Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease

Marco Lorenzi[a,b,*], Maurizio Filippone[c], Giovanni B. Frisoni[d,e], Daniel C. Alexander[f], Sebastien Ourselin[b], for the Alzheimer's Disease Neuroimaging Initiative[**]

[a] *Asclepios Research Project, Université Côte d'Azur, Inria, France*
[b] *Translational Imaging Group, Centre for Medical Image Computing, University College London, UK*
[c] *EURECOM, France*
[d] *Geneva Neuroscience Center, University Hospitals and University of Geneva, Switzerland*
[e] *IRCCS Fatebenefratelli, Brescia, Italy.*
[f] *POND group, Centre for Medical Image Computing, University College London, UK*

Abstract

Disease progression modeling (DPM) of Alzheimer's disease (AD) aims at revealing long term pathological trajectories from short term clinical data. Along with the ability of providing a data-driven description of the natural evolution of the pathology, DPM has the potential of representing a valuable clinical instrument for automatic diagnosis, by explicitly describing the biomarker transition from normal to pathological stages along the disease time axis. In this work we reformulated DPM within a probabilistic setting to quantify the diagnostic uncertainty of individual disease severity in an hypothetical clinical scenario, with respect to missing measurements, biomarkers, and follow-up information.

[*] Corresponding author

[**] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. 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

Email addresses: marco.lorenzi@inria.fr (Marco Lorenzi), maurizio.filippone@eurecom.fr (Maurizio Filippone), Giovanni.Frisoni@unige.ch (Giovanni B. Frisoni), d.alexander@ucl.ac.uk (Daniel C. Alexander), s.ourselin@ucl.ac.uk (Sebastien Ourselin)

We show that the staging provided by the model on 582 amyloid positive testing individuals has high face validity with respect to the clinical diagnosis. Using follow-up measurements largely reduces the prediction uncertainties, while the transition from normal to pathological stages is mostly associated with the increase of brain hypo-metabolism, temporal atrophy, and worsening of clinical scores. The proposed formulation of DPM provides a statistical reference for the accurate probabilistic assessment of the pathological stage of de-novo individuals, and represents a valuable instrument for quantifying the variability and the diagnostic value of biomarkers across disease stages.

Keywords: Alzheimer's disease, Diagnosis, Disease progression modeling, Gaussian process, Clinical trials

1. Introduction

Neurodegenerative disorders (NDDs), such as Alzheimer's disease (AD), are characterised by the progressive pathological alteration of the brain's biochemical processes and morphology, and ultimately lead to the irreversible impairment of cognitive functions [1]. The correct understanding of the relationship between the different pathological features is of paramount importance for improving the identification of pathological changes in patients, and for better treatment [2].

To this end, ongoing research efforts aim at developing precise models allowing optimal sets of measurements (and combinations of them) to uniquely identify pathological traits in patients. This problem requires the definition of optimal ways to integrate and jointly analyze the heterogeneous multi-modal information available to clinicians [3, 4, 5]. By consistently analyzing multiple biomarkers that to date have mostly been considered separately, we aim at providing a richer description of the pathological mechanisms and a better understanding of individual disease progressions.

Disease progression modeling (DPM) is a relatively new research direction for the study of NDD data [6, 7, 8, 9, 10, 11, 12, 13]. The main goal of DPM consists in revealing the natural history of a disorder from collections of imaging and clinical data by: 1) *quantifying* the dynamics of NDDs along with the related temporal relationship between different biomarkers, and 2) *staging* patients based on individual observations for diagnostic and interventional purposes. Therefore, this research domain is closely related to the exploitation of advanced statistical/machine-learning approaches for the joint modelling of the heterogeneous and information available to clinicians: imaging, biochemical, and clinical biomarkers. Differently from the several predictive machine-learning approaches proposed in the past in NDD research, disease progression models aim at explicitly estimating the temporal progression of the biomarkers from normal to pathological stages, to provide a better interpretation and understanding of the natural evolution of the pathology. For this reason it represents a very appealing modeling approach in clinical settings.

The main challenge addressed by DPM consists in the general lack a well-defined temporal reference in longitudinal clinical dataset of NDDs. Indeed, age or visit date information are biased time references for the individual longitudinal measurements, since the onset of the pathology may vary across individuals according to genetic and environmental factors [14]. This is a very specific methodological issue requiring the extension and generalization of the analysis approaches classically used in time-series analysis.

