Published in final edited form as:
Neuroimage. 2011 March 15; 55(2): 574–589. doi:10.1016/j.neuroimage.2010.10.081.

Predictive Markers for AD in a Multi-Modality Framework: An Analysis of MCI Progression in the ADNI Population

Chris Hinrichs[a,b,*], **Vikas Singh**[b,d,a,*], **Guofan Xu**[c,d], **Sterling C. Johnson**[c,d], and **the Alzheimers Disease Neuroimaging Initiative**[†]

Chris Hinrichs: hinrichs@cs.wisc.edu; Vikas Singh: vsingh@biostat.wisc.edu; Guofan Xu: gxu@medicine.wisc.edu; Sterling C. Johnson: scj@medicine.wisc.edu

[a] Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706

[b] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison Madison, WI 53705

[c] Geriatric Research Education and Clinical Center, Wm S. Middleton VA Hospital, Madison, WI 53705

[d] Wisconsin Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, WI 53705

Abstract

Alzheimer's Disease (AD) and other neurodegenerative diseases affect over 20 million people worldwide, and this number is projected to significantly increase in the coming decades. Proposed imaging-based markers have shown steadily improving levels of sensitivity/specificity in classifying individual subjects as AD or normal. Several of these efforts have utilized statistical machine learning techniques, using brain images as input, as means of deriving such AD-related markers. A common characteristic of this line of research is a focus on either (1) using a single imaging modality for classification, or (2) incorporating several modalities, but reporting *separate* results for each. One strategy to improve on the success of these methods is to leverage *all* available imaging modalities *together* in a single automated learning framework. The rationale is that some subjects may show signs of pathology in one modality but not in another – by combining all available images a clearer view of the progression of disease pathology will emerge. Our method is based on the Multi-Kernel Learning (MKL) framework, which allows the inclusion of an arbitrary number of views of the data in a maximum margin, kernel learning framework. The principal innovation behind MKL is that it learns an optimal combination of kernel (similarity) matrices while simultaneously training a classifier. In classification experiments MKL outperformed an SVM trained on all available features by 3% – 4%. We are especially interested in whether such markers are capable of identifying *early* signs of the disease. To address this question, we have examined whether our multi-modal disease marker (MMDM) can predict conversion from Mild Cognitive Impairment (MCI) to AD. Our experiments reveal that this measure shows significant group differences between MCI subjects who progressed to AD, and

[*] Corresponding authors. 5765 Medical Science Center, Madison, WI 53706, USA.

[†] Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database http://www.loni.ucla.edu/ADNI. 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. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI Manuscript Citations.pdf)

those who remained stable for 3 years. These differences were most significant in MMDMs based on imaging data. We also discuss the relationship between our MMDM and an individual's conversion from MCI to AD.

1 Introduction

A significant body of existing literature (Johnson et al., 2006; Whitwell et al., 2007; Reiman et al., 1996; Canu et al., 2010; Thompson and Apostolova, 2007) suggests that pathological manifestations of Alzheimer's disease begin many years before the patient becomes *symptomatic* – which is typically when cognitive tests can be used to make a diagnosis (Albert et al., 2001). Unfortunately, by this time significant neurodegeneration has already occurred. In an effort to identify AD-related changes early, a promising direction of ongoing research is focused on exploiting advanced imaging-based techniques to characterize prominent neurodegenerative patterns during the prodromal stages of the disease, when only mild symptoms of the disease are evident. A set of recent papers (Davatzikos et al., 2008a,b; Fan et al., 2008b; Vemuri et al., 2008) including work from our group (Hinrichs et al., 2009a,b) have demonstrated that this is indeed feasible by leveraging and extending state-of-the-art methods from Statistical Machine Learning and Computer Vision. Good discrimination (in identifying whether an image corresponds to a control or AD subject) has been obtained on classification tasks making use of MR *or* FDG-PET images (*i.e.*, *one* type of image data) (Davatzikos et al., 2008a,b; Fan et al., 2008b; Vemuri et al., 2008; Hinrichs et al., 2009a). A natural question then is whether we can exploit data from multiple modalities and biological measures (if available) *in conjunction* to (1) obtain improved accuracy, and (2) identify more subtle class differences (*e.g.*, sub-groups within MCI). This paper considers exactly this problem – *i.e.*, methods for systematic combination of multiple imaging modalities and clinical data for classification (*i.e.*, class prediction) at the level of individual subjects.

Recently, we have seen evidence that various aspects of AD-related neurodegeneration such as structural atrophy (Jack Jr. et al., 2005; deToledo-Morrell et al., 2004; Thompson et al., 2001), decreased blood perfusion (Ramírez et al., 2009), and decreased glucose metabolism (Hoffman et al., 2000; Matsuda, 2001; Minoshima et al., 1994) can be identified (in structural and functional images) in Mild Cognitive Impaired (MCI) and AD subjects, as well as at-risk individuals (Small et al., 2000; Querbes et al., 2009; Davatzikos et al., 2009). A number of groups have made significant progress by adapting well-known machine learning tools to the problem – this includes Support Vector Machines (SVMs), logistic regression, boosting, and other classification mechanisms. In the usual classification setting, a number of image acquisitions (training examples) are provided for which the subjects' clinical diagnosis is as certain as diagnostically possible. The objective is to choose a discriminating function which optimizes a statistical measure of the likelihood of correctly labeling 'future' examples. Such measures may be based on certain brain regions, (*e.g.*, the hippocampus or posterior cingulate cortex) for example. The function's output can then be used as a targeted disease marker in individuals that are not part of the training cohort. In the remainder of this section, we briefly review several interesting AD classification-focused research efforts, and lay the groundwork for introducing our contributions (*i.e.*, truly multi-modal analysis).

The machine learning, or classification approach has been used to provide markers for various neurological disorders including Alzheimer's disease (Davatzikos et al., 2008b; Klöppel et al., 2008; Vemuri et al., 2008; Duchesne et al., 2008; Arimura et al., 2008; Soriano-Mas et al., 2007; Shen et al., 2003; Demirci et al., 2008). These efforts have primarily utilized brain *images*, though some have also used other available biological measures. In (Fan et al., 2008b,a; Davatzikos et al., 2008a,b), the authors implemented a classification/pattern recognition technique using structural (sMR) images provided by the Baltimore Longitudinal Study of Aging (BLSA) dataset (Shock et al., 1984). The proposed methodology was to first segment the images into different tissue types, and then perform a non-linear warp to a common template space to allow voxel-wise comparisons. Next, voxels were selected to serve as "features" (using statistical measures of (clinical) group differences), used to train a linear Support Vector Machine (SVM) (Bishop, 2006). The reported accuracy was quite encouraging. The authors of (Klöppel et al., 2008) also used linear SVMs to classify AD subjects from controls using whole-brain MR images. An additional focus of their research was to separate AD cases from Frontal Temporal Lobar Degeneration (FTLD). The authors reported high accuracy (> 90%) on confirmed AD patients, and less where post-mortem diagnosis was unavailable. In related work, Vemuri *et. al.* (Vemuri et al., 2008) demonstrated a slightly different method of applying linear SVMs on another dataset obtaining 88 – 90% classification accuracy. More recently, the methods in (Fan et al., 2008a; Misra et al., 2008; Hinrichs et al., 2009a) have been applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, (http://www.loni.ucla.edu/ADNI/Data/) (Mueller et al., 2005) consisting of a large set of Magnetic Resonance (MR) and (18-fluorodeoxyglucose Positron Emission Tomography) FDG-PET images, giving accuracy measures similar to those reported in (Fan et al., 2008b,a; Davatzikos et al., 2008a,b). In (Hinrichs et al., 2009a), we proposed a combination of ℓ_1 sparsity and spatial smoothness bias, implemented via augmentation of the linear program used in training. The spatial bias lead to an increase in accuracy, and made the resulting images more interpretable. Steady increases in the levels of accuracy on this problem, *i.e.*, separating AD subjects from controls, have lead some researchers in the field to move towards the more challenging problem of making similar classifications on MCI subjects, with the expectation of extending such methods for identifying signs of the disease in its earlier stages. We provide a brief review of some preliminary efforts in this direction next.

Several recent studies (Schroeter et al., 2009; deToledo-Morrell et al., 2004; Dickerson et al., 2001; Hua et al., 2008) have shown that certain markers are significantly associated with conversion from MCI to AD. In (deToledo-Morrell et al., 2004; Dickerson et al., 2001), the authors show that traced volumes of the hippocampus and entorhinal cortex show significant group-level differences between converting and non-converting MCI subjects. We note that these studies show (in a *post-hoc* manner) that certain brain regions are correlated with AD histopathology; what we seek to do instead is to evaluate such markers in terms of their ability to classify novel examples. In (Hua et al., 2008) a large number of ADNI subjects were tracked longitudinally using Tensor-Based Morphometry (TBM). The authors compared conversion from MCI to AD over 1 year with atrophy in various regions, but a discussion of the predictive accuracy results was relatively limited (*i.e.*, included *p*-values of 0.02 between converters and non-converters). In (Davatzikos et al., 2009), the authors applied statistical techniques to both ADNI and BLSA subjects (Shock et al., 1984). A classifier was trained using ADNI subjects, and applied to MCI and control subjects (in the BLSA cohort) to provide a SPARE-AD disease marker. This procedure could successfully separate MCI and control subjects with high confidence (AUC of 0.885), and it was demonstrated that the MCI group had a larger increase in SPARE-AD scores longitudinally. However, the main focus in (Davatzikos et al., 2009) was *not* on predicting which MCI subjects would progress to AD, but rather on finding a marker for MCI itself. In (Querbes et al., 2009), cortical thickness measures were used on a large set of ADNI subjects to characterize disease progression in AD and MCI subjects. Freely available tools (FreeSurfer) were used to calculate cortical thickness values at points on the surface of each subject's brain (after warping to MNI template space) and then the thickness measures were agglomerated into 22 Regions of Interest (ROI), which the authors used as features (*i.e.*, covariates) in a logistic regression framework. Using age as a covariate, a set of AD and control subjects were used to train a logistic regression classifier for each subject, yielding a Normalized Thickness Index (NTI). It was found that this NTI was able to give 85% accuracy in separating AD subjects vs. controls, and had 73% accuracy (0.76 AUC) in predicting which MCI subjects would progress to full AD within 3 years. The latter objective is of special interest in the context of the techniques presented in this paper.

A common trend in the studies mentioned above is their focus on using a single scanning modality and processing pipeline. For instance, in a recent study (Schroeter et al., 2009), the authors surveyed 62 original research papers in a meta-analysis aimed at identifying which brain regions might make the most useful markers of AD-related atrophy, in a variety of different scanning modalities. A fundamental assumption is that the studies use only one scanning modality and analysis method in isolation, rather than combining the several available modalities into a single disease marker. However, each scanning modality and processing method can reveal information about different aspects of the underlying pathology. For instance, structural MR images may reveal patterns of gray matter atrophy, while FDG-PET images may reveal reduced glucose metabolism (Ishii et al., 2005), PIB imaging highlights the level of amyloid burden in brain tissue (Klunk et al., 2004), and SPECT imaging can allow an examination of cerebral blood flow (Ramírez et al., 2009); similarly, Voxel-Based Morphometry (VBM) shows gray matter density at baseline, while Tensor-Based Morphometry (TBM) shows longitudinal patterns of change (Hua et al., 2008). Another important issue one must consider is that as new types of biologically relevant imaging modalities become available, (*e.g.*, new tracers for use in PET scanners, or new pulse sequences in MRI scanners), it is desirable for the diagnostic process to incorporate such advances seamlessly. Further, since AD pathology is known to be heterogeneous, (Thompson et al., 2001) it may be advantageous to include multiple scanning modalities in a single classification framework. Indeed, a wide variety of markers may be available, and it is desirable to make the best use of *all* such information in a predictive setting. The main difficulty is that as the number of available input features grows, many machine learning algorithms may lose their ability to generalize to unseen examples, due to the disparity between the sample size and the increased dimensionality. To address this problem, we propose to employ a recent development in the machine learning literature, called Multi-Kernel Learning (MKL), which is designed to deal with multiple data sources while controlling model complexity. We have evaluated this method's performance on subjects from the ADNI data set, and report these results below. We have also applied the multi-modal classifier to MCI subjects, showing a promising ability to predict which subjects will convert from MCI to full AD in the ADNI sample.

The principal **contributions** of this paper are: **(1)** We propose a new application of Multi-Kernel Learning (MKL) to the task of classifying AD, MCI, and control subjects, which permits seamless incorporation of tens of imaging modalities, clinical measures, and cognitive status markers into a single predictive framework. The main ideas behind MKL are presented in Section 2.2; **(2)** We have conducted an extensive set of experiments using ADNI subjects, aimed at providing a rigorous evaluation of the method's ability to predict disease progression under conditions designed to match a clinical setting. We present these results in Section 4; **(3)** We employ our method to produce a Multi-Modality Disease Marker (MMDM) for MCI subjects, and present an analysis of its predictive value on rates of conversion from MCI to AD in Section 4.3. A discussion of our results is given in Section 5. [1]

[1] A preliminary conference version of this paper appeared as (Hinrichs et al., 2009b).

2 Algorithm

2.1 Support Vector Classification

In the following section, we present a brief overview of Support Vector Machines, (Cortes and Vapnik, 1995) illustrate the connection to Multi-Kernel Learning, and how this relates to the problem of disease classification from multiple modalities.

Machine learning methods are designed to find a classifier (*i.e.*, function) that correctly (or maximally) classifies a set of *n* training examples (*i.e.*, where class labels are known), while simultaneously satisfying some other form of *inductive bias* which will allow the algorithm to generalize, *i.e.*, correctly label future examples. Given a collection of points in a high dimensional space, SVM frameworks output a decision function separating classes (in a maximum margin sense) in that space; the 'bias' here is toward selecting functions with large margins. A linear decision boundary describes a *separating hyper-plane* – parameterized by a weight vector **w**, and an offset *b*. Classifying a new example **x** involves taking the inner product between **x** and **w** plus the offset *b*; the sign of this quantity indicates which side of the hyperplane **x** falls on (*i.e.*, its predicted class). In order to find the classifier, SVMs try not only to assign correct labels to each training example by placing them on the correct side of the hyperplane, but also attempt to place them some distance away. The measure of this distance is controlled by $\|\mathbf{w}\|_2$, or ℓ_2-norm of **w**. Thus, by rewarding the algorithm for reducing the magnitude of **w**, classifiers that correctly label the data (*and* have the widest margin) are selected, see (Schoelkopf and Smola, 2002) for details. SVMs choose an optimal classifier by optimizing the following primal/dual problem, whose solution **w** gives the separating hyperplane:

$$\begin{array}{c}\textbf{(primal)}\\ \min_{\mathbf{w},\xi}\frac{\left\|\mathbf{w}\right\|_2}{2}+C\sum_i \xi_i\\ \text{s.t. } y_i(\mathbf{w}^T x_i+b)\geq 1-\xi_i \quad \forall i\\ \xi_i\geq 0 \quad \forall i\end{array} \tag{1}$$

$$\begin{array}{c}\textbf{(dual)}\\ \min_{\alpha}\sum_i \alpha_i-\sum_{i,j}\alpha_i\alpha_j y_i y_j \underbrace{x_i^T x_j}_{\text{kernel}}\\ \text{s.t. } 0\leq\alpha_i\leq C \quad \forall i\\ \sum_i y_i\alpha_i=0 \quad \forall i\end{array} \tag{2}$$

In the primal problem (1), the *slack variables* ξ implement a *soft margin* objective. That is, for each example *i* that is not placed more than unit distance from the separating hyperplane, the slack variable ξ_i takes the value of the remaining distance from example *i* to the margin, which is then penalized in the objective. *C* is a constant parameter controlling the amount of emphasis on separating the data (if *C* is large,) vs. widening the margin (if *C* is small). Thus, the soft-margin objective allows for a trade-off between perfectly classifying every example, and widening the margin. The bias term *b* allows for separating hyperplanes ($\mathbf{w}^T x + b$) which do not pass through the origin. Class labels for each example are given as $y_i = \pm 1$, so that $y_i(\mathbf{w}^T x_i + b)$ will be positive iff $\mathbf{w}^T x + b$ gives x_i the correct sign specified by y_i.

Note that the hyperplane parameters **w** can be given as a linear combination of examples. It is a special property of the SVM formulation that the dual variables [2] α are exactly the coefficients of such a linear combination, *i.e.*, $\mathbf{w} = \Sigma_i \alpha_i y_i x_i$. For typical settings of C, the support of α will be sparse, giving rise to the term "Support Vector Machine".

