DEVELOPMENT OF DEMENTIA BIOMARKERS FROM HIPPOCAMPAL AND MEDIAL TEMPORAL LOBE SUBREGIONS VIA PENALIZED REGRESION

John Pluta

A THESIS

in

Biostatistics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Master of Science

2015

_____
Russell T. Shinohara, Assistant Professor of Biostatistics
Supervisor of Thesis

_____
Paul A. Yushkevich, Associate Professor of Radiology
Co-Supervisor of Thesis

_____
John H. Holmes, Professor of Medical Informatics and Epidemiology
Graduate Group Chairperson

Acknowledgment

I would like to thank my advisors, Dr. Taki Shinohara and Dr. Paul Yushkevich, for their contributions and insight. This work was truly a joint effort. I would also like to thank my parents for their unwavering support.

Abstract

Measurements of medial temporal lobe (MTL) structures from MR images, particularly whole hippocampal volumes, have emerged as promising biomarkers for the diagnosis and tracking of mild cognitive impairment (MCI), an intermediate stage of dementia that is often a precursor to Alzheimer's disease (AD). We evaluate a statistical method for creating a multivariate biomarker from a list of candidate subregions in the MTL, and compare this to two existing biomarkers: whole hippocampal volume, and volume of most affected subregion. Measurements were obtained using a fully automatic segmentation method. Whole hippocampal volumes and each subregion were analyzed as independent linear models, with greatest effect in BA35. Subregions were analyzed in a multivariate setting using the LASSO, a form of penalized regression. The LASSO solution path was traced using the LARS algorithm. Statistical testing was performed on the solution paths using the covariance test. Two subregions were found to be statistically significant: CA1 (p=0.004) and BA35 (p=0.044), and formed a multivariate biomarker. To evaluate performance, ROC curves were generated for each biomarker and were compared using the Venkatraman test. The univariate biomarker did not show a significant difference from whole hippocampal volume (AUC=0.75, 0.71, p=0.77), while the multivariate biomarker significantly better classifier (AUC=0.81, 0.71, p=0.04). This method is well suited for biomarker development as it identifies unique sources of

variance and scales to problems with a large number of variables, including the $p > n$ case.

Table of Contents

1. Introduction

The current model of Alzheimer's Disease (AD) views the disease as a process beginning with preclinical AD, progressing to mild cognitive impairment (MCI), and ending in dementia (McKhann et al., 2011; Jack et al., 2011). The intermediate MCI stage is characterized by evidence of cognitive decline, but to a small enough degree that the patient is still considered functionally independent. As MCI is the first stage to manifest symptoms, there is special interest in develoing biomarkers to assess this condition (Petersen et al., 2014). Histological and magnetic resonance image (MRI) analysis has shown that the medial temporal lobe (MTL) is affected early in the dementia process, and accordingly is a primary target for the development of dementia biomarkers (Schuff & Zhu, 2007; Visser et al., 1999). In particular, whole hippocampal volume, as measured by T1-weighted MRI, has emerged as a standard imaging biomarker for MCI (see Shi et al., 2009 for a review).

Despite being an established and reliable marker for dementia, there are some potential criticisms of this metric. It models the hippocampus as a homogenous structure, it is actually composed of histologically and functionally distinct subfields (Amaral & Lavenex, 2007; Duvernoy, 2005), and there is a body of literature showing that the neurodegenerative processes that lead to dementia affect these subfields in a selective manner (Bobinski et al., 1997, 1998; Braak & Braak, 1991; Galton et al., 2001; Jack et al., 2000; West et al., 1994). In this case, a biomarker that combined all subfields into a single summary measurement would

have less detection power than measuring an affected subfield alone. The second issue is that the hippocampus is not the only structure affected by the pathology underlying dementia; neighboring cortical structures, particularly the transentorhinal region of the peririhnal cortex, are also affected in the earliest stages of the disease (Braak & Braak, 1991; Galton et al., 2001; Lace et al., 2009). However, standard T1 structural imaging lacks sufficient detail to accurately capture all of these regions.

Recent work in imaging has focused on the use of a specialized sequence that uses high in-plane resolution, T2-weighting, and a particular orientation to visualize MTL structures with much greater detail than standard T1 structural imaging. This sequence has been used in several recent studies to delineate regions of the MTL and assess how they are affected by various conditions (La Joie et al., 2010; Malykhin et al., 2010; Mueller et al., 2007, 2008, 2010; Pluta et al., 2012; Wisse et al., 2012). Furthermore, Yushkevich et al., (2015) recently published a method utilizing these scans to automatically segment regions of MTL involved in memory processes, namely subfields of the hippocampus, entorhinal and perirhinal cortices (Brown & Aggleton, 2001; Squire et al., 2004; van Strien et al., 2009). A full list of the subregions generated by this process is presented in Table 1. Automatic segmentation removes the need for manual delineation of subregions, which is prohibitively time-consuming for even modestly sized datasets.

Analyzing this dataset presents several challenges. Subregions of the MTL are not independent entities, as they are functionally and structurally connected, and an

ideal statistical model would account for the covariance between regions to accurately portray how they are jointly affected in MCI. A multivariate model using typical regression techniques is not practical in this case, as subregions are highly collinear. A high degree of correlation between independent variables can inflate variance and cause coefficients estimates to be unstable (see Appendix A). Although collinearity is tractable with sufficiently large samples, imaging data is expensive and time consuming to acquire, and samples tend to be relatively small.

Another issue is that while we have many variables under consideration, it is likely that only a subset are truly informative. For example, under the segmentation protocol used in Yushkevich et al., (2015), the CA2 subregion had low reliability and a very small volume, and it is unlikely to contain any meaningful information. Similarly, some variables may be redundant. If two regions have a nearly identical pattern of atrophy, including both in a model adds little explanatory power. Thus, some method of removing spurious or redundant variables would be an important factor in creating a biomarker model. It is also important that the variable selection process is robust to collinearity, sample size, and dimensionality, which are notoriously problematic for traditional selection methods like stepwise regression (Hastie et al., 2001).

In this work, we create a model of MTL subregions that specifically accounts for their correlated nature by using penalized regression and performing a statistical selection process on the penalized model. We then leverage this model to develop a new MCI biomarker, and compare it to biomarkers from established

methods; the standard whole hippocampal volumes, and a model derived from the results of the marginal regressions. The framework for this selection process can be generalized to datasets with similar characteristics and with an even greater number of dependent variables, including the case where $p > n$.

