 DOI: 10.1111/biom.12343

Multiple Kernel Learning with Random Effects for Predicting Longitudinal Outcomes and Data Integration

Tianle Chen,[1] Donglin Zeng,[2] and Yuanjia Wang[3,*]

[1]Biogen, Cambridge, MA 02142, USA
[2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[3]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA
email: yw2016@columbia.edu

Summary. Predicting disease risk and progression is one of the main goals in many clinical research studies. Cohort studies on the natural history and etiology of chronic diseases span years and data are collected at multiple visits. Although, kernel-based statistical learning methods are proven to be powerful for a wide range of disease prediction problems, these methods are only well studied for independent data, but not for longitudinal data. It is thus important to develop time-sensitive prediction rules that make use of the longitudinal nature of the data. In this paper, we develop a novel statistical learning method for longitudinal data by introducing subject-specific short-term and long-term latent effects through a designed kernel to account for within-subject correlation of longitudinal measurements. Since the presence of multiple sources of data is increasingly common, we embed our method in a multiple kernel learning framework and propose a regularized multiple kernel statistical learning with random effects to construct effective nonparametric prediction rules. Our method allows easy integration of various heterogeneous data sources and takes advantage of correlation among longitudinal measures to increase prediction power. We use different kernels for each data source taking advantage of the distinctive feature of each data modality, and then optimally combine data across modalities. We apply the developed methods to two large epidemiological studies, one on Huntington's disease and the other on Alzheimer's Disease (Alzheimer's Disease Neuroimaging Initiative, ADNI) where we explore a unique opportunity to combine imaging and genetic data to study prediction of mild cognitive impairment, and show a substantial gain in performance while accounting for the longitudinal aspect of the data.

Key words: Disease prediction; Integrative analysis; Latent effects; Statistical learning.

1. Introduction

Accurate prediction of current and future clinical status of a patient based on subject-specific clinical and biological markers is an important goal for early diagnosis and monitoring disease progression. Modern technologies offer opportunities to collect data from heterogeneous sources such as genetic data, imaging data, and clinical data including electronic health records. Therefore, it is valuable to develop prediction rules that can accommodate heterogeneous sources of data to boost prediction power. Furthermore, many cohort studies on natural history and etiology of chronic diseases often span years and data may be collected at multiple visits. It is thus important to develop time-sensitive prediction rules that not only integrate data from multiple sources but also make use of the longitudinal nature of the data collected from the same subjects.

There is an extensive body of literature on longitudinal data analysis exploring the association between candidate predictors and outcomes measured repeatedly over time (e.g., Diggle et al., 2002). In these association analyses, primary goals are estimation and hypothesis testing of regression parameters which may not necessarily yield powerful prediction rules. The focus of the current work is on prediction of outcomes in future subjects or prediction of future observations on the same subject from longitudinal data with a potentially large number of predictors. For the purpose of prediction with longitudinal data, some previous research has focused on linear or quadratic discriminant analysis of longitudinal profiles or a sample of curves (e.g., James and Hastie 2001; Marshall and Baron 2000; Luts et al., 2013). These papers aim to classify a functional curve into two groups and rely on either linear mixed effects models (Verbeke and Leseffre, 1996; Marshall and Baron, 2000) or functional data analysis or their extensions (James and Hastie, 2001) to perform classification. In the past decades, there has been growing interest in using powerful machine learning methods to build effective predictive models for binary and continuous disease outcomes (Oquendo et al., 2012). Particularly, kernel-based methods such as support vector machine or support vector regression are proposed to classify longitudinal profile into groups (Pearce and Wand, 2009; Luts et al., 2012). However, disease outcomes in these approaches do not change with time so they are not applicable to classify clinical outcomes assessed repeatedly over time. Since most of the existing statistical learning methods assume the sample to be independent and identically distributed, there is a lack of literature on how to effectively incorporate within-subject dependence to improve prediction of future subjects' clinical outcomes or within-subject change especially when the clinical outcomes are binary.

In this paper, we introduce a novel statistical learning method to predict longitudinal binary outcomes in the

multiple kernel learning (Lanckriet et al., 2004; Bach and Lanckriet, 2004) framework. Our method not only uses observed feature variables but also introduces subject-specific unobserved latent variables to extract information from correlated outcomes and build time-sensitive prediction rules. More specifically, we use multiple additive kernels for observed feature variables, which can account for heterogeneous data sources taking advantage of the correlation within each data modality, while at the same time, we account for within-subject correlation of longitudinal measurements by introducing subject-specific short-term and long-term latent random effects modeled through a separate kernel. In many biomedical studies, the observed feature variables only explain some proportion of variability in outcomes, and the gain from using latent random effects to extract information from the remaining unexplained variability can be substantial. The weights used for each kernel are tuned based on minimizing the overall loss, therefore we optimally combine data across modalities in a data-driven fashion. In addition to methods for training model, we also develop methods for predicting future outcomes through observed features and unobserved latent effects when longitudinal training data are available.

On one hand, depending on the choice of kernels, the proposed method has some similarity to semiparametric or nonparametric mixed effect models for longitudinal data. However, unlike traditional mixed models, our proposed method aims at prediction accuracy, allows greater flexibility through the use of kernel machines, and is relatively easy to scale up for large dimensional data. On the other hand, using different kernels for feature variables and latent variables shares the same advantages with multiple kernel learning methods which have been developed to handle the challenges of integrating different data sources (Pavlidis et al., 2002; Lanckriet et al., 2004; Yu et al., 2010; Zhang and Shen, 2012). Specifically, the latter treats each data source component, for example, genetic data, imaging data or clinical data, as belonging to separate kernel spaces and finds an optimal way to combine them for prediction. The multiple kernel methods have been shown to yield much improved performance as compared to using one single kernel in various biomedical applications (Yu et al., 2010). Although our proposed method uses multiple kernel algorithms, one significant difference from the above literature is that separate kernels are also applied to unobserved latent variables.

The paper is structured as follows. In Section 2, we propose a learning method to predict longitudinal binary outcomes based on the support vector machine with multiple kernels. In Section 3, extensive simulation studies are conducted to illustrate small-sample performance of the proposed method and compare with some existing approaches. In Section 4, we apply the developed method to analyze the Alzheimer's disease neuroimaging initiative (ADNI) data, where a unique opportunity is presented to combine various modalities of imaging and genetic data to distinguish subjects with mild cognitive impairment (MCI) from subjects with Alzheimer's disease (AD), and we show a substantial gain in performance while accounting for the longitudinal correlation in the data. The proposed multiple kernel fusion with random effects proves to be effective in this application. Some remarks are provided in Section 5.

2. Multiple Kernel Fusion Learning for Longitudinal Data

We start by briefly introducing standard statistical learning through support vector machine with a single kernel, followed by incorporating a longitudinal component to the learning through fusing two kernels, and lastly we discuss integration of multiple data sources through fusing multiple heterogeneous kernels.

2.1. *Review of Support Vector Machine*

Let $\mathcal{X}$ denote a complete separable space for feature variables. The random feature variables $\mathbf{X}$ take values in $\mathcal{X}$, and the binary disease outcomes Y take values in $\mathbb{R}$. The goal of statistical learning is to train an optimal prediction function $f : \mathcal{X} \to \mathbb{R}$ to predict Y given $\mathbf{X}$ for any future subject, where the performance of prediction is quantified by the prediction error defined as $E[I(Yf(X) < 0]$. Due to the non-smoothness of $I(Yf(X) < 0)$, the optimal prediction function is usually obtained by minimizing the empirical version of some surrogate loss function. One such loss function most commonly used is the hinge loss, or the so called support vector machine (SVM, Vapnik, 1995), and it has been proven to be successful in a wide range of applications (Orru et al., 2012).