Note that in the dual problem (2), the examples only occur as inner products $\langle x_i, x_j \rangle$. These inner products can be captured in a single $n \times n$ matrix called a Gram matrix or kernel matrix, $\mathcal{K}$; see (Bishop, 2006). In practice, $\mathcal{K}$ is specified by the user and expresses some notion of similarity between the examples – that is, the magnitude of a kernel function of two examples expresses an inner product between corresponding points in an implicit Reproducing Kernel Hilbert Space $\mathcal{H}$. The translation from the original data space to $\mathcal{H}$ is commonly denoted as $\varphi(x)$; when the kernel function is modified, [3] the kernel space $\mathcal{H}$ and translation function $\varphi(x)$ are correspondingly modified. The kernel function can also be calculated analytically – among those commonly used are Linear, Polynomial, and Gaussian kernels. Briefly, a linear kernel function is simply the inner product of two examples in the original data space; thus, unmodified SVMs use a linear kernel. A polynomial kernel function is one in which each inner product is squared (or cubed etc.). Such kernels allow for polynomial decision boundaries, rather than simple hyperplanes. Finally, Gaussian kernels are based on the Euclidean distance between examples, by the formula

$$\exp\left(\frac{-\|x_i - x_j\|}{2\sigma}\right)$$

where σ is a bandwidth parameter and x_i and x_j may denote examples i and j. Gaussian kernel-based SVMs can be thought of as training a Gaussian mixture model as the pattern classifier. If a modified kernel function is used, corresponding to a non-linear transformation of the data, then the learned classifier is a linear function (*i.e.*, hyperplane) in the kernel space $\mathcal{H}$. Such a function typically maps back to a non-linear decision function in the original data space. A thorough treatment is given in (Bishop, 2006).

2.2 Multi-Kernel Pattern Classification

An extension of this idea is to combine many such functions of the data (*i.e.*, multiple kernels, each pertaining to one modality for example, or to different parameterizations of the kernel function, or to different sets of selected features), to create a single kernel matrix from which a better classifier can be learnt. Multi-kernel learning (MKL) (Lanckriet et al., 2004; Sonnenburg et al., 2006; Rakotomamonjy et al., 2008; Gehler and Nowozin, 2009; Mukherjee et al., 2010) formalizes this idea. This is achieved by adding a set of optimization variables called *subkernel weights* which are coefficients in a linear combination of kernels. The subkernel weights are chosen so that the resulting linear combination of kernel matrices (another kernel matrix) yields the best margin and separation on the training set, with additional regularization to reduce the chances of overfitting the data due to the increase in the degrees of freedom of the model.

[2] In linear and quadratic optimization, every primal problem has an associated dual problem; the optimal solution to one can be used to recover the optimal solution to the other.

[3] Any such modification must preserve the positive-definite property of the original kernel function.

$$\min_{\mathbf{w_k},\xi,\beta,b}\left(\sum_k \frac{\left\|\mathbf{w_k}\right\|_2}{\beta}\right)^2 + C\sum_i^N \xi_i + \left\|\beta_k\right\|_2^2$$
$$\text{s.t. } y_i\left(\sum_k \mathbf{w_k}^T\varphi_k(x_i)+b\right) \geq 1-\xi_i\ \forall i \qquad (3)$$

Here, β_k is the subkernel weight of the k-th kernel, and $\mathbf{w_k}$ is the set of weights for the k-th feature space, while ξ_i is a *slack variable* as described above. Regularization of the subkernel weights is accomplished by penalizing the squared 2-norm of β in the objective. Thus, in addition to minimizing the magnitude of each set of weights, the MKL algorithm also tries to minimize the magnitude of the subkernel weight vector. Thus as β_k grows larger, the corresponding $\mathbf{w_k}$ is penalized less, and therefore tends to have a larger contribution to the final classifier. The combined classifier is defined as $f(x) = \Sigma_k\, \mathbf{w_k}^T\varphi_k(x) + b$. Thus, the implicit kernel function is equal to $\Sigma_k\, \beta_k\varphi_k(x_i)^T\varphi_k(x_j)$. In the context of our application, it is helpful to think of the various kernel matrices as being derived from different sources of data (e.g., different modalities), different choice of kernel function or parameters, (*e.g.*, bandwidth parameter in a Gaussian kernel function,) or a different set of features. Their assigned weights can then be interpreted as their relative influence in learning a good classifier (*i.e.*, discriminative ability). Because there is a natural mechanism to control the greater complexity resulting from the increased dimensionality of multi-modality data, we believe that MKL is a preferable option rather than simply ‘concatenating’ all features together and using a regular SVM. Our proposed method then, is to calculate various kernel matrices from each available input modality – including brain images, cognitive scores and other characteristics, such as CSF assays or APOE genotype, and use MKL to train a optimal combined kernel and classifier.

Note that in the term $\left\|\beta_k\right\|_2^2$ the subkernel weights are penalized according to the Euclidean, or 2-norm. [4] A recent focus in MKL research has been to generalize this formulation to include other norms (Kloft et al., 2010), having different effects on the sparsity of the resulting vector of subkernel weights. For instance, the 1-norm is a sparsity inducing norm, while the 2-norm is not; norms between 1 and 2 allow a trade-off of emphasis between sparse and non-sparse solutions. When combining multiple imaging modalities for AD classification, it is preferable not to encourage sparsity, as the algorithm will be very likely to completely ignore some modalities.

3 Experimental Setup

3.1 Data

Data used in the evaluations of our algorithm were taken from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a \$60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to

[4] In general, the p-norm of a space $\mathcal{X}$ is given as $\|(\mathbf{x})\|_p = (\Sigma_i\, |x_i|^p)^p$, for $x \in \mathcal{X}$.

aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years.

Our data consisted of ADNI subjects for whom both MR and FDG-PET scans roughly 24 months apart were available (as of October 2009). For quality control purposes, several (16) subjects were removed due to motion artifacts (MR), reconstruction artifacts (FDG-PET) or other problems visible to an expert. All such evaluations were made *before* any classification experiments were conducted, so as not to unfairly bias the experimental results. Finally, we had data for 233 subjects (48 AD, 66 healthy controls, and 119 MCI subjects). Demographic data are shown in Table 1. Subject ID numbers are given in Tables 12 – 14. See Supplemental Materials.

3.2 Preliminary Image-processing

In order to apply SVM and MKL methods to imaging data, it is necessary to extract features which are common to all subjects. Using standard voxel-based morphometry methods, as described below, we warped the scans into a common template space, and used voxel intensities as features. That is, after extracting foreground voxels, (*i.e.*, those corresponding to brain tissue,) each subject can then be treated as a vector of fixed length.

T1-weighted MR images—Cross-sectional image processing of the baseline T1-weighted images was first performed using Voxel-Based Morphometry (VBM) toolbox in Statistical Parametric Mapping software (SPM, http://www.fil.ion.ucl.ac.uk/spm). The ADNI study provides repeated acquisitions of the MR scans, which we utilized by first performing an affine warp between duplicates, and then averaging them in order to boost the signal/noise ratio. We then segmented the original anatomical MR images into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) segments. Then by using the “DARTEL Tools” facility in SPM5, a study-cohort customized template was calculated based on all subjects’ baseline MR images with the registration results as well as all relevant flow fields (representing the transformations). All individual MR scans were subsequently warped to this new template. Modulated GM and WM segments were produced in the DARTEL template space, using both the original scans (Ashburner, 2007). Finally, the normalized maps were smoothed using an 8 mm isotropic Gaussian kernel to optimize signal to noise and facilitate comparison across participants. Analysis of gray matter volume employed an absolute threshold masking of 0.1 to minimize the inclusion of the white matter in analysis. Longitudinal MR image processing of baseline and 24-Month MR scans was performed with a tensor-based morphometry (TBM) approach in SPM5. We first co-registered the baseline and follow-up scans with rigid body affine transformation, and applied bias correction and intensity normalization to make both images comparable. Pre-processing TBM procedures are described in detail in a previous article (Kipps et al., 2005). Briefly, a deformation field was used to warp the corrected late image to match the early one within subject (Ashburner and Friston, 2000). The amount of volume change was quantified by taking the determinant of the gradient of deformation at a single-voxel level (*i.e.*, Jacobian determinant). Each subject’s Jacobian determinant map was normalized to the cohort-specific DARTEL template and smoothed using a 12 mm isotropic Gaussian kernel.

FDG-PET images—All FDG-PET images were first co-registered to each individual's baseline MR-T1 images and subsequently warped to the cohort-specific DARTEL template (see above). A mask of the Pons was manually drawn in the DARTEL template as the reference region. All of the normalized FDG-PET images were scaled to each individual's Pons average FDG uptake value and smoothed with a 12 mm isotropic Gaussian kernel.

Other biological and neurological data—In addition to MR and FDG-PET images, other biological measures and cognitive status measures are provided by ADNI for some subjects. These include CSF assays for certain compounds thought to be involved in neurodegeneration, such as AB1-42, Total Tau, and P-tau 181; NeuroPsychological Status Exam scores (NPSEs); and APOE genotype data. The complete list of biological measures, and their availability in the study population is shown in Tables 2 and 3.

3.3 Experimental Methodology

We performed two sets of classification experiments: **(1)** We first performed *multi-modal* classification experiments for separating AD and control subjects using baseline and longitudinal imaging data, (MR and FDG-PET), and other available cognitive/biological measures (CSF assays, NeuroPsychological Status Exams (NPSE), and APOE genotype). For comparison, we also present single-kernel experiments for each data modality (except APOE, since APOE genotype alone is not sufficient to diagnose AD), and on an SVM trained on the sum of all kernels, (or equivalently, the concatenation of all feature vectors). **(2)** Finally, we trained a classifier on the entire set of AD and control subjects and then applied it to the MCI population, giving a *Multi-Modality Disease Marker (MMDM)*. We compared this marker with NPSEs taken at 24 months, and examined its utility in predicting which MCI subjects would progress to AD, as opposed to remaining stable as MCI. Note that this is different from separating MCI subjects from AD/controls.

Kernel matrices—Kernel matrices used in our experiments were computed using a varying number of voxel-wise features, (*i.e.*, intensity values at each voxel,) and kernel functions i.e., linear, quadratic and Gaussian, for each imaging modality. For each fold, voxels were ranked by *t*-statistic between AD and control training subjects. That is, each voxel's intensity value can be thought of as a random variable, upon which we performed a *t*-test, and ranked the features by the resulting p-values. Separate kernels were computed using the top 250,000, 150,000, 100,000, 65,000, 25,000, 10,000, 5000 and 2000 features, respectively. These sets of features were chosen beforehand so as to give a reasonable coverage of the range of features available, while allowing the algorithm to choose a linear combination that leads to a discriminative kernel. In addition to performing an implicit feature selection step, this allows us to evaluate the MKL algorithm's ability to integrate tens to hundreds of kernels, as in the case when many more modalities are available. For each set of features, we constructed linear, quadratic, and Gaussian kernels, using a bandwidth parameter of 2 times the number of features for the Gaussian kernel. The Gaussian kernel bandwidth parameter should be chosen to be within the same order of magnitude as the majority of pairwise distances. Thus, when voxel-wise intensity values fall in the range [0, 1], a common choice for the bandwidth parameter is a small number times the number of features. By this process, we obtained 24 separate kernel matrices for each imaging modality. For non-imaging modalities, *i.e.*, CSF assays, NPSEs, and APOE genotype, all features were used, giving three kernels per modality. The biological measures used are shown in Table 2. Because only a subset of subjects had such measures available, we used zero values for those who did not. This means that kernel matrices had zero values where such data were missing, and therefore added nothing to the classification on those subjects. We chose a conservative approach to this problem, meaning that results can only improve if a statistical interpolation method were to be introduced. For computing the

MMDM for MCI subjects, all AD and CN subjects were used both in feature selection and training.

Before training a classifier using the kernels constructed as described above, it is necessary to perform some normalization; consider that the vector **w** which defines the separating hyperplane is a linear combination of examples. If the average magnitude of examples as implicitly represented by one kernel is orders of magnitude larger than that of another kernel, then for the same subkernel weights, one kernel will have a far greater contribution to **w**. In order to ensure that this is not the case, we adopted a standard approach to kernel normalization. The first step is to divide each kernel by the largest entry, so that all entries are in the range [0, 1]. Second, we re-centered the points in each kernel space by subtracting row and column mean values, and then dividing by the trace. See Bakir et al. (2007) for details. As a consequence of normalizing the kernels, the C parameter which controls the regularization trade-off can be set to a small integer. We therefore set $C = 10$; no fine tuning or model selection was necessary.

Recall that when longitudinal data are available, there is more than one way to perform spatial normalization of scans, and we treat them as different imaging modalities, because we expect different types of information to be revealed by each. From MR images, we have both baseline VBM, and TBM modalities; in FDG-PET we have baseline and 24 month scans, as well as the voxel-wise difference and ratio between scans at different time points. Kernels based on the longitudinal voxel-wise difference and ratio in FDG-PET images were found to have poor performance relative to the raw FDG-PET values (60% – 70% accuracy), and we did not make further use of them in our experiments.

ROC curves—We also computed Receiver Operator Characteristic curves (ROCs) for each set of experiments. Briefly, while a classification algorithm must output a ±1 group label, our algorithm can also output a ‘confidence’ level for each test subject which in this case is the signed output of the classifier. By ordering the confidence levels of the entire study population, and calculating a True Positive Rate (TPR or sensitivity) and False Positive Rate (FPR or 1 - specificity) for each level, an ROC curve qualitatively shows not only how many examples are misclassified, but provides a sense of how the classifier’s confidence relates to its correctness.

Cross-validated classification—For the first set of experiments, we performed AD vs. control classification experiments using 30 realizations of 10-fold cross-validation. That is, in each realization the study population was randomly divided into ten separate groups, or folds. Each fold was used as a “test” set, while the remaining data was used as a “training” set. Therefore, the algorithm was evaluated on AD and control examples which were unseen during the training process, while permitting us to use the entire dataset effectively. Various accuracy measures, such as test-set accuracy (% of test examples properly labeled as AD or control,) sensitivity, (% of AD cases labeled as such) and specificity (% of controls labeled as such), and area under ROC curves were computed by averaging over all 30 realizations. Using this methodology, we first evaluated each kernel function on its own, in an SVM framework. We then evaluated each modality in an MKL framework, by combining different kernel functions, all derived from the same modality and features. Finally, we combined all imaging modalities into a multi-modality MKL classification framework. We did the same for cognitive scores and biological measures, allowing for a comparison between different types of subject data in terms of their ability to identify signs of AD.

Comparison of subkernel weight vector regularization norms—Another interesting area of investigation is on the effect of different MKL norm regularizers, especially with regard to sparsity of the resulting classifier. Sparsity is often advantageous in the presence of non-informative or error-prone kernels, however an overly sparse combination can discard useful information, leading to a sub-optimal classifier. Thus, it is important to understand this trade-off. Using the cross-validation setup described above, we compared different subkernel norm regularizers, (1, 1.25, 1.5, 1.75, and 2), using all available kernel types, as shown in Tables 2 and 3. In order to demonstrate MKL's ability to combine fundamentally different sources of information, we also constructed additional kernels using subject age, APOE genotype, years of education, and geriatric depression scale as features. We expect that some of these additional kernels may or may not be as useful to the learning algorithm, so as to allow a meaningful assessment of the usefulness of applying sparsity in the kernel norm. For baseline comparison we trained an SVM on the sum of all kernels, which is equivalent to simply concatenating all feature vectors, by definition of the inner product of vectors.

MMDMs—Our next set of experiments were conducted to evaluate the ability of imaging-based markers to predict which subjects would convert from MCI to AD. In order to do this, we first trained an MKL classifier using all 114 AD and CN subjects, and then applied it to all 119 MCI subjects, giving an MMDM measure. This procedure was repeatedly performed using (a) imaging-based, (b) cognitive marker-based, and (c) biological measure-based kernels, so as to evaluate each type of data separately, and facilitated a better comparison among them. We also differentiated between baseline and longitudinal data.

To quantify the predictive value of the MMDMs, we separated the MCI subjects into three groups – those who had progressed to AD after three years, those who remained stable, and those who reverted to normal status – and calculated p-values of group differences using a *t*-test. We also computed ROC curves to quantitatively measure the degree of differentiation between the MCI groups as given by different types of biological measures. There are two ways to compute such ROCs: based on the differentiation between progressing and reverting MCI subjects, ignoring the stable MCI subjects; and based on the differentiation between progressing and non-progressing MCI subjects. In the former case, we treat stable MCI subjects as though their final status is not yet known, and thus the task is to predict whether a given subject will eventually revert, or progress. For our analysis, we calculated both kinds of ROC curves, and present results below.

Implementation—Our validation experiments and analysis framework were implemented in Matlab using an interface to the Shogun toolbox (Sonnenburg et al., 2006) (http://www.shogun-toolbox.org). The source code for this project and supplemental information will be made available at http://pages.cs.wisc.edu/~hinrichs/MKL_ADNI [upon publication].

4 Results and Analysis

We present here the results of our experiments on the ADNI data described in Section 3, and an analysis of the MKL algorithm in the context of MCI progression.

4.1 Separating AD subjects and Controls

As a first step, we separately evaluated the kernels produced by each modality by comparing their performance at classifying AD vs. control subjects using an MKL norm of 2.0, so as not to discard any useful information. Results of these experiments are shown in Figure 1. Note that the color scale is the same between all figures.

Our first set of multi-kernel experiments also focused on whether the algorithm could learn to separate AD subjects from controls. Our experimental method was to use 10-fold cross-validation repeated 30 times, using kernel matrices computed as described in 3.3. Accuracy, sensitivity, and specificity results are shown in Table 4. In order to compare the efficacy of imaging-based disease markers with other biological measures, we performed experiments **(1)** using only image-derived data, **(2)** using other biological measures, **(3)** using only NPSEs, and finally using all available data modalities.