2. Methods

2.1 Subjects

We use the same subjects as a previous study on MTL subregions in MCI (Yushkevich et al., 2015). Images were acquired from 92 subjects enrolled in a study of memory and dementia, conducted by the Penn Memory Center at the University of Pennsylvania. 7 subjects were excluded from the analysis due to poor image quality that prohibited segmentation, leaving a total of 85 participants (44 cognitively normal controls, 41 MCI) matched for age (p=0.386, by t-test) and gender (p=0.131, by χ^2 test). MCI status was diagnosed according to the criteria in (Petersen et al. 2004). All subjects provided informed consent, and all research was conducted in compliance with the guidelines set forth by the University of Pennsylvania Institutional Review Board and the National Institutes of Health.

2.2 Image Acquisition

All MR images were acquired on a 3T Siemens Trio scanner at the Hospital of the University of Pennsylvania. The scanning protocol acquired two kinds of images;

a standard T1-weighted (MPRAGE) structural image of the whole brain, and a specialized T2-weighted (TSE) image adapted from (Mueller et al., 2007a; Thomas et al., 2004; Vita et al., 2003). The TSE acquisition is oriented orthogonal to the main axis of the hippocampus and trades low out-of-plane resolution (2mm with a 0.6mm slice gap) for very high in-plane resolution (0.4 x 0.4 mm^2), which maximizes the visualization of hippocampal and medial temporal lobe subregions. The T1 sequence acquires scans with a resolution of 1.0 x 1.0 x 1.0 mm^3.

2.3 Subregion Processing

All subregion segmentations were obtained using the ASHS (Automatic Segmentation of Hippocampal Subfields) software package (http://www.nitrc.org/projects/ashs), as described in Yushkevich et al., (2015). This software uses multi-atlas segmentation with joint label fusion (Wang et al. 2013) and corrective learning (Wang et al., 2011) algorithms to obtain fully automatic and reproducible segmentations of hippocampal subfields and neighboring cortical structures of medial temporal lobe from T2-weighted TSE images (Figure 1). Subregions are measured in mm^3, and correspond to the physical size of the measured region, except for cortical regions BA35 and BA36. Due to limitations of the initial segmentation protocol, these regions are normalized by the number of slices spanned, and are thus measured in mm^2. See Yushkevich et al., (2015) for detail. Whole hippocampal volumes are measured in mm^3.

2.4 Whole Hippocampal Volumes

Whole hippocampal volumes were automatically obtained from the T1 images using FreeSurfer 5.1 software (Fischl, 2012). These volumes refer to the region composed of the hippocampus proper (Duvernoy, 2005): CA1, CA2, CA3, DG, and the lateral part of subiculum. Typical T1 scans lack the resolution and contrast to measure subregions of the hippocampus, and thus they are measured as a single structure.

2.5 Intracranial Volumes

Intracranial volume (ICV) refers to the overall size of the brain. As we expect the size of a given brain region to scale with the overall size of the brain, ICV is an important covariate in image analysis. ICV measurements were obtained from the T1 structural images using the Brain Extraction Tool (BET) software (Smith 2002).

3. Statistical Methods

3.1 LASSO

Standard multivariate regression techniques are unsuitable for this dataset, due to the high levels of correlation between independent variables. Penalized regression refers to a class of regression techniques that take the form of:

$$\hat{\beta} = \underset{\beta \in R^p}{\arg\min}(y - X\beta)^T (y - X\beta) + \phi$$

where ϕ is a penalty term that shrinks coefficient estimates, and uniquely determines how coefficient estimates are penalized. For example LASSO (least absolute shrinkge and selection operator; Tibshirani et al., 1996) uses an l_1-norm as a penalty, ridge regression uses an l_2-norm, and elastic net (Zou & Hastie, 2005) uses a combination of the two. While correlation between independent variables may inflate variance estimates, the penalty term serves to shrink them. In essence, penalized model reduce variance at the expense of admitting some bias.

In this work, we choose to employ the LASSO model because the l_1-norm forces certain coefficients to be shrunk exactly to 0, essentially enforcing variable selection. Variable selection is an important concern in our problem, where we expect the effect of interest to be constrained to a few regions, and we know a-priori that some regions may be mostly noise. The LASSO model takes the form of:

$$\hat{\beta} = \underset{\beta \in R^p}{\arg\min}(y - X\beta)^T (y - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i|,\ 0 \le \lambda \le \infty,$$

where λ is the tuning parameter that controls the degree of shrinkage of the coefficient estimates, y is the vector of outcomes, X is the matrix of independent variables, and β is the vector of coefficient estimates. For a large enough value of λ, all coefficients will be shrunken to 0, while at λ=0, there is no shrinkage (e.g., the usual least-squares estimates).

In the logistic regression setting with two classes, the LASSO model takes the form of:

$$\hat{\beta} = \max_{\beta_0, \beta} \left\{ \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where all terms are the same as in the linear case, except for *y* which is a binary indicator variable of category membership rather than the outcome. In logistic regression, the outcome is the log-odds of belonging to a given category, conditional on the covariates, e.g:

$$p(x) = P(Y = 1 \mid X = x)$$
$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + X\beta + \varepsilon$$

where

$$p(x) = \frac{e^{\beta_0 + X\beta}}{1 + e^{\beta_0 + X\beta}}. \qquad (1)$$

An important issue with using penalized regression is that while a penalized model may produce coefficient estimates with less variance**,** the objective function for the LASSO is non-linear and non-differentiable, because the derivative of the penalty term does not exist at 0. Consequently, we cannot use typical maximum likelihood methods to estimate parameters, and the usual inferential tools such as *p*-values and confidence intervals do not exist for the LASSO.

3.2 LARS

An alternative method of computing the LASSO solution is the LARS algorithm (Efron et al., 2004). While we can compute the solution at a particular value of λ or over a discrete set of values LARS traces the entire solution path across all values of λ. Having the full path is advantageous as it tells us the order at which variables enter the model, which provides some insight into the relative importance of each variable.

The algorithm works as follows. The model begins with all coefficients equal to zero, and the predictor with the greatest correlation to the outcome is added. The coefficient estimate of this predictor is increased in the direction of the sign of its correlation with the residuals until some other predictor has as much correlation with the current residuals as the original predictor does. Then, both of these predictors are increased in their joint ordinary least-squares direction until some third predictor is equally correlated with the updated residuals, and so on until the complete model is formed. LARS specifically accommodates the LASSO model by allowing for the addition or removal of a variable at each step. Specifically, if a coefficient estimate hits 0, it is removed from the active set. See section 3 of Efron et al., (2004) for more detail.