To tackle this problem, it is usually assumed that individual biomarkers are measured relatively to an underlying disease trajectory defined with respect to an absolute time axis describing the natural history of the pathology [7]. Each individual is thus characterized by a specific observation time that needs to be estimated in order to assess the individual pathological stage. According to this statistical setting, we therefore aim at estimating a *group-wise* disease model defined with respect to an absolute time scale, along with *individual* time re-parameterisation relative to the group-wise progression. This modeling paradigm has been implemented in a number of approaches proposed in the recent years, either by assuming continuous temporal trajectories of the biomarkers [7, 8, 9, 10, 11, 12, 13], or by modeling the disease progression as a sequence of discrete events [6, 15].

For example, in [8] the authors proposed to model the temporal biomarker trajectories through random effect regression, building on the theory of self-modeling regression [16], while the authors of [11] re-frame the random effect regression model in a geometrical setting, based on the assumption of a logistic curve shape for the average biomarker trajectories.

Continuous progression models have been recently extended to the modelling of brain images based on the time-reparameterization of voxel/mesh-based measures [9, 10, 13].

The use of disease progression models for diagnostic purposes is instead less investigated. Predictive models of patient staging were proposed within the setting of the Event Based Model [6], or still through random effect modeling [12]. However, the Event Based Model relies on the coarse binary discretization of the biomarker changes, and does not account for longitudinal observations, while the predictive models proposed in [12] and [17] require cohorts with known disease onset, and therefore lack flexibility while being prone to bias due to misdiagnosis and uncertainty of the conversion time.

Furthermore, these methods are generally not formulated in a probabilistic setting, which makes it difficult to account for uncertainties in biomarker progressions and diagnostic predictions. Indeed, the quantification of the variability associated with the biomarkers trajectories, as well as the assessment of the diagnostic uncertainty in *de-novo patients*, are crucial requirements for decision making in clinical practice [18].

Nonetheless, the ensemble of this research offers a sight of the potential of these approaches in representing a novel and powerful diagnostic instrument: in this study we thus aim at assessing the ability of DPM in providing a statistical reference for the transition from normal to pathological stages, for probabilistic diagnosis in the clinical scenario. To this end, we reformulate classical DPM within a Bayesian setting in order to allow the probabilistic estimate of the biomarker trajectories and the quantification of the uncertainty of predictions of the individual pathological stage. The resulting probabilistic framework is exploited in an hypothetical clinical scenario, for the estimation of the pathological stage in a de-novo cohort of testing individuals, by assessing the influence of missing observations, biomarkers, and follow-up information.

The manuscript is structured as follows. Section 2.1 formulates DPM based on Bayesian Gaussian Process regression [19], while Section 2.2 illustrates the validation of our model on clinical and multivariate imaging measurements from a cohort of 782 amyloid positive individuals extracted from the ADNI database.

2. Methods

2.1. Statistical setting

This section highlights the statistical framework employed in this study, based on the reformulation of self-modeling regression withing a Bayesian setting. This achieved by 1) defining a random effect Gaussian process regression model to account for individual correlated time series (section 2.1); 2) modeling individual time transformations encoding the information on the latent pathological stage (section 2.1.2); and 3) introducing a monotonicity information in order to impose a regular behaviour on the biomarkers trajectories (section 2.1.3). We finally illustrate in section 2.1.4 how the proposed framework leads to a probabilistic model of disease staging in de-novo individuals, naturally accounting for missing information. Further details on model specification and inference are provided in the Supplementary Section AppendixA.1, while the experimental validation on synthetic data is reported in Supplementary Section AppendixA.2.

2.1.1. Gaussian process-based random effect modeling of longitudinal progressions

In what follows, longitudinal measurements of N_b biomarkers $\{b_1, \ldots, b_{N_b}\}$ over time are given for N individuals.

We represent the longitudinal biomarker's measures associated with each individual j as a multidimensional array $(\mathbf{y}^j(t_1), \mathbf{y}^j(t_2), \ldots, \mathbf{y}^j(t_{k^j}))^\top$ sampled at k^j multiple time points $\mathbf{t} = \{t_1, t_2, \ldots, t_{k^j}\}$. Although different biomarkers may be in reality sampled at different time-points, for the sake of notation simplicity in what follows we will assume, without loss of generality, that the sampling time is common among them. The observations for individual j at a single time point t are thus a random sample from the following generative model:

$$
\begin{aligned}
\mathbf{y}^j(t) &= \left(y^j_{b_1}(t), y^j_{b_2}(t), \ldots, y^j_{b_{N_b}}(t)\right)^\top & (1)\\
&= \mathbf{f}(t) + \boldsymbol{\nu}^j(t) + \boldsymbol{\epsilon}, & (2)
\end{aligned}
$$