Note that the accuracy achieved using imaging-based MMDMs is nearly as good as that achieved using NPSEs. We believe this is promising, because NPSEs should be expected to perform better than imaging modalities when AD-related cognitive decline is present, even if the NPSEs were not used in making the diagnosis. This is because AD is currently diagnosed according to the patient's cognitive status, and while the NPSEs we utilized are *not* the same as those used in making a clinical diagnosis, they are nonetheless markers of detectable decline in cognition, and as such are not directly comparable to imaging-based markers. Rather, we include these experiments only to facilitate indirect comparison. Thus, for the imaging-based markers to be nearly as effective is quite promising.

The areas under each ROC curve (another measure of classification performance) are provided in Table 4. In terms of area under ROC curve, all modalities performed about as well as other accuracy measures would suggest. Again, we note that imaging modalities and cognitive scores performed very similarly under this measure.

In order to compare the effect of subkernel weight norms, we repeated the above experiments using all kernels and modalities available and MKL norms in the range of (1, 1.25, 1.5, 1.75, 2). These results are shown in Table 5. Note that among the MKL norms, accuracy increases slightly with MKL norm up to the point where sparsity is no longer strongly encouraged (at about 1.5), suggesting that overly sparse MKL norm regularizers do indeed lose information. We also note that the SVM's performance suffered significantly.

When using a 1-norm, out of the 72 available kernels, only 4 had non-zero weights: one TBM Gaussian kernel using 10,000 features, two VBM kernels, (one linear with 10,000 features, one quadratic with 25,000), none from the baseline FDG-PET scans, and one linear kernel with 2,000 features. In contrast, the subkernel weights chosen when using an MKL norm of 2 were *all* non-zero, and are shown in Figure 2. This means that in the context of AD classification, different modalities (and different representations of information from those modalities) contributed to in varying proportions to yield a discriminative classifier. It is perhaps interesting to note that most of the weight was placed on the VBM kernels, followed by the TBM and FDG-PET kernels.

4.2 Classifier brain regions

An important component of the evaluation of our method is an analysis of the brain regions selected by the algorithm. That is, if the algorithm is only given linear kernels from brain images, then the decision boundary itself can be interpreted as a set of voxel weights, using the formula $\mathbf{w}_m = \beta_m \Sigma_i \alpha_i \varphi_m(\mathbf{x}_i)$ where $\varphi_m(\mathbf{x})$ is the implicit (possibly non-linear) transform from the original data space to the kernel Hilbert space. An examination of these weights can reveal which brain regions were found to be most useful or discriminative (by the algorithm) in its predictions. Thus, the images of brain regions below are taken from the multi-modality classifier trained on all four imaging modalities used in our experiments, using *only* linear kernels. Note that from Figure 1, we can see that among the kernels derived from FDG-PET images, the most informative kernel used more than 65000 voxels, which implies that *classification strategies can benefit from using whole-brain images rather than examining small, localized brain regions, or ROIs* in FDG-PET imaging. The results are shown in Figures 3 – 6. Note that these weights were all calculated simultaneously in the MKL setting. These images can be interpreted as follows: image intensity in voxels showing a stronger red color contributes to a subject's healthy (positive) diagnosis, while intensity in voxels showing a stronger blue color contributes to a subject's diseased (negative) diagnosis, and intensity in yellow-, green- or cyan-colored voxels is essentially ignored. Note that these weights are purely relative, and thus have no applicable units. Each subject's final score is thus the difference between the weighted average intensity in the red and orange regions and the blue and cyan regions. We interpret this as meaning that red-orange (positive weighted) regions are those in which image intensity is a prerequisite of healthy status. For blue-cyan (negative weighted) regions, the literal interpretation is that the algorithm found higher intensity among the AD group than in the controls.

In some cases, we observe that negative weights are assigned in regions where higher image intensity is usually associated with positive status. There are several possible explanations for this, such as image normalization artifacts which artificially boost the intensity of these regions in some AD subjects. For instance in FDG-PET images, image intensity was normalized using a map of the Pons, and thus irregularities in this region could produce artificially inflated intensities in the rest of the image. Another possibility is brought up by (Davatzikos et al., 2009), which is that in MR images of gray matter, periventricular white matter may be mis-segmented as gray-matter, due to certain types of vascular pathology. A third possibility is that there is a small set of subjects whose characteristics is heterotypical of their group, and thus induce negative weights in regions which would otherwise have positive weights. Evidence of such a group was found in (Hinrichs et al., 2009a). In order to examine this possibility we found a set of subjects (5 subjects based on baseline FDG-PET scans, and 4 subjects based on baseline MR scans) who had unusually strong intensity in regions which had been assigned negative weights, and re-trained the MKL classifier without them. The resulting classifier was nearly free of such anomalous negative weights, which strongly suggests that these negative weights are entirely the result of the influence of a small group of outlier subjects, (9 out of 114). We have investigated this issue briefly in our previous work. (Hinrichs et al., 2009a) The weights assigned by this classifier can be seen in Figure 7. It is important to note that these subjects were removed *for visualization purposes only*, and were still used in computing accuracy and other performance estimates, and in the MCI analyses described below.

In Fig. 3, we can see that heteromodal, frontal, parietal regions and temporal lobes are given negative weights. The posterior cingulate cortex, lateral parietal lobules (bilaterally) and pre-frontal midline structures prerequisite of an indication of healthy status. The weights assigned to the FDG-PET scans taken at 24 months show a similar pattern, and are shown in Figure 4.

Among the MR-based kernels, the most informative kernels (as measured in a single-kernel setting,) used 5000 to 25000 voxels, implying that smaller regions, can be used to identify signs of AD-related gray matter atrophy. Thus, we expect to see a similar pattern in the multi-modality setting. Using the same interpretation of color as above, we can see that in the baseline GM density images, (VBM) hippocampal and parahippocampal regions are highlighted more clearly, consistent with the single-modality results which indicated that a small number of voxels are most informative in this modality. In the TBM-based images, we see that the hippocampal regions and parahippocampal gyri are highlighted, as well as middle temporal lobar structures bilaterally, indicating that longitudinal atrophy is concentrated in these regions, which is again consistent with the single kernel results, (and prior literature), (Braak et al., 1999) in which the top 25000 voxels produced the most informative classifier.

4.3 Correlations and predictions on the MCI population

For the second set of experiments, which involved MCI subjects, we trained a classifier on the entire AD and control population using MKL. This classifier was then applied to the MCI population, giving a MultiModality Disease Marker (MMDM). Using this methodology, only AD and control subjects were used to train the model, while MCI subjects were only used for evaluation, rather than other methodologies in which MCI subjects are used for training purposes. (Hua et al., 2008, 2009; Davatzikos et al., 2009) This process was repeated for each modality separately, as well as in groups of modalities. That is, all imaging modalities were combined, as were all NPSEs and biological measures. The outputs for each subject are shown in Figure 8. Subjects who remained stable are shown in blue; subjects who progressed to AD after 3 years or less are shown in red; subjects who reverted to normal cognitive status are shown in green. The four plots are divided between baseline (left) and longitudinal (right), and imaging-based (top) and NPSE-based (bottom) MMDMs. In each plot, a maximum accuracy cut-point is plotted as a solid black line. On the left we can see that neither of the baseline scans shows much differentiation between the groups, and the maximum accuracy separating line is essentially choosing the majority class. On the right, both the imaging-based and NPSE-based MMDMs provide better separation of the 2 groups. We also computed a set of MMDM scores based on CSF measures and APOE genetic markers, which did not show any ability to differentiate the 2 groups. An encouraging sign is that none of the reverting subjects were given negative scores.

In order to quantify these differences, we evaluated the degree of group-wise separation between progressing, reverting, and stable MCI subjects, under each of the available modalities, using a *t*-test. As shown in Table 6, the resulting *p*-values of the imaging-based MMDM (in separating progressing subjects from non-progressing) are several orders of magnitude lower than those based on NPSEs at 24 months, and two orders lower at baseline, suggesting that *imaging modalities offer a better view of future disease progression than current cognitive status*. We believe this is an interesting result of our analysis.

Area under ROC curve results are shown in Table 7; the corresponding ROC curves are shown in Figure 9. For ROCs showing separation between progressing and reverting subjects, the AUCs are very high, as we would expect. These curves are shown on the left in Figure 9. For comparison, we also computed ROC curves for single modalities, which are also shown in the figure. Of special relevance is the fact that *the MMDM based on imaging data alone outperformed all others*, both at baseline and at 24 months. The second comparison we made via ROC curves was between progressing subjects and all others. We accomplish this by using a different ground truth for computing the ROC curves. In this case, the task is to understand which of the MCI subjects will progress to AD in the near term (2–3 years), and which will remain stable or revert. These curves are shown on the right in Figure 9. In this case, the imaging-based MMDM, (shown in green) outperformed all others, most significantly at 24 months. The AUC for the image-based MMDM was 0.79, while that of the NPSE-based MMDM was 0.74. The highest leave-one-out accuracy achieved by the image-based MMDM was 0.723. For the NPSE the highest accuracy was 0.681 For the Biological measure-based MMDMs, it was not possible to achieve an accuracy greater than chance.

5 Discussion

We have shown in our experiments that our approach can offer a flexible means of integrating multiple sources of data into a single automated classification framework. As more types of information about subjects become available, either through new scanning modalities or new processing methods, they can simply be added to this framework as additional kernel matrices in a seamless manner. For instance, rather than choose whether to use TBM or VBM in our experiments, we used *both* by delegating the task of choosing the better (*i.e.*, more discriminative) view of the data to our model.

The principal novelty of this work is to introduce a new machine learning algorithm, Multi-Kernel Learning, to the application of discriminating different stages of AD using neuroimaging and other biological measures. Many existing works (Davatzikos et al., 2008a,b; Fan et al., 2008b,a; Vemuri et al., 2008; Duchesne et al., 2008; Davatzikos et al., 2009; Querbes et al., 2009; Klöppel et al., 2008; Ramírez et al., 2009; Kohannim et al., 2010; Walhovd et al., 2010), use either general linear models based on summary statistics, or machine learning algorithms such as SVMs, logistic regression, or AdaBoost, with extensive pre- and post-processing of imaging data which adapts these methods to the particular application. Of the machine learning methods mentioned here, all three are discriminative max-margin learning algorithms. Logistic regression uses a sigmoid function to approximate the hinge-loss function, and must be optimized via iterative methods. AdaBoost implicitly finds a margin by iteratively increasing the importance of examples which are misclassified, much the same way that examples inside the margin become support vectors in the SVM framework. Our method shares some commonalities in the sense that pre-processing of brain scans is also required before a classifier can be trained. However, by incorporating MKL, we can extend this framework to allow seamless integration of multiple sources of data while controlling the complexity of the resulting classifier without the need for creating summary statistics, (which discard a large amount of information).

We note that several studies have reported better *raw* performance at classifying AD and control subjects. There are several factors which can affect such results. First, there is the issue of the severity of the disease, and of the availability of gold-standard diagnosis. For instance, the authors of (Klöppel et al., 2008) reported that their accuracy suffered when autopsy data were not available due to the difficulty of diagnosing AD *in vivo*. The ADNI data set, on which our experiments were based, consists entirely of living subjects, having relatively mild AD. (See Table 1). Other studies have used ADNI subject data (Davatzikos et al., 2009; Querbes et al., 2009; Fan et al., 2008a), and while some have reported better performance than we have, issues such as image registration and warping, subject inclusion criteria (*e.g.*, image quality), or choice of feature extraction/representation might have a greater effect on final outcomes. A recent study, Cuingnet et al. (2010), addressed exactly these issues, finding that when these issues are controlled, the accuracy results are closer to those reported in this study. (See Table 4.) For example, if a pre-processing method is found to be particularly useful for discriminative purposes, that method can be swapped with our current pre-processing methods, or incorporated as additional kernels. The more important comparison is between single modality and multi-modality methods, *using the same data and pre-processing pipeline*. In addition, our experiments comparing MKL with a concatenated-features SVM show that MKL has advantages in the presence of non-informative kernels.

Single-modality results

Our experiments in single-modality AD classification give an indication of the relative merits of various scanning modalities. We note first that in FDG-PET scans, the top performing kernels are those which make use of at least 65,000 voxels, indicating that a performance gain of five percentage points or more can be made from using the *entire* brain volume, rather than using smaller selected regions. [5] That is, while most subjects can be identified by examining smaller regions, some subjects can only be identified by

[5] The authors of (Fan et al., 2008b) found similar results in FDG-PET images.

examination of whole-brain atrophy. This suggests that there is a small group of subjects having atypical disease progression (in the case of AD subjects) or that some control subjects may show early signs of disease. A somewhat surprising result is that longitudinal analysis of FDG-PET images did not have much discriminative power. Neither of the two methods we considered (voxel-wise temporal difference, and voxel-wise temporal ratio) had accuracy higher than about 65%. This is perhaps an indication that signs of atrophy in FDG-PET images accumulate slowly enough that changes over a 2-year period alone are not enough to distinguish AD with high accuracy.

In the MR-based modalities, we can see that in baseline VBM images, the highest performing kernels are those that focus on small brain regions of a few thousand voxels, while in TBM images, the best performance is obtained from larger regions of about 25,000 voxels. We interpret this to mean that (in classifying AD and control subjects,) the most indicative signs of atrophy already present at baseline can be found in hippocampal and para-hippocampal regions (not shown), but the atrophy occurring at the stage of full AD (*i.e.*, that which occurs in the two years following diagnosis), is more diffuse. This suggests that early signs of AD are more likely to be concentrated in smaller regions, such as the hippocampus, and other structures known to be affected by AD.

Secondly, we note that linear kernels performed as well as, or better than quadratic and polynomial kernels in all modalities examined, indicating that there are few quadratic or exponential effects which can be used for discriminative purposes. This can be interpreted that indications of pathology in each voxel contribute independently and cumulatively to the final diagnosis.

Multi-modality results

An interesting comparison which arose in our experiments was between the various imaging-based kernels *individually*, (see Figure 1), and the MKL experiments combining groups of modalities (see Table 4). MKL produces *linear* combinations of kernels, and therefore does not examine the interactions between them when evaluating new subjects. This means that the ideal situation is where the errors present in each kernel matrix are drawn randomly and independently. When combining modalities with strong similarities, it is therefore expected that some errors will cancel out, to the extent that those errors do not themselves arise from shared properties of both modalities. The rationale for combining modalities into groups for comparison is that while imaging modalities *are* expected to contain distinct (and useful) information about each subject, we expect that they will have some information in common. For instance, properties such as total inter-cranial volume or particular anatomical artifacts will be present in different scanning modalities, but not in other biological measures. Thus, we first examine MKL's ability to integrate groups of similar measures and modalities, before examining its ability to combine dissimilar sources of information.

First, we note that none of the individual kernels derived from imaging modalities achieved an accuracy greater than MKL when given the combination of imaging modalities. Moreover, when MKL was given the *entire* set of kernels from all available sources of information, it outperformed any of the groups of modalities, except for the NPSEs, where the differences were not significant. This is expected, because clinical diagnosis is already known, meaning that the disease has already reached a stage where cognitive status effects are measurable, in contrast to earlier stages, in which anatomical and physiological changes have begun to occur, but outward signs have not. Indeed, in the analysis of MCI progression (Tables 6 and 7), it is the imaging-based modalities which have the strongest performance. Finally, it is interesting that for the biological measures, such as CSF assays and APOE genotypes, while there is certainly some information contained in the kernels generated from these measures, by themselves they do not have nearly the discriminative power of either the imaging modalities, or the NPSEs. This may be due in part to the fact that these measures are not available for all subjects.

In Table 7 it may be surprising that the MMDM trained on all available modalities underperformed the one trained only on longitudinal imaging modalities. This is likely due to the fact that the training task and evaluation task were closely related, but slightly different. Thus, the subkernel weights estimated to give the optimal performance on the training task (AD vs. controls), may have been slightly less than optimal on the related task, (MCI progression). Despite this, the disparity in performance is small, and the MMDM using all combined modalities still outperformed all other MMDMs. It is also interesting to note that while the NPSEs dominated in the AD vs. control task of Section 4.1, in this task, the longitudinal NPSEs are roughly at parity with the baseline imaging modalities. (See Tables 6 and 7.) This suggests that signs of impending progression from MCI to AD are present in the imaging modalities approximately *two years* ahead of clinical psychological measures.

MKL-norm results

In our experiments with varying MKL norm, we found that norms which encouraged sparsity performed slightly worse than those which do not, suggesting that information is being needlessly discarded. The results in Table 5 show that above about 1.5, sparsity makes less of a difference, but at 1 or 1.25, sparsity is encouraged enough to affect MKL’s performance. In contrast, the concatenated-features SVM’s performance was significantly lower overall, as it has no mechanism for discarding non-informative kernels, especially when there are more kernels from many different sources. When given only kernels from a single modality, the SVM’s performance was closer to parity with MKL, however, this is expected, due to the relative ease of combining kernels from similar sources of information. Rather, it is when there is greater variety in the information content of the various kernels that MKL incrementally shows an advantage over the concatenated-features SVM. This demonstrates that regardless of the norm chosen, MKL has the ability to automatically detect and discard sets of features which do not contribute significantly to the optimal classifier. One could, in theory, manually select which features to include, and how to weight them, but this would essentially emulate the MKL process by hand using a regular SVM. With the proper construction of kernels, it is even conceivable that MKL could be used to automatically select ROIs.

