Prediction of longitudinal evolution of Alzheimer's Disease

Cagla Deniz Bahadir, Meenakshi Khosla, Yufu Mo (Alzheimer Police)

1 Introduction

In the TADPOLE project, we are provided with multi-modal data comprising PET, FMRI measures etc. for patients at high risk of developing Alzheimers disease. The patients are observed over a period of several years, so it is a longitudinal setting. The goal is to predict certain biomarkers of the disease (Normalized Ventricle Volume, MMSE, ADAS13 scores) as well as the disease state (healthy, MCI or Alzheimers) at future time points using these historic measurements.

We use algorithms designed especially for the longitudinal setting in our prediction framework. Multi-modality data and longitudinal prediction problems have been dealt with extensively in clinical literature [6] [1]. Young and Modat [5] developed the multi-kernel SVM method to achieve better performance on both classification and regression problems for Alzheimers disease. An interesting approach was proposed in [2], who put a prior on random effects and built a mixed effect model to tackle clinical longitudinal problem with non-linear functions like sigmoid functions. We experiment with a similar approach for the TADPOLE dataset. Another algorithm used in practice for repeated measures data is Functional Principal Component Analysis [3]. We use a version of FPCA that does not require pre-smoothing of observations on our clinical variables [4] and discuss the results and limitations of our approaches.

1.1 Data description and Preprocessing

The input data comprises bio-markers for 1737 patients evaluated at an interval of 6 months. However, patients frequently missed follow-ups and even during each visit, not all of the imaging data was necessarily collected. Thus, there is significant variance in the length of observations of each patient. Further, for each patient, the recorded observations are often at irregular intervals because of missed appointments.

2 Coding environment

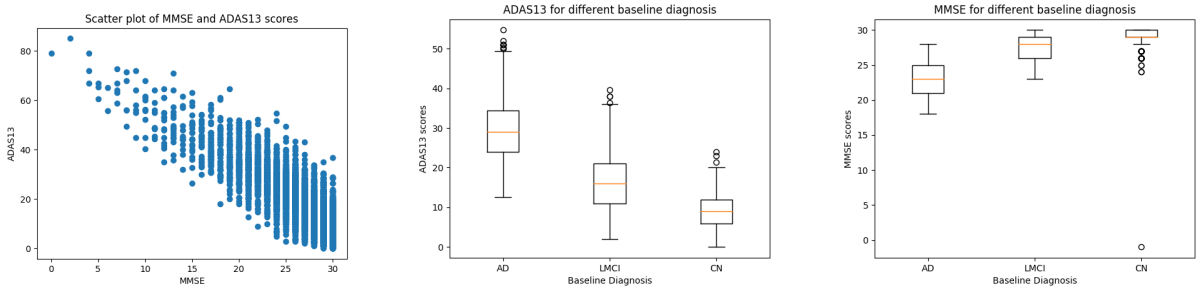
Python 3.5, MATLAB
Packages/toolbox used: NumPy, Pandas, scikit-learn, PACE, statsmodels.api, nlmm (MATLAB)

3 Visualization

As a preliminary step, we visualized the data to develop our prediction algorithms accordingly.

3.1 Correlations between prediction tasks

Figure 1 shows the strong correlations present between the variables to be predicted–the continuous variables (MMSE and ADAS13 scores) and categorical diagnosis variable (AD/MCI/CN) evaluated at baseline using our TADPOLE dataset as a preliminary analysis step. These strong correlations suggest that the clinical scores might be good predictors of the disease state. We exploit this fact later on in choosing the predictors in our classification algorithm.

