A Comparison of Strategies for Incorporating Nuisance Variables into Predictive Neuroimaging Models

Conference Paper · June 2015

DOI: 10.1109/PRNI.2015.28

8 authors, including:

Liana Catarina Lima Portugal
Universidade Federal Fluminense
20 PUBLICATIONS **269** CITATIONS
SEE PROFILE

Orlando Fernandes Jr
Federal University of Rio de Janeiro
6 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Leticia Oliveira
Universidade Federal Fluminense
94 PUBLICATIONS **1,397** CITATIONS
SEE PROFILE

Mirtes Garcia Pereira
Universidade Federal Fluminense
72 PUBLICATIONS **972** CITATIONS
SEE PROFILE

A Comparison of Strategies for Incorporating Nuisance Variables into Predictive Neuroimaging Models

Anil Rao*, Joao Monteiro *, John Ashburner†, Liana Portugal‡, Orlando Fernandes Jr‡,
Leticia De Oliveira‡, Mirtes Pereira‡, and Janaina Mourao-Miranda*
*Department of Computer Science, University College London, London, U.K. Email: a.rao@ucl.ac.uk
†Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London
‡Federal Fluminense University, Brazil

***Abstract*—In this paper we compare two different methods for dealing with so-called nuisance variables (NV) when training models to predict clinical/psychometric scales from neuroimaging data. In the first approach, the NV are used to adjust the imaging data by 'regressing out' their contribution to the image features. In the second approach, the NV are included as additional predictors in the model with a separate kernel that controls their contribution to the prediction function. We evaluate these methods using data from an fMRI and a structural MRI study, and discuss the results and interpretation of the two modelling approaches.**

***Keywords*-nuisance variables; gaussian processes ; regression;**

I. Introduction

There has been substantial interest in recent years in using multivariate regression models to predict clinical and psychometric scales from neuroimaging MRI [1]. There remains, however, some uncertainty as to how best to incorporate so-called 'nuisance variables' (NV) into the predictive models [2]. Such variables may include not only demographic variables such as age or gender but also those that indicate whether the subject is taking any medication.

One technique for accounting for the NV is to 'regress out' their contribution to the image data [3], [4]. Here, we fit a linear model for each image feature using the NV as predictors, and consider the residuals to be the image data after 'adjusting' for the NV. The NV-adjusted image data is then used as the input features in the predictive model. The aim is to remove variability in the image features associated with the NV, thereby improving predictions while also producing a model that can be interpreted as being driven solely by the image data. If the NV are associated with the clinical/psychometric scales, however, it is possible that the adjustment may remove information from the image data that is useful for prediction. As an alternative to using adjusted features, we can utilise the NV by including them as predictors along with the original image features during predictive modelling. The resulting model will then explicitly be a multivariate function of the image data and the NV, rather than one of the adjusted image data alone.

This paper explores these different approaches to accounting for NV in predictive modelling from neuroimaging data. To the best of our knowledge, there has been no earlier work directly comparing the effectiveness of the two methods. We apply both techniques to an fMRI study and a structural MRI study, and conclude with a discussion of the two approaches.

II. Materials

The first study consists of fMRI data from 34 volunteers (15 women) acquired on a 1.5-T Siemens (Magnetom Avanto) scanner. All subjects had no history of neurological or psychiatric illness. The study consisted of a simple detection task while the subjects viewed a set pictures displaying scenes of threat or neutral pictures. Three pictures of the same category were presented subsequently in a 15s block followed by 12s of rest period. The experimental session consisted of 28 blocks, pseudo randomized, divided into 4 runs. SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) was used for preprocessing, consisting of motion correction, spatial alignment of the functional images with the MNI template, and spatial smoothing with an 8-mm Gaussian filter (FWHM). For each subject and run a General Linear Model was used including the threat and the neutral pictures presentations as regressors. The contrast images between the threat and the control condition, considering all runs, provide the 158451 image features in the matrix $\mathbf{X}^{\mathrm{I}}$ used in the regression models for this dataset. The corresponding target variable $\mathbf{y}$ for prediction is the 'Negative Affect' (NA) which refers to the extent to which a person feels negative mood including anger, nervousness, fear, guilt, and sadness [5]. The NV we consider are those which are not pyschometric scales/assessments giving a list of 'Age', 'Gender', 'Educational Level', 'Weight', 'Sleep in hours', 'Medication'. All variables were continuous apart from the categorical variables 'Gender' (Male/Female), 'Educational Level' (High School/Undergraduate Student/Graduate/Postgraduate student), and 'Medication' (Not Taking/Taking). Each of these variables were coded as $0, 1 \ldots, k-1$ where k is the number of levels for the category. The NV provide the nuisance features $\mathbf{X}^{\mathrm{NV}}$ used in the regression models for this dataset. The univariate correlations and associated p-values of each of the NV with the target are shown in table I.