Brain regions selected

The classifier chosen by MKL consists of a set of kernel combination weights β, as well as a set of example combination weights α. These weights can be combined to give a single linear classifier based on voxel-wise features. The distribution of these voxel-weights chosen by the MKL algorithm therefore gives some insight into the relative importance of various brain regions, and we expect that a good classifier will place greater weight on regions known to be involved in AD.

It is well known that the Posterior Cingulate Cortex is involved in memory retrieval and related self referential processes (Northoff and Bermpohl, 2004; Piefke et al., 2003; Shannon and Buckner, 2004). As part of the limbic system, it has reciprocal connections with other memory areas including the dorsomedial and dorsolateral prefrontal cortex, the posterior parahippocampal cortex, presubiculum, hippocampus, entorhinal cortex, and thalamus (Mesulam, 2000). Previous imaging studies suggest the PCC is affected in AD even before clinical symptoms appear, consistent with the very early memory symptoms in AD (Xu et al., 2009; Ries et al., 2006). Interestingly, the earliest cerebral hypometabolism finding in

AD involves the PCC-precuneus rather than the hippocampus (Villain et al., 2008). Although the mechanism connecting cortical atrophy and hypometabolism in neurodegenerative disorders is not fully understood, intuitively, a positive relationship is expected. Both brain atrophy and cerebral hypometabolism reflect loss of neurons/synapses (Bobinski et al., 1999) and decrease in synaptic density/activity (Rocher et al., 2003). As mentioned in section 4.2, the brain regions selected by the MKL algorithm in FDG-PET images, as show in Figures 3 to 4, include the PCC and precuneus, the lateral parietal lobules, hippocampal and medial temporal regions, and the pre-frontal midline.

In MR longitudinal images (TBM, Figure 5), regions well-known to be atrophic in AD, such as the hippocampus, parahippocampal gyri, fusiform gyri and other middle temporal structures (Braak and Braak, 1991) are well highlighted. Expansion, (or reduced contraction) is associated with healthy status, and thus these regions are given positive weights, shown in red. Conversely, expansion in ventricles, and in the CSF surrounding the hippocampus is shown in blue. Expansion in these regions is correlated with AD pathology, and so these regions are given negative weights. In the baseline gray matter density images, (VBM, Figure 6) similar hippocampal and medial temporal regions are shown.

MCI conversion

The task of predicting conversion from MCI to full AD is known to be difficult, (Querbes et al., 2009; Davatzikos et al., 2009), and presents challenges *beyond* that of classifying AD and control subjects, or even that of classifying AD/control and MCI subjects. This difficulty arises largely from the “lag” between brain atrophy and cognitive decline. There are several interesting aspects of the MMDMs we have examined. First, we note that at baseline, neither NPSEs nor imaging modalities have a strong ability to detect which subjects will convert to AD. This may be a result of the ADNI selection criteria for MCI subjects – that is, MCI subjects are chosen so as to have very homogeneous cognitive characteristics at baseline, and so we expect that NPSEs will not be able to differentiate between progressing and stable MCI subjects very well. While the MMDM based on all combined imaging modalities does have a better AUC at baseline than the NPSEs, the improvement shown by the MMDM based on longitudinal imaging modalities suggests that a significant portion of the neurodegeneration responsible for the subjects’ conversion to AD takes place after MCI diagnosis. In addition, between baseline and 24 months, the imaging-based MMDM outperforms the NPSE-based MMDM by an even wider margin, as shown by the AUCs and *p*-values in Tables 6 and 7. This leads us to believe that while NPSEs can be a better marker for subjects who *already* are showing AD-related cognitive decline, the imaging modalities have slightly better predictive value for future decline. We expect that further progress can be made in adapting multi-kernel methods to work specifically with imaging data, allowing greater accuracy in identifying future patterns. Finally, we find it interesting that combining all imaging markers into a single MMDM offers a slight improvement over the best single imaging modality, which tends to be FDG-PET. This improvement is relatively stable over time, between baseline and 24 months.

6 Conclusion

In this paper we have presented a new application of recent developments from the machine learning literature to early detection of AD-related pathology. Using this measure of AD pathology, we constructed a predictive marker for MCI progression to AD. This method is fully *multi-modal* – that is, it incorporates all available sources of input relating to subjects, yielding a unified Multi-Modal Disease Marker (MMDM). Our results on the ADNI population indicate that this method has the potential to detect subtle changes in MCI subjects which may provide clues as to whether a subject will convert to AD, or remain stable. In particular, we have shown that imaging modalities have better ability to predict such outcomes than baseline neuropsychological scores, which is consistent with the view that neurological changes detected in neuroimages can *precede* clinically detectable declines in cognitive status. Our ongoing work focuses on further developing this method – which will permit even higher accuracy and sensitivity, and allow predictions at the level of individual subjects to be made with high confidence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by NIH grants R21-AG034315 (Singh) and R01-AG021155 (Johnson). Hinrichs is funded via a University of Wisconsin–Madison CIBM (Computation and Informatics in Biology and Medicine) fellowship (National Library of Medicine Award 5T15LM007359). Partial support for this research was also provided by the University of Wisconsin-Madison UW ICTR through an NIH Clinical and Translational Science Award (CTSA) 1UL1RR025011, a Merit Review Grant from the Department of Veterans Affairs, the Wisconsin Comprehensive Memory Program, and the Society for Imaging Informatics in Medicine (SIIM). The authors also acknowledge the facilities and resources at the William S. Middleton Memorial Veterans Hospital, and the Geriatric Research, Education, and Clinical Center (GRECC).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

The authors are grateful to Donald McLaren, Moo K. Chung and Sanjay Asthana for many suggestions and ideas.

References

Albert MS, Moss MB, Tanzi R, Jones K. Preclinical prediction of AD using neuropsychological tests. Journal of the International Neuropsychological Society 2001;7(05):631–639. [PubMed: 11459114]

Arimura H, Yoshiura T, Kumazawa S, Tanaka K, Koga H, Mihara F, Honda H, Sakai S, Toyofuku F, Higashida Y. Automated method for identification of patients with Alzheimer's disease based on three-dimensional MR images. Academic Radiology 2008;15(3):274–284. [PubMed: 18280925]

Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage 2007;38(1):95–113. [PubMed: 17761438]

Ashburner J, Friston KJ. Voxel-Based Morphometry - The Methods. Neuroimage 2000;11(6):805–821. [PubMed: 10860804]

Bakir, G.; Hofmann, T.; Schölkopf, B. Predicting structured data. The MIT Press; 2007.

Bishop, C. Pattern Recognition and Machine Learning. Springer; New York: 2006.

Bobinski M, De Leon MJ, Wegiel J, Desanti S, Convit A, Saint Louis LA, Rusinek H, Wisniewski HM. The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. Neuroscience 1999;95(3):721–725. [PubMed: 10670438]

Braak E, Griffin K, Arai K, Bohl J, Bratzke H, Braak H. Neuropathology of Alzheimer's disease: what is new since A. Alzheimer? European Archives of Psychiatry and Clinical Neuroscience 1999;249(9):14–22. [PubMed: 10654095]

Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta neuropathologica 1991;82(4):239–259. [PubMed: 1759558]

Canu E, McLaren DG, Fitzgerald ME, Bendlin BB, Zoccatelli G, Alessandrini F, Pizzini FB, Ricciardi GK, Beltramello A, Johnson SC, et al. Microstructural Diffusion Changes are Independent of Macrostructural Volume Loss in Moderate to Severe Alzheimer's Disease. Journal of Alzheimer's Disease. 2010

Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20(3):273–297.

Cuingnet R, Gérardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage. 2010

Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiology of Aging 2008a;29(4):514–523. [PubMed: 17174012]

Davatzikos C, Resnick SM, Wu X, Parmpi P, Clark CM. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 2008b;41(4):1220–1227. [PubMed: 18474436]

Davatzikos C, Xu F, An Y, Fan Y, Resnick SM. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 2009;132(8):2026–2035. [PubMed: 19416949]

Demirci O, Clark VP, Calhoun VD. A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia. Neuroimage 2008;39(4):1774–1782. [PubMed: 18396487]

deToledo-Morrell L, Stoub TR, Bulgakova M, Wilson RS, Bennett DA, Leurgans S, Wuu J, Turner DA. MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. Neurobiology of Aging 2004;25(9):1197–1203. [PubMed: 15312965]

Dickerson BC, Goncharova I, Sullivan MP, Forchetti C, Wilson RS¿, Bennett DA, Beckett LA, deToledo-Morrell L. MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. Neurobiology of aging 2001;22(5):747–754. [PubMed: 11705634]

Duchesne S, Caroli A, Geroldi C, Barillot C, Frisoni GB, Collins DL. MRI-Based Automated Computer Classification of Probable AD Versus Normal Controls. IEEE Transactions on Medical Imaging 2008;27(4):509–520. [PubMed: 18390347]

Fan Y, Batmanghelich N, Clark CM, Davatzikos C. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 2008a;39(4):1731–1743. [PubMed: 18053747]

Fan Y, Resnick SM, Wu X, Davatzikos C. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. Neuroimage 2008b;41(2):277–285. [PubMed: 18400519]

Gehler PV, Nowozin S. Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. Computer Vison and Pattern Recognition 2009:2836–2843.

Hinrichs C, Singh V, Mukherjee L, Xu G, Chung MK, Johnson SC. Spatially augmented LP Boosting for AD classification with evaluations on the ADNI dataset. NeuroImage 2009a;48(1):138–149. [PubMed: 19481161]

Hinrichs C, Singh V, Xu G, Johnson SC. MKL for Robust Multi-modality AD Classification. Medical Image Computing and Computer-Assisted Intervention 2009b;5762:786–794. [PubMed: 20428471]

Hoffman JM, Welsh-Bohmer KA, Hanson M, Crain B, Hulette C, Earl N, Coleman RE. FDG PET imaging in patients with pathologically verified dementia. Journal of Nuclear Medicine 2000;41(11):1920–1928. [PubMed: 11079505]

Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, Jack CR Jr, Weiner MW, Thompson PM. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage 2008;43(3):458–469. [PubMed: 18691658]

Hua X, Lee S, Yanovsky I, Leow AD, Chou YY, Ho AJ, Gutman B, Toga AW, Jack CR Jr, Bernstein MA, et al. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: An ADNI study of 515 subjects. NeuroImage 2009;48(4):668–681. [PubMed: 19615450]

Ishii K, Sasaki H, Kono AK, Miyamoto N, Fukuda T, Mori E. Comparison of gray matter and metabolic reduction in mild Alzheimers disease using FDG-PET and voxel-based morphometric MR studies. European Journal of Nuclear Medicine and Molecular Imaging 2005;32(8):959–963. [PubMed: 15800784]

Jack CR Jr, Shiung MM, Weigand SD, O'Brien PC, Gunter JL, Boeve BF, Knopman DS, Smith GE, Ivnik RJ, Tangalos EG, et al. Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. Neurology 2005;65(8):1227–1231. [PubMed: 16247049]

Johnson SC, Schmitz TW, Trivedi MA, Ries ML, Torgerson BM, Carlsson CM, Asthana S, Hermann BP, Sager MA. The influence of Alzheimer disease family history and apolipoprotein E varepsilon4 on mesial temporal lobe activation. Journal of Neuroscience 2006;26(22):6069–6076. [PubMed: 16738250]

Kloft, M.; Brefeld, U.; Sonnenburg, S.; Zien, A. Non-sparse regularization and efficient training with multiple kernels. 2010.

Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR, Ashburner J, Frackowiak RS. Automatic classification of MR scans in Alzheimer's disease. Brain 2008;131(3):681–689. [PubMed: 18202106]

Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, Bergström M, Savitcheva I, Huang GF, Estrada S, et al. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. Annals of neurology 2004;55(3):306–319. [PubMed: 14991808]

Kohannim O, Hua X, Hibar DP, Lee S, Chou YY, Toga AW, Jack CR, Weiner MW, Thompson PM. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiology of Aging. 2010

Lanckriet GRG, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI. Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research 2004;5:27–72.

Matsuda H. Cerebral blood flow and metabolic abnormalities in Alzheimer's disease. Annals of Nuclear Medicine 2001;15(2):85–92. [PubMed: 11448080]

Mesulam, MM. Principles of behavioral and cognitive neurology. Oxford University Press; USA: 2000.

Minoshima S, Foster NL, Kuhl DE. Posterior cingulate cortex in Alzheimer's disease. Lancet 1994;344(8926):895. [PubMed: 7916431]

Misra C, Fan Y, Davatzikos C. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. Neuroimage 2008;44(4):1415–1422. [PubMed: 19027862]

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

Mukherjee L, Singh V, Peng J, Hinrichs C. Learning Kernels for variants of Normalized Cuts: Convex Relaxations and Applications. Computer Vison and Pattern Recognition. 2010

Northoff G, Bermpohl F. Cortical midline structures and the self. Trends in Cognitive Sciences 2004;8(3):102–107. [PubMed: 15301749]

Piefke M, Weiss PH, Zilles K, Markowitsch HJ, Fink GR. Differential remoteness and emotional tone modulate the neural correlates of autobiographical memory. Brain 2003;126(3):650–668. [PubMed: 12566286]

Querbes O, Aubry F, Pariente J, Lotterie JA, Demonet JF, Duret V, Puel M, Berry I, Fort JC, Celsis P. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 2009;132(8):2036–2047. [PubMed: 19439419]

Rakotomamonjy A, Bach F, Canu S, Grandvalet Y. Simple MKL. Journal of Machine Learning Research 2008;9:2491–2521.

Ramírez J, Górrizand JM, Salas-Gonzalez D, Romero A, López M, Álvarez I, Gómez-Río M. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. Information Sciences. 2009

Reiman EM, Caselli RJ, Yun LS, Chen K, Bandy D, Minoshima S, Thibodeau SN, Osborne D. Preclinical Evidence of Alzheimer's Disease in Persons Homozygous for the ε4 Allele for Apolipoprotein E. New England Journal of Medicine 1996;334(12):752–758. [PubMed: 8592548]

Ries ML, Schmitz TW, Kawahara TN, Torgerson BM, Trivedi MA, Johnson SC. Task-dependent posterior cingulate activation in mild cognitive impairment. Neuroimage 2006;29(2):485–492. [PubMed: 16102979]

Rocher AB, Chapon F, Blaizot X, Baron JC, Chavoix C. Resting-state brain glucose utilization as measured by PET is directly related to regional synaptophysin levels: a study in baboons. Neuroimage 2003;20(3):1894–1898. [PubMed: 14642499]

Schoelkopf, B.; Smola, A. Learning from Kernels. MIT Press; 2002.

Schroeter ML, Stein T, Maslowski N, Neumann J. Neural correlates of Alzheimer's disease and mild cognitive impairment: A systematic and quantitative meta-analysis involving 1351 patients. NeuroImage 2009;47(4):1196–1206. [PubMed: 19463961]

Shannon BJ, Buckner RL. Functional-anatomic correlates of memory retrieval that suggest nontraditional processing roles for multiple distinct regions within posterior parietal cortex. Journal of Neuroscience 2004;24(45):10084–10092. [PubMed: 15537877]

Shen L, Ford J, Makedon F, Saykin A. Hippocampal shape analysis: surface-based representation and classification. Proceedings of SPIE 2003;5032:253–264.

Shock, N.; Greulich, R.; Andres, R., et al. Normal human aging: the Baltimore Longitudinal Study of Aging. Washington, DC: US Government Printing Office; 1984.

Small G, Ercoli LM, Silverman DH, Huang SC, Komo S, Bookheimer SY, Lavretsky H, Miller K, Siddarth P, Rasgon NL, et al. Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease. Proceedings of the National Aceademies of Science USA 2000;97(11): 6037–6042.

Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B. Large scale multiple kernel learning. Journal of Machine Learning Research 2006;7:1531–1565.

Soriano-Mas C, Pujol J, Alonso P, Cardoner N, Menchn JM, Harrison BJ, Deus J, Vallejo J, Gaser C. Identifying patients with obsessive-compulsive disorder using whole-brain anatomy. Neuroimage 2007;35(3)

Thompson PM, Apostolova LG. Computational anatomical methods as applied to ageing and dementia. British Journal of Radiology 2007;80(2):78–91.