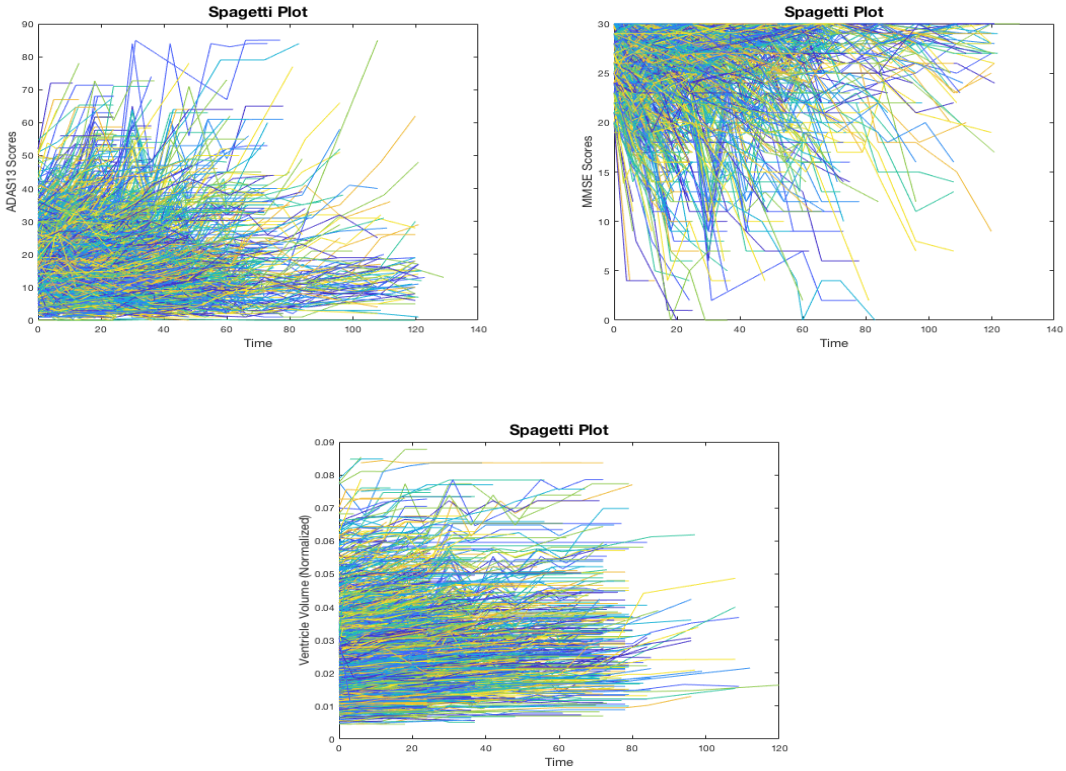


(a) (b) (c)

Figure 1: Mutual associations between the MMSE, ADAS13 scores and the diagnosis

3.2 Spaghetti Plots

Spaghetti plots are an important visualization tool for longitudinal data. They can help in visualizing within-subject (across time) and between-subject variability in the values of the variables of interest.



They are also very helpful in exploring the overall trends in the data. Looking at the plots, we were able to conclude that there is an overall tendency towards disease progression in the patients as time goes on, or equivalently as they age. This is evidenced by how the biomarker values changes across time for most patients. Ventricular enlargement has been shown to correlated with disease progression, and the spaghetti plots show that for most patients, the normalized ventricle volume actually increases with time. Similarly, higher values of ADAS13 and lower values of MMSE are correlated with advancement of the disease, which is the trend we also observe from the spaghetti plots.

Further, we can visually inspect the nature of the variability for different clinical variables. Normalized ventricles showed little variability across time for most subjects compared to high variability between

subjects. However, ADAS13 and MMSE scores showed both large within-subject and between-subject variability.

4 Baseline Algorithm

4.1 Implementation

The persistence model or the last-observation carry forward was implemented as the baseline algorithm. In this approach, we did not use the interpolated observation but instead used the last actually observed value of each clinical score for every patient.

4.2 Results

The performance of the persistence model was evaluated on the training and test data separately. Mean absolute error (MAE) and Mean squared error (MSE) were the performance metrics for the continuous regression variables. Classification accuracy, mAUC and balanced class accuracy are reported for the diagnosis classification.

- **Test Data**

	ADAS13	**MMSE**	**Ventricle Volume (Norm)**
MSE	42.663763	7.405498	0.000009
MAE	4.231707	1.721649	0.001767

Classification accuracy

	Total	**NL**	**MCI**	**Dementia**
Accuracy	0.808436	0.880342	0.725352	0.941176

Balanced Classification Accuracy:**0.882654**

- **Training Data**

	ADAS13	**MMSE**	**Ventricle Volume (Norm)**
MSE	49.193199	8.114596	0.000008
MAE	4.186921	1.716526	0.0017

