a MRI processing software, (Dale et al., 1999; Fischl et al., 2002; Fischl et al., 1999a; Fischl et al., 1999b). We use a widely studied longitudinal structural MRI dataset (from the Alzheimer's Disease Neuroimaging Initiative, or ADNI) to illustrate how these tools can be used for exploratory data visualization, model specification, model selection, parameter estimation, hypothesis testing, and statistical power analysis including sample

[1] In the LNI design we consider in this manuscript, each participant is scanned at potentially several time points and the imaging measurement of interest at each time point is a scalar, e.g., brain volume.

size estimation. We further perform a systematic empirical validation of the specificity, sensitivity and repeatability of the LME method and alternative approaches. Our results provide an objective quantification of the improvement in statistical detection afforded by the LME approach compared with competing methods. Finally, we assess the impact of including subjects with a single time point in the LME method.

The paper is organized as follows. Section 2.1 provides a discussion of the general characteristics of longitudinal data. Section 2.2 presents the LME method for the analysis of longitudinal data. Section 2.3 includes a brief description of alternative methods used in prior LNI studies. Section 2.4 offers a description of the data used in the experiments. In Section 3, we present experimental results that illustrate the proposed approach and compare it to benchmark methods. Finally, Section 4 provides a discussion of the main experimental findings and Section 5 closes with concluding remarks.

2 MATERIAL AND METHODS

2.1 The Characteristics of Longitudinal Data

In a longitudinal study, outcome variables are measured repeatedly on the same cohort of individuals at multiple time-points. The aim is to characterize changes in the individuals' measurements over time and their association with clinical, experimental or biological factors. Unlike cross-sectional studies, where the measurement is obtained at a single occasion, longitudinal studies allow direct assessment of within-subject changes across different time points, free of any between-subject variability. Changes in the mean measurement over time can then be estimated with greater precision and without confounding cohort effects (Fitzmaurice et al., 2011). Furthermore, more accurate predictions about an individual's measurement trajectory might be possible by pooling data across the population. This can be useful, for instance, to assess the effect of a drug in a specific individual in a pharmacological study.

In general, longitudinal data exhibit several distinctive characteristics. (1) Longitudinal measurements are ordered in time, reflecting the temporal trajectory of an underlying non-stationary continuous process. This is the major difference between vectors of repeated measures obtained in longitudinal studies and vectors of multivariate measurements from cross-sectional studies, where single measurements of multiple but

distinct variables are taken simultaneously. (2) Typically, serial measurements obtained for a single subject are positively correlated. This correlation is due to the smooth trajectory of the underlying biological process. In general, we expect pairs of repeated measures that are close in time to be more highly correlated than pairs of repeated measures further separated in time. (3) Between-subject variance is not usually constant over the duration of the study; instead, it might for example increase as a function of time due to the diverging trajectories of individuals and/or groups. (4) Finally, missing data and non-uniform timing is extremely common, particularly for longitudinal studies of larger duration.

2.2 Linear Mixed Effects Modeling for Longitudinal Data

There are two aspects of longitudinal data that require careful modeling: the mean measurement trajectory over time and the correlation structure among serial measurements. The models for the mean and covariance are *interdependent* because the vector of residuals (i.e., the observed minus fitted measurements) depends on the specification of the model for the mean.

LME models use the linear regression paradigm (Montgomery et al., 2007) to parsimoniously describe the average measurement and its temporal trajectory. In this approach, the mean measurement is expressed as a linear combination of a set of independent variables. The temporal trajectory is then determined by the contribution of time and/or time-varying variables. A major advantage of this approach is that the subjects in the study are not required to have a common set of measurement times (i.e., the data can be unbalanced).

Like any other statistical method, the selection of independent variables that models the mean measurement has to be made based on subject-matter grounds. On the other hand, without any additional knowledge, a useful strategy to model the mean trajectory is to simply assume it is linear in time. This is the default implementation in our toolkit. A justification for this strategy is due to the limited duration of studies, which typically can expose local and simple trends. More complex trajectories can be captured via piece-wise linear models or higher order (e.g. quadratic or cubic) polynomials. These models can be chosen by the user based on a graphical exploratory analysis of the data, such as by inspecting an illustration of smoothed measurements, e.g., a "*lowess*" plot (Cleveland,

1979). Then, as it is common in the linear regression paradigm, complex models can be compared to reduced models to determine whether they fit the data significantly better. For example, a quadratic model for the mean response over time can be compared to a linear model by testing the null hypothesis that the quadratic coefficient is zero.

There are generally three potential sources of variability influencing the correlation structure in longitudinal data: (1) between-subject variation, (2) inherent within-subject biological change, and (3) measurement error (Fitzmaurice et al., 2011). The first source of variability reflects natural variation in the individual's measurement trajectory. Some individuals' measurements are consistently higher than the population average, while others' are consistently lower. The inherent within-individual biological variation is a consequence of some subject-specific biological process that progresses gradually over time. Hence, random departures from an individual's modeled measurement trajectory are likely to be more similar when measurements are obtained close together in time. Finally, measurement error variance has a direct influence on the amount of correlation between serial measurements.

The LME method imposes structure on the covariance through the introduction of random effects. This approach provides both flexible and parsimonious models for the covariance and is particularly well suited to handling longitudinal data that are irregularly timed. A unique feature of these models is that they explicitly distinguish and allow the analysis of the between-subject and within-subject sources of variability. The following section provides the theoretical details of the LME framework.

2.2.1 Linear Mixed Effects Models: The Theory

Let us formally introduce the LME model for longitudinal data:

$$Y_i = X_i\beta + Z_i b_i + e_i, \qquad (2.1)$$

where Y_i is the $n_i \times 1$ vector of serial measurements for subject i (e.g. longitudinal MRI-derived thickness or volume measurements), n_i is the subject-specific number of serial measurements, X_i is the $n_i \times p$ subject design matrix for the fixed effects (including variables such as gender, education, clinical group, genotype and scan time), $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is a $p \times 1$ vector of unknown fixed effects regression coefficients, Z_i is the $n_i \times q, q \leq p$ design matrix for the random effects (e.g. scan time),

$b_i = \left(b_{i1}, b_{i2}, \ldots, b_{iq}\right)^T$ is a $q \times 1$ vector of random effects and $e_i = \left(e_{i1}, e_{i2}, \ldots, e_{\text{in}_i}\right)^T$ is a $n_i \times 1$ vector of measurement errors. Here Z_i links the vector of random effects b_i to Y_i and its columns are a subset of the columns of X_i. That is, any component of β can be allowed to vary randomly by simply including the corresponding column of X_i in Z_i. For example, in a model with only a randomly varying intercept Z_i is a $n_i \times 1$ vector composed of 1's. Note that all random effects other than the intercept need to be time varying, so that Z_i is not singular (In fact the measurement time itself is usually a random effect). The following common distributional assumptions are made:

$$b_i \sim N(0, \mathrm{D}),$$

$$e_i \sim N\left(0, \sigma^2 I_{n_i}\right),$$

where $N(0,\Sigma)$ denotes a zero mean multivariate Gaussian with covariance matrix Σ, I_{n_i} denotes the $n_i \times n_i$ identity matrix, and $b_1, \ldots, b_m, e_1, \ldots, e_m$ are independent with m being the number of subjects in the study. The components of b_i reflect how the subset of regression parameters for the i^{th} subject deviate from those of the population. The components of e_i represent random sampling or measurement errors.

The LME model provides an important distinction between the conditional and marginal means of Y_i. The conditional or subject-specific mean of Y_i, given b_i, is

$$E\left(Y_i \mid b_i\right) = X_i\beta + Z_i b_i,$$

while the marginal or population-averaged mean of Y_i is

$$E\left(Y_i\right) = X_i\beta.$$

Thus, in the LME model, the vector of regression parameters β (the fixed effects), is assumed to be the same for all individuals and have population-averaged interpretations, for example in terms of population mean trajectory. In contrast, the vector b_i (when summed with the corresponding fixed effects) makes up subject-specific regression coefficients, which describe the mean trajectory of the i^{th} individual.

We can also distinguish between the conditional covariance

$$Cov\left(Y_i \mid b_i\right) = Cov\left(e_i\right) = \sigma^2 I_{n_i},$$

and the marginal covariance of Y_i,

$$Cov(Y_i) = Cov(Z_i b_i) + Cov(e_i) = Z_i D Z_i^T + \sigma^2 I_{n_i},$$

which is *not* a diagonal matrix.

Thus by introducing random effects, correlations among the components of Y_i can be modeled. One can see that the model allows for the explicit analysis of between-subject (D) and within-subject (σ^2) sources of variation. Importantly, the marginal covariance of Y_i is expressed as a function of the time-varying random effects, which commonly includes measurement time itself.

Consider the following simple LME model, which has a randomly varying intercept and slope:

$$Y_{ij} = (\beta_1 + b_{11}) + (\beta_2 + b_{21}) t_{ij} + e_{ij}, \quad (2.2)$$

where Y_{ij} is the j^{th} measurement from subject i, t_{ij} is the time of measurement, and $j = 1, \ldots, n_i$. The model of (2.2) allows each individual's measurements to have his or her own unique linear mean trajectory.

2.2.2 Parameter Estimation

In this section we consider the problem of estimating the unknown coefficients β and model parameters σ and D. Given the distributional assumptions that have been made, the vector of measurements are distributed as

$$Y_i \sim N\left(X_i \beta, Z_i D Z_i^T + \sigma^2 I_{n_i}\right). \quad (2.3)$$

For given estimates $\hat{D}$ and $\hat{\sigma}$, we have a closed-form solution for the maximum likelihood (ML) estimate of β:

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i^T \hat{\Sigma}_i^{-1} X_i\right)^{-1} \sum_{i=1}^{m} X_i^T \hat{\Sigma}_i^{-1} y_i, \quad (2.4)$$

where $\hat{\Sigma}_i = Z_i \hat{D} Z_i^T + \hat{\sigma}^2 I_{n_i}$ and y_i is the realization of the random vector Y_i.

An unbiased estimate for $\hat{D}$ and $\hat{\sigma}$ can be obtained via maximizing the following restricted likelihood function (ReML procedure) (Verbeke and Molenberghs, 2000):

$$l_{ReML} = \frac{1}{2}\sum_{i=1}^{m} \log\left|\Sigma_i^{-1}\right| - \frac{1}{2}\sum_{i=1}^{m}\left(y_i - X_i\hat{\beta}\right)^T \Sigma_i^{-1}\left(y_i - X_i\hat{\beta}\right) - \frac{1}{2}\log\left|\sum_{i=1}^{m} X_i^T \Sigma_i^{-1} X_i\right|, \quad (2.5)$$

where $\Sigma_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$.