Thompson PM, Mega MS, Woods RP, Zoumalan CI, Lindshield CJ, Blanton RE, Moussail J, Holmes CJ, Cummings JL, Toga AW. Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. Cerebral Cortex 2001;11(1):1–16. [PubMed: 11113031]

Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, Boeve BF, Petersen RC, Jack CR Jr. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 2008;39(3):1186–1197. [PubMed: 18054253]

Villain N, Desgranges B, Viader F, de la Sayette V, Mezenge F, Landeau B, Baron JC, Eustache F, Chetelat G. Relationships between hippocampal atrophy, white matter disruption, and gray matter hypometabolism in Alzheimer's disease. Journal of Neuroscience 2008;28(24):6174–6181. [PubMed: 18550759]

Walhovd KB, Fjell AM, Brewer J, McEvoy LK, Fennema-Notestine C, Hagler DJ Jr, Jennings RG, Karow D, Dale AM. Combining MR Imaging, Positron-Emission Tomography, and CSF Biomarkers in the Diagnosis and Prognosis of Alzheimer Disease. American Journal of Neuroradiology 2010;31(2):347. [PubMed: 20075088]

Whitwell JL, Przybelski SA, Weigand SD, Knopman DS, Boeve BF, Petersen RC, Jack CR Jr. 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. Brain 2007;130(7):1777–1786. [PubMed: 17533169]

Xu G, Mclaren DG, Ries ML, Fitzgerald ME, Bendlin BB, Rowley HA, Sager MA, Atwood C, Asthana S, Johnson SC. The influence of parental history of Alzheimer's disease and apolipoprotein E {ε} 4 on the BOLD signal during recognition memory. Brain 2009;132(2):383. [PubMed: 18829694]

Figure 1a

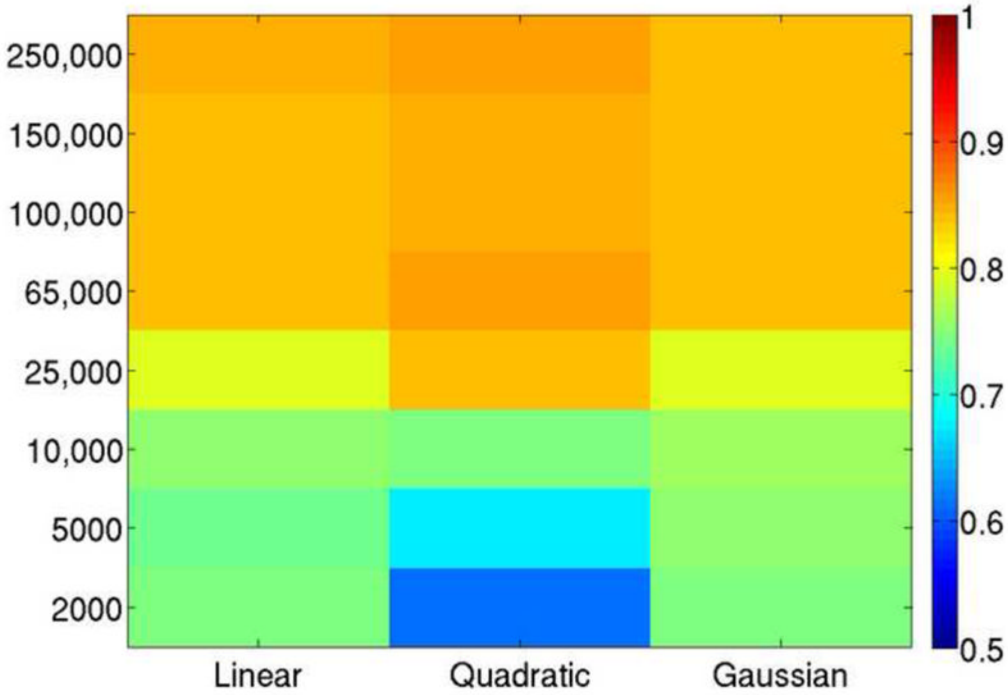
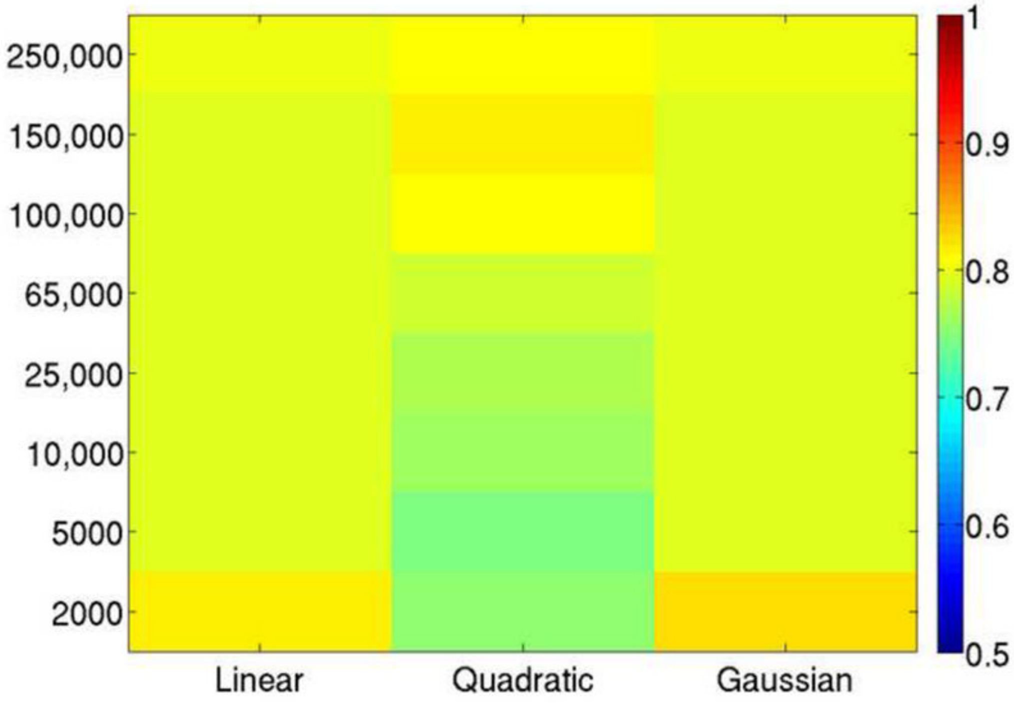


Figure 1b



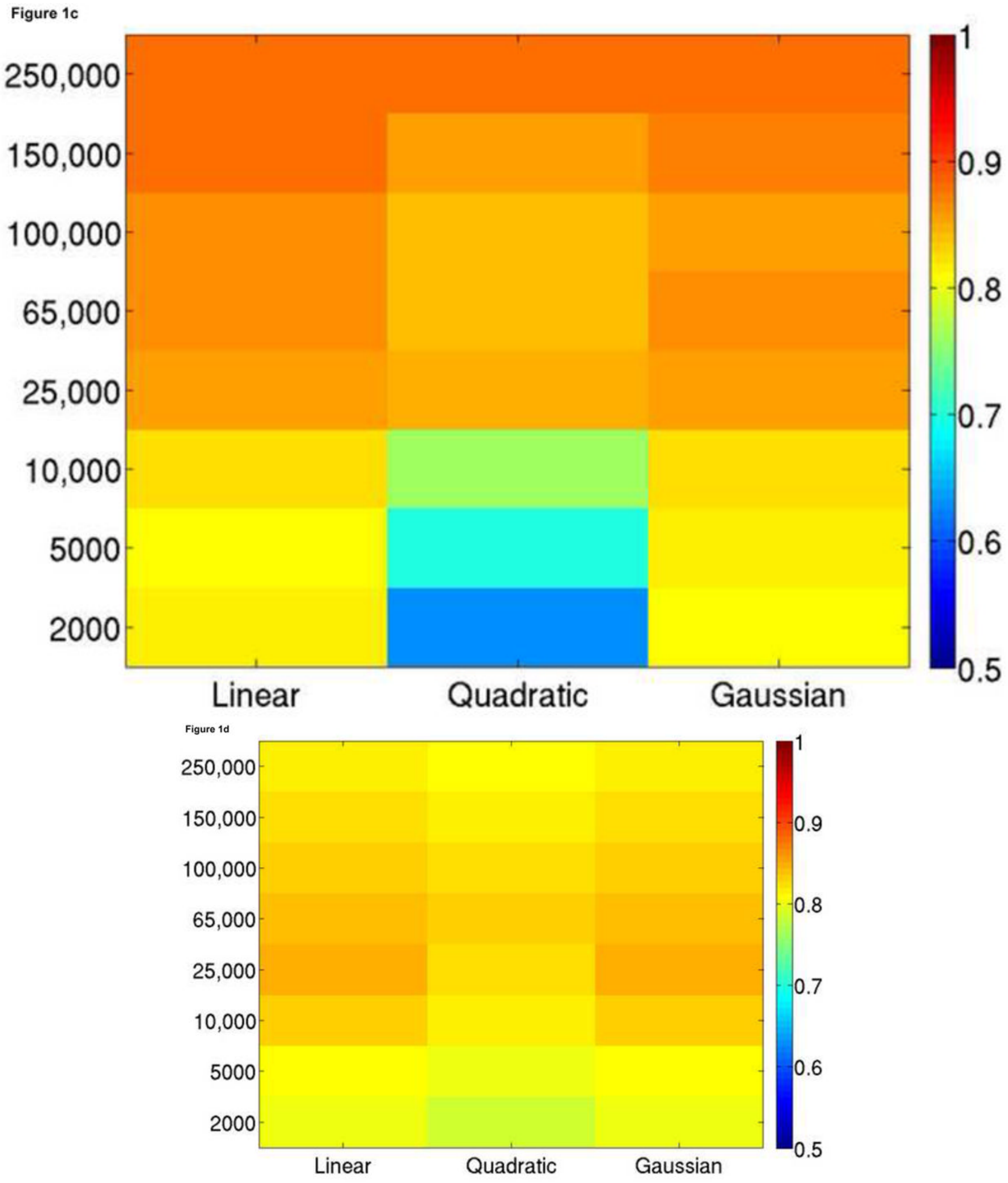


Figure 1.

Accuracies of single-kernel, single-modality methods. Color represents classification accuracy on unseen test data, ranging from blue (lowest, 50% accuracy,) to red (highest, 100% accuracy). The modalities used are, (a) FDG-PET scans at baseline, (b) VBM-processed MR baseline scans, (c) FDG-PET scans at 24 months, and (d) TBM-processed MR scans. See supplemental tables 8 – 11 for raw numbers.

Figure 2-a

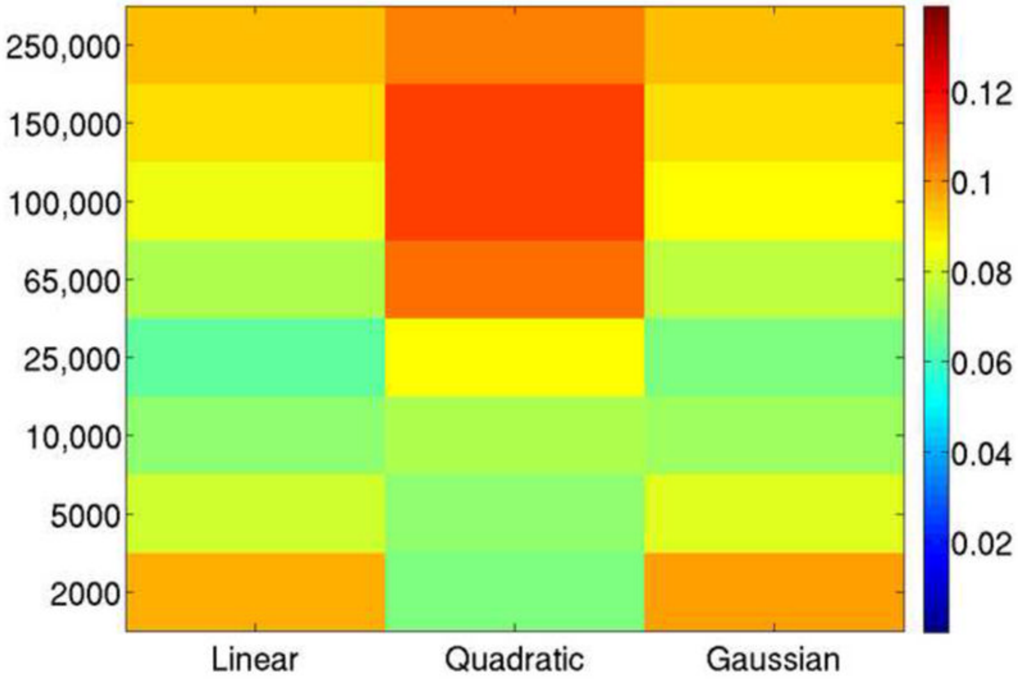
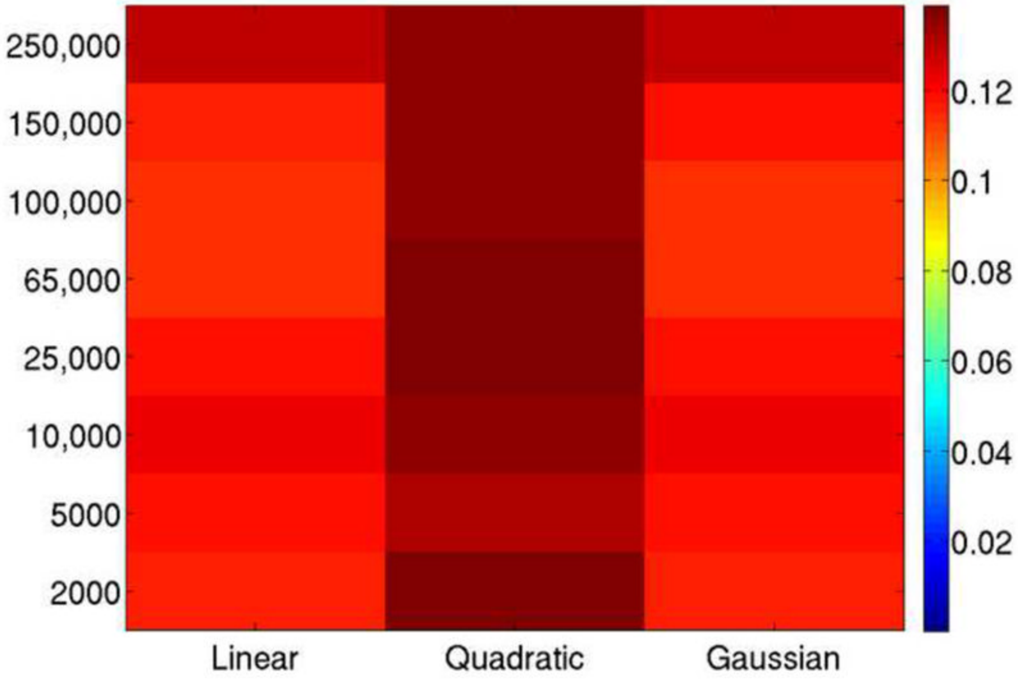


Figure 2-b



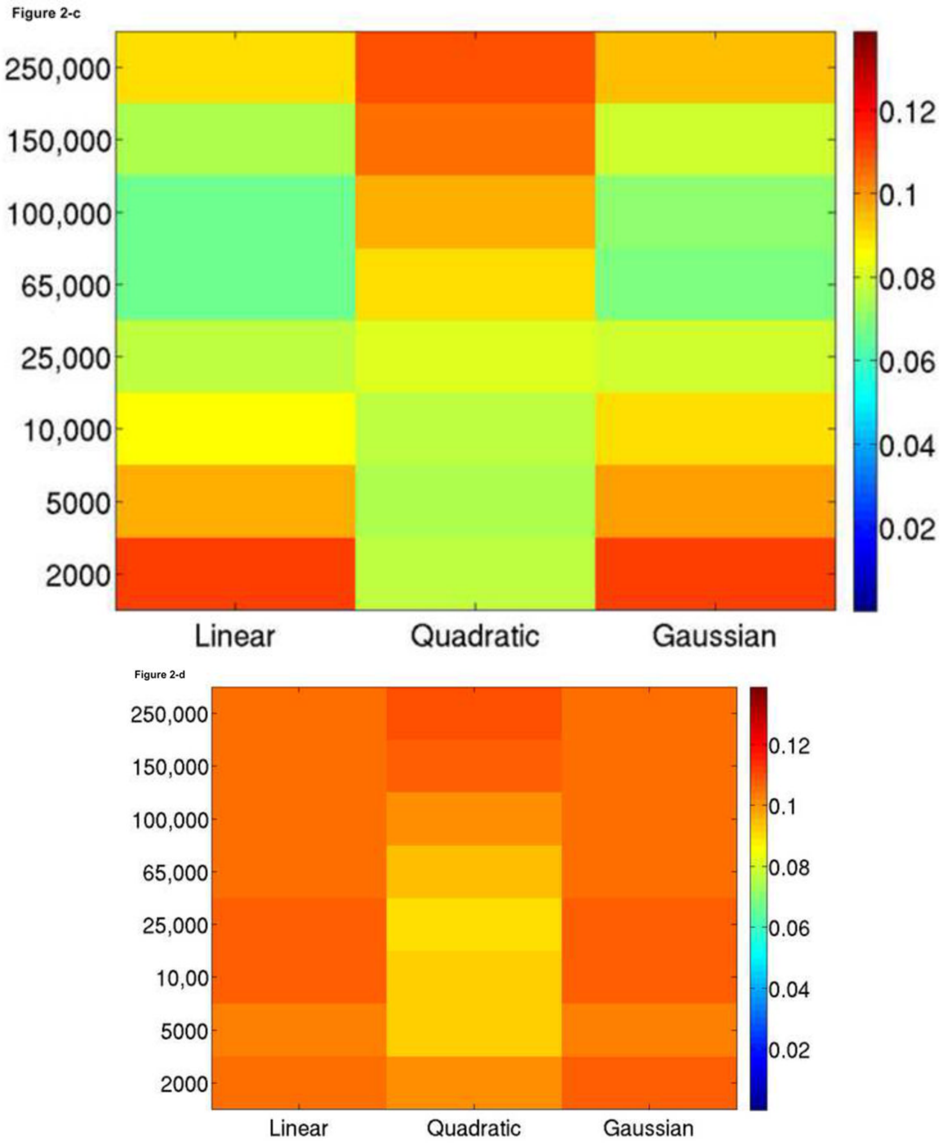
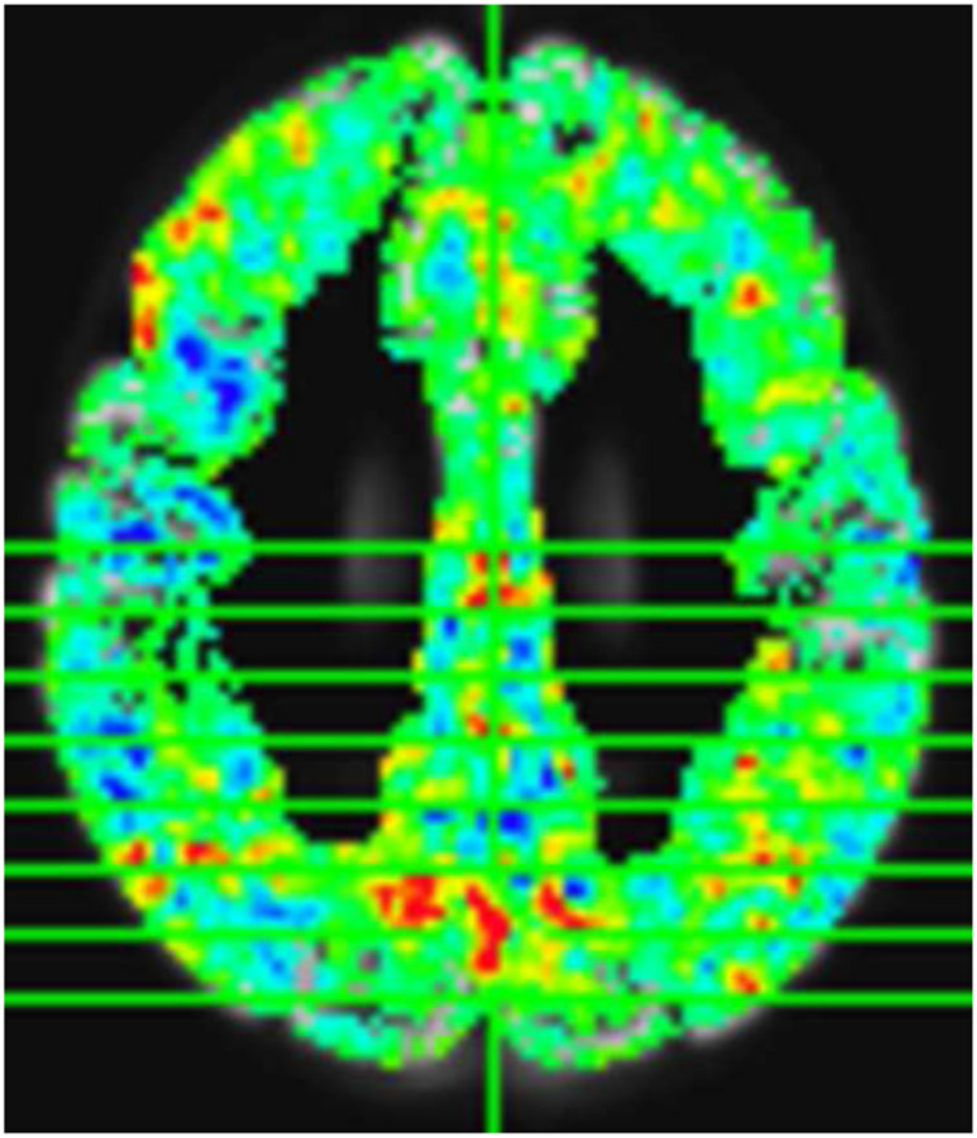