The second dataset consisted of the baseline T1-weighted MRI of 150 subjects from the longitudinal part of the

978-1-4673-7145-2/15 $31.00 © 2015 IEEE
DOI 10.1109/PRNI.2015.28

TABLE I
FMRI DATA, UNIVARIATE CORRELATIONS (P-VALUES) OF NUISANCE VARIABLES WITH NA

Age	Gender	Education	Weight	Sleep	Medication
-0.32 (0.07)	0.23 (0.20)	-0.33 (0.05)	-0.01 (0.96)	-0.08 (0.64)	0.15 (0.41)

TABLE II
STRUCTURAL MRI DATA, UNIVARIATE CORRELATIONS (P-VALUES) OF NUISANCE VARIABLES WITH MMSE

Age	Gender	Education	SES
-0.03 (0.73)	-0.20 (0.02)	0.20 (0.02)	-0.18 (0.03)

OASIS study (www. oasis-brains.org). Preprocessing was performed using SPM12b (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) and consisted of averaging session repeat scans, grey matter segmentation and group-wise registration using Dartel to a study-specific template. The aligned images were transformed to the 2mm MNI template and smoothed with a Gaussian kernel of 8mm FWHM. A mask was applied to select voxels that had a probability of being grey matter above 0.1, giving an initial set of images that provide the 170663 image features in the matrix $\mathbf{X}^{\mathrm{I}}$ used in the regression models for this dataset. Here, the target variable for prediction $\mathbf{y}$ is the 'Mini-Mental State Examination' (MMSE) which is commonly used to diagnose and assess dementia. An initial list of NV was obtained from the OASIS website at http://www.oasis-brains.org/pdf/oasis_longitudinal.csv. We firstly discard measures that are clinical scales or have zero-variance across the data. In addition, we remove measures related to brain volume/head size, as those were determined using a different preprocessing to the one in this work. This gives a final list of 'Age', 'Gender', 'No. of Years Education', 'Socioeconomic Status', and reduces $\mathbf{X}^{\mathrm{I}}$ to 142 subjects due to missing data. All variables were continuous apart from the categorical variables 'Gender' and 'Socioeconomic Status' (Upper/Upper-middle/Middle/Lower-middle/Lower) which were coded in the same fashion as for the fMRI study. These NV provide the nuisance features $\mathbf{X}^{\mathrm{NV}}$ used in the regression models for this dataset. The univariate correlations and associated p-values of each of the NV with the target are shown in table II.

III. METHODS

A. Gaussian Process Regression

We use Gaussian Process Regression (GPR) for the predictive modelling in our experiments. Gaussian Processes have recently gained popularity for building predictive neuroimaging models for regression and classification [6], [7], [8].

Given a set of n observations $\mathbf{x}_i \in \mathbb{R}^p$ with associated target variables y_i, Gaussian processes impose a multivariate Gaussian prior on a set of latent variables f_i, where the mean and covariance of the prior are functions of the inputs $\mathbf{x}_i$ [9]:

$$\begin{aligned} E(f_i) &= m(\mathbf{x}_i) \\ Cov(f_i, f_j) &= k(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{1}$$