Assume that we have n independent observations $(\mathbf{x}_i, y_i), i = 1, ..., n$. With a linear prediction function $f(\mathbf{x}_i) = \langle \mathbf{x}_i, \mathbf{w} \rangle + d$, where the inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T\mathbf{b}$, the primal optimization problem of the SVM has the form (e.g., Hastie et al., 2009)

$$\min_{\mathbf{w}\in\mathcal{X}, d\in\mathbb{R}} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \xi_i \right\} \tag{1}$$

subject to the constraints with slack variables ξ_i

$$y_i(\langle \mathbf{x_i}, \mathbf{w} \rangle + d) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \text{ for all } i = 1, ..., n.$$

To accommodate nonlinear boundary, a Mercer kernel $k(\cdot, \cdot)$ is defined such that $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$, where $\Phi(\cdot)$ is the mapping from the input space to a higher dimensional feature space, and $\langle \cdot, \cdot \rangle$ is the inner product defined in the reproducing kernel Hilbert space (RKHS, Wahba 1990). The corresponding dual form becomes

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right\},$$

leading to the decision functions of the form $d(\mathbf{x}) = \text{sign}\,(\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + d)$. Note that one advantage of solving the optimization from the dual form is that the explicit form of $\Phi(\cdot)$ does not need to be known as long as the kernel function $k(\cdot, \cdot)$ is well defined (Kimeldorf and Wahba, 1970).

2.2. *Proposed Multiple Kernel Learning for Longitudinal Data*

For longitudinal biomedical data, outcome measures on the same subjects are correlated after accounting for the observed fixed effects feature variables. Taking advantage of such correlation is expected to lead to improved prediction. Classical longitudinal analysis divides into two camps: estimating

the marginal population-average effect, and estimating the subject-specific effect given the random effects. For the former view, correlation among repeated measures is treated as nuisance parameter, while for the latter it is modeled through subject-specific random effects. In our setting, subject-specific classifications are of interest instead of population average effects, therefore we introduce random effects to the SVM framework to improve prediction in our proposed approach.

Assume that we have n independent subjects and the ith subject has n_i visits. Let y_{ij} denote the disease outcome for the ith subject at the jth visit coded as "1" for diseased subjects and "−1" for non-diseased subjects. Let $\mathbf{x}_{ij}$ denote a vector of feature variables collected at the same visit. We introduce two latent random effects for subject i, a time-invariant effect a_{ij}, which aims to capture the long-term latent effect across all the visits from the same subject, and a time-varying effect b_{ij}, which attempts to account for short-term latent effect or local influence from recent history that depends on the time interval between visits. Therefore, for a subject with feature variables $\mathbf{x}_{ij}$ at time t_{ij}, a prediction rule with subject-specific random effects can be expressed as

$$\text{sign}\{f(\mathbf{x}_{ij}, a_{ij}, b_{ij})\},$$

where the prediction function has the form

$$f(\mathbf{x}_{ij}, a_{ij}, b_{ij}) = \langle \Phi_x(\mathbf{x}_{ij}), \mathbf{w}\rangle + w_a\Phi_a(a_{ij}) + w_b\Phi_b(b_{ij}). \quad (2)$$

Here, $\Phi_x(\mathbf{x})$ consists of some mapping from the input space $\mathcal{X}$ to a higher-order feature space (e.g., the basis function associated with some reproducing kernel Hilbert space) and both $\Phi_a(a)$ and $\Phi_b(b)$ are nonlinear transformation of the latent effects which will be induced by some kernel functions defined for a_{ij} and b_{ij}, respectively in Section 2.3. For identifiability, we also assume that a_{ij} and b_{ij} are standardized random variables with mean zero and variance one. Clearly, since $\mathbf{a}$ and $\mathbf{b}$ are unobserved random variables, conventional SVM techniques cannot be directly applied.

When including the random effects into the model, the single kernel SVM becomes a multi-kernel SVM with one kernel for fixed effects and two kernels for random effects. Following the multiple kernel learning framework, a weight parameter θ is then assigned to each kernel and a fused kernel is formed as a linear combination of kernels under an L_2-norm regularization constraint on the weight parameters. The weights are chosen in a data-driven way to minimize the loss function under the fused kernels. Thus, the primal form in the feature space becomes

$$\begin{aligned}
&\min_{\mathbf{w}_x\in\mathcal{X}} \ \frac{1}{2}\left(\frac{\mathbf{w}_x^T\mathbf{w}_x}{\theta_x} + \frac{w_a^2}{\theta_a} + \frac{w_b^2}{\theta_b}\right) + C\sum_{i,j}^{n,n_i}\xi_{ij} \qquad (3)\\
&\text{subject to} \ \ y_{ij}\left(\sqrt{\theta_x}\langle\Phi_x(\boldsymbol{x}_{ij}), \mathbf{w}_x\rangle + \sqrt{\theta_a}w_a\Phi_a(a_{ij})\right.\\
&\qquad \left. + \sqrt{\theta_b}w_b\Phi_b(b_{ij})\right) \geq 1-\xi_{ij}\\
&\qquad \xi_{ij}\geq 0, \ i=1,...,n, \ \text{and} \ j=1,...,n_i,\\
&\qquad \theta_x^2+\theta_a^2+\theta_b^2=1, \quad \theta_x,\theta_a,\theta_b\geq 0.
\end{aligned}$$

As a remark, comparing the optimization problem for longitudinal data (3) with the original standard SVM primal form (1), we observe that the objective function for the former is a conic combination of the separate objective functions for the latter with a quadratic constraint. Furthermore, the resemblance with multiple kernel learning allows easy generalization to accommodate data from heterogeneous sources by using separate kernels for observed feature variables from each source. Such method incorporates prior knowledge on each source while performing integration. Contrary to concatenating all variables in a single kernel, using separate ones reflects prior knowledge that the feature variables from the same source have stronger correlations than with variables from difference sources. For example, assuming there are P data sources of fixed effects and two kernels for random effects, the corresponding primal form is

$$\begin{aligned}
&\min_{\mathbf{w}\in\mathcal{X},\theta_p,\theta_a,\theta_b\in\mathbb{R}} \ \frac{1}{2}\left(\sum_{p=1}^{P}\frac{\mathbf{w}_p^T\mathbf{w}_p}{\theta_p} + \frac{w_a^2}{\theta_a} + \frac{w_b^2}{\theta_b}\right) + C\sum_{i,j}^{n,n_i}\xi_{ij}\\
&\text{subject to} \ \ y_{ij}\left(\sum_{p=1}^{P}\sqrt{\theta_p}\langle\Phi_p(\boldsymbol{x}_{ijp}), \mathbf{w}_p\rangle + \sqrt{\theta_a}w_a\Phi_a(a_{ij})\right.\\
&\qquad \left. + \sqrt{\theta_b}w_b\Phi_b(b_{ij})\right) \geq 1-\xi_{ij}\\
&\qquad \xi_{ij}\geq 0, \ i=1,...,n, \ \text{and} \ j=1,...,n_i,\\
&\qquad \sum_{p=1}^{P}\theta_p^2+\theta_a^2+\theta_b^2=1, \quad \theta_p,\theta_a,\theta_b\geq 0, \quad p=1,\cdots,P.
\end{aligned}$$