Figure 2.

Subkernel weights (*β*) chosen by the MKL algorithm with 2-norm regularization. Weights are relative, and have no applicable units. The modalities used are, (a) FDG-PET scans at baseline, (b) VBM-processed MR baseline scans, (c) FDG-PET scans at 24 months, and (d) TBM-processed MR scans.

Figure 3a



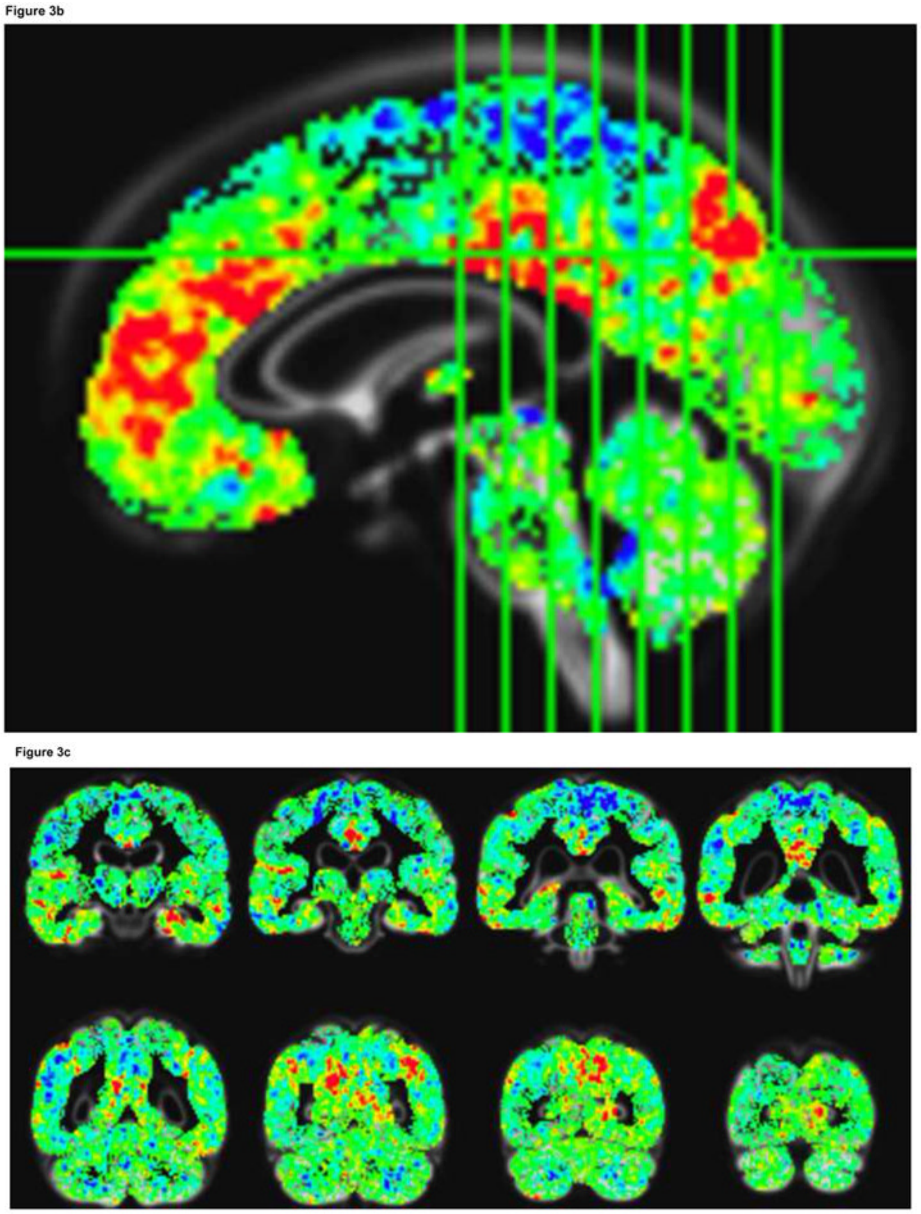
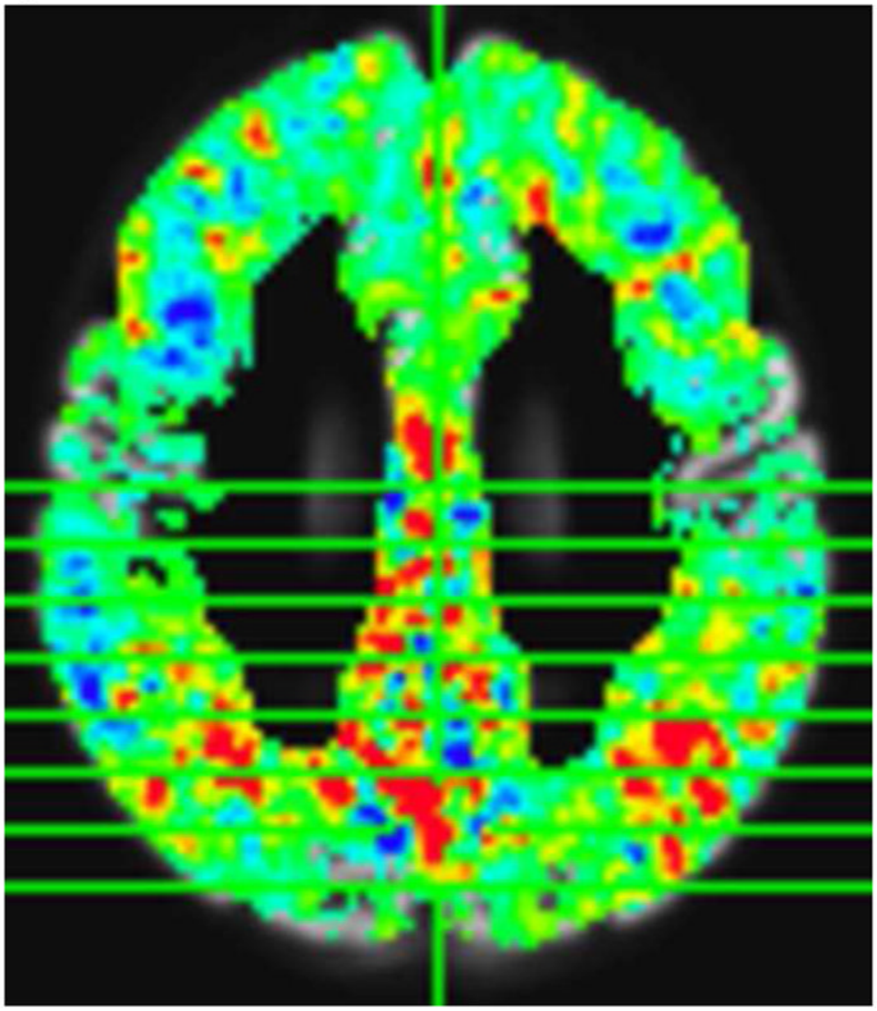


Figure 3.

Voxels used in the classifier for FDG-PET baseline images. Weights are relative, and have no applicable units. Blue indicates negative weights, associated with AD, while green indicates zero or neutral weight, while red indicates positively weighted regions associated with healthy status. Green bars in the axial and sagittal views correspond to coronal slices.

Figure 4a



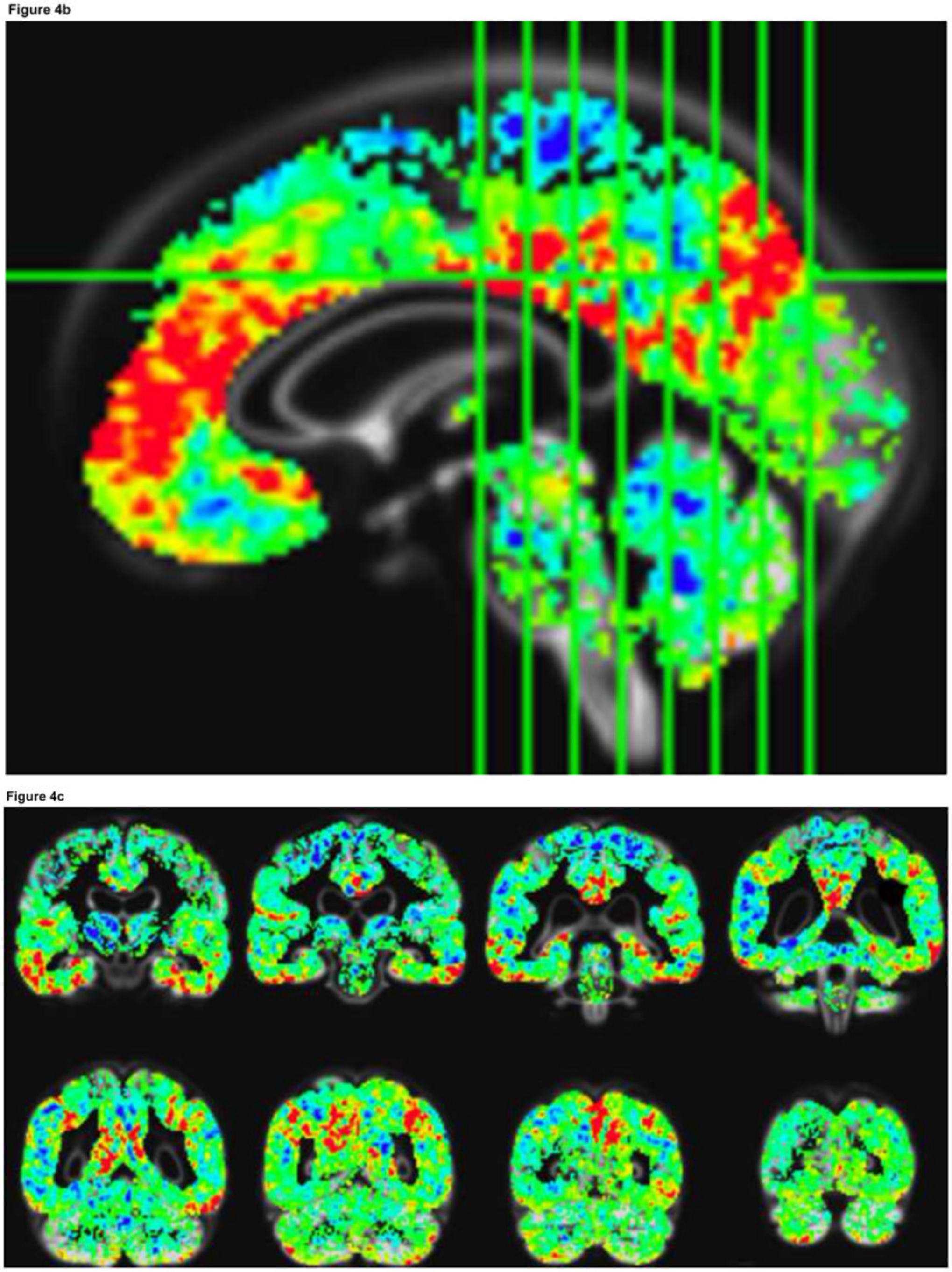


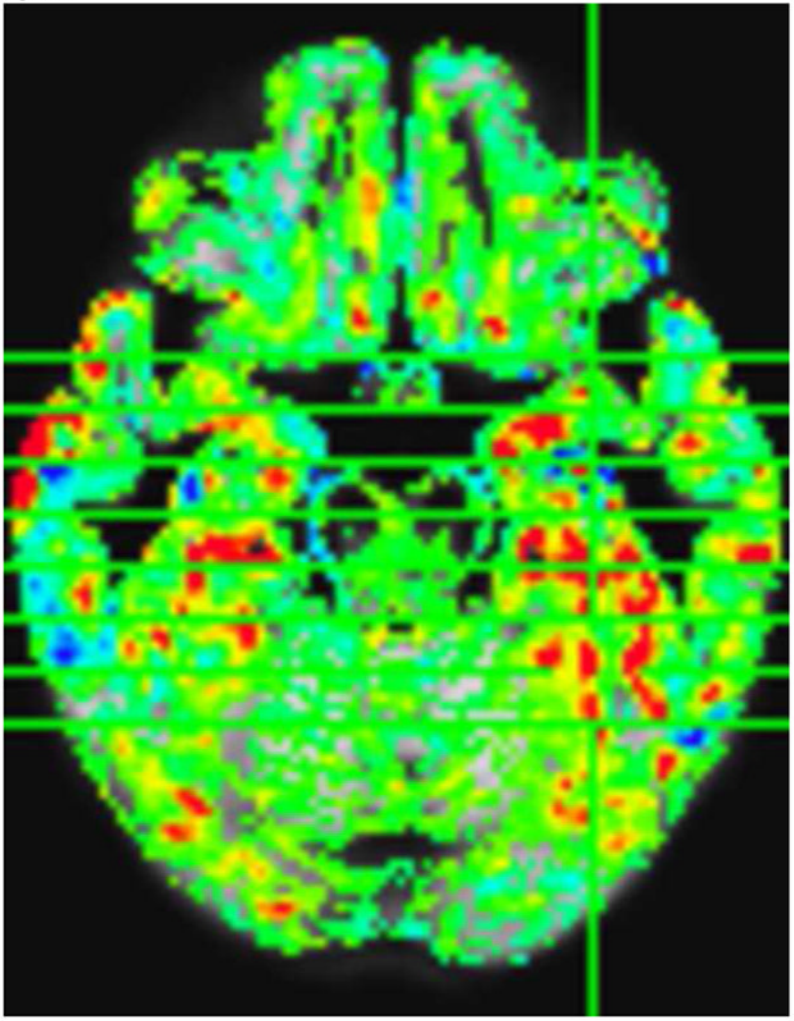
Figure 4b

Figure 4c

Figure 4.

Voxels used in the classifier for FDG-PET images at 24 months. Weights are relative, and have no applicable units. Blue indicates negative weights, associated with AD, while green indicates zero or neutral weight, while red indicates positively weighted regions associated with healthy status. Green bars in the axial and sagittal views correspond to coronal slices.