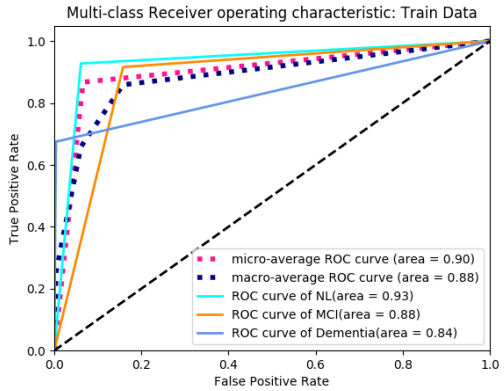
Classification accuracy

	Total	**NL**	**MCI**	**Dementia**
Accuracy	0.867403	0.890017	0.815657	0.979675

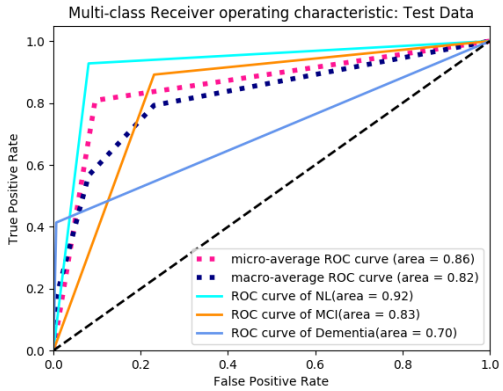
Balanced Classification Accuracy:**0.819235**

5 Proposed algorithms for Regression variables

The key observation in our TADPOLE data-set is the non-independence of data points. Because each subject contributes multiple time points to the data, these observations are not independent. Hence, to account for this, we propose to use mixed effects models to make our predictions. We can introduce patient-specific trends or random effects in our model. In LMMs, we adopt priors on random effects and they are assumed to be normally distributed. Further, we can also use Non-linear mixed effect models such as sigmoidal function which might be better suited for biological variables.



(a) ROC: Train



(b) ROC: Test

5.1 Linear Mixed effect models

The LMMs are designed especially for longitudinal setting. In our model, we have i=1,..N subjects and each subject i has n_i number of time-points or observations. The response in our LMM η_{ij} is the clinical variable we are trying to predict (ADAS13, MMSE, Ventricle Volume) for individual subject i at time-point j. The fixed effect is initially just time, which is also a random effect. This is because we trying to model both group-specific and individual-specific trends in our model. Mathematically, we can express the model as,

$$\eta_{ij} = x_{ij}^T \beta + z_{ij}^T \mu_i$$

Here, x_{ij} represents the fixed effect and β corresponds to the fixed effect parameter. z_{ij} is the vector of variables having random effects (column of ones added to include random intercepts) and the random effect μ_i is assumed to follow a normal distribution with zero and covariance structure given by Σ_v.

5.1.1 Results

The following table summarizes the Mean squared error (MSE) results obtained using a very simple linear mixed effect model with just time as a fixed effect. For the MMSE and Ventricle Volume, two random effects were included for each subject (both slope and intercept). The ADAS13 score showed worse performance with added slope effect and hence only the random effect of intercept was included in our preliminary analysis. For an initial comparison, the model was fit on the Input Data and performance was calculated on the training and test set separately. However, in future, we would develop the model on the Input and Training dataset and only use the test set for assessing the performance of our model.

- **Mean squared error**

	Training	**Testing**	**Validation**
ADAS13	48.33786	46.69634	53.2818
MMSE	7.160737	7.846385	6.4264
Ventricle Norm	8.93E-06	8.44E-06	6.31E-06

- **Mean absolute error**

	Training	**Testing**	**Validation**
ADAS13	4.7762	4.9011	4.9889
MMSE	1.768	1.8756	1.6724
Ventricle Norm	1.79E-03	1.62E-03	1.65E-03

As can be seen in the results above, some of these measures outperform the baseline results. For example, the MSE for MMSE in the training dataset predicted using our proposed algorithm outperforms the

corresponding results for the baseline algorithm. Further, the normalized ventricles volume in the validation dataset is better than the results obtained through baseline. However, this can be further improved and in the next section, we propose strategies for better predictions.