There is no closed-form solution to the optimization of (2.5) and numerical iterative solvers need to be used. We have implemented three widely used optimization methods: The Expectation Maximization (EM) algorithm (Laird et al., 1987) and two Newton-Raphson based procedures using either the Hessian or the expected information matrix of the restricted log-likelihood. The forms for the first and second partial derivatives of l_{ReML} can be found in (Lindstrom and Bates, 1988). When the expected information matrix is used in the optimization procedure the algorithm is commonly referred to as the Fisher's scoring scheme. Formulas for the expected information matrix can be found in (Kenward and Roger, 1997). Finally, we note that we do not impose any structure on D, other than it needs to be positive definite. To achieve this constraint, we parameterize it via its Cholesky decomposition.

2.2.3 Selection of Random Effects

In the LME approach, given a model for the mean, the covariance structure is determined by the choice of random effects. One good strategy to identify the appropriate set of random effects is via the likelihood ratio test, where the likelihood of nested models can be compared.

Here, one can start with a "basic model", which would only include the bias as a random effect. Once the model parameters and coefficients are estimated for the basic model, the corresponding restricted maximum likelihood value can be computed. One would then proceed to add random effects to the basic model. For example, time-varying variables can be added to the basic model as additional random effects one by one in a greedy fashion, where the variable that produces the highest increase in the restricted likelihood function will be added, only if this increase is statistically significant. The significance of a likelihood increase in nested models can be assessed based on a chi-square mixture statistic (Fitzmaurice et al., 2011).

2.2.4 Hypothesis Testing

In conducting hypothesis tests, we will use $\hat{\beta}$ and its estimated asymptotic covariance matrix

$$\hat{Cov}_{asymptotic}\left(\hat{\beta}\right)=\left(\sum_{i=1}^{m} X_i^T \hat{\Sigma}_i^{-1} X_i\right)^{-1},$$

where $\hat{\Sigma}_i$ is the ReML estimator of Σ_i.

In general, for a given contrast matrix L, the two competing hypotheses are

$H_0: \quad L\beta = 0 \quad$ and $\quad H_A: \quad L\beta \neq 0.$

Under the null hypothesis, it can be shown that the following F-distribution holds:

$$F = \frac{\left(L\hat{\beta}\right)^T \left(L\hat{Cov}_{asymptotic}\left(\hat{\beta}\right)L^T\right)^{-1} L\hat{\beta}}{rank(L)}. \qquad (2.6)$$

However, determining the degrees of freedom associated with the above F-test is challenging and several approximations have been proposed, e.g. (Satterthwaite, 1946). In particular, we have implemented a Satterthwaite-based approximation for the following scaled F-statistic:

$$F = \kappa\frac{\left(L\hat{\beta}\right)^T \left(L\hat{Cov}_{KR}\left(\hat{\beta}\right)L^T\right)^{-1} L\hat{\beta}}{rank(L)}, \qquad (2.7)$$

where $\hat{Cov}_{KR}\left(\hat{\beta}\right)$ is a small-sample bias corrected estimate of the covariance matrix of $\hat{\beta}$. This procedure allows the covariance among the ReML covariance parameter estimates to be taken into account when estimating the effective degrees of freedom of the F-test and thus different contrasts will exhibit different degrees of freedom. Details on the computation of κ, $\hat{Cov}_{KR}\left(\hat{\beta}\right)$ and the effective degrees of freedom can be found in (Kenward and Roger, 1997).

2.2.5 Sample Size Estimation and Statistical Power Analysis

Sample size and power calculations are more complex for longitudinal designs than for the simpler cross-sectional setting. The major challenge is missing data, which has a direct effect on power. In our toolbox, we have implemented two approximate methods for performing power calculations. The first method is intended for the planning phase, i.e., before data are collected, and can be used to obtain approximate estimates of the required sample size or the power to detect a particular effect size for a given sample size. The second method has a different purpose: namely, to provide an estimate of the power of a realized study, i.e., after the data have been collected.

The first method is based on a simple extension of the sample size and power formulae for a cross-sectional study with a univariate measurement (Fitzmaurice et al., 2011). In a two-group study, the approximate sample size N per group is:

$$N = \frac{\left(z_{(1-\alpha/2)} + z_{(1-\gamma)}\right)^2 2\varphi^2}{\delta^2}, \quad (2.8)$$

where $1-\gamma$ is the power of the test, α is the significance level, $z_{(1-\alpha/2)}$, $z_{(1-\gamma)}$ denote the $(1-\alpha/2)\times 100\%$ and $(1-\gamma)\times 100\%$ percentiles of a standard normal distribution, δ is the effect of interest, which for example can be any element of the vector β considered as a mixed effect (e.g., intercept or slope) and φ^2 is the corresponding diagonal element of the following covariance matrix

$$C = \sigma^2\left(Z_c^T Z_c\right)^{-1} + D,$$

with Z_c denoting the subject-level common random effects design matrix for the subjects in the study (i.e., assuming a balanced study).

Equation (2.8) can be re-arranged to determine the power of the planned study given a sample size:

$$z_{(1-\gamma)} = \sqrt{N\frac{\delta^2}{2\varphi^2}} - z_{(1-\alpha/2)}. \quad (2.9)$$

Finally, a conservative approach for adjusting for possible missing data is to inflate the required sample size N in each group to account for the expected proportion of subjects who will drop out before the completion of the study. E.g. if the rate of attrition is expected to be 10% in each group, the sample size in each group should be $N/0.9$.

The second method for power calculations allows a more precise approximation of the power of a realized (retrospective) experiment (given the actual unbalanced data over time with the missing data pattern). It is based on a non-central F-approximation to the distribution of the F-statistic in equation (2.6) under the alternative hypothesis (Helms, 1992). The degrees of freedom of the non-central F-distribution are $c = rank(L)$ and $v_e = \sum_{i=1}^{m} n_i - rank([XZ])$, with $X = \left[X_1^T X_2^T \ldots X_m^T\right]^T$ and $Z = Diag\left([Z_1, Z_2, \ldots, Z_m]\right)$

being the full fixed and random effects design matrices of the study. The non-centrality parameter is given by $nc = \left(L\hat{\beta}\right)^T \left(LC\hat{o}v_{asymptotic}\left(\hat{\beta}\right)L^T\right)^{-1} L\hat{\beta}$.

This non-central F-distribution can be used to perform power computations for tests of fixed effect hypotheses. The approximate power is

$$1-\gamma = 1-F\left(cv;c,v_e,nc\right), \quad (2.10)$$

where $F\left(cv;c,v_e,nc\right)$ is the cumulative distribution function of the non-central F-distribution evaluated at the critical value $cv = F^{-1}\left(1-\alpha;c,v_e\right)$, which is the inverse of the cumulative distribution function of the central F-distribution with c,v_e degrees of freedom evaluated at $1-\alpha$.

2.3 Alternative Methods for Analyzing Longitudinal Neuroimaging Data

Barring notable exceptions that use appropriate LME models, e.g. (Davatzikos and Resnick, 2002; Driscoll et al., 2011; Lau et al., 2008; Lerch et al., 2005; Shaw et al., 2008; Thambisetty et al., 2010; Tosun et al., 2010; Whitwell et al., 2011), there are two alternative methods that have been widely used to analyze LNI data in a large number of prior studies. The first approach is repeated measures (or within-subject) ANOVA, e.g. (Asami et al., 2011; Blockx et al., 2011; Bonne et al., 2001; Giedd et al., 1999; Ho et al., 2003; Kaladjian et al., 2009; Mathalon et al., 2001; Pantelis et al., 2003; Resnick et al., 2010; Sidtis et al., 2010; Sluimer et al., 2009), which can be shown to be equivalent to a linear model with at most a single random effect. Here, measurement occasions are treated as levels of a within-subject factor and time is not modeled as a continuous variable. Hence the method is only well suited for balanced longitudinal data with a small number of serial measurements. Furthermore, the correlation among repeated measurements, if modeled, is supposed to arise from the additive contribution of an individual-specific random effect, namely a random intercept. This imposes a particular covariance structure known as compound symmetry:

$$Var\left(Y_{ij}\right) = \sigma_b^2 + \sigma_e^2$$

$$Corr\left(Y_{ij},Y_{ik}\right) = \frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2} \quad (i=1,\ldots,N);\ (j,k=1,\ldots,n), j \neq k$$

where σ_b^2 and σ_e^2 are the variance of the random effect and the measurement error respectively. This structure for the covariance has some justification in certain designs. For example, in an fMRI experiment where the within-subject factor is randomly allocated to subjects, compound symmetry can hold. However, the constraint on the correlation among repeated measurements is not appropriate for longitudinal data, where the correlations are expected to decay with increasing separation in time. Also, the assumption of constant variance across time is often unrealistic.

Another common approach to the analysis of LNI data reduces the sequence of repeated measures for each individual to summary values (e.g. the annualized difference between two measures, the slope of a regression line, or deformation tensors), e.g. (Desikan et al., 2011; Fotenos et al., 2005; Fouquet et al., 2009; Frings et al., 2011; Hedman et al., 2011; Holland et al., 2009; Hua et al., 2010; Hua et al., 2009; Jack Jr et al., 2009; Josephs et al., 2008; Kalkers et al., 2002; Kasai et al., 2003; Paviour et al., 2006; Sabuncu et al., 2011; Sluimer et al., 2008; Whitwell et al., 2007). These summary measures are then submitted to standard parametric or non-parametric statistical methods for cross-sectional analysis. Such an approach is not appropriate when the data are unbalanced over time, since summary measures will not be drawn from the same distribution (e.g. will have different variance), violating a fundamental assumption made by standard statistical methods. In addition, as our experiments demonstrate there can be a significant loss in statistical power due to ignoring the correlation among the repeated measures and omitting subjects with a single time-point.

2.4 Longitudinal ADNI Data

In our experiments presented in the following section, we analyzed longitudinal brain MRI data (T1-weighted, 1.5 Tesla) from the Alzheimer Disease Neuroimaging Initiative (ADNI). The data were processed with FreeSurfer (version 5.1.0, http://surfer.nmr.mgh.harvard.edu) and its new longitudinal processing pipeline (http://surfer.nmr.mgh.harvard.edu/fswiki/LongitudinalProcessing) (Reuter and Fischl, 2011; Reuter et al., 2010; Reuter et al., 2012). The FreeSurfer processing pipeline is fully automatic and includes steps to compute a representation of the cortical surface between white and gray matter, a representation of the pial surface, a segmentation of white matter

from the rest of the brain; to perform skull stripping, bias field correction, nonlinear registration of the cortical surface of an individual with a stereotaxic atlas, labeling of regions of the cortical surface, and labeling of sub-cortical brain structures. Furthermore, for each MRI scan, FreeSurfer *automatically* computes subject-specific thickness measurements across the entire cortical mantle and within anatomically defined cortical regions of interest (ROIs) such as the entorhinal cortex, volume estimates of a wide range of sub-cortical structures such as the hippocampus, and estimates of the intra-cranial volume (ICV). In all subsequent analyses, we summed the volumes of the two hippocampi to obtain the total hippocampal volume and averaged thickness measurements from the bilateral entorhinal cortex ROIs to compute the mean thickness within the entorhinal cortex.