Figure 5a



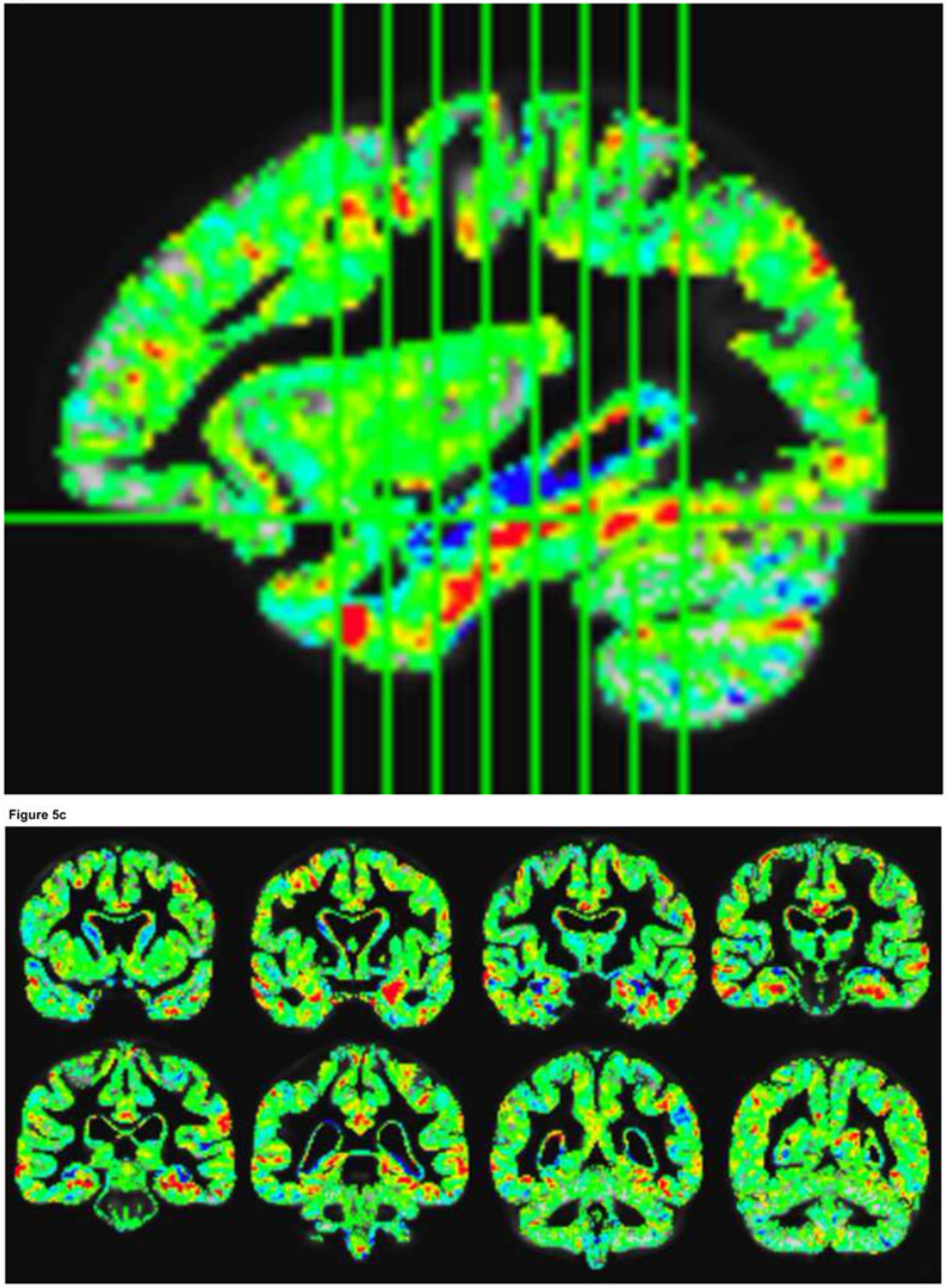


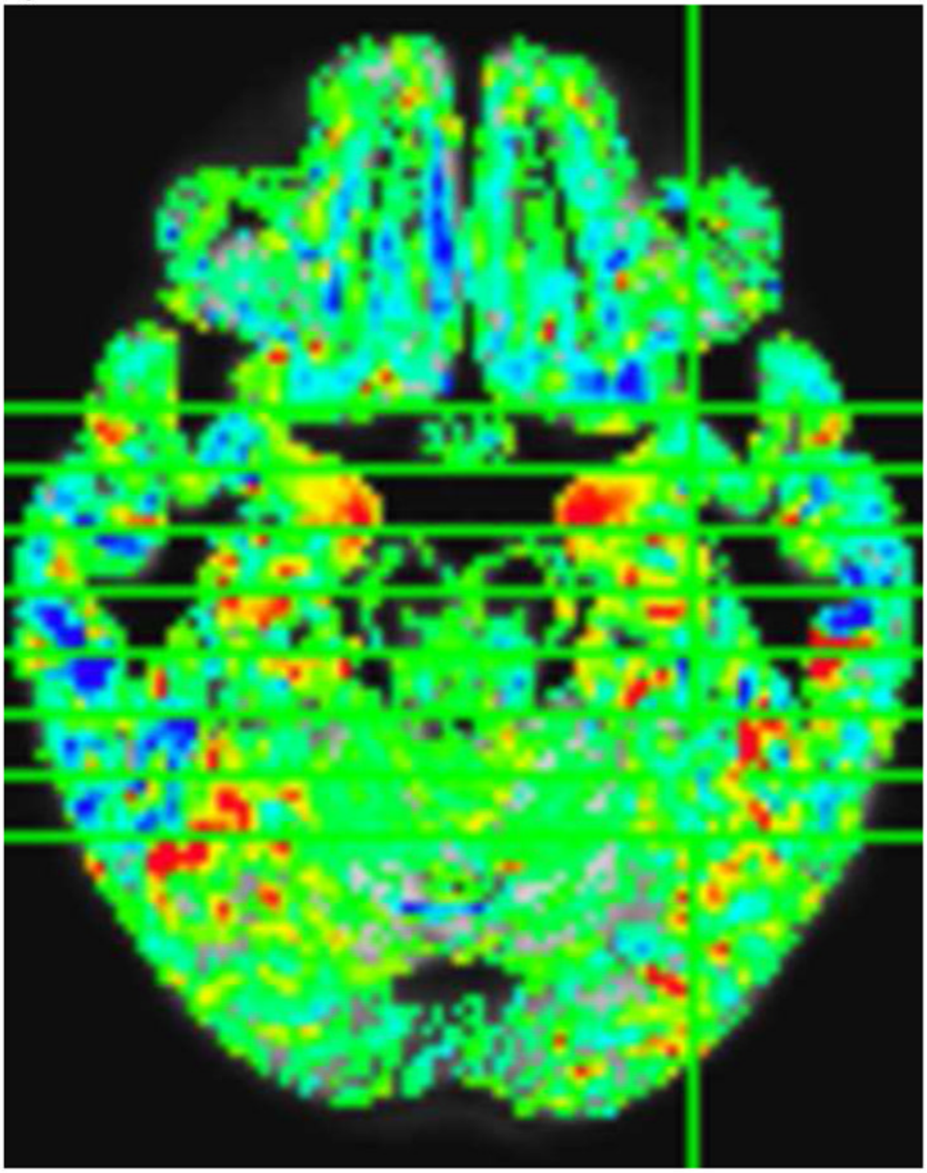
Figure 5b

Figure 5c

Figure 5.

Voxels used in the classifier for TBM-processed MR images. Weights are relative, and have no applicable units. Blue indicates negative weights, associated with AD, while green indicates zero or neutral weight, while red indicates positively weighted regions associated with healthy status. Green bars in the axial and sagittal views correspond to coronal slices.

Figure 6a



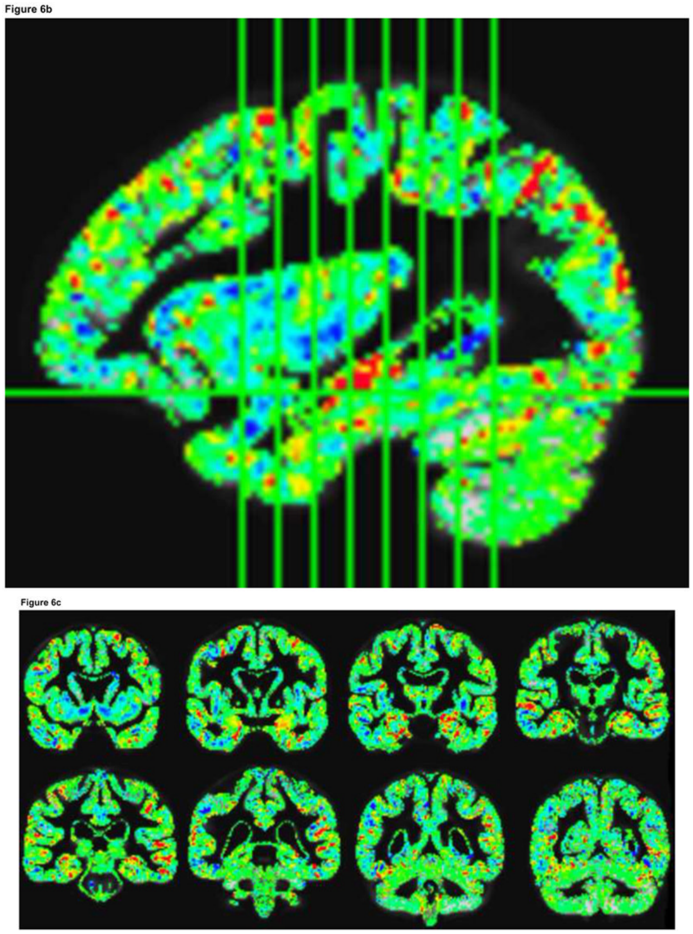


Figure 6.
Voxels used in the classifier for VBM-processed (GM density) MR images. Weights are relative, and have no applicable units. Blue indicates negative weights, associated with AD, while green indicates zero or neutral weight, while red indicates positively weighted regions associated with healthy status. Green bars in the axial and sagittal views correspond to coronal slices.

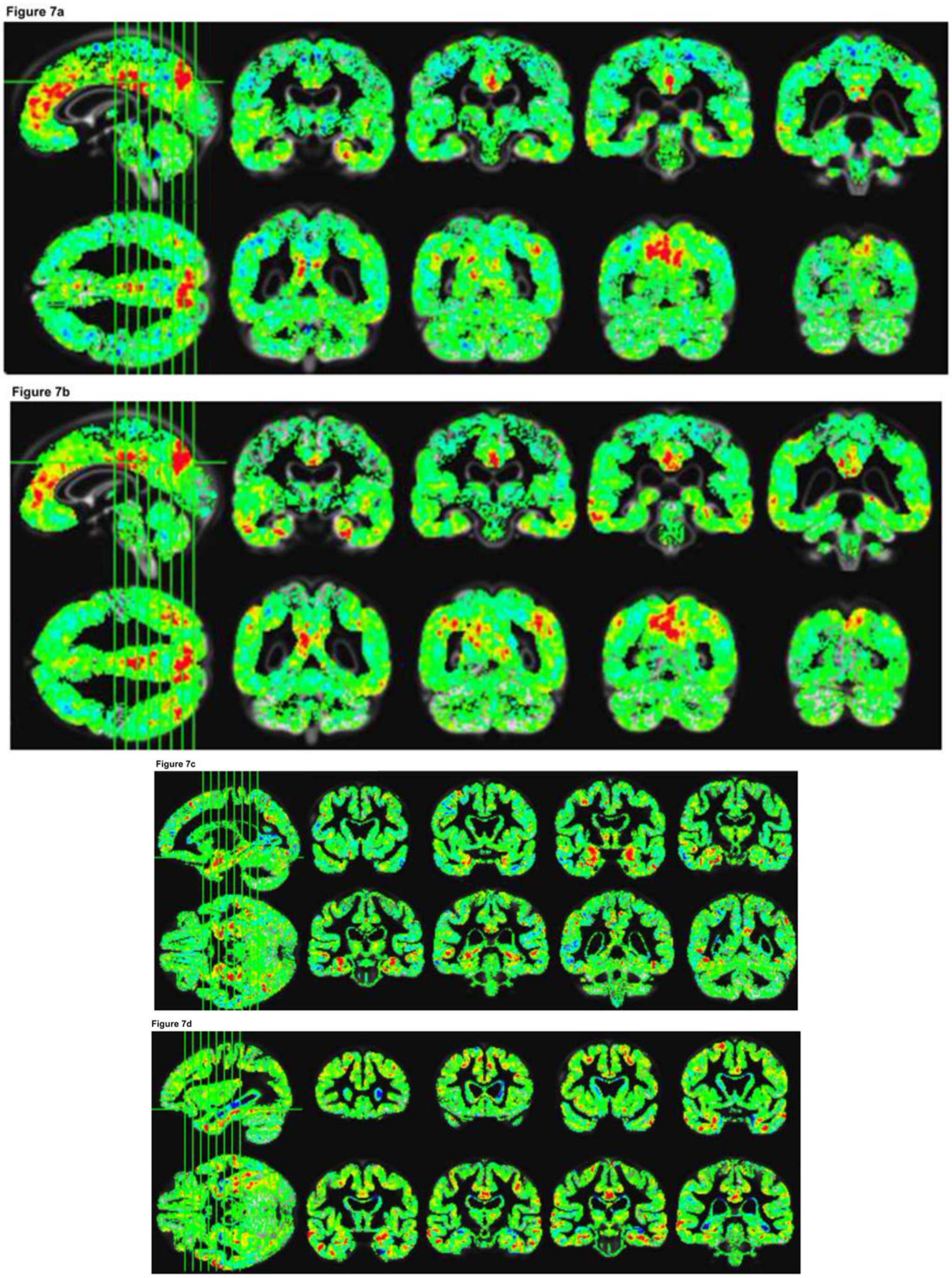


Figure 7.

Voxel weights assigned by the MKL classifier when the outlier subjects were removed. (a) FDG-PET baseline images; (b) FDG-PET images at 24 months; (c) VBM-processed baseline MR images; (d) TBM-processed longitudinal MR scans.