where $\mathbf{f}(t) = (f_{b_1}(t), f_{b_2}(t) \ldots, f_{b_{N_b}}(t))^\top$ is the fixed effect function modelling the biomarker's longitudinal evolution, $\boldsymbol{\nu}^j(t) = (\nu^j_{b_1}(t), \nu^j_{b_2}(t), \ldots, \nu^j_{b_{N_b}}(t))^\top$ is the individual random effect, and $\boldsymbol{\epsilon} = (\epsilon_{b_1}, \epsilon_{b_2}, \ldots, \epsilon_{b_{N_b}})^\top$ is time-independent observational noise. The group-wise evolution is modelled as a GP, $\mathbf{f} \sim \mathcal{GP}(0, \Sigma_G)$, the individual random effects are assumed to be correlated perturbations $\boldsymbol{\nu}^j \sim \mathcal{N}(0, \Sigma_S)$, while the observational noise is assumed to be a Gaussian heteroskedastic term $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma_\epsilon)$, where Σ_ϵ is a diagonal matrix $\mathrm{diag}[\boldsymbol{\sigma}^2_{b1}, \boldsymbol{\sigma}^2_{b2}, \ldots, \boldsymbol{\sigma}^2_{b_{N_b}}]$.

Fixed Effect Process

The covariance function Σ_G describes the biomarkers temporal variability, and is represented as a block-diagonal matrix $\Sigma_G(\mathbf{f}, \mathbf{f}) = \mathrm{diag}[\Sigma_{b_1}(\mathbf{f}_{b_1}, \mathbf{f}_{b_1}), \Sigma_{b_2}(\mathbf{f}_{b_2}, \mathbf{f}_{b_2}), \ldots, \Sigma_{b_{N_b}}(\mathbf{f}_{b_{N_b}}, \mathbf{f}_{b_{N_b}})]$, where each block represents the within-biomarker temporal covariance expressed as a negative squared exponential function $\Sigma_b(\mathbf{f}_b(t_1), \mathbf{f}_b(t_2)) = \eta_b \exp\left(-\frac{(t_1 - t_2)^2}{2\, l_b^2}\right)$, and where the parameters η_b and l_b are the marginal variance and length-scale of the biomarker's temporal evolution, respectively.

Individual Random Effects

The random covariance function Σ_S models the individual deviation from the fixed effect, and is represented as a block-diagonal matrix $\Sigma_S(\boldsymbol{\nu}^j, \boldsymbol{\nu}^j) = \text{diag}[\,\Sigma^j_{b_1}(\boldsymbol{\nu}^j_{b_1},\ \boldsymbol{\nu}^j_{b_1}), \Sigma^j_{b_2}(\boldsymbol{\nu}^j_{b_2}, \boldsymbol{\nu}^j_{b_2}), \ldots, \Sigma_{b^j_{N_b}}(\boldsymbol{\nu}^j_{b_{N_b}}, \boldsymbol{\nu}^j_{b_{N_b}})]$, where each block Σ^j_b corresponds to the covariance function associated with the individual process $\boldsymbol{\nu}^j_b(t)$. Thanks to the flexibility of the proposed generative model, any form of the random effect covariance Σ_S can be easily specified in order to model the subject-specific biomarkers' progression. In what follows we will use a linear covariance form $\Sigma^j_b(\boldsymbol{\nu}^j_b(t_1), \boldsymbol{\nu}^j_b(t_2)) = (\sigma^j_b)^2 \left((t_1 - \bar{\mathbf{t}})(t_2 - \bar{\mathbf{t}})\right)$, where $\bar{\mathbf{t}}$ is the average observational time for individual j, when more than 4 measurements are available, and i.i.d. Gaussian covariance form $\Sigma^j_b(\boldsymbol{\nu}^j_b(t_1), \boldsymbol{\nu}^j_b(t_2)) = (\sigma^j_b)^2$ when 2 or 3 measurements are available, while assigning it to 0 otherwise (thus by accounting only for the observational noise $\boldsymbol{\sigma}^2_b$). This choice is motivated by stability concerns, in order to keep the model complexity compatible with the generally limited number of measurements available for each individual.

2.1.2. Individual time transformation

The generative model (1) is based on the key assumption that the longitudinal observations across different individuals are defined with respect to the same temporal reference. This assumption may be invalid when the temporal alignment of the individual observations with respect to the common group-wise model is unknown, for instance in the typical scenario of a clinical trial in AD where the patients' observational time is relative to the common baseline, and where the disease onset is a latent event (past or future) which is not directly measurable. This modeling aspect is integrated by assuming that each individual measurement is made with respect to an absolute time-frame τ through a time-warping function $t = \phi^j(\tau)$ that models the time-reparameterization with respect to the common group-wise evolution. Model (1) can thus be reparameterized as

$$\mathbf{y}^j(\phi^j(\tau)) = \mathbf{f}(\phi^j(\tau)) + \boldsymbol{\nu}^j(\phi^j(\tau)) + \boldsymbol{\epsilon}. \tag{3}$$

The present formulation allows the specification of any kind of time transformation, and in what follows we shall focus on the modelling of a linear reparameterization of the observational time $\phi^j(\tau) = \tau + d^j$. This modeling assumption is mostly motivated by the choice of working with a reasonably limited number of parameters, compatibly with the generally short follow-up time available per individual (cfr. Table 2). Within this setting, the time-shift d^j encodes the disease stage associated with the individual relatively to the group-wise model.