The computation of the multiple kernel learning is essentially a quadratically-constrained quadratic programming (QCQP) problem

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}}\min_{\boldsymbol{\theta}} \ \sum_{ij}^{n,n_i}\alpha_{ij} - \frac{1}{2}\sum_{i,k}^{n}\sum_{j,l}^{n_i}\alpha_{ij}\alpha_{kl}y_{ij}y_{kl}\{\sum_{p=1}^{P}\theta_p k_p(\mathbf{x}_{ijp}, \mathbf{x}_{klp})\\
&\qquad + \theta_a k_a(a_{ij}, a_{kl}) + \theta_b k_b(b_{ij}, b_{kl})\} \qquad (4)\\
&\text{subject to} \quad 0\leq\alpha_{ij}\leq C, \ i,k=1,\ldots,n, \ j,l=1,\ldots,n_i, \sum_{i,j}^{n,n_i}\alpha_{ij}y_{ij}=0,\\
&\qquad \sum_{p}^{P}\theta_p^2+\theta_a^2+\theta_b^2=1, \ \theta_p,\theta_a,\theta_b\geq 0, \ p=1,\cdots,P.
\end{aligned}$$

where $k_p(\mathbf{x}_{ijp}, \mathbf{x}_{klp}) = \langle\Phi_p(\mathbf{x}_{ijp}), \Phi_p(\mathbf{x}_{klp})\rangle$ is the kernel for the reproducing kernel Hilbert space for $\mathbf{x}_{ijp}$, and $k_a(a_{ij}, a_{kl}) = \langle\Phi_a(a_{ij}), \Phi_a(a_{kl})\rangle$ and $k_b(b_{ij}, b_{kl}) = \langle\Phi_b(b_{ij}), \Phi_b(b_{kl})\rangle$ are kernel functions for some inner products defined for latent effects we discuss next.

2.3. *Choice of Kernel Functions for Latent Effects*

Here we introduce kernels to model the two random effects a_{ij} and b_{ij} respectively. Recall kernel matrix measures similarity between two observations, a natural choice of kernel function is the covariance structure of the random effects which can also be considered as the inner product with respect to

its distribution function. Thus, we assume that the similarity between the latent effects from independent subjects is zero, the similarity between the long term random effects on the same subjects is a constant ρ, and the similarity between local short term random effects depends on the time interval between the two measurements. Specifically, to account for the long-term latent effects, we can consider a_{ij} to represent the common random effect shared across visits plus an independent random error component, and therefore the commonly shared random effect will contribute to prediction at each visit. Equivalently, construct elements in a kernel matrix as $k_a(a_{ij}, a_{kl}) = 1$ if $i = k$, $j = l$; $k_a(a_{ij}, a_{kl}) = \rho$ if $i = k$, $j \neq l$; and $k_a(a_{ij}, a_{kl}) = 0$ if $i \neq k$. Next, in order to account for short term latent random effects, we assume an exponential covariance structure for $\mathbf{b}_i$. Thus, $k_b(b_{ij}, b_{kl}) = \exp\{-\alpha|t_{ij} - t_{il}|\}$ if $i = k$; and $k_b(b_{ij}, b_{kl}) = 0$ if $i \neq k$. The kernel function for n_i long-term random effects $\mathbf{a_i} = (a_{i1}, \cdots, a_{in_i})^T$ and short term random effects $\mathbf{b}_i = (b_{i1}, \cdots, b_{in_i})^T$ with measurement time points $(t_{i1}, \cdots, t_{in_i})^T$ are defined as

$$\mathbf{K}_{\mathbf{a}_i} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots \\ \rho & \cdots & \rho & 1 \end{pmatrix}_{n_i \times n_i}$$

$$\mathbf{K}_{\mathbf{b}_i} = \begin{pmatrix} 1 & e^{-\alpha|t_{i1}-t_{i2}|} & \ldots & e^{-\alpha|t_{i1}-t_{in_i}|} \\ e^{-\alpha|t_{i1}-t_{i2}|} & 1 & \ldots & e^{-\alpha|t_{i2}-t_{in_i}|} \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}_{n_i \times n_i}$$

where α is a scale parameter.

Under the above choice of kernels, we can optimize the dual form (4) using the quadratic programming. Earlier work suggests exhaustive search at given values of $\boldsymbol{\theta}$ and treating the fused kernels as a new kernel in a standard SVM optimization problem. However, the computational burden is high. A computationally efficient algorithm for solving the optimization problem (4) was proposed in Yu et al. (2010) to solve for weights $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ simultaneously. Specifically, the dual form (4) is solved under the Cauchy–Schwarz inequality as

$$\min_{t,\boldsymbol{\alpha}} \ \frac{1}{2}t - \sum_{i,j}^{n,n_i} \alpha_{ij}$$

$$\text{subject to } \sum_{i,j}^{n,n_i} \alpha_{ij} y_{ij} = 0,\ 0 \leq \alpha_{ij} \leq C,\ t \geq \|\boldsymbol{\gamma}\|_2,$$

where $\boldsymbol{\gamma} = \left\{\boldsymbol{\alpha}^T \boldsymbol{Y}\boldsymbol{K}_1\boldsymbol{Y}\boldsymbol{\alpha}, ..., \boldsymbol{\alpha}^T \boldsymbol{Y}\boldsymbol{K}_P\boldsymbol{Y}\boldsymbol{\alpha}, \boldsymbol{\alpha}^T \boldsymbol{Y}\boldsymbol{K}_a\boldsymbol{Y}\boldsymbol{\alpha}, \boldsymbol{\alpha}^T \boldsymbol{Y}\boldsymbol{K}_b\boldsymbol{Y}\boldsymbol{\alpha}\right\}^T$, and the optimal weight parameters for the pth kernel is $\theta_p^* = \boldsymbol{\alpha}^T \boldsymbol{Y}\boldsymbol{K}_p\boldsymbol{Y}\boldsymbol{\alpha}/\|\boldsymbol{\gamma}\|_2$.

2.4. *Prediction of Future Observations*

For a longitudinal study, we distinguish two types of prediction of interest. We define type A prediction as predicting outcome for a new subject with the observed feature variables $\mathbf{x}$ only and no prior history information, for example, prediction for a new subject at the baseline visit. We define type B prediction as predicting outcomes at future follow-up time points for an existing subject with observed prior visit outcomes and feature variables $\mathbf{x}$. One of the main components of our proposed learning is to extract information from existing correlated outcomes to improve future prediction. For each type of the prediction, we discuss a different strategy in predicting the outcomes.

For type A prediction on a new subject with feature variables $\mathbf{x}_i$, directly using designed kernel functions and the fitted prediction function (2) is equivalent to using fixed effects only to predict the outcome and set the random effects at their mean level, zero. This is because the designed kernel functions $\mathbf{k}_a$ and $\mathbf{k}_b$ for random effects have non-zero values only between two visits on the same subject. In type A problem, the existing subjects and the new subject are independent, and therefore the fitted score from solving the dual form (4) do not involve random effects, which corresponds to using the population mean value for all subjects with fixed effects $\mathbf{x}_i$ to perform prediction.

To include random effects for type A prediction, we repeatedly draw independent random effects a_i and b_i from a working Gaussian distribution. For each random draw, we computed the predictive function as in (2) and classify the outcome using the sign of $f(\mathbf{x}_i, a_i, b_i)$. The final predicted outcome is based on a majority vote: if more than 50% of random draws lead to positive predicted outcomes, the final predicted outcome would be positive, and otherwise negative.