In the logistic setting the solution path is solved using the algorithm presented in Park and Hastie (2007), which is analogous to LARS for generalized linear models. However, the method of computing steps and estimates is quite different. In ordinary linear regression, the outcome *y* is a known quantity and only

the coefficients need to be estimated. In the logistic regression setting, the model to fit is defined as in (1), and both *p(x)* and the coefficient estimates are unknown quantities that are directly related to each other. Accordingly, they are estimated iteratively in a simultaneous fashion using the Newton-Raphson method (see Agresti 2002). The GLM version of LARS uses a similar iterative 'Predictor-corrector' method to approximate step size and coefficient estimates at each step.

3.3 Covariance Test

In the standard linear regression setting, we often want to compare two models and see which better explains the data. For instance, we may test a model with some set of covariates versus the intercept-only model, to see if the covariates are explain a significant amount of the variance in the outcome. A common way to do this is through the likelihood-ratio test. We can express this test mathematically as:

$$\Lambda(x) = -2\log\left(\frac{\sup\{L(\theta;x) : \theta \in \Theta_0\}}{\sup\{L(\theta;x) : \theta \in \Theta\}}\right)$$

$$\Lambda(x) \xrightarrow{d} \chi^2_p$$

where $p = \dim(\theta)$. The likelihood-ratio test would be useful for LASSO models since it measures the impact of a variable on the overall model rather than drawing inference on the point estimates of parameters. Recall that the LARS algorithm adds variables to the model sequentially; then, a likelihood-ratio test conducted as each variable enters the model could assess whether the addition of a particular variable

had a significant change in likelihood, which would provide some statistical guarantee on the importance of a variable while circumventing the problematic nature of interpreting LASSO coefficients.

However, there are two problems with this approach. The first is that the objective function of the LASSO is non-differentiable and therefore violates the regularity assumptions of the likelihood-ratio test, necessitating an alternative form. Secondly, the asymptotic distribution of the likelihood-ratio test does not account for the adaptive nature of the model selection process in LARS. One possible approach would be to select the model through sequential testing on half of the data, and then test the resulting model versus the null model on the other half. This method would effectively cut the sample size in half, and is only feasible for large datasets.

Recent work by Lockhart et al., (2014) proposes a test, entitled the covariance test, which aims to provide some inference into LASSO models using the likelihood-ratio framework. This test uses the entirety of the data, and explicitly accounts for the adaptive nature of the LARS algorithm in generating the solution path. The covariance test is analogous to performing a likelihood-ratio test at each step of the LARS solution path. Recall that the LARS solution path is formed by a series of knots, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$, and A is the set of variables included in the model just before λ_k, called the active set. The jth variable enters the active set at λ_k, but has a value of zero upon entry. Thus we evaluate at λ_{k+1}, where the jth variable

has had its full effect on the data, and test two competing models: one with the active set including j and the other without. Stated mathematically, we have:

$$\hat{\beta}(\lambda_{k+1}) = \underset{\beta \in R^A}{\arg\min} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda_{(k+1)}\|\beta\|_1, A \cup \{j\} \in p$$

$$\tilde{\beta}_A(\lambda_{k+1}) = \underset{\beta \in R^A}{\arg\min} \frac{1}{2}\|y - X\beta_A\|_2^2 + \lambda_{(k+1)}\|\beta_A\|_1, A \in p$$

The test statistic follows an intuitive form, essentially measuring the drop in covariance (hence the name) between the model with the active set *A* versus $A \cup \{j\}$:

$$T_k = \left(\left\langle y, X\hat{\beta}(\lambda_{k+1})\right\rangle - \left\langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\right\rangle\right)/\sigma^2 \qquad (2)$$

The test is chosen to have this specific form because

$$T_k \xrightarrow{d} Exp(1),$$

and the *Exp(1)* distribution specifically accounts for the adaptive nature of LARS. In simulations, Lockhart et al., (2014) demonstrate that performing adaptive model selection under the usual χ^2 distribution resulted in a greatly inflated type-I error rate of 38%, whereas the *Exp(1)* distribution was much closer to the desired error rate of 5%.

In practice, the variance is usually unknown, so the typical least-squares estimate is used:

$$\hat{\sigma}^2 = \left\|y - X\hat{\beta}^{LS}\right\|_2^2/(n - p)$$

which yields a new statistic, F_k, where:

$$F_k = \left(\left\langle y, X\hat{\beta}(\lambda_{k+1}) \right\rangle - \left\langle y, X_A \tilde{\beta}_A(\lambda_{k+1}) \right\rangle \right) / \hat{\sigma}^2$$

$$F_k \xrightarrow{d} F_{2,n-p}$$

Note that although the variables this model contains highly correlated variables, this only effects the variance of the coefficients and not the variance of the overall model.

For a logistic regression with binomial outcomes, the covariance test statistic has the same form as in (2). In this case, the *y* values represent the log-odds of being classified in one of the two categories. Although the form is the same, the method by which the parameters are estimated is quite different, and is based on the algorithm described in Park & Hastie, (2007). An important note is that the theory for the logistic case has not been rigorously examined. However, simulations in Lockhart et al., (2014) demonstrate that the *Exp(1)* distribution seems adequate, and perhaps even conservative, for the logistic case.

The traditional approach to forming a LASSO model would be to obtain a value of lambda that optimizes some parameter (such as AUC, mean-squared error, or likelihood) through cross-validation (Tibshirani et al., 1996), and then solving at that value. LARS expands on this idea by providing some frame of reference for inference on the variable selection process of the LASSO, in that the order that variables enter the model is a reflection of their relationship to the outcome. The covariance test expands on LARS and formalizes the selection process into something statistically interpretable. Essentially, it selects which variables truly explain independent sources of variance in the outcome.

4. Data Analysis

4.1 Marginal Models

A series of logistic regression models was created, with one model for each subregion. The models all had the following form:

$$\log\left(\frac{P(MCI_i)}{P(CTL_i)}\right) = \beta_0 + \beta_1 SR_i + \beta_2 Age_i + \beta_3 Gender_i + \beta_4 ICV_i + \varepsilon_i \qquad (3)$$

where *SR* is the subregion, and age, gender, and ICV are nuisance variables. The purpose of the marginal models is to characterize the overall pattern of atrophy in MCI. The region that shows the greatest difference between MCI and controls will be selected for evaluation as a biomarker.

4.2 LASSO Models

Two logistic LASSO models were created, one for each MTL. Each model included all eight subregions as independent variables, with the log-odds of being classified as MCI as the dependent variable. The framework of the traditional LARS algorithm does not support nuisance variables. Despite not being highly correlated with the outcome, these parameters are important sources of variance and must be accounted for. Some path solution algorithms do support nuisance variables by adding them to the model at step 0, and leaving them unpenalized. Since this is not

an option for the LARS function used by the covariance test, we instead adjust all subregion data before entering it in the model. That is, for each subregion, we construct an ordinary linear model with subregion as the dependent variable and age, gender, and ICV as the dependent variables, and use the residuals in subsequent analysis. We compare these results to that of a model with subregions as penalized terms and nuisance variables as unpenalized terms, and note that the results are quite similar. After adjustment, variables were transformed to a standard normal distribution to put the variables on the same scale (recall, subregions of the hippocampal formation are measures of volume while cortical subregions are a measure of average area).