2.1.3. Monotonic constraint in random-effect multimodal GP regression

Due to the non-parametric nature of Gaussian process regression, we need an additional constraint on model (3) in order to identify a unique solution for the time transformation. By assuming a steady temporal evolution of biomarkers from normal to pathological values, we shall assume that the biomarker trajectories described by (3) follow a (quasi) monotonic behaviour. This requirement can be implemented by imposing a prior positivity constraint on the derivatives of the GP function. Inspired by [20], we impose a monotonicity constraint by assuming a probit-likelihood for the derivative measurements $\mathbf{m}(t)$ associated with the derivative process $\dot{\mathbf{f}}(t) = \frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t}$ at time t:

$$p(\mathbf{m}(t)|\dot{\mathbf{f}}(t)) = \Phi\left(\frac{1}{\lambda}\dot{\mathbf{f}}(t)\right), \tag{4}$$

with $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(x|0,1)\,dx$. The quantity $\lambda > 0$ is an additional model parameter controlling the degree of positivity enforced on the derivative process, with values approaching zero for stronger monotonicity constraint. In what follows, the monotonicity of each biomarker is controlled by placing 10 derivative points equally spaced on the observation domain, and by fixing the N_b derivative parameters $\{\lambda_{b_k}\}_{k=1}^{N_b}$ to the value of 1e-6. The position of the derivative points was updated at each iteration, according to the changes of the GP domain.

By following a similar construction, we could equally enforce a monotonic behavior to the random effects associated with the individual trajectories. This additional constraint would however come with a cumbersome increase of the model complexity, since it would introduce an additional layer of virtual derivative parameters (with associated location) per individual. Moreover, while we are interested in modeling a globally monotonic biomarker trajectory on the fixed parameters, we relax this constraint at the individual level, since some subjects may be characterised by non strictly monotonic time-series due to specific clinical conditions.

Model likelihood and parameters

Given the sets of individual biomarker measurements $\mathbf{y} = \{(\mathbf{y}^j(t_i))_{i=1}^{k^j}\}_{j=1}^{N}$, and of D control derivatives $\mathbf{m} = \{m_{b_k}(t'_l)\}_{l=1}^{D}$ at points $t' = \{t'_l\}_{l=1}^{D}$ for the progression of each biomarker b_k, the random effect GP model posterior is:

$$\begin{aligned} p\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}, \mathbf{m}\right) &= \frac{1}{Z} p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu}|t) p(\mathbf{y}|\mathbf{f}, \boldsymbol{\nu}) p(\mathbf{m}|\dot{\mathbf{f}}) \\ &= p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu}|t) p(\mathbf{y}|\mathbf{f}, \boldsymbol{\nu}) \\ & \prod_k \prod_l \Phi\left(\frac{1}{\lambda} \dot{f}_{b_k}(t'_l)\right), \end{aligned} \quad (5)$$

where $\boldsymbol{\nu} = \{\nu^j\}_{j=1}^{N}$. Due to the non-Gaussianity of the derivative term Φ, the direct inference on the posterior is not possible due to its analytically intractable form. For this reason, we employ an approximate inference scheme based on classical approaches to Gaussian process with binary activation functions [21] (AppendixA.1).

Overall, model (3) is identified by $(N_j+3)N_b+N_j$ parameters, represented by the fixed effects and noise $\boldsymbol{\theta}_G = \{\eta_{b_k}, l_{b_k}, \epsilon_{b_k}\}_{k=1}^{N_b}$, by the individual random effects parameters $\boldsymbol{\theta}_G^j = \{\sigma_{b_k}^j\}_{k=1}^{N_b}$ and by the time-shifts d^j.

In what follows, the optimal parameters are obtained by maximising the approximated log-marginal likelihood derived from the posterior (5) through conjugate gradient descent, via alternate optimization between the hyper-parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_G^j$, and the individuals' time-shifts d^j. Regularization is also enforced by introducing Gaussian priors for the parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_G^j$.