For type B prediction, we use an existing subject's predictors and outcomes at prior visits to predict their future follow up outcomes. We can then directly compute the random effects for the same subject at a future time t^* using the designed kernel matrices $\mathbf{K}_a$ and $\mathbf{K}_b$, and the fitted predictive function is obtained from the solutions to (4).

3. Simulation Studies

In this section, we conducted simulation studies to compare the empirical performance of multi-kernel SVM with several standard alternatives for analyzing longitudinal data. We started with a setting where we generated data from a single data source. The first simulation setting and results are summarized in the Supplementary Materials Section A.1. In order to mimic the real data application where the data are complex and from heterogeneous sources, in simulation setting 2, we generated the dichotomous outcomes from the following model:

$$Y_{ij} = \text{sign}\{\beta_0 T_{ij} + \boldsymbol{\beta}_1^T \boldsymbol{Z}_i + \boldsymbol{\beta}_2^T \boldsymbol{X}_{ij}^* + \beta_3 W_i^* + a_{ij} + b_{ij} + \epsilon_{ij}\},$$

where T_{ij} is the age of the ith subject at the jth visit. The age was simulated from a uniform distribution ranging from 10 to 70 years old, and the two subsequent visits from a subject were generated to be approximately 3 years apart. Here $\boldsymbol{Z}_i$ is a vector of time-invariant binary markers of the ith subject which remain the same at each visit; $\boldsymbol{X}_{1i}$ is a vector of time-invariant continuous markers of ith subject uniformly ranging from -2 to 2; and $\boldsymbol{X}_{2ij}$ is a vector of time-varying continuous markers with a correlation $\rho(X_{2ij}, X_{2ik}) = \exp(-\alpha|t_{ij} - t_{ik}|)$ with $\alpha = 1$ between the jth and kth visits of the ith subject. Vector $\boldsymbol{X} = (\boldsymbol{X}_1^*,\ \boldsymbol{X}_2^*)$ are the mapping of $(\boldsymbol{X}_1,\ \boldsymbol{X}_2)$ in

the new feature space corresponding to a polynomial kernel with degree 2, for example, the inner product $< u^*, v^* >$ in the feature space equals $K(u, v)$ in the original space, where K is a polynomial kernel with degree 2. In the Supplementary Materials Figure A2 we demonstrated a typical set of $\boldsymbol{X}$ when its dimension is 2. The boundary for the two groups is nonlinear in the original space (top panel), while in the new three-dimensional feature space the boundary becomes a separating plane which is linear (bottom panel). Markers $\boldsymbol{W}_i$ is a time-invariant three-dimensional vector randomly located either on the outer sphere with a radius equal to 2 or on the inner sphere with a radius equal to 1 (with equal probability, and each radius has a small random error) (Supplementary Materials Figure A1). A single radial kernel SVM can generate a sphere-shaped boundary and perfectly separate the two groups of $\boldsymbol{W}$'s. Therefore, the corresponding oracle kernels to use for the fixed effects in this setting are a linear kernel for T, a linear kernel for $\boldsymbol{Z}$, a polynomial kernel with degree 2 for $\boldsymbol{X}$, and a radial kernel for $\boldsymbol{W}$.

Subject-specific latent effects $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ are subject-specific random effects. $\boldsymbol{a_i}$ is generated from $MVN(\mathbf{0}, \Sigma_a)$, where Σ_a is a correlation matrix with compound-symmetric structure ($\rho = 0.5$), and $\boldsymbol{b_i}$ is generated from $MVN(\mathbf{0}, \Sigma_b)$, where Σ_b is a correlation matrix with exponential correlation structure, for example, $\rho_{j,k} = \exp(-\alpha|t_j - t_k|)$ with $\alpha = 1$. ϵ_{ij} are normally distributed random errors of the i^{th} subject at the j^{th} visit.

We conducted two types of prediction for different purposes. In type A prediction we generated samples with a size of $n = 500$ subjects, each having 4 visits. Two-thirds of the subjects are included in the training set and the remaining one-third as the testing set. In all the simulations here, we used cross-validation to choose parameter ρ in fitting the long term random effects kernel introduced in Section 2.3 by grid search. It takes about 60 minutes to run one simulation round of the multiple kernel SVM (a sample size of 500, 4 visit time points and 6 kernels for different data sources) with fivefold CV to select from 4 candidate values for a parameter on a workstation (Intel Xeon 2.30 Ghz CPU). We present the results in Figure 1. In the top panel we compared a total of six methods, a logistic regression, a generalized mixed effects regression, and four different SVMs, including a fixed-effects single radial kernel SVM (concatenate all feature variables in a single radial kernel; "fixed-effects" refers to ignoring random effects in both model fitting and prediction), a fixed-effects multiple radial kernel SVM (one separate radial kernel for each group of variables from one source), a fixed-effects multiple oracle kernel SVM, and a mixed-effects multiple oracle kernel SVM ("mixed effects" refers to including kernels for random effects in both model fitting and prediction). In this case, the logistic regression and the generalized mixed effects regression perform substantially worse than the SVM based methods in terms of all fit indices: accuracy (1-misclassification rate), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). In addition, the variability of the former two approaches are much larger than the latter, indicating that the SVM based methods provide more stable predictions.

When comparing the four SVM-based approaches, the single radial kernel SVM performs the worst (results for the single linear or polynomial kernel are even worse than using radial kernel, so they are not shown here), indicating the advantage for using separate kernels for fixed effects when data are heterogenous. Using multiple oracle kernels greatly improves the performance comparing to using multiple radial kernels (same type of kernels), which confirms the importance of using appropriate kernels for data from different sources. When comparing the performance of fixed-effects versus mixed-effects multiple oracle kernel SVM, we see that including kernels for random effects reduces variability for all the fit indices and improves or maintains their mean values. A paired t-test comparing fixed-effects versus mixed-effects multiple oracle kernel SVM shows a significant decrease in misclassification rate ($p < 0.001$).

In type B prediction we generated samples with a size of $n = 500$ subjects, each having 6 visits. The first 3 visits of each subject are used as the training set and the rest 3 visits as the testing set. We predicted the subject-specific outcomes for the last 3 visits for each subject. The bottom panel of Figure 1 compares the performance of multiple oracle kernel SVM with or without random effects to logistic regression and generalized mixed effects regression. A paired t-test comparing fixed-effects versus mixed-effects multiple oracle kernel SVM shows a significant decrease in misclassification rate ($p < 0.001$). The magnitude of improvement is greater than that in type A prediction, suggesting that the developed method is more powerful when predicting subject-specific outcomes when some outcomes on the prior visits of the same subject are available.