5.2 Non-Linear Mixed Models

The assumption of a linear trend for biological variables or clinical scores is often not accurate. Sigmoidal functions might be more suited to model the temporal progression of such variables, assuming that the values tend to flatten out or reach an asymptote for large times. Therefore, our next approach was to exploit non-linear mixed models for making predictions at future time points. In this approach, our random effects included the parameters of the sigmoidal function: the asymptotic response value, the mid-point (or inflection point) and the scale. So instead of including the fixed effect of time as a linear term, we instead used the following function for time,

$$f(x_{ij}) = \frac{\phi_1}{1 + exp(-\frac{x_{ij} - \phi_2}{\phi_3})}$$

where ϕ_1, ϕ_2, ϕ_3 are the fixed effect parameters of our model and a subset or all of them would also be included as random effects.

5.2.1 Maximum Likelihood

We estimate the parameters of a non-linear mixed-effect model by maximizing a likelihood function,

$$p(y|\theta, \sigma^2, \Psi) = \int p(y|\theta, \sigma^2, \eta) p(\eta|\Psi) d\eta$$

where y is the response data, θ is the vector of fixed effects, σ^2 is the error variance, Ψ is the covariance matrix for random effects, and η is the vector of unobserved random effects. $p(y|\theta, \sigma^2, \Psi)$ is the marginal density of y, $p(y|\theta, \sigma^2, \eta)$ is the conditional density of y given the random effects η, and the prior distribution of η is $p(\eta|\Psi)$. This integral contains a non-linear function of the fixed effects and variance parameters needed to maximize. For non-linear models, the integral does not have a closed form, and needs to be solved numerically, which involves simulating the function at each time step of an optimization algorithm. Therefore, the estimation can take a long time for complex models, and initial values of parameters might play an important role for successful convergence.

5.2.2 Hand-picked features

The fixed covariates we plan to use subsequently for experimenting with our model would be a subset of the handpicked predictive features listed on the TADPOLE website. These include age, APOE4,RAVLTimmediate, Hippocampus, WholeBrain, Entorhinal, MidTemp volumes, FDG, AV45 etc. Further, since some of these are themselves a function of time, we would try two different approaches: (1) Use their baseline value as constant features for each subject (2) Model them through mixed effect models and predict their values at future time points concurrently using a multi dimensional output or response variable.

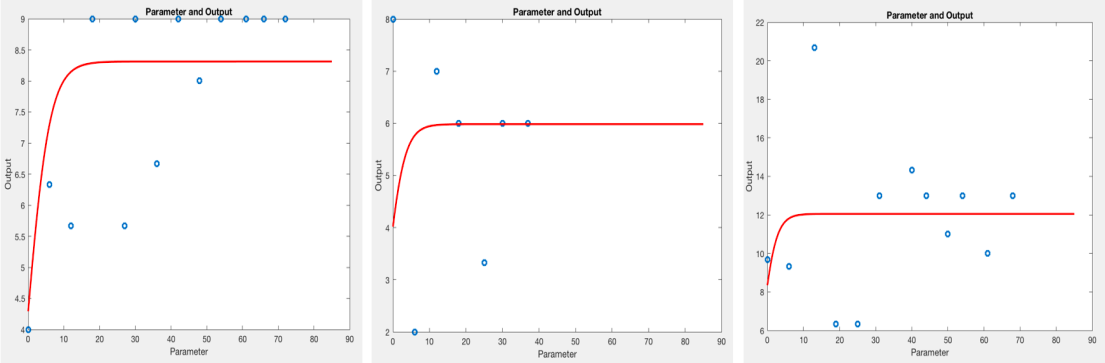
For initial analysis, we included Age and APOE4 as fixed effects. Thus, our non-linear mixed effect models was as follows,

$$f(x_{ij}) = \frac{\phi_{i1} + \phi_4 Age + \phi_5 APOE4}{1 + exp(-\frac{x_{ij} - \phi_{i2}}{\phi_{i3}})}$$

5.2.3 Results

The NLME model had several issues with convergence. We present the results for the ADAS13 score, for which the model coverged to a stable solution with the chosen initialization.