2.1.4. Prediction of observations and individual staging

Gaussian processes naturally allow for probabilistic predictions given the observed data. At any given time point t^*, the posterior biomarker distribution has the Gaussian form $p(\mathbf{f}^*|t^*, \mathbf{y}, t, \mathbf{m}, t') \sim \mathcal{N}(\mathbf{f}^*|\boldsymbol{\mu}^*, \Sigma^*)$:

$$\begin{aligned}
\boldsymbol{\mu^*} &= \Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t))(\Sigma_{\text{joint}} + \tilde{\Sigma}_{\text{joint}})^{-1}\tilde{\boldsymbol{\mu}}_{\text{joint}} \qquad (6)\\
\Sigma^* &= \Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t^*)) - \Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t))\\
&\quad (\Sigma_{\text{joint}} + \tilde{\Sigma}_{\text{joint}})^{-1}\Sigma_G(\mathbf{f}(t), \mathbf{f}(t^*)), \qquad (7)
\end{aligned}$$

where the matrix $\left(\Sigma_{\text{joint}} + \tilde{\Sigma}_{\text{joint}}\right)$ is the joint covariance resulting from the inference scheme detailed in Supplementary Section AppendixA.1 [20].

We also derive a probabilistic model for the individual temporal staging given a set of biomarker observations $\mathbf{y}^*$, thanks to the Bayes formula:

$$\begin{aligned} p(t^*|\mathbf{y}^*, \mathbf{y}, t, \mathbf{m}, t') &= p(\mathbf{y}^*|t^*, \mathbf{y}, t, \mathbf{m}, t') \\ & \quad p(t^*)/p(\mathbf{y}^*|\mathbf{y}, t, \mathbf{m}, t'), \end{aligned} \tag{8}$$

which we compute by assuming an uniform distribution on t^*, and by noting that $p(\mathbf{y}^*|t^*, \mathbf{y}, t, \mathbf{m}, t') \sim \mathcal{N}(\boldsymbol{\mu}^*, \Sigma^* + \Sigma_\epsilon)$. In particular, the covariance form $\Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t^*))$ can be specified in order to account for incomplete data, and thus generalizes the GP model for predictions in presence of *missing biomarker observations*. The posterior distribution (8) quantifies the *confidence* of the model about the individual disease staging, and thus is a valuable information about the precision of the diagnosis. We will also compute the *expectation* of the distribution $p(t^*|\mathbf{y}^*, \mathbf{y}, t, \mathbf{m}, t')$, which provides a scalar value that can be used in subsequent classification methods.

2.2. Meterials and Methods

2.2.1. Study Participants

Data used in the preparation of this article were obtained from the ADNI database (`http://adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. 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 MCI and early AD. For up-to-date information, see `www.adni-info.org`.

2.2.2. Data Processing

We collected longitudinal measurements for the ADNI individuals with baseline values of cerebrospinal fluid (CSF) Aβ amyloid lower than the nominal val-

Group	NL	NL converted	MCI stable	MCI converted	AD
Training data					
N	67	5	0	53	75
Age	73 (6)	81.4 (5.2)	/	72 (7.7)	73 (8.5)
Sex (% females)	61	0	/	43	45
Education (yrs)	16.2 (2)	17.2 (3)	/	15.8 (2.6)	16 (2.4)
ADAS13	8.8 (4.5)	13.8 (2.4)	/	22.6 (6.7)	31.3 (8.5)
FAQ	0.2 (0.6)	0.4 (0.5)	/	5.2 (4.5)	12.9 (7)
RAVLT learning	5.6 (2.6)	2.2 (1.9)	/	3.2 (2.5)	1.8 (1.7)
Entorhinal (cm^3)	3.9 (0.6)	3.7 (0.5)	/	3.2 (0.7)	2.9 (0.6)
Hippocampus (cm^3)	7.5 (0.9)	6.7 (0.7)	/	6.2 (0.9)	6 (9.3)
Ventricles (cm^3)	36 (20)	57 (26)	/	42 (21)	47 (22)
Whole brain (cm^3)	1057 (105)	1106 (116)	/	1040 (107)	1013 (113)
FDG	6.6 (0.5)	6.1 (0.65)	/	5.7 (0.6)	5.2 (0.64)
AV45	1.2 (0.2)	1.3 (0.09)	/	1.4 (0.2)	1.4 (0.2)
Testing data					
N	74	17	243	106	145
Age	75.3 (5.9)	76.5 (4)	73.3 (7)	73.6 (7.3)	75 (7.9)
Sex (% females)	55	41	39	40	39
Education (yrs)	16 (2.9)	16.2 (2.6)	16 (2.8)	16 (3)	15.3 (3.1)
ADAS13	9.8 (4)	11.7 (3.4)	15.7 (6)	21 (6.1)	29.4 (8.2)
FAQ	0.5 (1.3)	0.6 (1.6)	2.7 (3.5)	5.1 (4.7)	12.9 (6.8)
RAVLT learning	5.6 (2.2)	5.6 (2.7)	4.3 (2.5)	2.8 (2.2)	1.8 (1.9)
Entorhinal (cm^3)	3.8 (0.4)	3.6 (0.7)	3.6 (0.7)	3.1 (0.7)	2.7 (0.7)
Hippocampus (cm^3)	7.2 (0.7)	7.2 (0.8)	6.9 (1)	6 (0.8)	5.7 (0.1)
Ventricles (cm^3)	33 (15)	44 (21)	39 (23)	41 (23)	49 (24)
Whole brain (cm^3)	1019 (102)	1055 (93)	1056 (100)	992 (110)	972 (124)
FDG	6.5 (0.62)	6.4 (0.7)	6.3 (0.7)	5.9 (0.6)	5.4 (0.7)
AV45	1.21 (0.19)	1.4 (0.2)	1.3 (0.19)	1.4 (0.2)	1.4 (0.2)