4.3 Biomarkers

We propose to examine three biomarkers for MCI; whole hippocampal volumes; the best performing subregion from the marginal models and the multivariate subregion model selected by the covariance test. Biomarker models follow the exact format as (3). For whole hippocampal volumes, SR is the whole hippocampus. For the model on the covariance test, SR is the group of subregions chosen as statistically significant by the covariance test, which could possibly range from 0 to 8.

Once the logistic models were created for each case, their classification ability was determined through receiver operating characteristic (ROC) curve analysis. ROC curves are a method of quantifying the predictive performance of a

classification model. They plot the false positive rate versus the true positive rate as a function of the classifier threshold. For example, in this study we are interested in classifying subjects as control or MCI. From the logistic regression models, we can get probabilities that each subject belongs to the MCI group, and by varying the discrimination threshold (that is, the minimum probability at which someone is classified as an MCI), we can compare the models predictions to the true values. The overall performance of a model can be concisely summarized by the area under the ROC curve (AUC). The AUC is an intuitive statistic that can be interpreted as the probability of the model to correctly classify a given subject into one of two groups, ranging between 0.5 (prediction is purely chance) to 1 (prediction is perfectly accurate). However, assessing the AUC on the same data that we used to fit the model is clearly a biased approach, and results in inflated AUC values. Since no secondary dataset is currently available, and the sample size is too small for data splitting, leave-one-out cross-validation was used to assess model performance. *K*-fold Cross-validation splits a data set into *k* groups, with *k*-1 forming a training set and the last group forming a testing set; a model is fit on the training set and the fit is assessed on the testing set. The groups are randomly permuted, and the procedure is repeated *k* times. In leave-one-out cross-validation, $k=n$. This method prevents overfitting, and provides performance results that better reflect how the model would perform on an independent dataset.

Finally, AUC curves were compared against each other using the Venkatraman test (Venkatraman & Begg, 1996). We use this test based on the recommendation of Bandos et al., (2005), which found through simulation that the

permutation based Venkatraman test was more powerful than the non-parametric test of Delong et al., (1988) for data with large AUCs, moderate correlation, and relatively small sample size. In practice, both tests produced similar results.

4.4 Statistical software

All analysis was done using the R programming language (R Development Core Team, 2008). The solution path for the LASSO model (Figure 2) was calculated using the generalized LARS algorithm via the `lars.glm` function, as implemented in the `covTest` package (Lockhart et al., 2013). The covariance test (also implemented in the same package) was run via the `covTest` function on each group. ROC curve analysis was performed with the pROC package (Robin et al., 2011).

5. Results

Results from the marginal models are presented in Table 2. The marginal models reveal a diffuse pattern of atrophy in the left medial temporal lobe, with many subregions being affected, whereas in the right hemisphere, subregions were generally preserved except for CA1. Based on these results, there appears to be an overall greater degree of AD pathology present in the left MTL, and we focus on this region for generating biomarkers.

Table 3 presents the results of the covariance test for the left and right MTL. BA35 and CA1 are significant variables in the left MTL, while CA1 alone is significant in the right. Note that the magnitudes of the test statistic are not comparable across the two models; CA1 right has almost twice the drop in covariance as CA1 left, but these terms do not have meaning outside of their respective models. Since CA1 right is the only region affected in the right MTL, it explains nearly all of the variance in the model. Additionally, the solution paths (and subsequently, the tuning parameters) are different for both models, which also changes the magnitude of the covariance statistic.

The results of the marginal models show that BA35 is the most affected region in the presence of MCI, and as such is selected as a univariate biomarker. The covariance test found CA1 and BA35 to both be significant variables, and they are used jointly to create a multivariate biomarker. Stated mathematically, we now have 3 competing models to be assessed as biomarkers:

The whole hippocampal volume model:

$$\log\left(\frac{P(MCI_i)}{P(CTL_i)}\right) = \beta_0 + \beta_1 WholeHippVolume_i + \beta_2 Age_i + \beta_3 Gender_i + \beta_4 ICV_i + \varepsilon_i,$$

The univariate model, chosen from the marginal results:

$$\log\left(\frac{P(MCI_i)}{P(CTL_i)}\right) = \beta_0 + \beta_1 BA35_i + \beta_2 Age_i + \beta_3 Gender_i + \beta_4 ICV_i + \varepsilon_i,$$

and the multivariate model, chosen from the results of the covariance test:

$$. \log\left(\frac{P(MCI_i)}{P(CTL_i)}\right) = \beta_0 + \beta_1 CA1_i + \beta_2 BA35_i + \beta_3 Age_i + \beta_4 Gender_i + \beta_5 ICV_i + \varepsilon_i$$

ROC curves and corresponding AUCs, as computed via cross-validation, are displayed in Figure 3. The Venkatraman test shows no difference between the univariate model and the whole hippocampal model (AUCs=0.75, 0.71, p=0.77), and a significant difference between the multivariate model and the whole hippocampal model (AUCs=0.81, 0.71, p=0.04).

6. Discussion

In this work, we describe a method for developing early AD biomarkers by modeling data using penalized regression, performing post-selection inference on those models using the covariance test, and then entering the selected variables into a new regression model. This approach is data-driven and applicable even to relatively small datasets as there is no data splitting or resampling involved. It also avoids the usual problems presented by penalized regression (lack of inferential tools) by returning the final result in the form of a standard regression. The variable

selection attribute of the LASSO, as well as the ordering process of LARS, virtually ensures that any significant independent variables will be at most moderately correlated.

To elaborate further, the covariance test can be thought of as a series of sequential tests; at each step, we are asking if the addition of a variable significantly improves the model. For example in the left MTL, we first ask if BA35 improves the fit over the intercept-only model. At the second step, we ask if jointly modeling BA35 and CA1 improves the fit over modeling BA35 alone. Essentially, the covariance is testing for independent sources of variance. For example, the marginal results show that in the left MTL, CA1 and DG have comparable degrees of atrophy, yet DG is the last variable to be entered into the LASSO model, and offers no improvement in predictive power. This is because the pattern of atrophy is so similar to that of CA1 that any information in it is redundant, and thus it explains no additional variance. BA35 however, explains variance in the data that is not accounted for by CA1.