Sample plots obtained by applying non-linear mixed effect models on the ADAS13 score are shown in the following figure,



Results for ADAS13:

- Simple Model

	Test	**Validation**
MAE	4.7833	4.9261

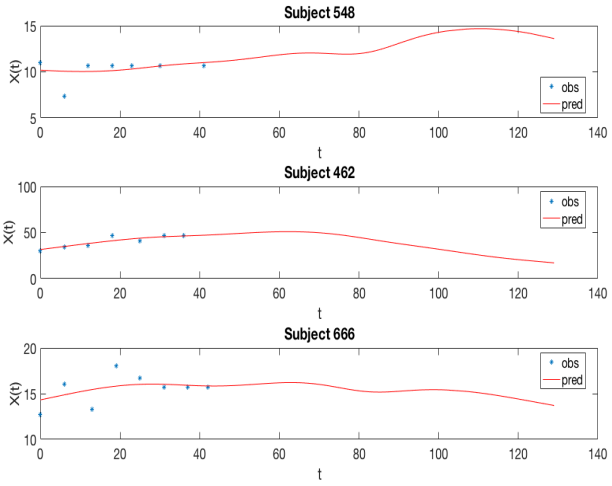
- Added fixed effects

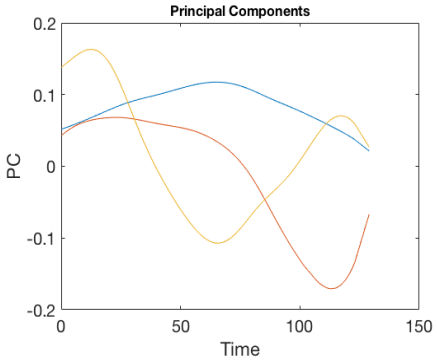
	Test	**Validation**
MAE	5.8772	5.4321

5.3 Functional Principal Component Analysis

We implement another algorithm for improving our results and smoothing out noisy observations: Functional PCA. It works very similar to multivariate PCA, except for the fact it is designed especially for functional data. Functional data refers to data where each observation is a curve instead of a point or finite dimensional vector and so the data is highly correlated and the collection of points display a certain smoothness property. Thus, longitudinal data essentially comes in the category of functional data because each patient displays a curve (for each clinical score). However, whatever we observe are discrete realizations from this curve for all subjects. The goal of FPCA is to find dominant modes of variation in the data and express each curve (corresponding to each subject) as a linear combinations of certain eigenfunctions (basis curves).

For implementing FPCA, we used the toolbox PACE - Principal Analysis by Conditional Estimation. Since we have very sparse observations for each subject, this algorithm is more suited because it doesn't require obtaining smooth estimates of the curve before identifying principal components. The number of principal components are chosen to explain a fixed proportion of variance in the data (set to 0.9). Sample fitted FPCA curves are shown in the following figure.





5.3.1 Results

Overall, Functional PCA showed improved results for MMSE and ADAS13. However, linear mixed-effect models significantly outperformed functional FPCA for the normalized Ventricles volumes. The results for FPCA are summarized in the table below.

- Validation Results

	ADAS13	**MMSE**	**Ventricle Volume (Norm)**
MAE	4.70693577	1.77283048	0.00554082

- Test Results

	ADAS13	**MMSE**	**Ventricle Volume (Norm)**
MAE	4.3155178	1.7578174	0.0049772

6 Classification

6.1 Implementation of Decision Trees

The TADPOLE project includes designing a classification algorithm alongside the regression models to be able to predict the disease state of the patients in their future visits. As explained in the previous sections the regression values show correlations within each other and with the disease states. Therefore, our classification model utilizes these relations and the previous disease states as inputs.