In order to examine the effect of kernel function misspecification, we conducted two sensitivity analyses with mild and moderate misspecification of $\mathbf{K_b}$. For these sensitivity analyses, to save computational burden, we adopted a practical approach where we used cross-validation to select ρ in $\mathbf{K_a}$ in the first few replications and fixed their values at the chosen ones for other replications. Based on our experience for simulations in Figure 1, the results do no differ substantially between the practical approach (results not shown) and full scale cross-validation (identical up to 2 decimal places). Simulations with full scale cross-validation are expected to be similar or better. The short-term time-invariant random effects were generated to follow an AR-1 structure

$$\mathbf{K_{b_i}} = \begin{pmatrix} 1 & \rho_b & \dots & \rho_b^{n_i-1} \\ \rho_b & 1 & \dots & \rho_b^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}_{n_i \times n_i},$$

while when fitting the model we misspecify the kernel matrix as the exponential structure for $\mathbf{K_b}$ in Section 2.3. When the autocorrelation parameter for AR-1 structure $\rho_b = 0.85$, $\mathbf{K_b}$ is more similar to that under exponential structure for the model fitting, which we considered as a mild misspecification. When $\rho_b = 0.5$, the difference in $\mathbf{K_b}$ between AR-1 and exponential structure is larger, which we considered as a moderate misspecification. We compared 4 SVM models for type A prediction: a fixed-effects multiple radial kernel SVM, a fixed-effects multiple fused kernel SVM ("fused" means all the kernels in the model were correctly specified except for $\mathbf{K_b}$, if random-effects included), a population-mean level mul-

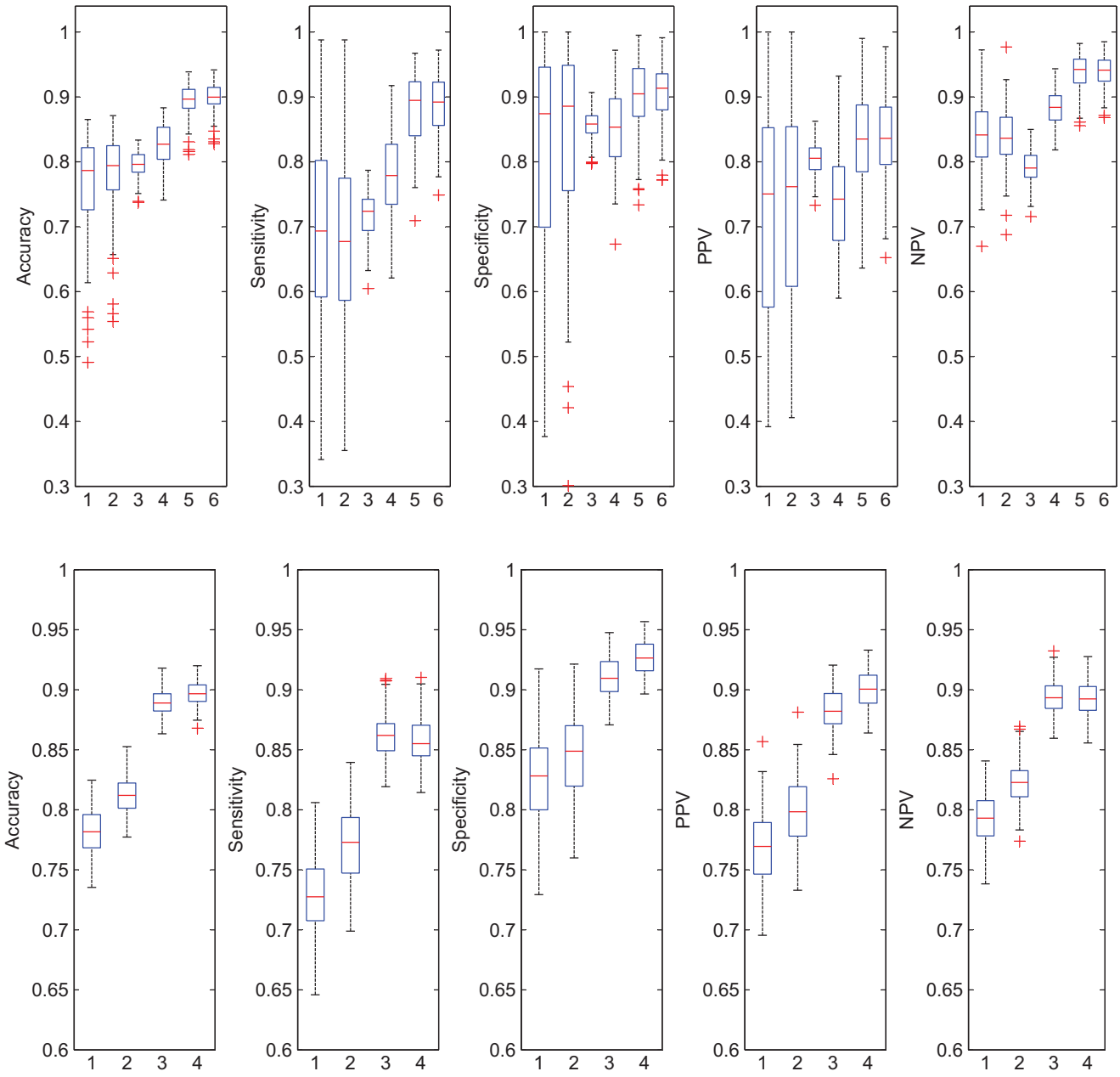

Figure 1. Simulation setting 2 (multiple data sources). Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2-generalized mixed effects regression, 3-single radial kernel SVM (fixed-effects), 4-multiple radial kernel SVM (fixed-effects), 5-multiple oracle kernel SVM (fixed-effects), 6-multiple oracle kernel SVM (mixed-effects). Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-logistic regression, 2-generalized mixed effects regression, 3-multiple oracle kernel SVM (fixed-effects), 4-multiple oracle kernel SVM (mixed-effects).

tiple fused kernel SVM ("population-mean level" where we included kernels for random effects in only model fitting but not the prediction), and a mixed-effects multiple fused kernel SVM. We compared 2 SVM models for type B prediction (a fixed-effects and a mixed-effects multiple fused kernel SVM). The results are shown in Figures 2 and 3. We can see that the results under both mild and moderate misspecification are pretty similar to those without misspecification, indicating that the performance (especially type A prediction) is not sensitive to the choice of kernel function as long as the tuning parameters are chosen in a data-driven way.

4. Application to ADNI Data

Data used in the preparation of this article were obtained from the Alzheimer's disease neuroimaging initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging , the National Institute of Biomedical Imaging and Bioengineering, the Food and