A distinct strength of this approach is that the framework of penalized regression followed by variable selection inference through the covariance test is scalable to higher dimensional problems. For example, the hippocampus is structurally quite different in the anterior portion (the head), the central portion (the body), and the posterior portion (the tail), and imaging results verify that these regions are functionally distinct as well (Chen et al., 2010; Das et al., 2012; Greicius et al., 2003; Milne et al., 2012). Further dividing hippocampal subregions into head-

body-tail components (e.g., CA1-head, CA1-body, and CA1-tail) could lead to even more sensitive biomarkers, at the expense of three times as many independent variables. In a normal regression setting, adding this many variables could be prohibitive, but LASSO models can accommodate even cases where $p >> n$. Furthermore, the marginal results are somewhat subjective in that what regions we declare significant is a function of the method of multiple comparisons correction used. In a higher dimensional case, only the strongest effects would survive multiple comparisons correction. While the covariance test does not fully identify all of the changes associated with the disease process, it does unambiguously identify those that are most important for classification.

The marginal results show the greatest effect in BA35, and LARS chooses it to enter the model first. However, the covariance test shows that CA1 is a far more important predictor. This discrepancy is motivated by the fact that while the difference in means between groups is comparable for BA35 and CA1, the variance of CA1 is much larger than in BA35. For any basic test of means across groups (such as Welch's t-test or Wald test), a larger variance implies less certainty, and thus a smaller test statistic. The same argument applies for correlation, which explains why BA35 is chosen first by LARS. There is an unintuitive relationship between correlation and drop in covariance as measured by the covariance test. CA1 is less correlated with the outcome than BA35 because of its greater variance, but including it is necessary to explain that variance. This result is an important example of the utility of the covariance test, in that accurate prediction is about fully characterizing the data, rather than finding the largest differences between groups.

It is difficult to say whether the difference in variances between groups in CA1 and BA35 has a physiological meaning. AD pathology originates in the transentorhinal region and then continues on to the hippocampus (Braak & Braak, 1985, 1991). This could explain the greater stability in BA35 measurements and higher variance in CA1; all MCI subjects have at least some atrophy in BA35, but it is not clear how soon after CA1 begins to be affected. However, another explanation could be that CA1 is a measure of volume, while BA35 is a measure of average area. Then we are essentially looking at a distribution of averages, which will have a lower variance than a distribution of volumes.

In the ordinary regression setting, interpreting the results of the model is straightforward. However in this work, we build an OLS model based on a selection process from the LARS algorithm; thus, the results are conditional on the selection process (see Berk et al., (2013) for a description of some of the problems inherent with this process). Post-selection inference is an active area of research, and there is a growing body of literature specifically dedicated to inference after using the LASSO or LARS as a selection method (Lee et al., 2014; Taylor et al., 2014). The covariance test can be considered a form of post-selection inference, as it provides a setting for statistical testing of variables from some set chosen by the tuning parameter lambda. The exact process used in this paper (e.g., drawing inference on an OLS model generated by the selection process of the covariance test) has not yet been explored in terms of its inferential properties. However, the covariance test provides justification for retaining the selected variables and similar work by Lee et al., (2014) suggested that for a post-selection OLS model, coefficient estimates were

essentially unchanged by the selection process for larger effect sizes. Thus, we believe that our results are reasonable, though more work on the theoretical implications of selection and inference is warranted.

7. Tables

7.1 Table 1: Subregions of the medial temporal lobe used in analysis

Subregion	**Description**	**Location**
CA1	Cornu Ammonis, region 1	Hippocampal Formation
CA2	Cornu Ammonis, region 2	Hippocampal Formation
CA3	Cornu Ammonis, region 3	Hippocampal Formation
DG	CA4/Dentate Gyrus	Hippocampal Formation
SUB	Subiculum	Hippocampal Formation
ERC	Entorhinal Cortex	Hippocampal Formation/Parahippocampal Cortex
BA35	Brodmann area 35; transentorhinal region	Perirhinal Cortex
BA36	Brodmann area 36	Perirhinal Cortex

Table 1: Subregions of the medial temporal lobe used in the analysis, using the nomenclature of (Amaral et al., 2007).

7.2 Table 2: Results of the marginal models

Subregion	**Left MTL**		**Right MTL**	
	Z-Score	**p-value**	**Z-Score**	**p-value**
CA1	*-3.75*	*<0.001*	*-3.58*	*<0.001*
CA2	-1.14	0.255	-2.41	0.016
CA3	-1.15	0.250	-1.79	0.073
DG	*-3.33*	*<0.001*	-2.91	0.004
SUB	-2.41	0.020	-2.53	0.011
ERC	*-3.03*	*0.002*	-2.17	0.030
BA35	*-4.24*	*<0.001*	-2.14	0.032
BA36	-2.83	0.005	-2.83	0.005

Table 2: Results of the marginal regression models. Statistically significant results, after adjusting for multiple comparisons, are italicized.

7.3 Table 3: Results of the covariance test

Left MTL				**Right MTL**		
Predictor	**Drop in Cov**	**p-value**		**Predictor**	**Drop in Cov.**	**p-value**
BA35	*3.313*	*0.044*		**CA1**	*10.253*	*<0.001*
CA1	*5.505*	*0.004*		**BA35**	0.185	0.831
BA36	1.022	0.360		**CA2**	0.181	0.835
CA3	0.782	0.457		**CA3**	1.039	0.354
CA2	0.374	0.688		**BA36**	0.034	0.966
ERC	0.158	0.854		**DG**	0.113	0.893
SUB	-0.075	1.000		**SUB**	0.125	0.882
DG	0.363	0.695		**ERC**	0.023	0.977

Table 3: Results of covariance test for left and right MTL.

8. Figures

8.1 Figure 1: Coronal slice of the medial temporal lobe with subregion segmentations

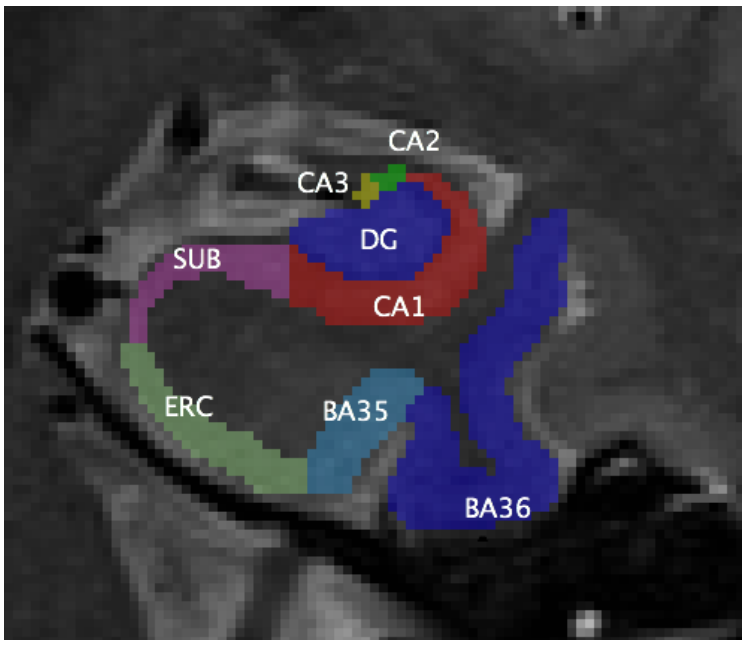

Figure 1: A coronal slice of the medial temporal lobe, with subregions segmented by ASHS.