Figure 8a

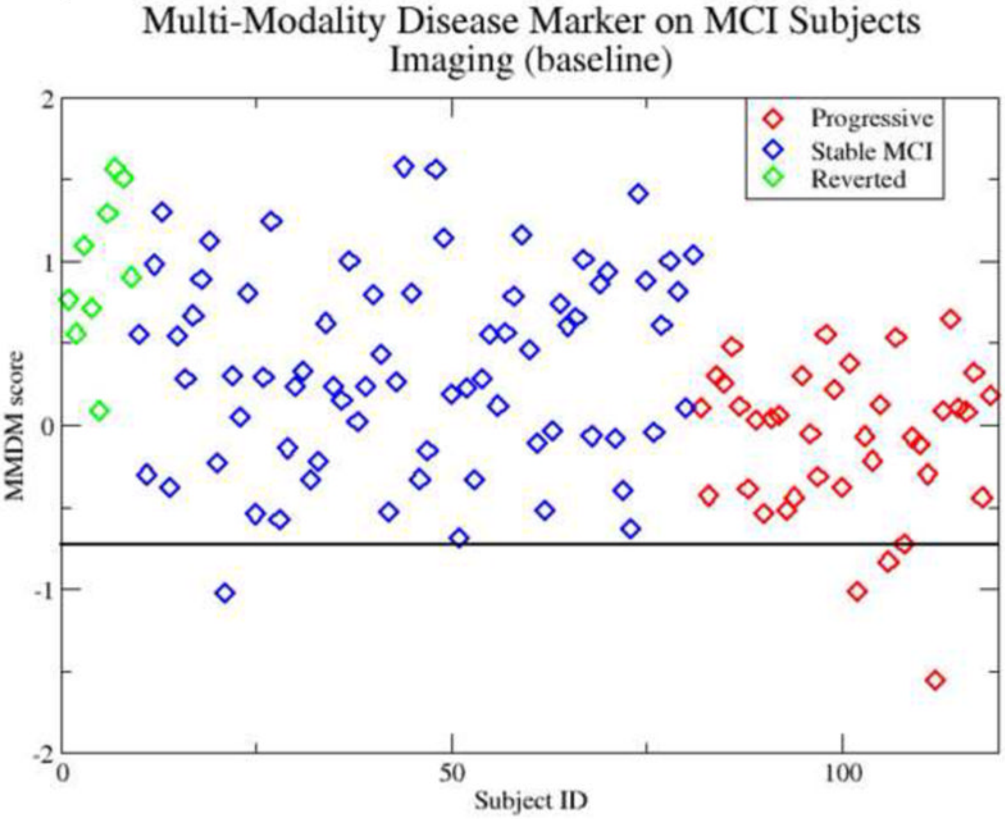


Figure 8b

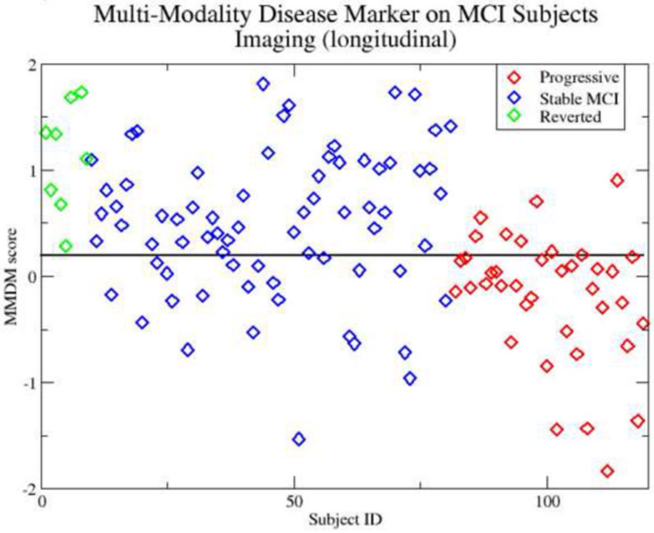


Figure 8c

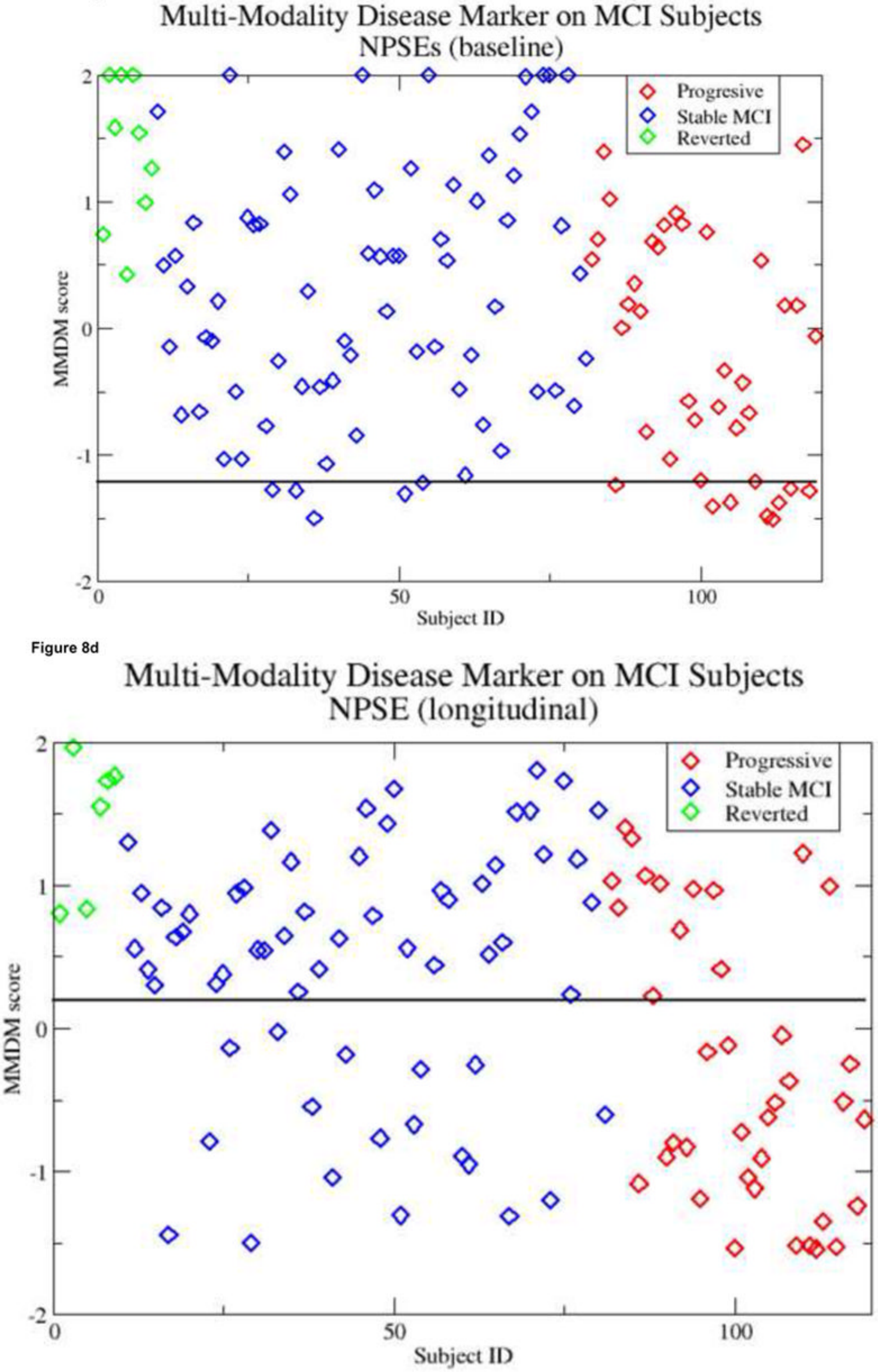


Figure 8.

MMDMs applied to the MCI population. Subjects which remained stable are shown in blue; subjects which progressed to AD are shown in red; subjects which reverted to normal cognitive status are shown in green. In each figure, a line giving maximal post-hoc accuracy is shown. Note that in some cases, the best accuracy can be achieved by simply labeling all subjects as the majority class. In some cases, MMDM scores were truncated to ±2 so as to preserve the relative scales. On the left (a,c) are shown MMDMs based on information available at baseline. Note the homogeneity of the groups, leading to poor separability. Imaging-based MMDMs are shown a the top (a), while MMDMs based on NPSEs are shown below (c). On the right (b,d) are shown MMDMs based on all modalities available at 24 months. Note the improved separability between the progressing (red) and stable (blue) MCI subjects. Note that the imaging-based marker above (b) shows slightly greater separation of the 2 groups.

Figure 9a

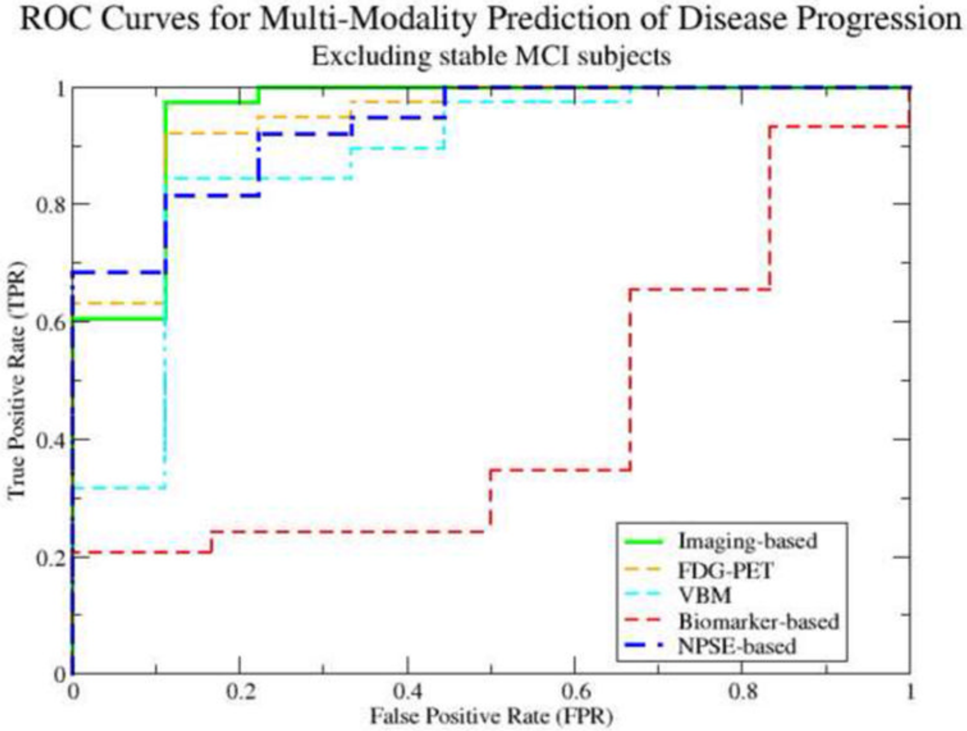


Figure 9b

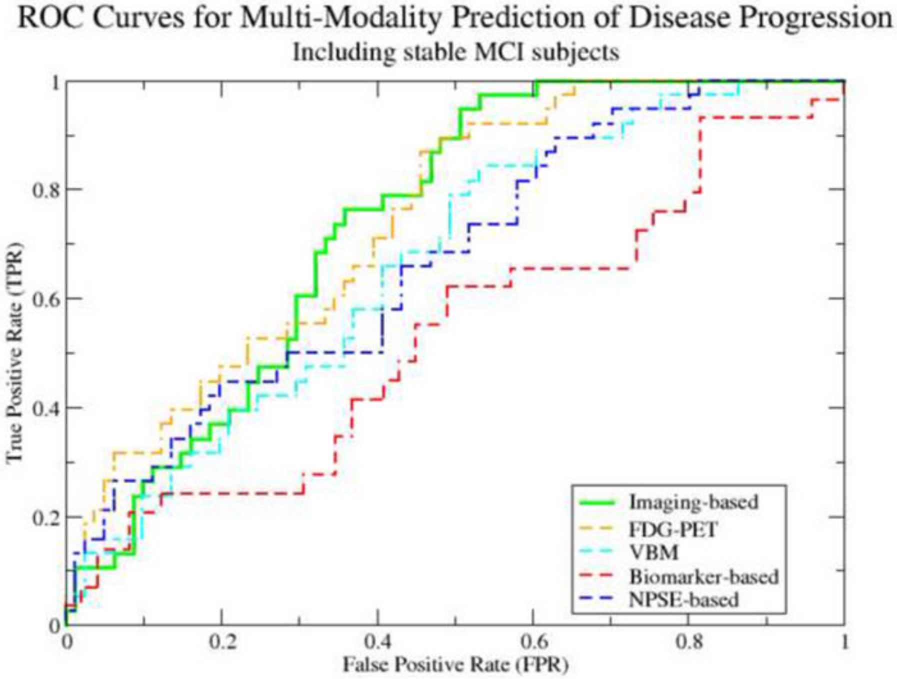


Figure 9c

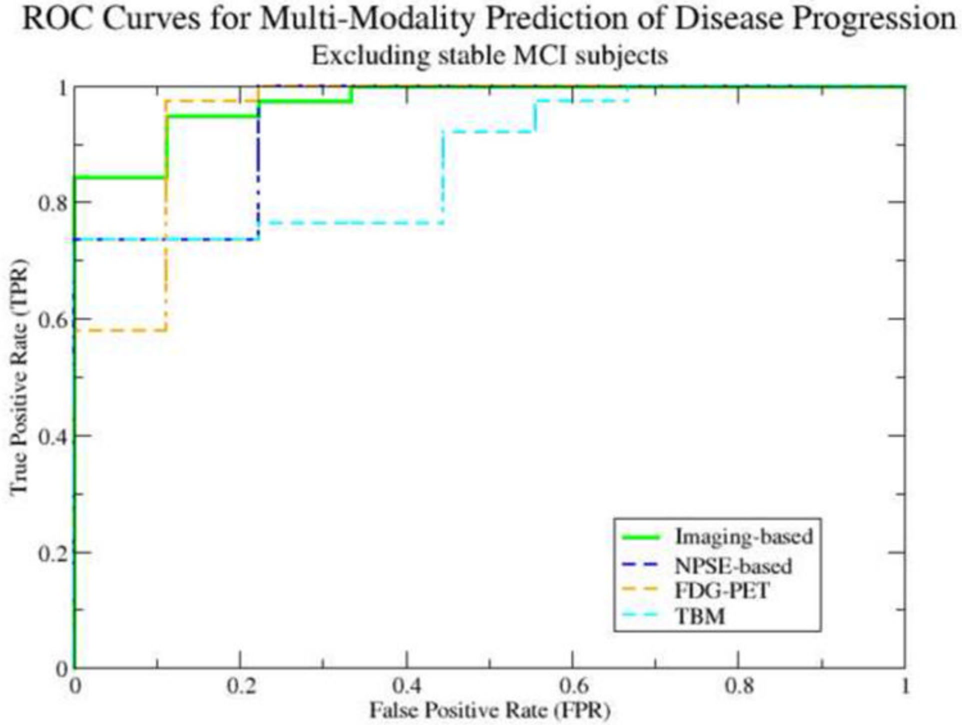


Figure 9d

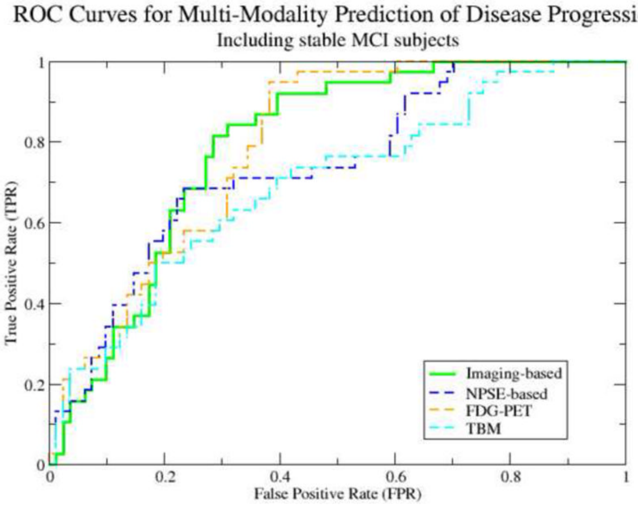


Figure 9.

ROC curves for multi-modality learning on disease progression of MCI subjects using various disease markers. The ROC curves for separating progressing and reverting MCI subjects on the left (a,c). The ROC curves for separating progressing MCI subjects from all others are shown on the right, (b,d). The top row (a,b) shows the curves derived from information available at baseline, while those on the bottom (c,d) were derived from scans and markers taken at both baseline and 24-months.

TABLE 1

Study population demographics

Demographic and neuropsychological characteristics of the study population.

	controls (mean)	controls (s.d.)	MCI (mean)	MCI (s.d.)	AD (mean)	AD (s.d.)
Age at baseline	76.2	4.59	75.1	7.44	76.6	6.28
Gender(M/F)	40/26	–	79/40	–	25/23	–
APOE carriers	17	–	63	–	37	–
MMSE at Baseline	29.17	0.85	27.18	1.64	23.50	1.92
MMSE at 24 months	28.67	3.73	25.54	4.84	18.98	6.60
ADAS at baseline	9.94	4.27	17.26	6.13	28.27	9.80
Years of Education	16.15	3.02	15.73	2.82	14.60	3.17
Geriatric Depression	0.97	1.35	1.40	1.28	1.71	1.47

TABLE 2

Biological measures data used in kernel functions

Non-imaging biological measures used to construct kernels for experiments. Cerebro-Spinal Fluid (CSF) assays and APOE genotype data were utilized.

Type	Subjects available
Tau	130
Amyloid-Beta 142	130
P-Tau 181P	130
T-Tau	130
APOE Genotype	233

TABLE 3
Cognitive markers used in kernel functions

Non-imaging cognitive markers used to construct kernels for experiments.

Cognitive measure	Subjects available
Rey auditory/verbal 1–5 scores	233
Rey auditory delayed recall scores	233
Category Fluency scores	233
Trail-making A & B	233
Digit-span scores	233
Boston Naming scores	233
ANART errors	233

TABLE 4
Accuracy results of validation experiments using 2-norm MKL

Comparison of 2-norm MKL with different types of input data modalities.

Modalities used	Accuracy	Sensitivity	Specificity	Area under ROC
Imaging modalities	0.876	0.789	**0.938**	0.944
Biological measures	0.704	0.581	0.794	0.767
Cognitive scores	**0.912**	**0.892**	0.926	**0.983**
All modalities	0.924	0.867	0.966	0.977

TABLE 5

Comparison of different MKL norms with the SVM trained on concatenated-features

Comparison of different MKL norms in the presence of uninformative kernels, and an SVM trained on a concatenation of all features for comparison.

MKL norm used	Accuracy	Sensitivity	Specificity	Area under ROC
1.0	0.914	0.867	0.949	0.977
1.25	0.916	0.865	0.954	0.980
1.5	0.921	0.874	0.956	0.982
1.75	0.923	0.872	0.961	0.982
2.0	0.922	0.870	0.959	0.981
SVM (concatenated features)	0.882	0.844	0.910	0.970

TABLE 6

t-statistic p-values for comparisons between MMDMs of stable MCI subjects, progressing subjects, and reverting subjects

Significance of group-level differences in MMDM scores assigned to MCI subjects. There are 3 groups of MCI subjects - those who reverted to normal status, those who remained stable for 3 years, and those who progressed to full AD in 3 years.

Modalities used	Reverting vs. rest	Progressing vs. rest
Biological measures (baseline)	0.65	0.58
Imaging Data (baseline)	1.31×10^{-3}	1.78×10^{-6}
Imaging Data (longitudinal)	5.69×10^{-4}	3.29×10^{-7}
NPSEs (baseline)	2.63×10^{-3}	5.51×10^{-4}
NPSEs (longitudinal)	2.44×10^{-4}	2.19×10^{-6}

TABLE 7

Area Under ROC results for different classes of MMDMs in predicting MCI progression to AD

Area under ROC curves for predicting whether MCI subjects will progress to AD or not. In the left column are AU ROCs for the task of separating only progressing subjects from reverting subjects, while ignoring stable MCI subjects. On the right are AU ROCs for separating progressing subjects from all other subjects.

Modalities used	Progressing vs. Reverting	Progressing vs. Rest
Biological measures (baseline)	0.4368	0.5292
Imaging Data (baseline)	0.9532	0.7378
Imaging Data (longitudinal)	**0.9737**	**0.7911**
NPSEs (baseline)	0.9298	0.6693
NPSEs (longitudinal)	0.9415	0.7385
All Modalities	0.9708	0.7667