The longitudinal stream in FreeSurfer (Reuter et al., 2012) utilizes an unbiased subject-specific template (Reuter and Fischl, 2011), which is created by co-registering scans from each time-point using a robust and inverse consistent registration algorithm (Reuter et al., 2010). Several steps in the processing of the serial MRI scans (e.g., skull stripping, atlas registration, etc.) are then initialized with common information from the subject-specific template. This strategy has been shown to lead to increased statistical power and better separation of groups based on atrophy rates (Reuter et al., 2012). Note that the publicly distributed version of FreeSurfer's longitudinal stream does not handle subjects with a single MRI scan (i.e., single visit), which traditionally have been processed using cross-sectional tools. Since the cross-sectional image processing steps are different from the longitudinal stream, inclusion of single time point measurements in subsequent statistical analysis can introduce a bias, as demonstrated in our supplementary analysis. See also (Reuter et al., 2012) where a similar bias was quantified by processing the first time point cross-sectionally and the second longitudinally (initializing it with results from the first) in a test-retest study with no expected structural change. To address this issue we modified FreeSurfer's longitudinal framework to process subjects with a single time point in the following manner: we created a pose normalized (upright) version of the input images by symmetrically registering it with its left-right reversed image into a mid-space (Reuter et al., 2010), we then processed it as the subject-specific template and used it for the initialization of subsequent image processing steps, such as skull

stripping. This ensures the input image from a subject with a single scan undergoes the same processing and interpolation steps as serial images in the longitudinal stream and thus makes results comparable (see Supplementary Material).

Tables 1 and 2 provide descriptive statistics of the analyzed sample. We subdivided the subjects into five clinical groups. (1) Stable healthy control (HC): those who were clinically healthy throughout the follow-up period. (2) Converter HC (cHC): those who were clinically healthy at baseline but converted to Mild Cognitive Impairment (MCI, a transitional phase between healthy and dementia) (Gauthier et al., 2006) or dementia stage of Alzheimer's disease (AD) within the follow-up period. (3): Stable MCI (sMCI): those who were categorized MCI at baseline and remained so throughout the study. (4) Converter MCI (cMCI): those who were MCI at baseline and progressed to the dementia phase of AD during follow-up. (5) AD patients: those who were diagnosed with dementia of the Alzheimer type at baseline.

In our experiments, we only focused on two biomarkers, namely mean thickness within the entorhinal cortex (averaged across hemispheres; ECT) and total hippocampal volume (HV), since these are two classical MRI-derived markers that are known to be strongly associated with early AD (Dickerson et al., 2001; Jack Jr et al., 1997). These measurements were automatically computed using FreeSurfer.

-----Tables 1 and 2 about here-----

ADNI is a multi-site study, where the MRI data were collected using a range of scanner types. Although a significant amount of effort was put into matching the imaging protocol and quality across sites (via phantom and subject scans), there is still a chance that the coil type has an effect on the analysis. We conducted a supplementary analysis to assess this effect. Our results indicate that there were two coil types that had a significant influence on the measurement of hippocampal volume (see Supplementary Table S3), but our general conclusions about longitudinal changes were not altered. Since there was a significant number of subjects for which coil type information were not provided (and therefore these subjects were omitted from the supplementary analysis), we decided to drop coil type information from all our subsequent analyses in order to boost sample size.

Unless specified otherwise, all analyses included the following independent variables as fixed effects: time from baseline, clinical group membership (HC was the reference group and there were indicator variables for all remaining groups. E.g., for the sMCI indicator, the value was one if the subject was clinically categorized as sMCI and zero otherwise), the interaction between clinical group indicators and time from baseline, baseline age, sex, APOE genotype status (one if e4 carrier and zero if not), the interaction between APOE genotype status and time (of scan) from baseline (note that this variable was included based on the evidence that e4 accelerates atrophy during the prodromal phases of AD (Jack Jr et al., 2008)), and education (in years). Furthermore an estimate of intra-cranial volume (ICV) (Buckner et al., 2004) was included as a fixed effect for the analysis of HV, but not ECT since there was no significant association with the latter. Random effects were determined via a likelihood ratio test as explained above. In all analyses *both* intercept and time were included in the final model as random effects. This suggests that compound symmetry did not hold for HV and ECT in the longitudinal ADNI.

In general, longitudinal studies are conducted to assess group differences between the trajectories of variables of interest. Therefore, we constrained our analysis to the association between the group-time interaction (i.e., group-specific atrophy rate) for the two biomarkers: HV and ECT.

3 RESULTS

3.1 Comparing rates of atrophy across four clinical groups

In our first experiment, we excluded converter HC subjects, since this is the smallest group (N=17) and little has been reported on this group in prior work. Our goal here is to illustrate the LME methodology for characterizing well-known differences between four well-studied clinical groups: HC, stable MCI, converter MCI and AD patients (see Tables 1 and 2 and previous section). Figure 1 shows the *lowess* plots for the two biomarkers (HV and ECT) in these four clinical groups. These plots reveal that a linear model is likely to be sufficient to capture follow-up trends and there is no need for including higher order terms for time.

---- Figure 1 about here ----

The hypotheses we tested and the inference results (F-value, degrees of freedom –DF-, and uncorrected p-value) are as follows. Note that somewhat unusually, the DF depends on the contrast, because of the Satterthwaite-based approximation we use (see Equation 2.7). We include exact expressions for these hypotheses in the Supplementary Material.

H1) Is there any difference in the rate of change among the four groups (HC, sMCI, cMCI, and AD)?
HV: F value = 43.7, DF = [3 645.3], p = 0
ECT: F value = 40.4, DF = [3 632.9], p = 0

H2) Is there any difference in the rate of change between HC and sMCI?
HV: F value = 13.8, DF = [1 552.9], p = 2.3e-4
ECT: F value = 14.6, DF = [1 526.7], p = 1.5e-4

H3) Is there any difference in the rate of change between sMCI and cMCI?
HV: F value = 28.3, DF = [1 578.3], p = 1.5e-7
ECT: F value = 30.3, DF = [1 554.3], p = 5.5e-8

H4) Is there any difference in the rate of change between cMCI and AD?
HV: F value = 5.1, DF = [1 798.8], p = 0.02
ECT: F value = 1.4, DF = [1 830.6], p = 0.22

Figure 2 shows the retrospective power (Equation 2.8) for comparing the rates of atrophy between sMCI and cMCI using the ADNI data. ECT provides slightly more power than HV in detecting longitudinal group differences. Table 3 provides sample size estimates (based on Equation 2.9) for prospective studies that compare atrophy rates between sMCI versus cMCI and AD versus HC. Effect sizes and dropout rates were computed based on the ADNI sample.

---- Figure 2 and Table 3 about here ----

3.2 Comparing rates of atrophy between HC and converter HC

Our second experiment focused on the converter HC (cHC) subjects (N=17), who were clinically healthy at baseline yet progressed to MCI or AD over the course of the study. Mean time for conversion was 2.6 years from baseline (with a standard deviation of 1.1 years). We compared HV and ECT atrophy rates between cHC and HC subjects. Figure 3 shows the corresponding lowess plots. For entorhinal cortex, the lowess plot

suggests that cHC subjects exhibit a nonlinear trajectory, which can be captured with the following piecewise linear model:

$$\beta_1 + \beta_2 t + \beta_3 (t - 1.2)_+, \quad (3.1)$$

where t is time (in years) from baseline, and $(x)_+$ is only nonzero and equal to x if x is positive and zero otherwise. We note that the term 1.2 in Equation (3.1) comes from the visual inspection of Figure 3b that reveals a breakpoint in the trajectory of ECT around 1.2 years. For the hippocampus, we adopted a simple linear model as we did in the previous experiment.

---- Figure 3 about here ----

The hypotheses we tested and the inference results are as follows.

H5) Is there any difference between the trajectories of cHC and HC?
HV: F value = 8.8, DF = [1 218.0], p = 0.0034
ECT[2]: F value = 4.3, DF = [2 392.7], p = 1.5e-4

H6) Is there any difference between the first and second slopes of the piecewise linear model in cHC subjects?
ECT: F value = 4.5, DF = [1 622.3], p = 0.034

H7) Is there any difference in the first slopes of HC and sHC subjects?
ECT: F value = 0.0, DF = [1 685.4], p = 0.97

H8) Is there any difference in the second slopes of HC and sHC subjects?
ECT: F value = 7.6, DF = [1 514.2], p = 0.006

Figure 4 shows the retrospective power (Equation 2.10) for comparing the rates of atrophy between HC and cHC using the ADNI data. Here, HV provides slightly more power than ECT in detecting longitudinal group differences.

---- Figure 4 about here ----

3.3 Comparison of LME to alternative methods

In the third experiment, our goal was to provide an objective comparison of the LME approach with the two widely-used alternative methods, namely repeated measures

[2] Note that the inference involves two parameters corresponding to the two slopes in the piecewise model of (3.1).

ANOVA (rm-ANOVA) and cross-sectional analysis of the slope (x-slope), i.e. annualized rate of atrophy estimated for each individual. We implemented rm-ANOVA via a LME model with a single random effect for the intercept. As we discuss above, this imposes a compound symmetry structure on the covariance between repeated measures – a model that is unlikely to be appropriate for typical LNI data. For the second benchmark, we estimated each subject's slope using the best-fit line (in the least square sense) to its longitudinal measurements. Then we conducted a standard least-square regression (GLM) with the same independent variables as the other two methods.

We were interested in assessing the specificity, sensitivity and reliability of the three methods in a realistic longitudinal design. To achieve this, we conducted two-group comparison analyses on the rates of HV loss in HC subjects and AD patients, using an empirical strategy inspired by (Thirion et al., 2007). There were two main reasons for our particular choice of biomarker and groups. Firstly, from prior work we were confident that there is a significant difference between the HV atrophy rates of HC and AD groups (Jack Jr et al., 2010). Secondly, our sample size estimates (see Table 3) indicated that with a relatively small number of subjects, we had a good chance of detecting the difference in atrophy rates. Hence, we could draw a relatively large number of pseudo-independent subsamples (with say N = 10-30 subjects from each group) from the entire ADNI sample to conduct our analyses.