8.2 Figure 2: LARS solution path for left-MTL subregions

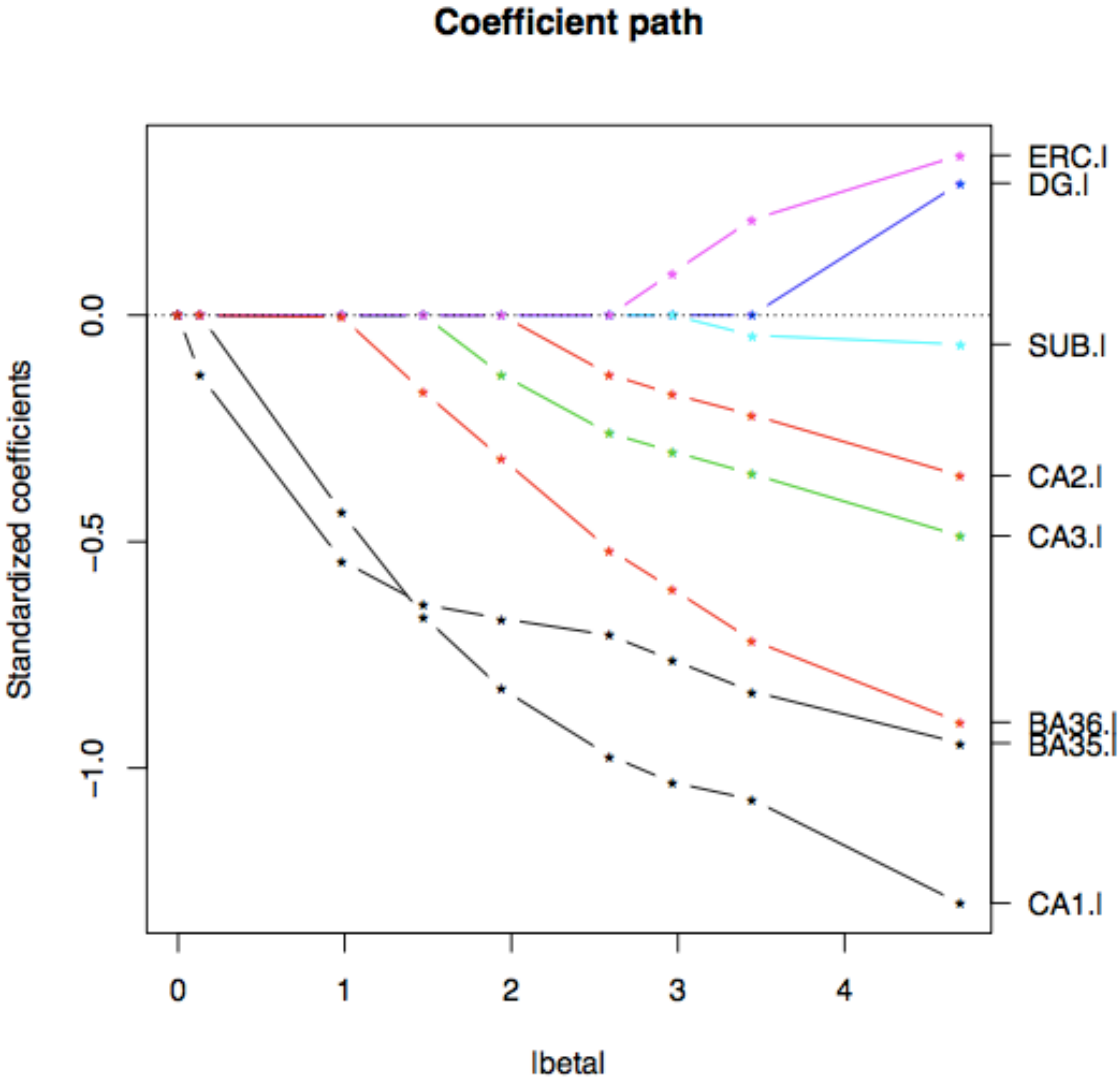

Figure 2: LARS solution path for left-MTL subregions. Asterisks represent knots in the solution path, points where a variable enters the model. As the amount of regularization decreases, more variables are entered into the model and coefficient estimates grow.

8.3 Figure 3: ROC curves for competing biomarker models

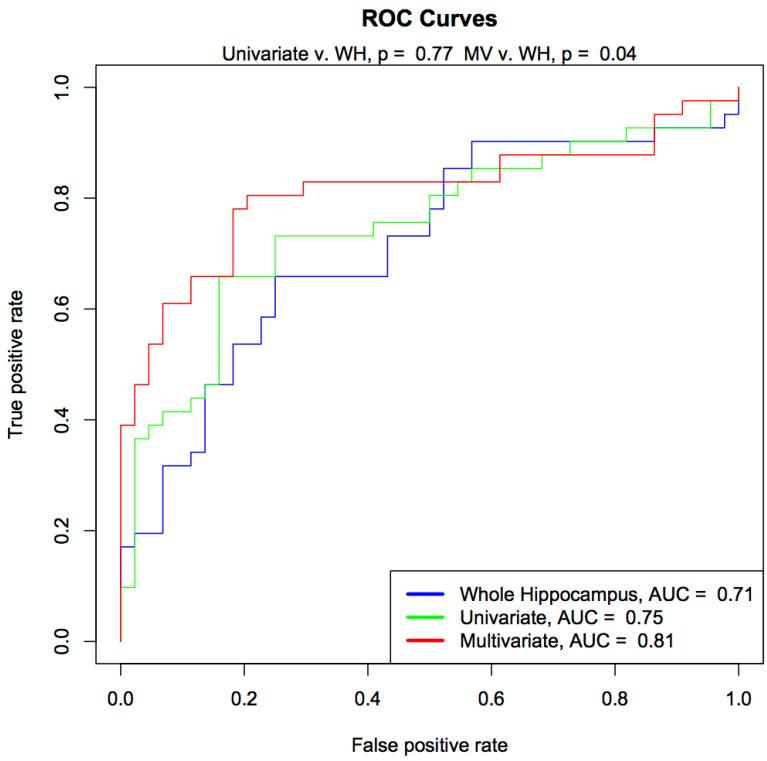

Figure 3: ROC curves for competing biomarker models.