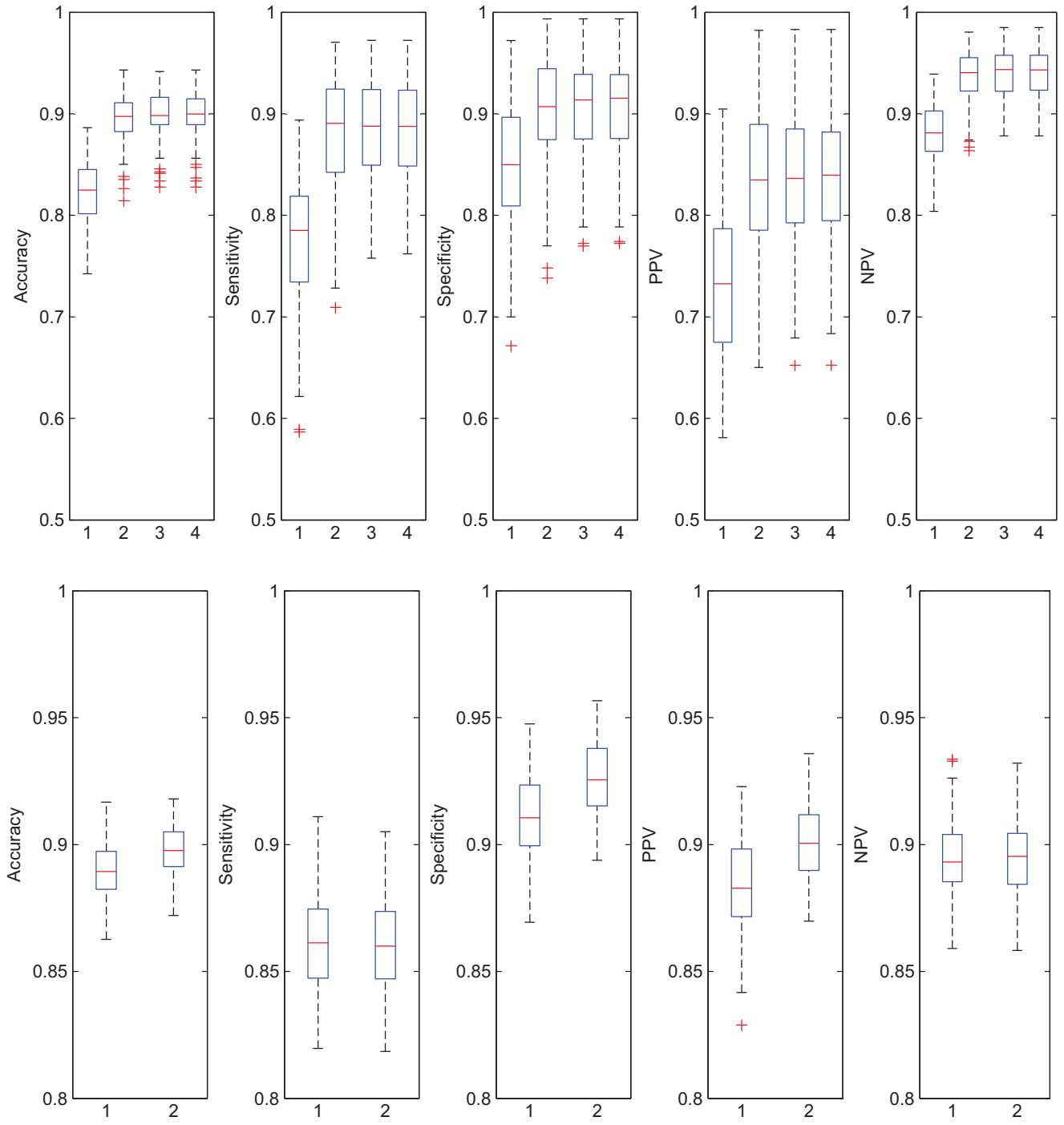

Figure 2. Sensitivity analysis (multiple data sources): mild misspecification for random effects. Top panel presents type A prediction of new subjects (left to right): 1-multiple radial kernel SVM (fixed-effects), 2-multiple fused kernel SVM (fixed-effects), 3-multiple fused kernel SVM (pop. mean), 4-multiple fused kernel SVM (mixed-effects). Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-multiple fused kernel SVM (fixed-effects), 2-multiple fused kernel SVM (mixed-effects).

Drug Administration (FDA), private pharmaceutical companies and non-profit organizations. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

In 2009, efforts to integrate genetic research related to ADNI biomarkers were planned and carried out to assess genes beyond ApoE, the largest known genetic risk factor for

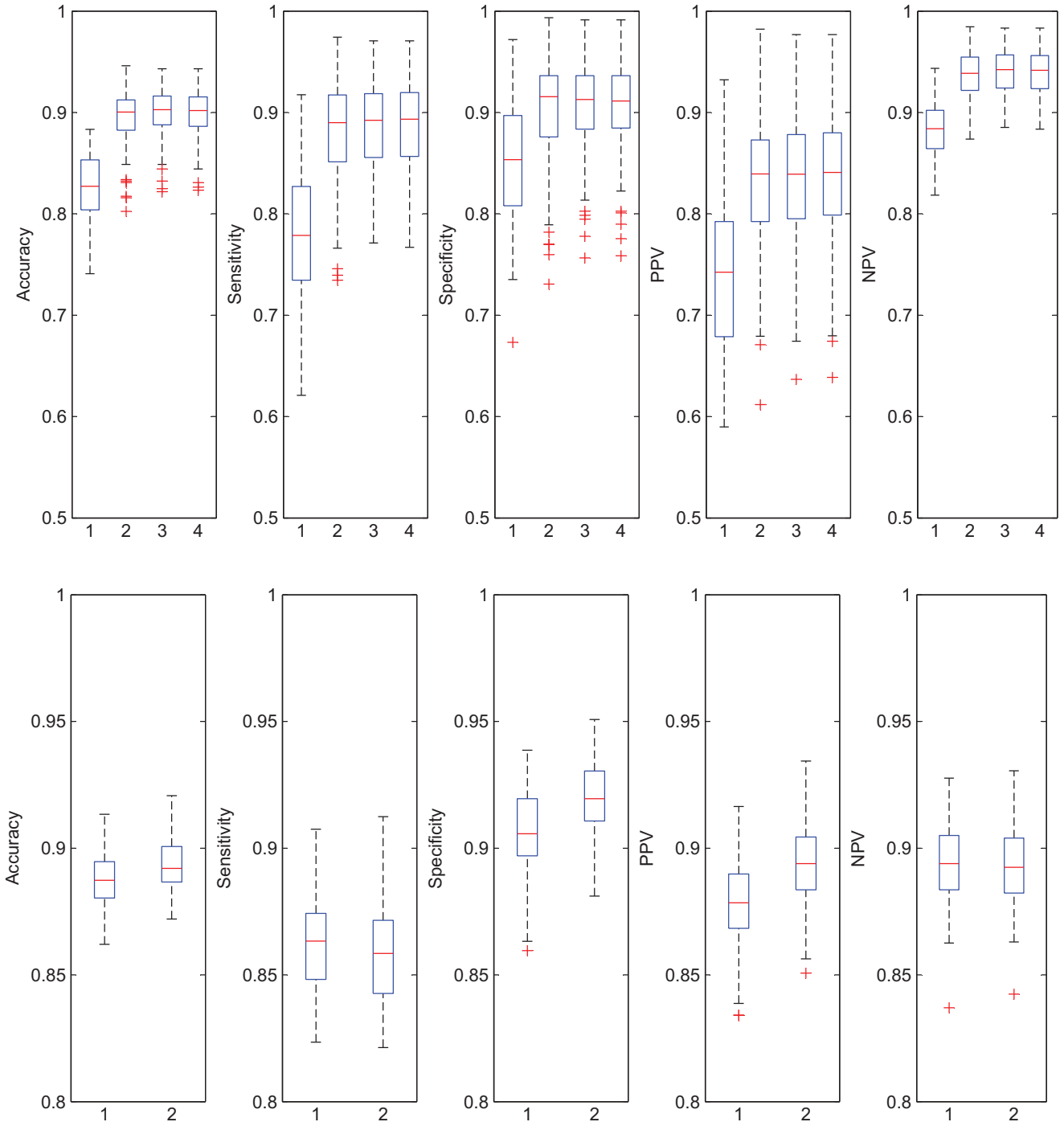

Figure 3. Sensitivity analysis (multiple data sources): moderate misspecification for random effects. Top panel presents type A prediction of new subjects (left to right): 1-multiple radial kernel SVM (fixed-effects), 2-multiple fused kernel SVM (fixed-effects), 3-multiple fused kernel SVM (pop. mean), 4-multiple fused kernel SVM (mixed-effects). Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-multiple fused kernel SVM (fixed-effects), 2-multiple fused kernel SVM (mixed-effects).