We assume a zero mean function $m(\mathbf{x}_i) \equiv 0$ throughout this work. The targets y_i are related to the latent variables f_i through the likelihood function. We use a Gaussian Likelihood for all models:

$$P(y_i \mid f_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - f_i)^2}{2\sigma^2}} \tag{2}$$

where $\sigma > 0$ is the standard deviation of the noise. The posterior distribution of the target y_* of a test point $\mathbf{x}_*$, given the training data, then has the closed form

$$\begin{aligned} y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_* &\sim \mathcal{N}(\bar{y}_*, Var(y_*)) \text{ where} \\ \bar{y}_* &= \mathbf{k}_*(K + \sigma^2 I)^{-1}\mathbf{y} \\ Var(y_*) &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*(K + \sigma^2 I)^{-1}\mathbf{k}_*^T + \sigma^2 \end{aligned} \tag{3}$$

in which K is the $n \times n$ matrix of training set covariances and $\mathbf{k}_*$ is the n-dimensional row vector of test-training covariances. In practice the mean of the posterior $\bar{y}_*$ is taken to be the prediction of the target at test point $\mathbf{x}_*$.

The predictive distribution for y_* has hyperparameters consisting of the parameters $\boldsymbol{\theta}$ of the covariance function $k(\cdot, \cdot)$, and the likelihood parameter σ. We estimate these by maximising the log marginal likelihood $\log \mathcal{Z}$, which is the log probability of the training data given the model. This is given by:

$$\begin{aligned} \log \mathcal{Z} = &-\frac{1}{2}\mathbf{y}^T(K(\boldsymbol{\theta}) + \sigma^2 I)^{-1}\mathbf{y} \\ &-\frac{1}{2}\log\left|K(\boldsymbol{\theta}) + \sigma^2 I\right| - \frac{n}{2}\log 2\pi \end{aligned} \tag{4}$$

We now describe the different GPR models, ie. input features $\mathbf{x}_i$ and kernel functions $k(\cdot, \cdot)$, used in this work.

B. Models Tested

1) Images Only: The baseline model is one where only the image features are used for prediction, so each input $\mathbf{x}_i \equiv \mathbf{x}_i^{\mathrm{I}}$. In this case, we use a linear kernel plus bias for training and prediction:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\mathrm{I}} {\mathbf{x}_j^{\mathrm{I}}}^T}{l^2} + b^2 \tag{5}$$

2) Adjusted Images Only: We produce NV-adjusted images by fitting a linear model to each image feature in turn using the NV, and then subtracting the fitted values from the original images. This can be performed very efficiently for all image features simultaneously to give the set of NV-adjusted images $\mathbf{X}^{\mathrm{IA}}$:

$$\mathbf{X}^{\mathrm{IA}} = \mathbf{X}^{\mathrm{I}} - \mathbf{X}^{\widehat{\mathrm{NV}}}({\mathbf{X}^{\widehat{\mathrm{NV}}}}^T \mathbf{X}^{\widehat{\mathrm{NV}}})^{-1} {\mathbf{X}^{\widehat{\mathrm{NV}}}}^T \mathbf{X}^{\mathrm{I}} \tag{6}$$

where $\mathbf{X}^{\widehat{\mathrm{NV}}}$ is the matrix $\mathbf{X}^{\mathrm{NV}}$ augmented with a column of ones, and the ith row of $\mathbf{X}^{\mathrm{IA}}$ corresponds to a single adjusted image $\mathbf{x}_i^{\mathrm{IA}}$. The above corresponds exactly with the 'residual matrix-forming' framework described in [10]. Note that the image adjustment is performed using the images of all subjects rather than specific cohorts of the data as in [3], [4]. This is because we aim to explore the impact of a straightforward adjustment of the image data without using any additional hypotheses. GPR training and prediction is then performed

using the kernel described in section III-B1, using inputs $\mathbf{x}_i \equiv \mathbf{x}_i^{\text{IA}}$:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\text{IA}} \mathbf{x}_j^{\text{IA}\,T}}{l^2} + b^2 \tag{7}$$