Table 1: Baseline sociodemographic and clinical information for training and testing study cohort. NL: normal individuals, MCI: mild cognitive impairment, AD: Alzheimer's patients. ADAS13: Alzheimer's Disease Assessment Scale-cognitive subscale, 13 items; FAQ: Functional Assessment Questionnaire; RAVLT learning: Rey Auditory Verbal Learning Test, learning item; FDG: (18)F-fluorodeoxyglucose positron emission tomography (PET) imaging; AV45: (18)F-florbetapir Amyvid PET imaging.

ues of 192 pg/ml. The information was extracted from the ADNIMERGE[1] R package[22] (MEDIAN field of the UPENNBIOMK_MASTER table). This preliminary selection is aimed to validate the model on a clinical population likely to represent the whole disease time-span.

The model was trained on a group of 200 randomly selected individuals including healthy volunteers, mild cognitive impairment subjects converted to AD (MCI conv), and AD patients having at least one measurement for each of the

[1] adni.bitbucket.io/adnimerge.html

following biomarkers: *volumetric measures* (hippocampal, ventricular, entorhinal, and whole brain volumes), *glucose metabolism* (average normalized FDG uptake in prefrontal cortex, anterior cingulate, precuneus and parietal cortex), *brain amyloidosys* (average normalized AV45 uptake in frontal cortex, anterior cingulate, precuneus and parietal cortex), and *functional, neuropsychological and cognitive function* measured by common scores (ADAS13, RAVLT learning, and FAQ)[2]. The testing set was composed of the remaining 582 subjects, including a subgroup of MCI non converted to AD during the observational time (MCI stable). The image-derived measures used in the study (volumetric MRI and average uptake values for AV45- and FDG-PET) were the scalar estimates reported in the ADNIMERGE package (adnimerge table). The volumetric measures were scaled by the individual total intracranial volume, and all the biomarkers measurements were converted into quantile scores (0 to 1 for normal to abnormal values), with respect to the biomarkers distribution of the *training set*. This latter modeling precaution is aimed to avoid spurious correlation between training and testing data due to the combined normalization of the values.

The modeling results were evaluated with respect to the baseline diagnostic information reported in the ADNI database, assessed according to the WMS and NINCDS/ADRDA AD criteria [23]. Conversion to MCI or AD was established according to the last follow-up information. Moreover, the MCI group was composed by 138 individuals with baseline diagnosis of early MCI, assessed through the Wechsler Memory Scale Logical Memory II. Among these subjects, 14 of them were in the training group (26% of the total MCI training set size),

[2] ADAS13: Alzheimer's Disease Assessment Scale-cognitive subscale, 13 items; FAQ: Functional Assessment Questionnaire; RAVLT learning: Rey Auditory Verbal Learning Test, learning item; FDG: (18)F-fluorodeoxyglucose positron emission tomography (PET) imaging; AV45: (18)F-florbetapir Amyvid PET imaging.

while the remaining 124 were in the testing set (35% of the total MCI testing set size).

Table 1 shows baseline clinical and sociodemographic information of the individuals used respectively in training and testing set, while in Table 2 we report the average follow-up time and the ratio of missing data of the pooled sample. Supplementary Section AppendixA.2.6 reports the R code used for data pre-processing.

2.3. Longitudinal modelling of Alzheimer's disease progression

2.3.1. Model training

The model was applied in order to estimate the temporal biomarker evolution and the disease stage associated with each individual in the training set. The plausibility of the model was assessed by group-wise comparison of the predicted time-shift, and by correlation with respect to the time to AD diagnosis for the MCI individuals subsequently converted to AD. For sake of comparison we also correlated the progression modelled with our approach with respect to the one estimated with the method proposed in [8]. The method was applied to the training data by using the standard parameters defined in the R package GRACE[3] (see Supplementary Material AppendixA.2.2 for further details).