AD (Ashford, 2004). Since then, genetic and imaging data are available to contribute to the understanding of biological etiology of AD and MCI. The proposed multiple kernel framework exploits this unique opportunity to combine imaging and genetic data to predict the progression of MCI and early AD. Previous studies showed that some imaging biomarkers are important in predicting conversion from MCI to AD and early AD progression (Devanand et al., 2008; Hampel et al., 2008). It is conceivable that imaging variables are more correlated with each other than with genetic markers. If both types of

data are concatenated in a single kernel, for instance, a polynomial kernel, unnecessary polynomial correlation will be imposed between imaging and genetic markers. In a multiple kernel learning with separate kernels, however, such correlation is reduced, avoiding overfitting and unwanted complexity. In our framework, one could use existing kernels designed for imaging data and genetic data separately. Such analyses has not been reported in ADNI literature before.

Our analysis goal is to distinguish the subjects who have MCI and the subjects who have dementia using demographic, clinical, imaging, and genetic markers. Our further inclusion criteria of samples were: subject's disease status being MCI or dementia, having 4 or more follow-up records, and having complete imaging and genetic data. The sample used in our analysis contains 213 participants from all 3 phases with 1055 longitudinal follow-up records. The key data were merged from various case report forms and biomarker lab measures across the ADNI protocols (http://www.adni-info.org/).

The feature variables we used include demographic variables (age, gender, and education level), clinical variables (clinical dementia rating sum of boxes scores, the Alzheimer's Disease Assessment Scale (11 and 13), mini-mental state examination, Rey Auditory Verbal Learning Test and functional assessment questionnaire), imaging markers (volume measures of ventricles, hippocampus, entorhinal cortex, and intra-cranial volume), and genetic markers (ApoE4 and 16 SNPs on the PICALM gene). The PICALM gene was reported to be a causal gene for AD (Harold et al., 2009), and therefore the SNPs in this gene were included in our analyses. We used four separate kernels for each source of variables in the multiple fused kernel SVM: a polynomial kernel with degree two for age at each visit, a radial kernel for demographic variables and clinical variables, a linear kernel for imaging variables, and an identity-by-state (IBS) kernel for genetic markers. The IBS kernel is specially designed to measure the similarity between two subjects' SNPs based on their identity by state information and has been proven to be useful in genome-wide association studies (Wu et al., 2010). The other kernel types were selected by small scale cross-validation. The kernels for the short-term and long-term latent effects were specified as in Section 2.3, where α and ρ were selected by small scale cross-validation.

The top panel of Figure 4 summarized the results of a logistic regression, a fixed-effects single radial kernel SVM, a fixed-effects multiple fused kernel SVM, and a mixed-effects multiple fused kernel SVM for type A prediction. All the feature variables and the pairwise interaction terms for demographic and clinical variables were included in logistic regression. The performance of multiple kernel SVMs improves upon the logistic regression in terms of all the fit indices, and upon the single radial kernel SVM in terms of accuracy, specificity, and PPV. Sensitivity of the single kernel SVM is slightly better than multiple kernel SVMs. The inclusion of latent random effects to a multiple fused kernel SVM makes little difference in terms of type A prediction. The bottom panel of Figure 4 compares the fixed-effects and mixed-effects multiple fused kernel SVM for type B prediction. In this case, accounting for random effects in the multiple fused kernel SVM leads to a substantial gain in accuracy, sensitivity and NPV, which reflects the ability of using the latent random effects kernel matrix to extract correlated similarity information of the outcomes on the same subject (within-subject outcomes are often similar to some extent). In this example, the fixed-effects feature variables explained some proportion of variability while the latent effects improve prediction by extracting information from the unexplained variability in type B prediction. Specificity and PPV for the mixed-effects SVM is slightly lower, however, to a much lesser extent.

Another real data example based on PREDICT-HD study (Paulsen et al., 2008) can be found in Supplementary Materials Section A.2.

5. Discussion

In this work, we present new statistical learning methods for longitudinal data. While we adopted a MKL algorithm similar to Yu et al. (2010), one significant novelty of our approach is to construct kernel functions via latent random effects which can account for both long-term and short-term dependence. Conventional approaches for analyzing longitudinal data include generalized estimating equations which aim at estimating population average effects, and generalized linear mixed effects model regression which aim at estimating subject-specific effects. These regression-based methods focus on estimating the association between the outcome and the predictors, while the large margin-based statistical learning approaches directly focus on classification and prediction. We compared our methods to generalized mixed effects regression since our goal is the subject-specific prediction of disease status. Our proposed kernel-based learning method offers an effective alternative especially when the number of predictors is large and it can be easily scaled up. A key feature is to embed correlation of longitudinal observations into kernel matrices and take advantage of multiple kernel learning methodologies. With a single data source and a relatively small amount of predictors, the conventional approaches may perform adequately. However, when there are multiple heterogeneous data sources, the improvement of the proposed method is more evident. Making connections to multiple kernel learning allows the proposed method to enjoy easy integration of heterogeneous data sources to boost information while accounting for the longitudinal data structure. We have shown through our simulation and real data analyses that when prior scientific knowledge suggests distinct distribution of feature variables, treating each component with a separate appropriate kernel and then combining in an optimal way allows substantial information gain.

To account for the longitudinal feature of data, we discuss two types of novel prediction procedures here (type A and type B prediction) to utilize latent effects in the prediction. We show that by extracting information on the distributions of the random effects, we improve prediction both for future subjects and for future outcomes on the same subject given feature variables and past outcomes. However, for longitudinal studies, the type B problems are more commonly encountered in applications where the outcome at a follow-up visit for the same subject is desirable, and our learning method is more effective than ignoring correlation among observations. When the interest is on predicting outcomes for a new subject at the baseline time-point, conventional approaches may work