3) Images & NV: The NV are incorporated into the predictive model by appending them to the image features, so that each input to the GPR is $\mathbf{x}_i \equiv [\mathbf{x}_i^{\text{I}}, \mathbf{x}_i^{\text{NV}}]$. We use a kernel that is the sum of the kernel in section III-B1 and a linear Automatic Relevance Determination (ARD) kernel applied to the NV only:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\text{I}} \mathbf{x}_j^{\text{I}\,T}}{l^2} + b^2 + \mathbf{x}_i^{\text{NV}} \Lambda^{\text{ARD}} \mathbf{x}_j^{\text{NV}\,T} \tag{8}$$

where Λ^{ARD} is a diagonal matrix with entries $\frac{1}{l_1^2}, \ldots, \frac{1}{l_{d_{NV}}^2}$ in which d_{NV} is the number of nuisance variables. The hyperparameters $l_1^2, \ldots, l_{d_{NV}}^2$ scale the NV so that their contribution to the kernel, and hence, the decision function, is controlled.

4) NV Only: Here we use only the NV as predictors in the model, so that each input is $\mathbf{x}_i \equiv \mathbf{x}_i^{\text{NV}}$. In this case we take the kernel to be the sum of a bias term and the ARD kernel described in section III-B3:

$$k(\mathbf{x}_i, \mathbf{x}_j) = b^2 + \mathbf{x}_i^{\text{NV}} \Lambda^{\text{ARD}} \mathbf{x}_j^{\text{NV}\,T} \tag{9}$$

C. Model Evaluation and Significance Testing

We evaluate the models described in section III-B for both datasets. In order to explore the effects of using different sets of NV during model building, we train GPRs using a minimal set of NV consisting of just 'Age' and 'Gender', in addition to the full sets described in section III. In what follows, we refer to the minimal and complete set of NV as NV_0 and NV_1, respectively.

Model performance was determined using leave-one-out cross validation for the fMRI dataset and 10-fold cross-validation for the structural MRI dataset. In each case all features were standardised to mean zero and unit variance using the training folds. We calculate the mean-squared error (MSE), the Pearson correlation coefficient ρ, and Kendall's coefficient τ of the target variables and the predictions in each case. Permutation tests with a total of 100 labellings were used to determine if the models were performing better than chance for these measures, and the p-values were corrected for multiple comparisons using the False Discovery Rate. In addition, the optimised Log Marginal Likelihood, $\log \mathcal{Z}$, was determined using the whole dataset for training, and we correct it for model complexity to enable model comparison using the finite sample size corrected Aikake Information Criterion and the Bayesian Information Criterion:

$$\begin{aligned} \log \mathcal{Z}_{\text{AIC}} &= \log \mathcal{Z} - p - \frac{p(p+1)}{n-p-1} \\ \log \mathcal{Z}_{\text{BIC}} &= \log \mathcal{Z} - \frac{p}{2} \log(n) \end{aligned} \tag{10}$$

Here n is the size of the dataset, and p is the number of model hyperparameters (equal to the number of kernel parameters plus 1 likelihood parameter). [1] Bayes factors, which describe how favoured one model is over another can then be determined using the ratio of $\log \mathcal{Z}_{\text{AIC}}$ or $\log \mathcal{Z}_{\text{BIC}}$ of the two models.

IV. Results and Discussion

Table III summarises the results for the fMRI study. We can see that the only models that achieve statistical significance for all the measures are those that include the original images: None of the models that use the adjusted images ('Adj. Im.' in the table) achieve significance for any of the measures. This is true even for the model that uses images adjusted by just 'Age' and 'Gender', neither of which is significantly correlated to the target. In addition, we cannot perform better than chance if we only use the NV as predictors. In comparison, the model that uses the image data & NV_1 performs best with regards to MSE, ρ, and τ, implying that the NV, when combined with the image data, have been useful as predictors of the target. The log of the BIC-Corrected Bayes factor comparing this model to the just NV_1 model is 23, which indicates that including the image data gives a much more likely model than NV_1 alone.

For illustration, figure 1 shows the unnormalised image weight vectors for the model using image data only, image data adjusted by NV_1, and image data & NV_1, when training using the complete dataset. The colour map is stretched to occupy the full range of the image vectors, which is given for each model. The model with adjusted images has a range that is relatively close to zero, which gives a prediction function that is closer to taking the mean value of the targets. This results in the poorer predictive accuracies when using the adjusted images.