For each sample size value (e.g. N = 15 per group), we randomly selected two sets of independent AD+HC samples, (i.e., two independent samples of 2N) from the eligible portion of the ADNI sample (all ADNI HC and AD subjects). There was no overlap between the two independent samples and each sample contained the same number of AD and HC subjects. We repeated this procedure 200 times to obtain 200 random pairs of independent AD+HC samples of a certain size (that is, 400 random AD+HC samples in total).

For each sample, we used the three methods (LME, rm-ANOVA and x-slope) to compute parametric p-values for the difference between the rates of atrophy of the two clinical groups (AD vs. HC). Next, we conducted a permutation test (Good, 2000; Nichols and Holmes, 2002) for each sample by shuffling the clinical group memberships and repeating the inference (2000 permutations). A non-parametric p-value was

computed for each sample and each method based on the ranking (with respect to the 2000 permutations) of the corresponding parametric p-values. The permutation approach relies on assumptions that are weaker than those required for the parametric p-values and is known to yield an accurate assessment of the probability of false positive (type 1 error, p-value, or equivalently specificity) when the number of permutations is large (Nichols and Holmes, 2002).

Thus, we considered the agreement between the parametric and non-parametric p-values as a measurement of the accuracy of the parametric p-values, or the specificity of the parametric model. Figure 5 shows the mean (averaged across the 400 random AD+HC samples) absolute difference between the parametric and non-parametric p-values for different sample sizes and different methods. These results revealed that both LME and x-slope provided significantly higher specificity than rm-ANOVA for modest sample sizes (2N less than 50).

---- Figure 5 about here ----

To assess sensitivity, we computed the detection (true positive) rate across the 400 samples (200 pairs) for a range of p-value (alpha) thresholds and 2N=20 (see Figure 6). Here we assumed that the underlying ground truth was that there is a difference between hippocampal atrophy rates of HC versus AD subjects. Instances where the p-value was less than an alpha threshold were considered a “detection” and remaining cases were treated as a false negative. The true positive rate (or sensitivity) was quantified as the fraction of detections. Our results indicate that LME yields significantly higher sensitivity than the two alternative approaches. Note that, these results indicate we have about 70% power with the threshold (alpha) set to 0.05 and 2N = 20. This is in agreement with the approximate sample size estimate computed for 80% power (Table 3).

---- Figure 6 about here ----

Finally, we were interested in quantifying repeatability, by comparing results between the two independent samples obtained at each random draw (200 pairs). Figure 7 shows

the rate at which each method was able to detect the difference in *both* samples for a range of p-value thresholds (alpha values). These results suggest that LME yields longitudinal findings that are more likely to be repeatable in an independent sample.

---- Figure 7 about here ----

3.4 Assessing the effect of including subjects with a single time point

In this final experiment, our goal was to quantify the effect of including subjects with a single time-point into the LME-based analysis of longitudinal data. The theoretical expectation is that data from subjects with a single visit may contain valuable information about between-subject variability, which can in turn improve our inference on the remaining longitudinal measurements. In practice, most studies choose to exclude these subjects in their analyses, because their methods cannot handle these cases and/or they are cautious of introducing a bias into the analysis, since there might be inter-group differences in dropout rates. However, the LME approach recommends to include all scans from all time-points into the analysis (Fitzmaurice et al., 2011).

As an objective assessment, we conducted the following experiment. We first established a sample of 50HC+50AD subjects from the ADNI data, in which each subject has four repeated measurements (MRI-derived hippocampal volume). We call this the "full sample." We then performed 1000 simulations. In each simulation we randomly selected 20 subjects from the AD group (20% of the full sample) to remove their last three repeated measures from the data (therefore leaving only their baseline HV measurements). Thus, for each simulation we had a "reduced sample," which consisted of a group of 50HC+30AD completers (i.e., they had all four repeated measures) and 20 AD subjects with a single measurement ("dropouts"). We then fit two LME models with the same independent variables as above: one model was based on the reduced sample excluding the dropouts (i.e., only 50HC+30AD completers). The second model was computed based on the entire reduced sample, which included the 20 AD dropouts. We then compared these model fits with that obtained on the full sample. Figure 8 shows the difference between the fixed effect coefficient estimates obtained on the reduced sample (with and without the dropouts) and full sample. These results suggest that including

subjects with a single time-point (dropouts) increases the accuracy of the model fit, and would thus lead to improved inference.

---- Figure 8 about here ----

4 DISCUSSION

Linear Mixed Effects (LME) models offer a more powerful and versatile framework for the analysis of longitudinal data than many other popular methods (Fitzmaurice et al., 2011). The LME approach elegantly handles unbalanced data (with variable missing rates across time-points and imperfect timing), makes use of subjects with a single time-point to characterize inter-subject variation, and provides a parsimonious way to represent the group mean trajectory and covariance structure between serial measurements. Yet, its use in neuroimaging seems to be limited to a small number of studies, which represent a minority in the rapidly growing LNI literature. We found that many prior LNI studies used sub-optimal approaches that at best offer reduced power to detect effects and at worst can lead to incorrect inferences. Our goal in this work was to advocate the use of LME models for LNI data analysis by providing the theoretical background and the implementation of an array of computational tools that build on the LME framework. We intended to illustrate the proper use of these tools using a well studied, real-life longitudinal dataset. Finally and most importantly, we provided a validation of our tools and an objective comparison with two popular alternative methods via analyses on these data.

In the first experiment, we applied the LME model to a well-known pair of AD biomarkers (hippocampal volume –HV- and entorhinal cortex thickness –ECT-) and obtained results that were in agreement with prior work. The lowess plots revealed that a linear model was suitable to characterize the longitudinal trajectories in the follow-up period. Our inferences indicated that there was a significant difference between the HV and ECT atrophy rates across HC, sMCI, and cMCI subjects. This difference diminished (and became statistically insignificant for ECT) when comparing cMCI subjects and AD patients.

In the second experiment, we compared atrophy rates between HC subjects and converter HC subjects, who were clinically healthy at baseline but progressed to MCI or

clinical AD at follow-up. The lowess plots revealed an intriguing, nonlinear trajectory of entorhinal cortex thickness in the cHC group, which could be captured via a piece-wise linear model with a knot at 1.2 years. Our LME-based inference further confirmed that this was an appropriate model, since the two slopes of the piece-wise linear model were statistically significantly different. Intriguingly, the knot (or elbow) of the piece-wise linear model (at around 1.2 years) was on average about 1.4 years prior to the event of clinical conversion, suggesting that atrophy rates accelerate prior to the beginning of clinical symptoms. Furthermore, our inferences confirmed that in the cHC group both HV and ECT exhibited an overall longitudinal trajectory that was statistically significantly different from the controls. For ECT this difference was driven by the apparently sudden acceleration of atrophy in the cHC subjects at around the end of the first year of the study. For HV, there was no such nonlinearity that was discernible in the group trajectories.

In the third experiment, our goal was to provide an objective assessment of the three competing methods widely used to analyze longitudinal data. We focused on HV, a well-established marker of AD, which also has a relatively large effect size. This enabled us to interrogate a large number of random sub-samples of relatively small size, where the effect of interest was detectable and average across these random experiments. The ADNI data, with its variable missing data pattern, imperfect follow-up timing, and multi-site nature, provided a perfect example of a realistic LNI study, in which we can objectively quantify the performance of the different methods. Our results supplied evidence supporting our theoretical expectations: the LME approach provides more sensitivity in a realistic LNI setting than repeated measures ANOVA or the analysis of summary metrics such as annualized atrophy rates, with good control on specificity. Furthermore, the resulting findings are more likely to be replicated in an independent study.

Finally, in a fourth experiment, we aimed to quantify the improvement in model fit afforded by the LME method by including subjects with a single time-point. To achieve this, we first established a full dataset with 50 AD and 50 HC subjects, all of which had four scans. Then we simulated 1000 random subsets of this sample, where 20 AD patients dropped out after the first visit. Our results, once again, were in line with the theoretical

expectations: including subjects with a single time point can dramatically improve the accuracy of the model fit in the LME approach.

The present study focused on the univariate analysis, where the correction for "multiple comparisons" is not an issue. In future work, we intend to extend the LME framework and our computational tools to the mass-univariate setting, where one interrogates effects across a large number of pixels/voxels. This will be the topic of an upcoming follow-up paper.

5 CONCLUSIONS

The Linear Mixed Effects (LME) approach provides a powerful and flexible framework for the analysis of LNI data. We have implemented and validated these computational tools, which will be made freely available within FreeSurfer to complement its longitudinal image-processing pipeline.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the National Institutes of Health (NIH) (grant U01 AG024904). The ADNI is funded by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott Laboratories, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation Plc, Genentech Inc, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson Services Inc, Eli Lilly and Company, Medpace Inc, Merck and Co Inc, Novartis International AG, Pfizer Inc, F. Hoffman-La Roche Ltd, Schering-Plough Corporation, CCBR-SYNARC Inc, and Wyeth Pharmaceuticals, as well as nonprofit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the US Food and Drug Administration. Private sector contributions to the ADNI are facilitated by the Foundation for the NIH. The grantee organization is the Northern California Institute for Research and Education Inc, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. The

ADNI data are disseminated by the Laboratory for NeuroImaging at the University of California, Los Angeles.

Support for this research was provided in part by the National Center for Research Resources (P41-RR14075), the National Institute for Biomedical Imaging and Bioengineering (R01EB006758), the National Institute on Aging (AG022381), the National Center for Alternative Medicine (RC1 AT005728-01), the National Institute for Neurological Disorders and Stroke (R01 NS052585-01, 1R21NS072652-01, 1R01NS070963, 2R01NS042861-06A1, 5P01NS058793-03), the National Institute of Child Health and Human Development (R01-HD071664), and was made possible by the resources provided by Shared Instrumentation Grants 1S10RR023401, 1S10RR019307, and 1S10RR023043. Additional support was provided by The Autism & Dyslexia Project funded by the Ellison Medical Foundation, and by the NIH Blueprint for Neuroscience Research (5U01-MH093765), part of the multi-institutional Human Connectome Project. Dr. Sabuncu received support from a KL2 Medical Research Investigator Training (MeRIT) grant awarded via Harvard Catalyst, The Harvard Clinical and Translational Science Center (NIH grant #1KL2RR025757-01 and financial contributions from Harvard University and its affiliated academic health care centers), and an NIH K25 grant (NIBIB 1K25EB013649-01).

Finally, the authors would like to thank Nick Schmansky and Louis Vinke for their efforts in downloading and processing the ADNI MRI scans.