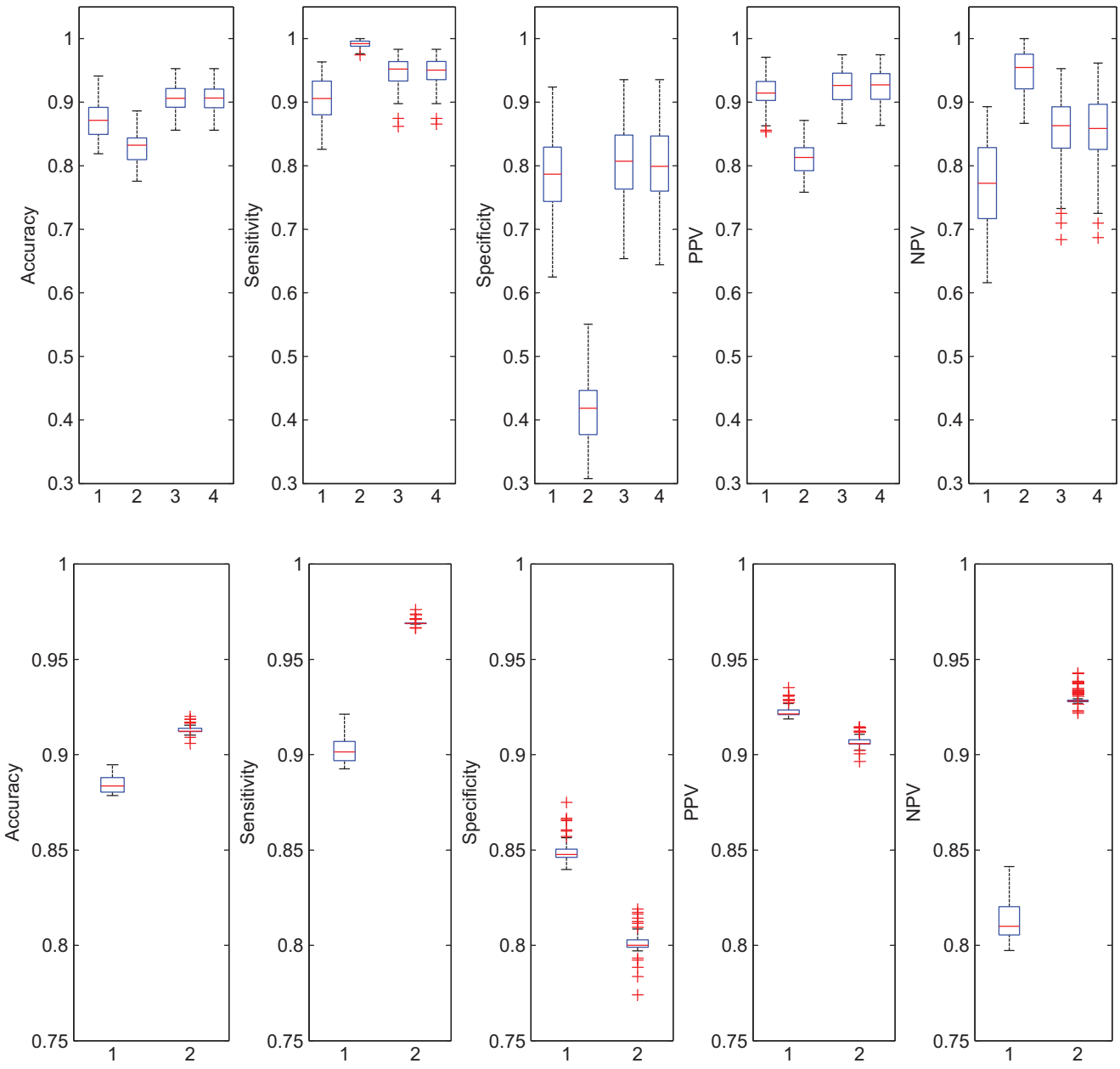

Figure 4. ADNI study. Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2-single radial kernel SVM (fixed-effects), 3-multiple fused kernel SVM (fixed-effects), 4-multiple fused kernel SVM (mixed-effects). Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-multiple fused kernel SVM (fixed-effects), 2-multiple fused kernel SVM (mixed-effects).

as well. The choice of covariance structure and the choice of appropriate kernel functions is related to the choice of the best representation of the kernel space. There is no consensus on these issues in the current literature which warrants future study on these matters.

We adopt the use of L_2-norm kernel fusion which leads to a non-sparse integration of multiple data sources, which may be more appealing in biomedical applications where it is believed there is no clear "winner" and each data modality contributes partial information to the prediction. Besides the L_2-norm on weights θ_p, other regularization, such as L_1-norm and L_∞-norm, can also be imposed in the kernel fusion. L_1-norm generates a sparse integration, which can be used for data source selection when the number of data sources is large and no prior information on which source is more predictive is available. L_∞-norm assigns the dominant weight parameter to only one kernel, which can be used when there is the need for a unique data source competition.

Daemen and Moor (2009) proposed a kernel function for clinical variables which computes the rescaled similarity. Our proposed algorithm is different from Daemen and Moor (2009) in that their final kernel matrix is a simple average of individ-

ual kernels, while our algorithm finds the optimal weight for each kernel matrix in a data-driven way. The kernels for random effects we proposed are based on subject-specific latent effects so they capture the temporal similarity of the observations for the same subject, while Daemen and Moor (2009) did not handle longitudinal data.

In our proposed method, the decision function takes an additive structure of the feature variables and the latent effects. A natural extension will be to include the interactions between them in the prediction rule. The proposed algorithm can be easily modified to handle this issue through tensor products of kernel matrices. Here we do not assume a distribution for random effects, but uses kernel functions to capture correlation. Although in practice the optimal kernel types to use may be unknown, a pragmatic solution might be to consider several different combinations of kernel types and choose the one with the smallest misclassification rate. Lastly, the kernel matrices for $\mathbf{a}_i$ and $\mathbf{b}_i$ may be misspecified so that it will be interesting to further study the robustness of the prediction rule to the specification of these matrices.

6. Supplementary Materials

Appendices referenced in Sections 3 and 4 with additional simulation and real data analysis results as well as sample Matlab code are available at the *Biometrics* website on Wiley Online Library.

Acknowledgements

This work is supported NS073671, NS082062, P01CA142538, NUL1 RR025747, Alzheimer's Disease Neuroimaging Initiative (ADNI) (U01 AG024904, DOD ADNI, W81XWH-12-2-0012). Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The authors acknowledge the investigators within the ADNI who contributed to the design and implementation of ADNI. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. Additional acknowledgements are in Supplementary Materials.

References

Ashford, J. W. (2004). Apoe genotype effects on alzheimers disease onset and epidemiology. *Journal of Molecular Neuroscience* **23**, 157–165.

Bach, F. R. and Lanckriet, G. R. G. (2004). Multiple kernel learning, conic duality, and the smo algorithm. In *In Proceedings of the 21st International Conference on Machine Learning (ICML)*.

Daemen, A. and Moor, B. D. (2009). Development of a kernel function for clinical data. *31st Annual International Conference of the IEEE EMBS, September 2-6* pages 5913–5917.

Devanand, D. P., Liu, X., Tabert, M. H., Pradhaban, G., Cuasay, K., Bell, K., de Leon, M. J., Doty, R. L., Stern, Y., and Pelton, G. H. (2008). Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biological Psychiatry* **64**, 871–879.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford: Oxford University Press.

Hampel, H., Brgerb, K., Teipelb, S. J., Bokdea, A. L., Zetterbergc, H., and Blennow, K. (2008). Core candidate neurochemical and imaging biomarkers of alzheimers disease. *Dementia* **4**, 38–48.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., and Hamshere, M. L. (2009). Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. *Nature Genetics* **41**, 1088–1093.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **65**, 533–550.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* **41**, 495–502.

Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* **5**, 27–72.

Luts, J., Molenberghs, G., Verbeke, G., Van Huffel, S., and Suykens, J. A. (2012). A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis* **56**, 611–628.

Oquendo, M. A., Baca-Garcia, E., Artes-Rodriguez, A., Perez-Cruz, F., Galfalvy, H. C., Blasco-Fontecilla, H., and Duan, N. (2012). Machine learning and data mining: strategies for hypothesis generation. *Molecular Psychiatry* **17**, 956–959.

Orru, G., Pettersson-Yeoa, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioral Reviews* **36**, 1140–1152.

Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., et al. (2008). Detection of huntington's disease decades before diagnosis: the predict-hd study. *Journal of Neurology, Neurosurgery & Psychiatry* **79**, 874–880.

Pavlidis, P., Cai, J., Weston, J., and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *Journal of Computational Biology* **9**, 401–411.

Pearce, N. and Wand, M. (2009). Explicit connections between longitudinal data analysis and kernel machines. *Electronic Journal of Statistics* **3**, 797–823.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* **86**, 929–942.

Yu, S., Falck, T., Daemen, A., Tranchevent, L.-C., Suykens, J. A., Moor, B. D., and Moreau, Y. (2010). L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics* **11**, 309.

Zhang, D. and Shen, D. (2012). Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PloS one* **7**, e33182.

Received May 2014. *Revised March* 2015. *Accepted April* 2015.