TABLE III
PREDICTION OF NA FOR THE DIFFERENT MODELS. NV_0={AGE, GENDER}, NV_1={AGE, GENDER, EDUCATION, WEIGHT, SLEEP, MEDICATION }. * INDICATES BETTER PERFORMANCE THAN CHANCE, $p < 0.05$.

MODEL	MSE	ρ	τ	$\log \mathcal{Z}$	$\log \mathcal{Z}_{\text{AIC}}$	$\log \mathcal{Z}_{\text{BIC}}$
Images Only	27.15*	0.45*	0.30*	-85.16	-88.56	-90.45
Adj. Im., NV_0	38.00	0.03	0.13	-106.28	-109.68	-111.57
Adj. Im., NV_1	35.48	0.03	0.06	-107.02	-110.42	-112.31
Images & NV_0	28.33*	0.42*	0.27*	-85.02	-91.09	-93.84
Images & NV_1	22.41*	0.58*	0.44*	-81.75	-94.50	-97.62
NV_0 Only	35.62	0.02	0.03	-107.22	-111.91	-114.27
NV_1 Only	40.39	-0.09	-0.03	-106.70	-117.58	-120.81

Table IV gives the results for the structural MRI data. Here we see that all approaches perform better than chance across the measures when predicting the MMSE score. The log of the BIC-Corrected Bayes factor comparing the image data & NV_1 model to the NV_1 only model is 15, which indicates that including the image data gives a much more strongly favoured model than NV_1 alone. In contrast to the fMRI study, the models that use adjusted images perform relatively well even though the NV are correlated with the targets.

[1] We adopt the conservative approach of counting all ARD hyperparameters in the AIC, BIC calculations, rather than excluding those with a negligible contribution to the kernel function.

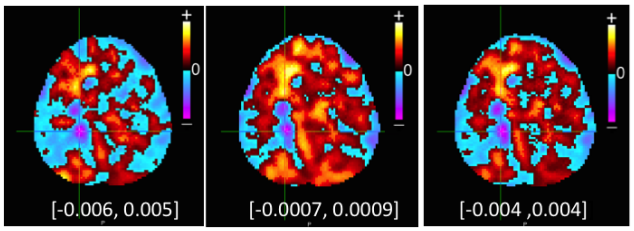

Fig. 1. This figure shows an axial slice through the weight vectors for the fMRI dataset, with the weight range given in brackets (a) Image data only, (b) Image data adjusted by all NV, (c) Image data and all NV

TABLE IV
PREDICTION OF MMSE FOR THE DIFFERENT MODELS. NV_0={AGE, GENDER}, NV_1={AGE, GENDER, EDUCATION, SES}. * INDICATES BETTER PERFORMANCE THAN CHANCE, $p < 0.05$.

MODEL	MSE	ρ	τ	$\log\mathcal{Z}$	$\log\mathcal{Z}_{\text{AIC}}$	$\log\mathcal{Z}_{\text{BIC}}$
Images Only	7.51*	0.43*	0.21*	-339.93	-343.02	-347.37
Adj. Im., NV_0	7.29*	0.43*	0.18*	-339.91	-343.00	-347.35
Adj. Im., NV_1	7.42*	0.41*	0.16*	-343.76	-346.85	-351.20
Images & NV_0	7.20*	0.46*	0.26*	-336.59	-341.81	-348.98
Images & NV_1	7.61*	0.45*	0.25*	-335.40	-342.82	-352.75
NV_0 Only	8.76*	0.12*	0.06*	-354.95	-359.10	-364.87
NV_1 Only	8.63*	0.18*	0.13*	-352.91	-359.23	-367.78

Given that gender is significantly correlated with the target, it is also important to consider the accuracies for each gender alone. We find that, for subjects with an MMSE of 30, the images adjusted by NV_0 model gives a smaller difference between accuracies for females and males (MSE: 5.64/4.57), than the images & NV_0 (MSE: 3.22/10.07) and images alone (MSE: 4.33/8.00) models. This implies that including the NV_0 as predictors appears to focus the model on predicting the females rather than the males for this value of MMSE. Figure 2 shows the unnormalised image weight vectors for the models using image data only, image data adjusted by NV_1, and image data & NV_1, when training using the complete dataset. We can see that there are small differences between the weight vectors of each of the approaches.