TABLES

Table 1. Longitudinal ADNI sample characteristics

Variable	Stable HC	Converter HC	Stable MCI	Converter MCI	AD	*p*-value
Number of subjects	210	17	227	166	188	
Baseline age	75.9 ± 5 [60-90]	76.7 ± 5.1 [63-84]	74.8 ± 7.7 [55-90]	74.7± 7.1 [55-89]	75.2 ± 7.5 [55-91]	0.3464
Female %	48.1	47.1	33.48	38.6	47.3	<0.01[a]
APOE-ε4 Carriers %	25.7	41.2	43.2	67.5	66	<0.0001[a]
Education	16.1 ± 2.8 [6-20]	16.1 ± 2.8 [12-20]	15.6 ± 3.1 [4-20]	15.7 ± 2.9 [6-20]	14.7 ± 3.2 [4-20]	<0.001

Baseline age (in years) and education values are in mean ± standard deviation; Ranges are listed in square brackets; p-values indicate effects across the groups

Key: Converter MCI, mild cognitive impairment subjects who convert to Alzheimer's disease; Converter HC, healthy controls who convert to either MCI or Alzheimer's disease.

[a] Using Fisher's exact test; ANOVA-derived p-values were used in the other cases.

Table 2. Number and timing of scans per time point by clinical group (Stable HC, N=210; Converter HC, N=17; Stable MCI, N=227; Converter MCI, N=166; AD, N=188).

Time point	Stable HC	Converter HC	Stable MCI	Converter MCI	AD	Time from baseline
baseline	210	17	227	166	188	0
year 0.5 (month 6)	197	17	194	161	166	0.58 ± 0.07 [0.21-0.94]
year 1	183	17	177	153	150	1.08 ± 0.07 [0.68-1.38]
year 1.5	0	0	153	136	0	1.59± 0.08 [1.26-1.92]
year 2	129	14	108	106	96	2.09 ± 0.10 [1.58-2.88]
year 3	115	6	68	70	0	3.09 ± 0.09 [2.52-3.45]
year 4	11	0	3	10	0	4.12 ± 0.09 [3.98-4.38]
Total	845	71	930	802	600	

Time from baseline (in years) is in mean ± standard deviation; Ranges are listed in square brackets.

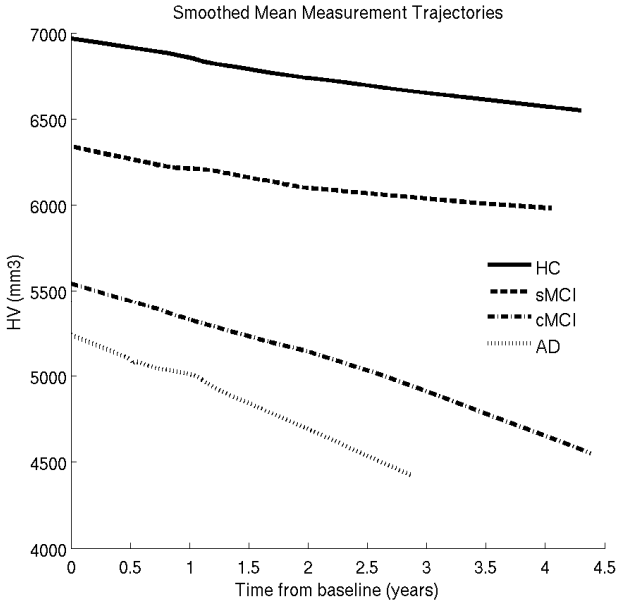
Table 3. Conservative estimates of total sample size (2N, where N is the number of subjects in each group) for two prospective longitudinal studies (two-year studies with 5 serial scans obtained every six months from baseline) comparing Alzheimer patients (AD) vs healthy controls (HC) and stable MCI (sMCI) vs converter MCI (cMCI) groups, respectively. The power is set to 80% and the effect size (rate of change per year) is set to the slope regression coefficient estimated by the analysis of the ADNI data. Sample size estimates were *inflated* by a factor of 1.84 based on the drop out rate observed in the ADNI data (45.5% of subjects dropped out at the end of 2 years).

Prospective longitudinal studies	Effect size (per year)	Total sample size
AD vs HC / HV	-131.94 mm^3	30
AD vs HC / ECT	-0.1 mm	32
cMCI vs sMCI / HV	-62.99 mm^3	162
cMCI vs sMCI / ECT	-0.05 mm	146

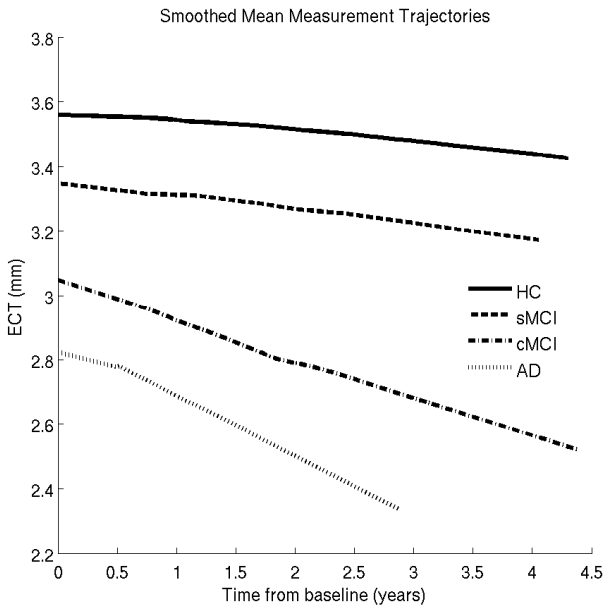
Key: HV, total hippocampal volume; ECT, average entorhinal cortical thickness

FIGURES

Figure 1. Locally weighted smoothed mean measurement trajectory (lowess plot) for each of the four clinical groups. This method produces a smooth curve by centering a window of fixed size at each time-point and fitting a straight line to the data within that window. The lowess estimate of the mean at a time-point is simply the predicted values at that time-point from the fitted regression line. In this plot, the fraction of the total number of data points included in the sliding window was set to 0.7. HC: healthy control; sMCI: stable MCI; cMCI: converter MCI; AD: Alzheimer patients. (A) Hippocampal volume (HV). (B) Entorhinal cortex thickness (ECT).



(A)



(B)

Figure 2. Statistical power versus alpha (false positive rate) to discriminate the atrophy rates of stable and converter MCIs. HV: hippocampal volume. ECT: entorhinal cortex thickness.

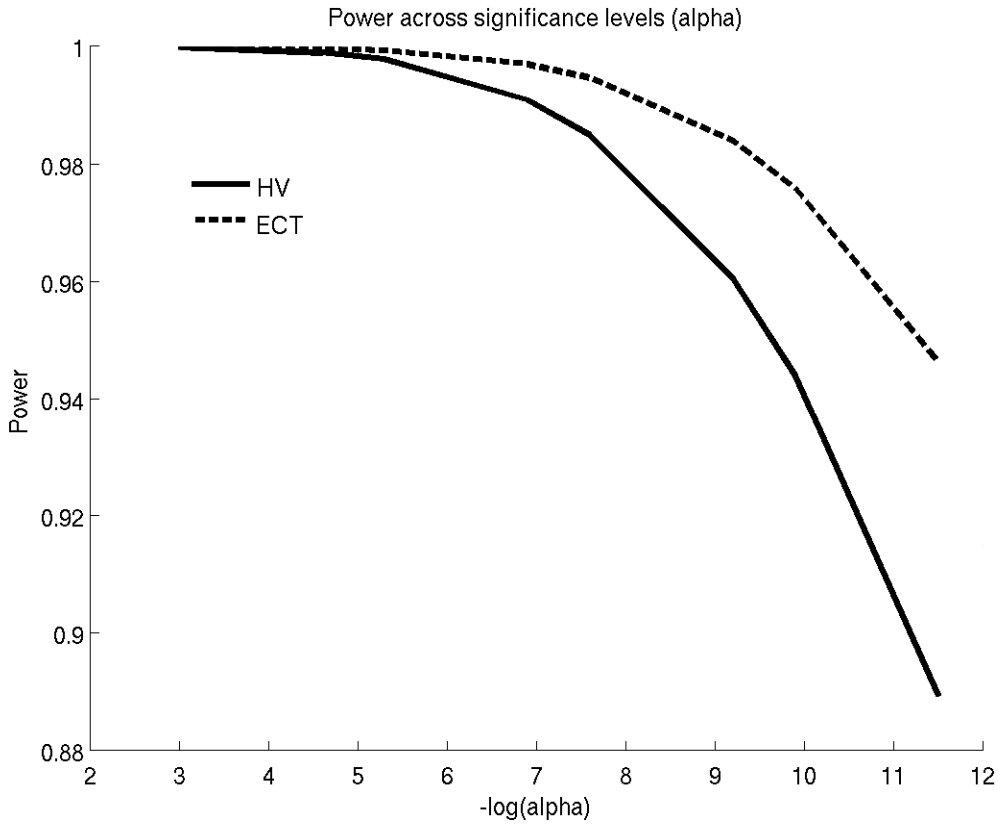
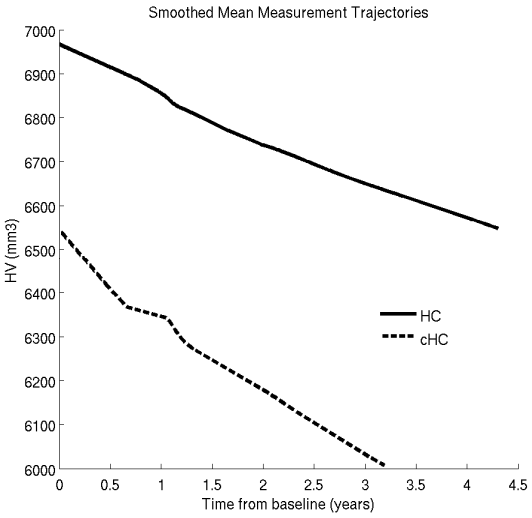
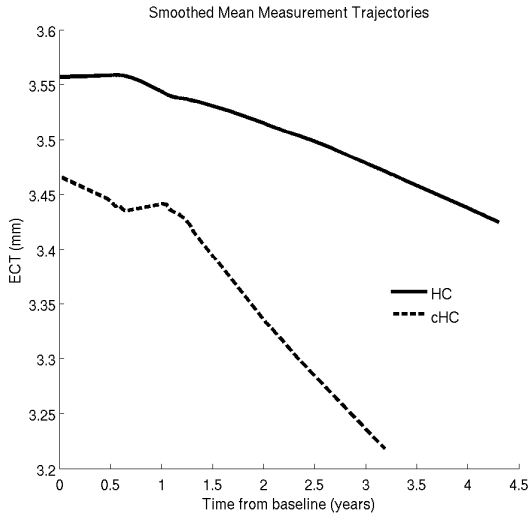


Figure 3. Locally weighted smoothed mean measurement trajectory (lowess plot) for two groups. This method produces a smooth curve by centering a window of fixed size at each time-point and fitting a straight line to the data within that window. The lowess estimate of the mean at a time-point is simply the predicted values at that time-point from the fitted regression line. In this plot, the fraction of the total number of data points included in the sliding window was set to 0.7. HC: healthy controls who remained so throughout the study; and cHC: converter HCs, who were healthy at baseline but progressed to MCI or AD during follow-up. Mean time to progression was 2.6 years from baseline. (A) Hippocampal volume (HV). (B) Entorhinal cortex thickness (ECT).