9. Appendices

9.1 Appendix A: *Example of how correlation between independent variables inflates variance estimates.*

Suppose we have a regular linear model with two predictors, vectors x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

or in matrix form,

$$y = X\beta + \varepsilon,\ X = [1\ x_1\ x_2],\ \beta = [\beta_0\ \beta_1\ \beta_2]$$
$$y \sim N(X\beta, \sigma^2)$$

For simplicity, assume that both predictors are standard normal variables:

$$x_1, x_2 \sim N(0,1)$$

The least squares estimate for $\hat{\beta}$ is then

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\text{var}(\hat{\beta}) = \text{var}\left((X'X)^{-1}X'y\right) = (X'X)^{-1}X'X(X'X)^{-1}\sigma^2 = (X'X)^{-1}\sigma^2$$

We can express the variance as:

$$\text{var}(\hat{\beta}) = (X'X)^{-1}\sigma^2 = \left(\begin{bmatrix}1\\ x_1\\ x_2\end{bmatrix}\begin{bmatrix}1 & x_1 & x_2\end{bmatrix}\right)^{-1}\sigma^2$$
$$= \begin{pmatrix}1'1 & 1'x_1 & 1'x_2\\ x_1'1 & x_1'x_1 & x_1'x_2\\ x_2'1 & x_2'x_1 & x_2'x_2\end{pmatrix}^{-1}\sigma^2 = \begin{pmatrix}n & \sum x_{1i} & \sum x_{2i}\\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i}\\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2\end{pmatrix}^{-1}\sigma^2$$

$$= \begin{pmatrix} 1 & \frac{\sum x_{1i}}{n} & \frac{\sum x_{2i}}{n} \\ \frac{\sum x_{1i}}{n} & \frac{\sum x_{1i}^2}{n} & \frac{\sum x_{1i}x_{2i}}{n} \\ \frac{\sum x_{2i}}{n} & \frac{\sum x_{1i}x_{2i}}{n} & \frac{\sum x_{2i}^2}{n} \end{pmatrix}^{-1} \frac{\sigma^2}{n}$$

Noting that:

$$E[X] = \frac{\sum x_i}{n}, \ \text{var}[X] = E[X^2] - E[X]^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2$$

$$cor(X,Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_x \sigma_y},$$

we have:

$$\frac{\sum x_{1i}}{n} = E[x_1] = 0, \ \frac{\sum x_{1i}^2}{n} = \text{var}[x_1] = 1, \ \frac{\sum x_{1i}x_{2i}}{n} = cor(x_1,x_2) = r$$

which yields:

$$\text{var}(\hat{\beta}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & r \\ 0 & r & 1 \end{pmatrix}^{-1} \frac{\sigma^2}{n} = \frac{1}{1-r^2} \begin{pmatrix} 1-r^2 & 0 & 0 \\ 0 & 1 & -r \\ 0 & -r & 1 \end{pmatrix} \frac{\sigma^2}{n}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{n(1-r^2)}$$

As the correlation *r* between x_1 and x_2 grows, the denominator shrinks and inflates the estimate of the variance. However, the sample size *n* has a similar effect, and can counteract strong correlation.

10. References

Agresti, A. (2002). Categorical Data Analysis. Wiley & Sons, Inc. Hoboken, New Jersey, United States.

Amaral, D., Lavenex, P. (2007). Hippocampal neuroanatomy. In: The Hippocampus Book. Andersen, P., Morris, R., Amaral, D., et al. Oxford University Press, New York, pp. 37-219.

Bandos, A., Rockette, H., Gur, D. (2005). A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. Statistics in Medicine, 24:2873-2893.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. (2013). Valid post-selection inference. Ann Statist, 41:802:-837.

Bobinski, M., de Leon, MJ, Tarnawski, M., Wegiel, J., Bobinski, M., Reisberg, B., Miller, D., Wisniewski, H. (1998). Neuronal and volume loss in CA1 of the hippocampal formation uniquely predicts duration and severity of Alzheimer disease. Brain Research, 805:267-269.

Bobinski, M., Wegiel, J., Tarnawski, M., Bobinski, M., Reisberg, B., de Loen, M., Miller, D., Wisniewski, H. (1997). Relationships between regional neuronal loss and neurofibrillary changes in the hippocampal formation and duration and severity of Alzheimer disease. J Neuropathol Exp Neurol, 56:414-420.

Braak, H., Braak, E. (1991). Neuropathological staging of Alzheimer-related changes. Acta Neuropathol, 82:239-259.

Braak, H., Braak, E. (1985). On areas of transition between entorhinal allocortex and temporal isocortex in the human brain. Normal morphology and lamina-specific pathology in Alzheimer's disease. Acta Neuropathol, 68(4):325-32.

Brown, M., Aggleton, J. (2001_. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? Nat Rev Neurosci, 2(1):51-61.

Chen, K., Chuah, L., Sim, S., Chee, M. (2010). Hippocampal region-specific contributions to memory performance in normal elderly. Brain and Cognition, 72:400-407.

Das, S., Pluta, J., Mancuso, L., Kliot, D., Orozco, S., Dickerson, B., Yushkevich, P. (2012). Increased functional connectivity within medial temporal lobe in mild cognitive impairment. Hippocampus, 23(1):1-6.

DeLong, E., Delong, D., Clarke-Pearson, D. (1988). Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44:837-845.

Duvernoy, H. (2005). The Human Hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI. Berlin, Germany: Springer.

Efron, B., Hastie, T., Johnstone, Iain, Tibshirani, R. (2004). Least angel regression. The Annals of Statistics, 32(2):407-499.

Fischl, B. (2012). Freesurfer. Neuroimaging, 62:774-781.

Galton, C., Patterson, K., Graham, K., Lambon-Ralph, M., Williams, G., Antoun, N., Sahakian, B., Hodges, J. (2001). Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. Neurology, 57:216-225.

Greicius, M., Krasnow, B., Boyett-Anderson, J., Eliez, S., Schatzberg, A., Reiss, A., Menon, V. (2003). Regional analysis of hippocampal activation during memory encoding and retrieval: fMRI study. Hippocampus, 13:164-174.

Haste, T., Tibshirani, R., Friedman, J. (2001). The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

Hebert, L., Weuve, J., Scherr, P., Evans, D. (2013). Alzheimer disease in the United States (2010-2050) estimated using the 2010 Census. Neurology, 80(19):1778-83.