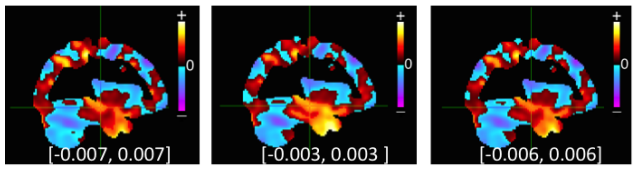

Fig. 2. This figure shows a sagittal slice through the weight vectors for the structural MRI dataset, with the weight range given in brackets (a) Image data only, (b) Image data adjusted by all NV, (c) Image data and all NV

V. CONCLUSION

In this work we have evaluated different approaches to incorporating nuisance variables into neuroimaging models for prediction of clinical/psychometric scales. We found that, while the use of adjusted images had little impact on predictive accuracy for the structural MRI dataset, it gave models that performed no better than chance for the fMRI data. In comparison, all approaches that incorporated the original imaging data and optionally included the nuisance variables, achieved statistical significance for all measures across both datasets. The variability in the results when using adjusted images suggests that the outcome of the procedure is quite data dependent, although it is possible that performing the adjustment in a different way would recover some predictive accuracy.

It is important to consider the differences between the motivation and interpretation of the models produced by each approach. Adjusting the images aims to, for example, transform the image features of subjects with different ages into a new set of features with the 'same' age. The resulting predictive model is then intended to be agnostic to the age of the subjects. In contrast, including age as a predictor aims to explain the variance in the target variable with multivariate relationships between it, the imaging data and age. The ultimate choice of approach may therefore depend not only on whether we wish to 'remove' the effects of a NV, but also on whether it is appropriate given the clinical/neuroscientific question we wish to answer. Future work will focus on further investigation of the properties of each approach.

ACKNOWLEDGMENT

The OASIS study was funded by grants P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584. AR and JMM were supported by the Wellcome Trust under grant no. WT102845/Z/13/Z.

REFERENCES

[1] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease," *NeuroImage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[2] M. R. G. Brown, G. S. Sidhu, R. Greiner, N. Asgarian, M. Bastani, P. H. Silverstone, A. J. Greenshaw, and S. M. Dursun, "ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements," *Frontiers in Systems Neuroscience*, vol. 6, no. September, pp. 1–22, 2012.

[3] J. Dukart, M. L. Schroeter, and K. Mueller, "Age correction in dementia–matching to a healthy brain." *PloS one*, vol. 6, no. 7, p. e22193, Jan. 2011.

[4] A. Abdulkadir, O. Ronneberger, S. J. Tabrizi, and S. Kloppel, "Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI," *Proceedings - 2014 International Workshop on Pattern Recognition in Neuroimaging, PRNI 2014*, pp. 1–4, 2014.

[5] L. a. Clark and D. Watson, "Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications." *Journal of abnormal psychology*, vol. 100, no. 3, pp. 316–336, 1991.

[6] A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen, and J. Mourão Miranda, "Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes." *NeuroImage*, vol. 49, no. 3, pp. 2178–89, Mar. 2010.

[7] O. M. Doyle, J. Ashburner, F. O. Zelaya, S. C. R. Williams, M. a. Mehta, and a. F. Marquand, "Multivariate decoding of brain images using ordinal regression." *NeuroImage*, vol. 81, pp. 347–57, Nov. 2013.

[8] J. Young, M. Modat, M. J. Cardoso, A. Mendelson, D. Cash, and S. Ourselin, "Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment," *NeuroImage: Clinical*, vol. 2, pp. 735–745, 2013.

[9] C. E. Rasmussen, *Gaussian processes for machine learning*. MIT Press, 2006.

[10] C. Chu, Y. Ni, G. Tan, C. J. Saunders, and J. Ashburner, "Kernel regression for fMRI pattern prediction," *NeuroImage*, vol. 56, no. 2, pp. 662–673, 2011.