[3] https://mdonohue.bitbucket.io/grace/

Ventr	Hippo	Ent	Whole Brain	ADAS13	FAQ	RAVLT	AV45	FDG
Training data								
2.3 (0)	2.3 (0)	2.3 (0)	2.3 (0)	3 (0)	3.3 (0)	3.3 (0)	1.9 (0)	1.6 (0)
Testing data								
3.4 (11)	3.4 (11)	3.4 (11)	3.4 (11)	3.9 (0)	3.9 (0)	3.9 (0)	3.8 (43)	3 (19)

Table 2: Average follow-up years and percentage of individuals with missing data (in parenthesis).

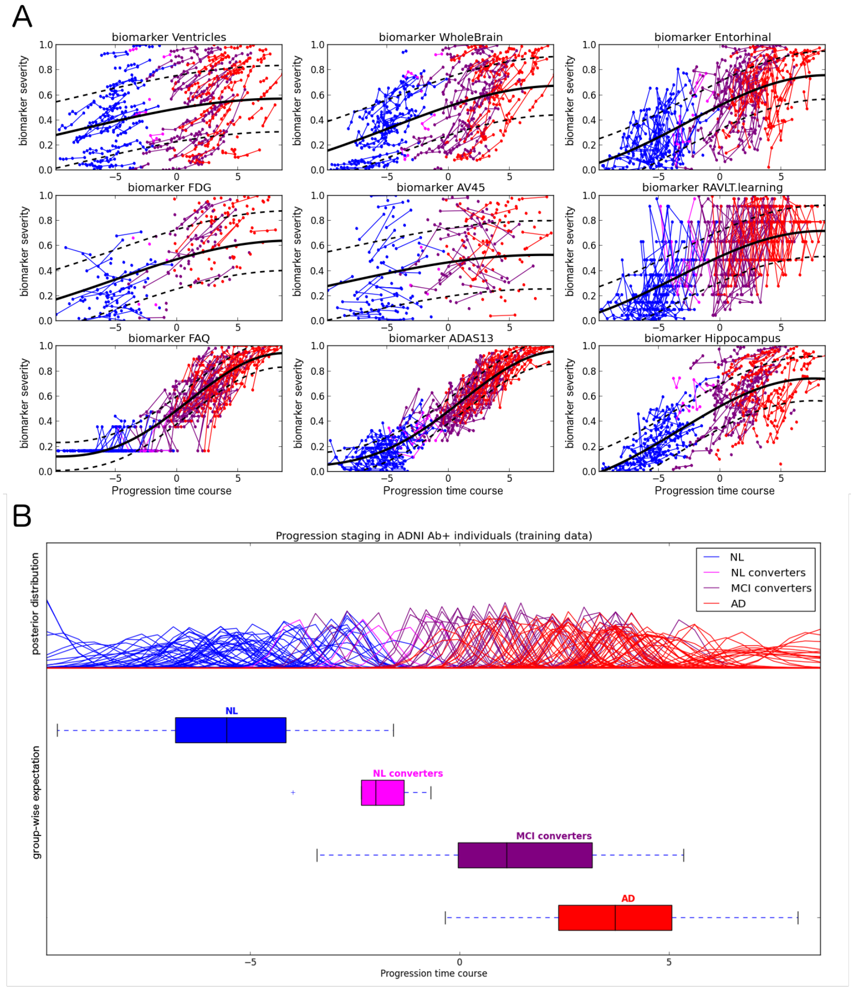


Figure 1: A) Modelled biomarker progression in the training set of 200 Aβ amyloid positive individuals (solid/dashed lines: mean ± sd). B) Posterior prediction for the individual time shift in training data (top: individual time-shift distribution; bottom: group-wise boxplot of the expected time-shift). Healthy individuals are generally displaced at the early stages of the pathology, while the predictions for MCI and AD patients are associated with respectively intermediate and late progression stages. NL: normal individuals, MCI: mild cognitive impairment, AD: Alzheimer's patients.

2.3.2. Model testing on de-novo individuals

The estimated probabilistic disease progression model provides a valuable clinical reference, as it can be used to predict an individual pathological stage, as well as to quantify the biomarkers predictive value, or the influence of missing data. To this end, we estimated the predictive performance of the model in assessing the individual pathological stage with respect to follow-up assessments and missing biomarkers. This was done by estimating the predictive accuracy of the group-wise separation obtained via increasing thresholds of the estimated temporal progression.