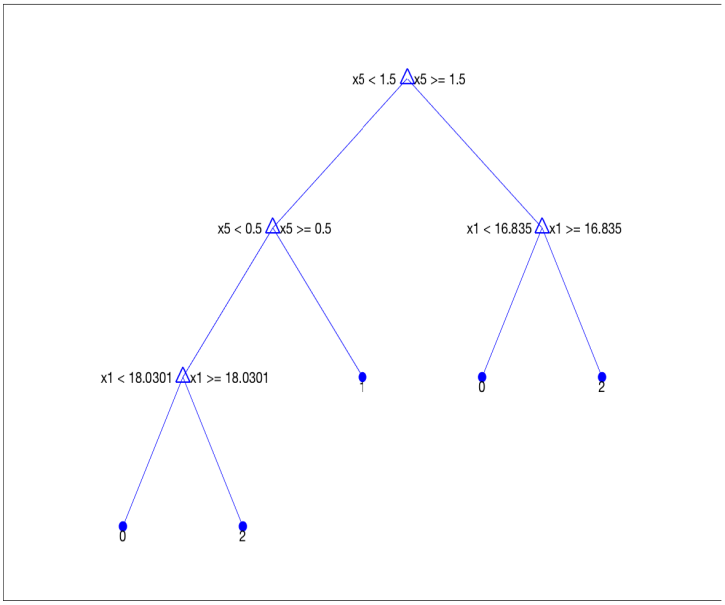
Since the visualizations showed different correlation trends between the regression values, the first approach to the classification problem was using k-means clustering. The algorithm presented several advantages such as being very straightforward to implement and requiring a significantly short training time. Although several distance measures were used (eg. Square Eucledian, cosine and correlation etc.), the values were not linearly separable in 2D or 3D space (combinations of several variables). This required using kernel functions, which had to be manually evaluated.

The non-promising results with the k-means clustering algorithm and the strong correlation between the last disease state and the future disease states (referencing to the accuracy in the persistence baseline algorithm) created a need for an algorithm that could use the classification parameters with the regression parameters. Therefore, we decided to implement Decision Trees. The Decision Tree algorithm was also straightforward to implement and required very little training time.

The algorithm was implemented on MATLAB, trained on the input data with ten folds cross-validation method and evaluated on the test and the validation data (Appendix section: Classification). During training, only the last observed regression values: MMSE, ADAS13, Ventricles, Age and the previous disease state were utilized for each patient for classifying the last observed disease state. The depth of the Decision Tree was limited to four splits, which gave better results than more complex structures due to increased flexibility and mitigated over-fitting problem.

6.2 Results

The resultant tree with four splits is shown in the following figure. The Decision Tree puts the most emphasis on the ADAS13 score and the last disease state which are depicted as x1 and x5.



	Test Data	Validation Data
Accuracy for all data	81 percent	72 percent
Maximum accuracy bu disease state	90 percent (MCI)	86 percent (AD)

As previously mentioned the Decision Tree was evaluated for the test and the validation datasets. The results are shown in the Table. The Tree gives 81 percent accuracy on the test dataset with the highest accuracy of 90 percent in detecting MCI patients. The accuracy for the validation data is 72 percent and the highest accuracy of 86 percent is in detecting AD patients. The results prove that the Decision Tree is a suitable classification model for the TADPOLE dataset.

7 Limitations

The aforementioned algorithms couldn't produce a very significant improvement from the baseline. This might have occurred due to the following limitations of our approaches:

- Linear models are not the best models for biological variables as they can be inaccurate if we record data for patients over long intervals. One should expect the values for these variables to saturate over time, a phenomena that is not captured by linear models.
- Finding the optimal solution for non-linear models is difficult because it is very sensitive to the initialization. Better initializations can help in avoiding local maxima and improve convergence, however the approach overall is brute-force.
- Functional PCA works well but representing each subject as a linear combination of a very few eigenfunctions might be tricky because there is a great variability in trends between subjects.
- Decision Trees used the forecasts of the clinical variables to predict the disease state. However, these themselves are not completely accurate and hence could significantly impact the performance of the classification algorithm.

8 Contribution of individual members

All team members contributed equally to the pre-processing, baseline implementation and visualization. Cagla worked on classification of disease state whereas Meenakshi and Yufu worked on algorithms for prediction of continuous/regression variables at future time points.

9 Acknowledgements

We thank Professor Mert Sabuncu for his useful suggestions and comments throughout the course of the project.

References

[1] Robert D. Gibbons, Donald Hedeker, and Stephen DuToit. Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, 6(1):79–107, 2010. PMID: 20192796.

[2] Mary J. Lindstrom and Douglas M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687, 1990.

[3] F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *ArXiv Mathematics e-prints*, March 2006.

[4] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[5] Jonathan Young, Marc Modat, Manuel J. Cardoso, Alex Mendelson, Dave Cash, and Sebastien Ourselin. Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2(Supplement C):735 – 745, 2013.

[6] Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, and Jieping Ye. Modeling disease progression via multi-task learning. *NeuroImage*, 78(Supplement C):233 – 248, 2013.