Jack, C., Albert, M., Knopman, D., McKhann, G., Sperling, R., Carrillo, M., Thies, B., Phelps, C. (2011). Introduction to the recommendations from the National Institute on Aging – Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 7(3):257-262.

Jack, C.R., Petersen, R.C., Xu Y., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Boeve, B.F., Tangalos, E.G., Kokmen, E. (2000). Rates of hippocampal atrophy correlation with change in clinical status in aging and AD. Neurology, 55:484-489.

La Joie, R., Fouquet, M., Mezenge, F., Landeau, B., Villain, N., Mevel, K., Pelerin, A., Eustache, F., Desgranges, B., Chetelat, G. (2010). Differential effect of age on hippocampal subfields assessed using a new high-resolution 3T MR sequence. Neuroimage, 53(2):506-14.

Lace, G., Savva, G., Forster, G., de Silva, R., Brayne, C., Matthews, F., Barclay, F., Dakin, L., Ince, P., Wharton, S. (2009). Hippocampal tau pathology is related to neuroanatomical connections: an ageing population-based study. Brain, 132;1324-1334.

Lee, J. Sun, D. Sun, Y., Taylor, J. (2014). Exact post-selection inference with the lasso. arXiv:1311.6238 [math, stat].

Lockhart, R., Taylor, J., Tibshirani, R., Tibshirani, R. (2014). A significance test for the lasso. Annals of Statistics, 42(2):413-468.

Lockhart, R., Taylor, J., Tibshirani, R., Tibshirani, R. (2013). covTest: Computes covariance test for adaptive linear modeling. R package version 1.02. http://CRAN.R-project.org/package=covTest.

Malykhin, N., Lebel, R., Coupland, N., Wilman, A., Carter, R. (2010). In vivo quantification of hippocampal subfields using 4.7T fast spin echo imaging. Neuroimage, 49(2):1224-30.

McDonald, C., McEvoy, L., Gharapetian, L., Fennema-Notestine, C., Hagler, D., Hollad, D., Koyama, A., Brewer, J., Dale, A., and the Alzheimer's Disease Neuroimaging

Initiative. (2009). Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. Neurology, 73:457-65.

McKhann, G., Knopman, D., Chertkow, H., Hyman, B., Jack, C., Kawas, C., Klunk, W., Koroshetz, W., Manly, J., Mayeux, R., Mohs, R., Morris, J., Rossor, M., Scheltens, P, Carillo, M., Thies, B., Weintraub, S., Phelps, C. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging – Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 7(3):263-269.

Milne, A., MacQueen, G., Hall, G. (2012). Abnormal hippocampal activation in patients with extensive history of major depression: an fMRI study. J Psychiatry Neurosci, 37(1):28-36.

Mueller, S., Schuff, N., Raptentsetsang, S., Elman, J., Weiner, M. (2008). Neuroimage, 42(1):42-8.

Mueller, S., Schuff, N., Yaffe, K., Madison, C., Miller, B., Weiner, M. (2010). Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease. Hum Brain Mapp, 31(9):1339-47.

Mueller, S., Stables, L., Du, A., Schuff, N., Truran, D., Cashdollar, N., Weiner, M. (2007). Measurements of hippocampal subfields and age related changes with high resolution MRI at 4T. Neurobiol Aging, 28:719-726.

Park, M., Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. J R Statist Soc, 69(4):659-677.

Petersen, R. (2004). Mild cognitive impairment as a diagnostic entity. J Intern Med, 256:183-194.

Petersen, R., Caraccilo, B., Brayne, C., Gauthier, C., Jelic, V., Fratiglioni, L. (2014). Mild cognitive impairment: a concept in evolution. Journal of Internal Medicine, 275:214-228.

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12:77.

Schuff, N., Zhu, X. (2007). Imaging of mild cognitive impairment and early dementia. The British Journal of Radiology, 80(2):109-114.

Shi, F. Liu, B., Zhou, Y. Chunshui, Y., Jiang, T. (2009). Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: meta-analyses of MRI studies. Hippocampus, 19:1055-1064.

Smith, S. (2002). Fast robust automated brain extraction. Hum Brain Mapp, 17:143-55.

Squire, L., Stark, C., Clark, R. (2004). The medial temporal lobe. Annu Rev Neurosci, 27:279-306.

Taylor, J., Lockart, R., Tibshirani, R., Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. arXiv:1401.3889 [math, stat].

Thomas, D., Vita, E., Roberts, S., Turner, R., Yousry, T., Ordidge, R. (2004). High-resolution fast spin echo imaging of the human brain at 4.7 T: implementation and sequence characteristics. Magn Reson Med, 51:1254-1264.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J R Statist Soc, 58(1):267-288.

van Strien, N., Cappaert, N., Witter, M. (2009). The anatomy of memory: An interactive overview of the parahippocampal-hippocampal network. Nat Rev Neurosci, 10:272-282.

Venkatraman, E., Begg, C. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika, 83:835-848.

Visser, P., Scheltens, P., Verhey, F., Schmand, B., Launer, L., Jolles, J., Jonker, C. (1999). Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. J Neurol, 246(6):477-85.

Vita, E., Thomas, D., Roberts, S., Parkes, H., Turner, R., Kinchesh, P., Shmueli, K., Yousry, T., Ordidge, R. (2003). High resolution MRI of the brain at 4.7 Tesla using fast spin echo imaging. Br J Radiol, 76:631-637.

Wang, H., Das, S., Suh, J., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P., Alzheimer's Disease Neuroimaging Initiative. (2011). A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. Neuroimage, 55:968-985.

Wang, H., Suh, J., Das, S., Pluta, J., Craige, C., Yushkevich, A. (2013). Multi-atlas segmentation with joint label fusion. IEEE Trans Pattern Anal Mach Intell, 35:611-613.

West, M.J., Coleman, P.D., Flood, D.G., Troncoso, J.C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. Lancet, 344:769-772.

Whitwell, J., Przybelski, S., Weigand, S., Knopman, D., Boeve, B., Petersen, R, Jack, C. (2007). 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. Brain, 130:1777-86.

Wisse, L., Gerritsen, L., Zwanenburg, J., Kuijf, H., Luijten, P., Biessels, G., Geerlings, M. (2012). Subfields of the hippocampal formation at 7T MRI: in vivo volumetric assessment. Neuroimage, 61(4):1043-9.

Yushkevich, P., Pluta, J., Wang, H., Xie, L., Ding, S., Gertje, E., Mancuso, L., Kliot, D., Das, S., Wolk, D. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. Human Brain Mapping, 36:258-287.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. J R Statist Soc B, 67(2):301-320.