(A)



(B)

Figure 4. Statistical power versus alpha (false positive rate) to discriminate the atrophy rates of stable and converter healthy controls (HC). HV: hippocampal volume. ECT: entorhinal cortex thickness.

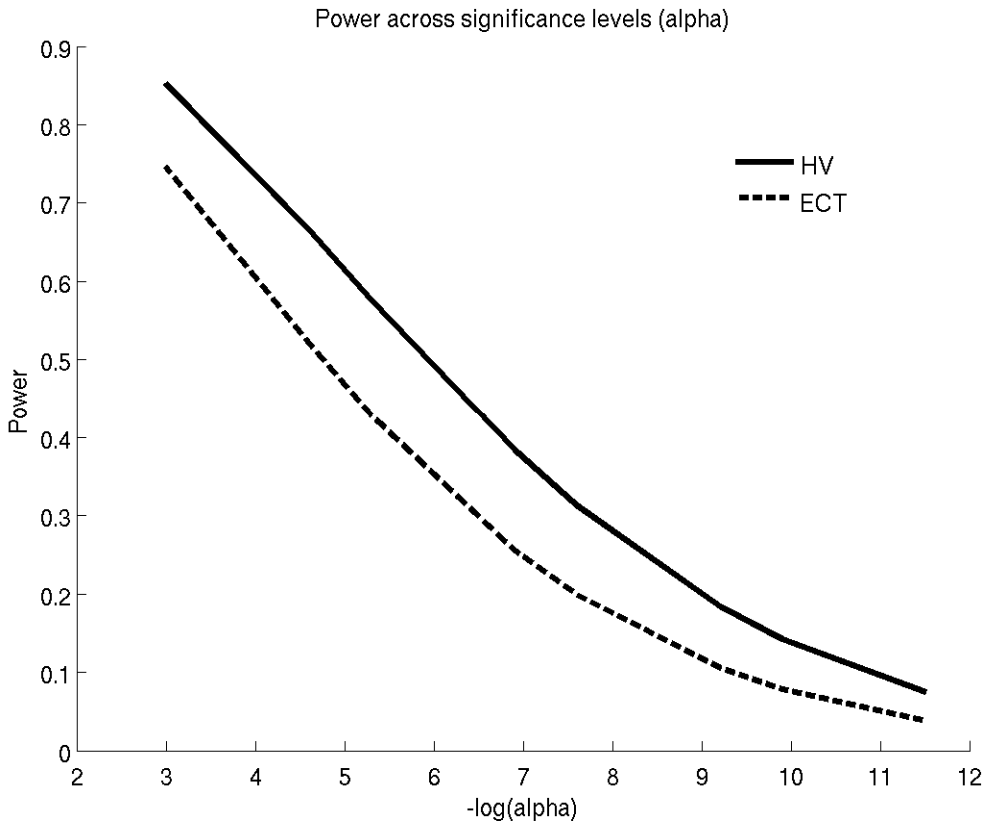


Figure 5. The mean absolute difference between non-parametric and parametric p-values for three statistical methods in comparing hippocampal volume loss rates between healthy controls (HC) and Alzheimer patients (AD) (Experiment 3) as a function of total sample size. LME: Linear Mixed Effects model with random intercept and slope. Rm-ANOVA: random effects ANOVA. X-Slope: GLM-based cross-sectional analysis of annualized rate of atrophy (slope).

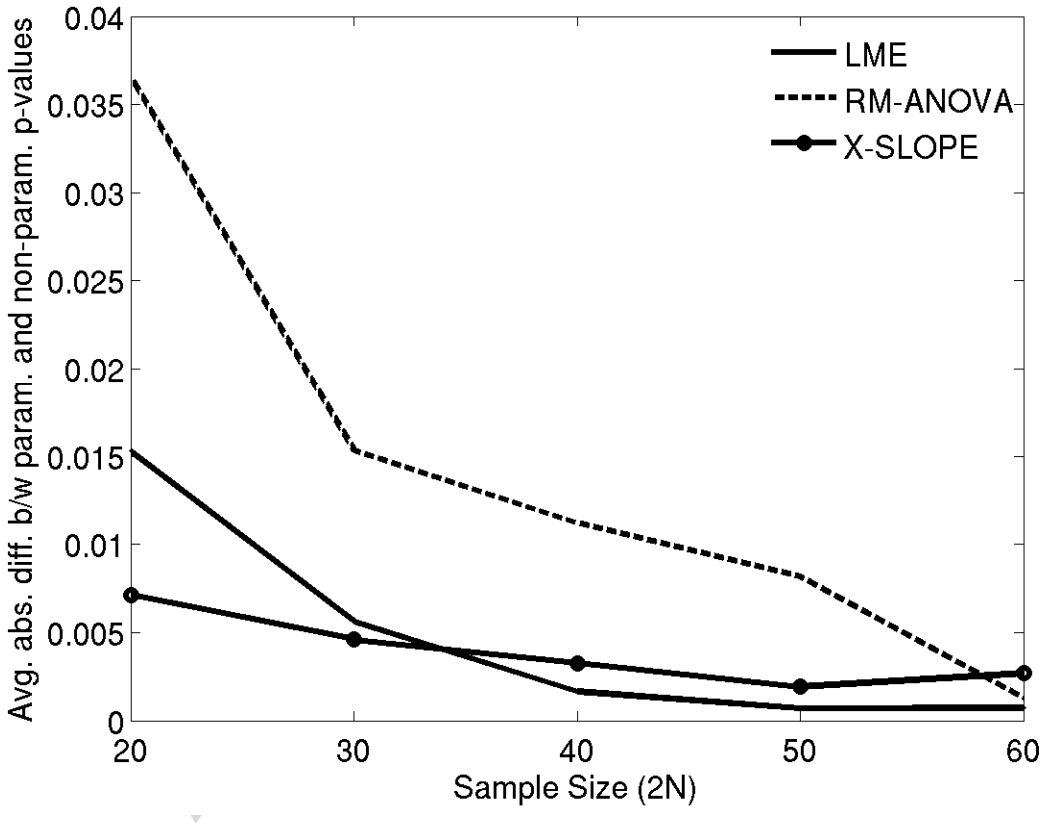


Figure 6. Detection rate (the frequency of true positives) in differentiating hippocampal volume loss rates between healthy controls and AD patients (Experiment 3), as a function of alpha (p-value threshold) with 2N=20 subjects. LME: Linear Mixed Effects model with random intercept and slope. Rm-ANOVA: random effects ANOVA. X-Slope: GLM-based cross-sectional analysis of annualized rate of atrophy (slope).

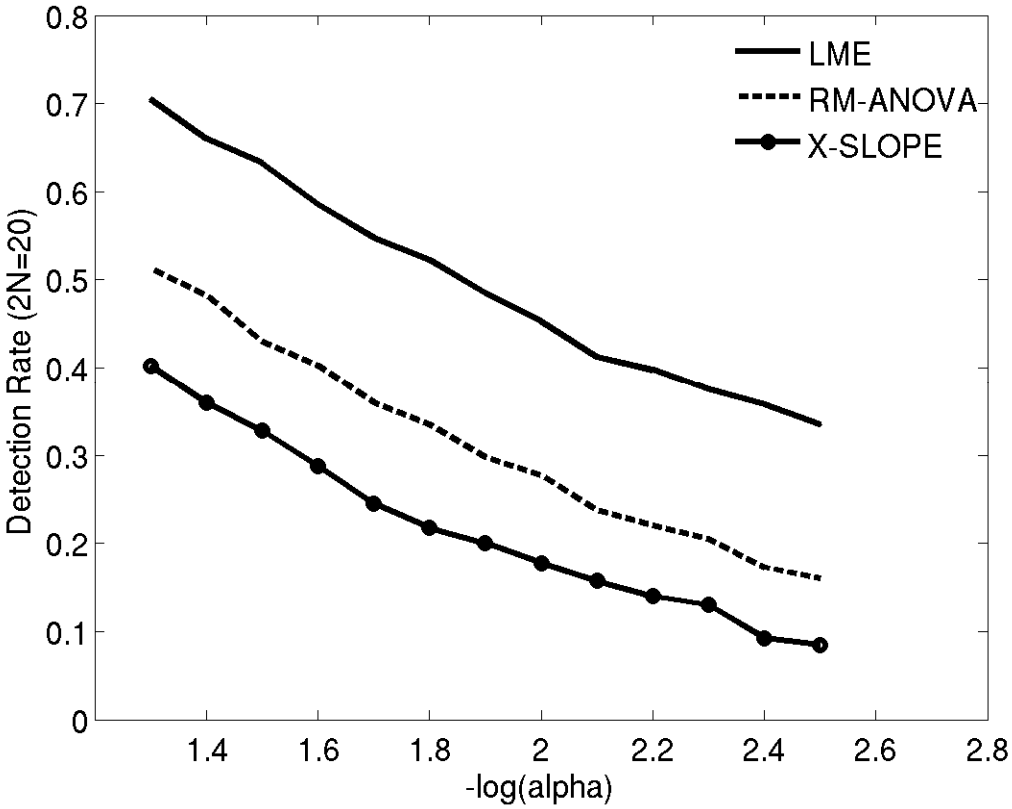


Figure 7. Repeatability (the frequency at which a method differentiates hippocampal volume loss rates between healthy controls and AD patients in *two independent* samples of 2N= 20) versus alpha (p-value threshold) (Experiment 3). LME: Linear Mixed Effects model with random intercept and slope. Rm-ANOVA: random effects ANOVA. X-Slope: GLM-based cross-sectional analysis of annualized rate of atrophy (slope).

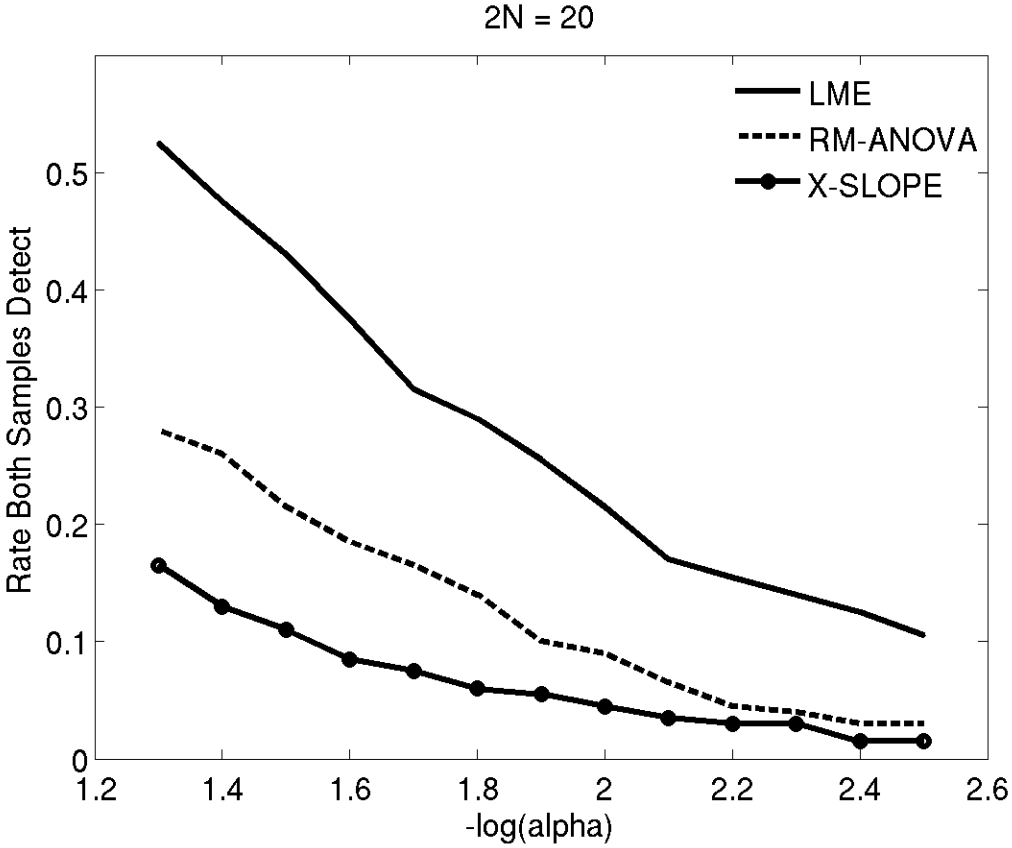
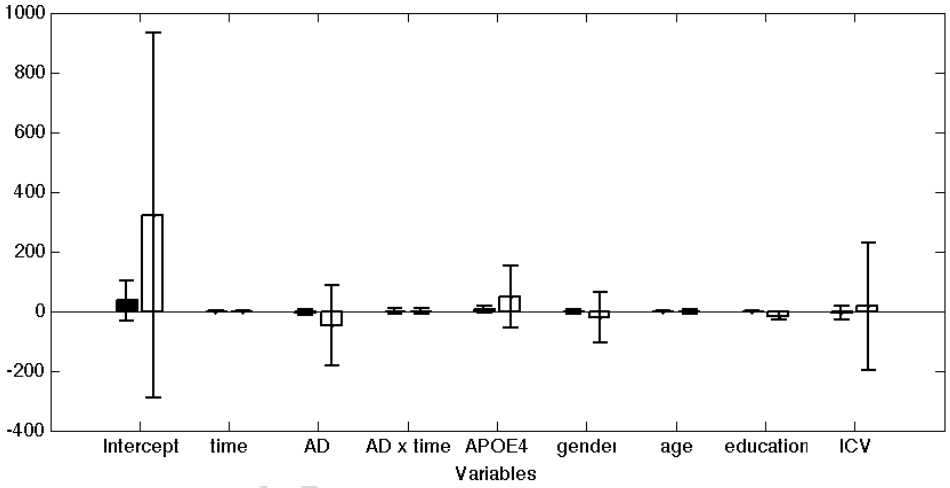


Figure 8. The influence of including subjects with a single time-point on LME-based inference results. MRI-derived total hippocampal volume was the dependent variable. The full sample contained 50 HC and 50 AD subjects, all with 4 visits (scans). We had 1000 random simulations, in which a reduced dataset was generated, by treating 20 random AD subjects as dropouts and discarding their last three scans. The y-axis shows the average *difference* between the coefficient estimates obtained on the reduced sample by including (black bars) or discarding (white bars) the 20 dropout AD patients, and the coefficients from the full sample. The error bars show the standard deviations across 1000 random simulations. These results suggest that including the subjects with a single time-point increases the accuracy of the model fit and introduces minimal bias.



REFERENCES

Asami, T., Bouix, S., Whitford, T.J., Shenton, M.E., Salisbury, D.F., McCarley, R.W., 2011. Longitudinal loss of gray matter volume in patients with first-episode schizophrenia: DARTEL automated analysis and ROI validation. Neuroimage.

Blockx, I., Van Camp, N., Verhoye, M., Boisgard, R., Dubois, A., Jego, B., Jonckers, E., Raber, K., Siquier, K., Kuhnast, B., Dolle, F., Nguyen, H.P., Von Horsten, S., Tavitian, B., Van der Linden, A., 2011. Genotype specific age related changes in a transgenic rat model of Huntington's disease. Neuroimage 58, 1006-1016.

Bonne, O., Brandes, D., Gilboa, A., Gomori, J.M., Shenton, M.E., Pitman, R.K., Shalev, A.Y., 2001. Longitudinal MRI study of hippocampal volume in trauma survivors with PTSD. The American journal of psychiatry 158, 1248.

Buckner, R.L., Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., Snyder, A.Z., 2004. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. NeuroImage 23, 724-738.

Chetelat, G., Landeau, B., Eustache, F., Mezenge, F., Viader, F., de La Sayette, V., Desgranges, B., Baron, J.C., 2005. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. Neuroimage 27, 934-946.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 829-836.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical Surface-Based Analysis* 1:: I. Segmentation and Surface Reconstruction. NeuroImage 9, 179-194.

Davatzikos, C., Resnick, S.M., 2002. Degenerative age changes in white matter connectivity visualized in vivo using magnetic resonance imaging. Cerebral cortex 12, 767-771.

Davis, C.E., Jeste, D.V., Eyler, L.T., 2005. Review of longitudinal functional neuroimaging studies of drug treatments in patients with schizophrenia. Schizophrenia research 78, 45-60.

Desikan, R.S., McEvoy, L.K., Thompson, W.K., Holland, D., Roddey, J.C., Blennow, K., Aisen, P.S., Brewer, J.B., Hyman, B.T., Dale, A.M., 2011. Amyloid β associated volume loss occurs only in the presence of phospho tau. Annals of neurology.

Dickerson, B.C., Goncharova, I., Sullivan, M., Forchetti, C., Wilson, R., Bennett, D., Beckett, L., deToledo-Morrell, L., 2001. MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. Neurobiology of aging 22, 747-754.

Dickerson, B.C., Sperling, R.A., 2005. Neuroimaging biomarkers for clinical trials of disease-modifying therapies in Alzheimer's disease. NeuroRx 2, 348-360.

Driscoll, I., Beydoun, M.A., An, Y., Davatzikos, C., Ferrucci, L., Zonderman, A.B., Resnick, S.M., 2011. Midlife obesity and trajectories of brain volume changes in older adults. Human Brain Mapping.

Fischl, B., 2012. Freesurfer. NeuroImage.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341-355.

Fischl, B., Sereno, M.I., Dale, A.M., 1999a. Cortical Surface-Based Analysis* 1:: II: Inflation, Flattening, and a Surface-Based Coordinate System. NeuroImage 9, 195-207.

Fischl, B., Sereno, M.I., Tootell, R.B.H., Dale, A.M., 1999b. High-resolution intersubject averaging and a coordinate system for the cortical surface. Human Brain Mapping 8, 272-284.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2011. Applied longitudinal analysis. Wiley.

Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., 2009. One-year brain atrophy evident in healthy aging. The Journal of Neuroscience 29, 15223-15231.

Fotenos, A.F., Snyder, A., Girton, L., Morris, J., Buckner, R., 2005. Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology 64, 1032-1039.

Fouquet, M., Desgranges, B., Landeau, B., Duchesnay, E., Mézenge, F., De La Sayette, V., Viader, F., Baron, J.C., Eustache, F., Chételat, G., 2009. Longitudinal brain metabolic changes from amnestic mild cognitive impairment to Alzheimer's disease. Brain 132, 2058-2067.

Frings, L., Mader, I., Landwehrmeyer, B.G., Weiller, C., Hüll, M., Huppertz, H.J., 2011. Quantifying change in individual subjects affected by frontotemporal lobar degeneration using automated longitudinal MRI volumetry. Human Brain Mapping.

Friston, K.J., 2007. Statistical parametric mapping: the analysis of functional brain images. Academic Press.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R.C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., 2006. Mild cognitive impairment. The Lancet 367, 1262-1270.

Ge, Y., Grossman, R., Udupa, J., Fulton, J., Constantinescu, C., Gonzales,ÄìScarano, F., Babb, J., Mannon, L., Kolson, D., Cohen, J., 2000. Glatiramer acetate (Copaxone) treatment in relapsing-remitting MS. Neurology 54, 813-817.

Giedd, J.N., Blumenthal, J., Jeffries, N.O., Castellanos, F.X., Liu, H., Zijdenbos, A., Paus, T., Evans, A.C., Rapoport, J.L., 1999. Brain development during childhood and adolescence: a longitudinal MRI study. Nature neuroscience 2, 861-862.

Girden, E.R., 1992. ANOVA: Repeated measures. Sage Publications, Inc.

Good, P.I., 2000. Permutation tests. Wiley Online Library.

Hedman, A.M., van Haren, N.E.M., Schnack, H.G., Kahn, R.S., Hulshoff Pol, H.E., 2011. Human brain changes across the life span: A review of 56 longitudinal magnetic resonance imaging studies. Human Brain Mapping.

Helms, R.W., 1992. Intentionally incomplete longitudinal designs: I. Methodology and comparison of some full span designs. Statistics in Medicine 11, 1889-1913.

Ho, B.C., Andreasen, N.C., Nopoulos, P., Arndt, S., Magnotta, V., Flaum, M., 2003. Progressive structural brain abnormalities and their relationship to clinical outcome: a

longitudinal magnetic resonance imaging study early in schizophrenia. Archives of General Psychiatry 60, 585.

Holland, D., Brewer, J.B., Hagler, D.J., Fennema-Notestine, C., Dale, A.M., Weiner, M., Thal, L., Petersen, R., Jack Jr, C.R., Jagust, W., 2009. Subregional neuroanatomical change as a biomarker for Alzheimer's disease. Proceedings of the National Academy of Sciences 106, 20954-20959.

Holland, D., McEvoy, L.K., Dale, A.M., 2011. Unbiased comparison of sample size estimates from longitudinal structural measures in ADNI. Human Brain Mapping.

Hua, X., Hibar, D.P., Lee, S., Toga, A.W., Jack Jr, C.R., Weiner, M.W., Thompson, P.M., 2010. Sex and age differences in atrophic rates: an ADNI study with n= 1368 MRI scans. Neurobiology of aging 31, 1463-1480.

Hua, X., Leow, A.D., Levitt, J.G., Caplan, R., Thompson, P.M., Toga, A.W., 2009. Detecting brain growth patterns in normal children using tensor based morphometry. Human Brain Mapping 30, 209-219.

Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology 9, 119-128.

Jack Jr, C.R., Lowe, V.J., Weigand, S.D., Wiste, H.J., Senjem, M.L., Knopman, D.S., Shiung, M.M., Gunter, J.L., Boeve, B.F., Kemp, B.J., 2009. Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. Brain 132, 1355-1365.

Jack Jr, C.R., Petersen, R.C., Xu, Y.C., Waring, S.C., O'Brien, P.C., Tangalos, E.G., Smith, G.E., Ivnik, R.J., Kokmen, E., 1997. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. Neurology 49, 786-794.

Jack Jr, C.R., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Lowe, V., Kantarci, K., Bernstein, M.A., Senjem, M.L., Gunter, J.L., 2012. Shapes of the Trajectories of 5 Major Biomarkers of Alzheimer Disease. Archives of Neurology, archneurol. 2011.3405 v2011.

Jack Jr, C.R., Weigand, S.D., Shiung, M.M., Przybelski, S.A., O'Brien, P.C., Gunter, J.L., Knopman, D.S., Boeve, B.F., Smith, G.E., Petersen, R.C., 2008. Atrophy rates accelerate in amnestic mild cognitive impairment. Neurology 70, 1740-1752.

Josephs, K.A., Whitwell, J.L., Ahmed, Z., Shiung, M.M., Weigand, S.D., Knopman, D.S., Boeve, B.F., Parisi, J.E., Petersen, R.C., Dickson, D.W., 2008. β amyloid burden is not associated with rates of brain atrophy. Annals of neurology 63, 204-212.

Kaladjian, A., Jeanningros, R., Azorin, J.M., Nazarian, B., Roth, M., Anton, J.L., Mazzola Pomietto, P., 2009. Remission from mania is associated with a decrease in amygdala activation during motor response inhibition. Bipolar disorders 11, 530-538.

Kalkers, N.F., Ameziane, N., Bot, J.C.J., Minneboo, A., Polman, C.H., Barkhof, F., 2002. Longitudinal brain volume measurement in multiple sclerosis: rate of brain atrophy is independent of the disease subtype. Archives of neurology 59, 1572.

Kasai, K., Shenton, M.E., Salisbury, D.F., Hirayasu, Y., Onitsuka, T., Spencer, M.H., Yurgelun-Todd, D.A., Kikinis, R., Jolesz, F.A., McCarley, R.W., 2003. Progressive decrease of left Heschl gyrus and planum temporale gray matter volume in first-episode schizophrenia: a longitudinal magnetic resonance imaging study. Archives of General Psychiatry 60, 766.

Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 983-997.

Laird, N., Lange, N., Stram, D., 1987. Maximum likelihood computations with repeated measures: application of the EM algorithm. Journal of the American Statistical Association, 97-105.

Lau, J.C., Lerch, J.P., Sled, J.G., Henkelman, R.M., Evans, A.C., Bedell, B.J., 2008. Longitudinal neuroanatomical changes determined by deformation-based morphometry in a mouse model of Alzheimer's disease. Neuroimage 42, 19-27.

Lerch, J.P., Pruessner, J.C., Zijdenbos, A., Hampel, H., Teipel, S.J., Evans, A.C., 2005. Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. Cereb Cortex 15, 995-1001.

Lindstrom, M.J., Bates, D.M., 1988. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association, 1014-1022.

Mathalon, D.H., Sullivan, E.V., Lim, K.O., Pfefferbaum, A., 2001. Progressive brain volume changes and the clinical course of schizophrenia in men: a longitudinal magnetic resonance imaging study. Archives of General Psychiatry 58, 148.

Ment, L.R., Kesler, S., Vohr, B., Katz, K.H., Baumgartner, H., Schneider, K.C., Delancy, S., Silbereis, J., Duncan, C.C., Constable, R.T., 2009. Longitudinal brain volume changes in preterm and term control subjects during late childhood and adolescence. Pediatrics 123, 503-511.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2007. Introduction to linear regression analysis. John Wiley & Sons.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human brain mapping 15, 1-25.

Pantelis, C., Velakoulis, D., McGorry, P.D., Wood, S.J., Suckling, J., Phillips, L.J., Yung, A.R., Bullmore, E.T., Brewer, W., Soulsby, B., 2003. Neuroanatomical abnormalities before and after onset of psychosis: a cross-sectional and longitudinal MRI comparison. The Lancet 361, 281-288.

Paviour, D.C., Price, S.L., Jahanshahi, M., Lees, A.J., Fox, N.C., 2006. Longitudinal MRI in progressive supranuclear palsy and multiple system atrophy: rates and regions of atrophy. Brain 129, 1040-1049.

Resnick, S., Sojkova, J., Zhou, Y., An, Y., Ye, W., Holt, D., Dannals, R., Mathis, C., Klunk, W., Ferrucci, L., 2010. Longitudinal cognitive decline is associated with fibrillar amyloid-beta measured by [11C] PiB. Neurology 74, 807-815.

Reuter, M., Fischl, B., 2011. Avoiding asymmetry-induced bias in longitudinal image processing. NeuroImage 57, 19-21.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: A robust approach. NeuroImage 53, 1181-1196.

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

Sabuncu, M.R., Desikan, R.S., Sepulcre, J., Yeo, B.T.T., Liu, H., Schmansky, N.J., Reuter, M., Weiner, M.W., Buckner, R.L., Sperling, R.A., 2011. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. Archives of neurology 68, 1040.

Satterthwaite, F.E., 1946. An approximate distribution of estimates of variance components. Biometrics bulletin 2, 110-114.

Schumann, C.M., Bloss, C.S., Barnes, C.C., Wideman, G.M., Carper, R.A., Akshoomoff, N., Pierce, K., Hagler, D., Schork, N., Lord, C., 2010. Longitudinal

magnetic resonance imaging study of cortical development through early childhood in autism. The Journal of Neuroscience 30, 4419-4427.

Shaw, P., Kabani, N.J., Lerch, J.P., Eckstrand, K., Lenroot, R., Gogtay, N., Greenstein, D., Clasen, L., Evans, A., Rapoport, J.L., 2008. Neurodevelopmental trajectories of the human cerebral cortex. The Journal of Neuroscience 28, 3586-3594.

Sidtis, J.J., Strother, S.C., Naoum, A., Rottenberg, D.A., Gomez, C., 2010. Longitudinal cerebral blood flow changes during speech in hereditary ataxia. Brain and language 114, 43-51.

Sluimer, J.D., Van Der Flier, W.M., Karas, G.B., Fox, N.C., Scheltens, P., Barkhof, F., Vrenken, H., 2008. Whole-Brain Atrophy Rate and Cognitive Decline: Longitudinal MR Study of Memory Clinic Patients1. Radiology 248, 590-598.

Sluimer, J.D., Van Der Flier, W.M., Karas, G.B., Van Schijndel, R., Barnes, J., Boyes, R.G., Cover, K.S., Olabarriaga, S.D., Fox, N.C., Scheltens, P., 2009. Accelerating regional atrophy rates in the progression from normal aging to Alzheimer's disease. European radiology 19, 2826-2833.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23, S208-S219.

SPM, http://www.fil.ion.ucl.ac.uk/spm/.

Sullivan, E.V., Pfefferbaum, A., Rohlfing, T., Baker, F.C., Padilla, M.L., Colrain, I.M., 2011. Developmental change in regional brain structure over 7 months in early adolescence: Comparison of approaches for longitudinal atlas-based parcellation. Neuroimage.

Thambisetty, M., An, Y., Kinsey, A., Koka, D., Saleem, M., Kraut, M., Ferrucci, L., 2011. Plasma clusterin concentration is associated with longitudinal brain atrophy in mild cognitive impairment. Neuroimage.

Thambisetty, M., Wan, J., Carass, A., An, Y., Prince, J.L., Resnick, S.M., 2010. Longitudinal changes in cortical thickness associated with normal aging. Neuroimage 52, 1215-1223.

Thirion, B., Pinel, P., Meriaux, S., Roche, A., Dehaene, S., Poline, J.B., 2007. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. NeuroImage 35, 105-120.

Thompson, W.K., Hallmayer, J., O'Hara, R., Alzheimer's Disease Neuroimaging, I., 2011. Design considerations for characterizing psychiatric trajectories across the lifespan: application to effects of APOE-epsilon4 on cerebral cortical thickness in Alzheimer's disease. Am J Psychiatry 168, 894-903.

Tosun, D., Schuff, N., Truran-Sacrey, D., Shaw, L.M., Trojanowski, J.Q., Aisen, P., Peterson, R., Weiner, M.W., 2010. Relations between brain tissue loss, CSF biomarkers, and the ApoE genetic profile: a longitudinal MRI study. Neurobiology of aging 31, 1340-1354.

Verbeke, G., Molenberghs, G., 2000. Linear mixed models for longitudinal data. N.Y.: Springer.

Whitwell, J., Weigand, S., Gunter, J., Boeve, B., Rademakers, R., Baker, M., Knopman, D., Wszolek, Z., Petersen, R., Jack Jr, C., 2011. Trajectories of brain and hippocampal atrophy in FTD with mutations in MAPT or GRN. Neurology 77, 393-398.

Whitwell, J.L., Jack Jr, C.R., Parisi, J.E., Knopman, D.S., Boeve, B.F., Petersen, R.C., Ferman, T.J., Dickson, D.W., Josephs, K.A., 2007. Rates of cerebral atrophy differ in different degenerative pathologies. Brain 130, 1148-1158.

HIGHLIGHTS

- We discuss Linear Mixed Effects (LME) models in the context of longitudinal imaging
- We contrast LME with widely used methods in longitudinal imaging
- We illustrate, validate and benchmark LME-based computational tools
- These tools will be freely available in FreeSurfer