3. Results

3.1. Model plausibility

The estimated biomarker progression (Figure 1-A) shows a biologically plausible description of the pathological evolution, compatible with previous findings in longitudinal studies in familial AD [24], and with the hypothetical models of AD progression [2, 25]. The progression is defined on a time scale spanning roughly 20 years, and is characterized at the initial stages by high-levels of AV45, followed by the abnormality of ventricles volume, of FDG uptake, and of the whole brain volume. These latter measures are however heterogeneously distributed across clinical groups, and with rather large variability. The evolution is further characterized by increasing abnormality of the volumetric measures (especially hippocampal volume), and by the steady worsening of neuropsychological scores such as FAQ. The model thus shows that the transition from normal to pathological levels is essentially characterized by increase of hypometabolism, followed by the pronounced temporal brain atrophy. Moreover, the worsening of the neuropsychological and functional scores closely (almost linearly) follows the progression in the advanced clinical stages. The joint visualization of the temporal progression of the biomarkers with temporal derivative of the modelled average trajectories is shown in Supplementary figure A.10. The illustration confirms that ADAS13 and FAQ are characterised by very similar longitudinal profiles, and show the largest changes in the latest stages of the pathology (peak of the derivative at t>0). On the contrary, the change in hippocampal volume is more strongly associated with the earlier stages of the pathology. AV45 and ventricles volumes are the least informative and are associated with the lowest changes.

Figure 1-B (top) shows the posterior time-shift distributions associated with the individuals. The distributions denote the confidence of the model in associating to each individual a temporal staging with respect to the global pathological progression. The boxplot of Figure 1-B (bottom) reports the group-wise expectation of the individual time-shifts. Healthy individuals (blue) are associated with the early stages of the pathology in both training and testing data, while MCI (purple) and AD patients (red) are characterized by respectively intermediate and late predicted progression stages. The group-wise comparison between the expected time-shifts was statistically significant between each group pairs (ANOVA, p <1e-6). Moreover, the time to conversion to AD in the MCI group was significantly correlated with the disease staging quantified by the expectation of the individual time distributions ($R^2 = -0.4$, $p = 3.8e - 4$).

Finally, when applying [8] to the training data we measured a strong agreement between the resulting progression and the one obtained with our method, resulting in a correlation between the corresponding individual time-shifts of 0.94 (p <1e-6) (Supplementary Material AppendixA.2.2).

3.2. Assessing diagnostic uncertainty in testing data: an illustrative example.

This section illustrates the use of the model represented in Figure 1 for the quantification of diagnostic uncertainty in testing individuals. We consider the hypothetical scenario where the baseline values for different biomarkers are measured for a given patient, namely FAQ, hippocampal and ventricle volumes. We assume that the biomarkers values correspond to the 20th percentiles with

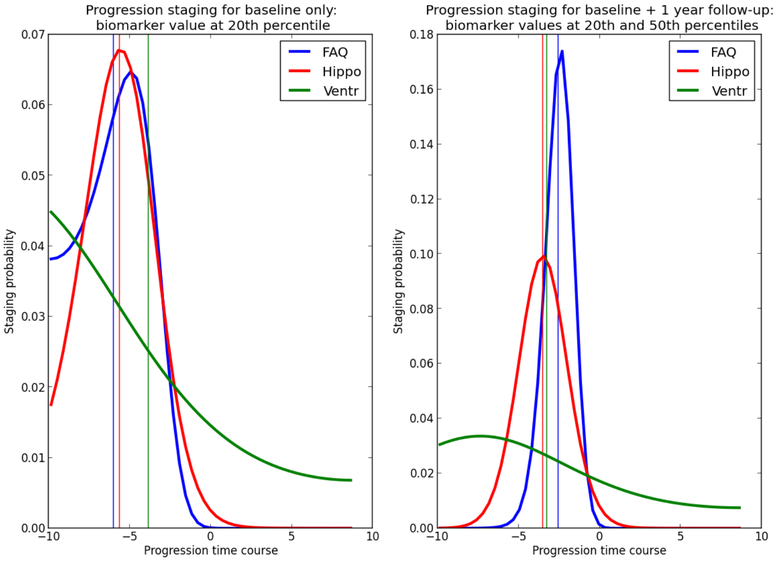


Figure 2: Illustrative example: posterior prediction of disease staging for a testing individual based on baseline (left), and baseline + follow-up (right) information for three biomarkers: FAQ, hippocampal and ventricles volume. The biomarkes values correspond to the 20th and 50th percentiles of the training-group distribution for respectively baseline and follow-up measures. Adding the follow-up information leads to increased estimates of the disease staging and to generally lower prediction uncertainty. Although the distributions associated with different biomarkers generally lead to similar expectations, FAQ and hippocampal volume lead to the lowest diagnostic uncertainty. Vertical lines: expectation for each posterior distribution.

respect to the biomarkers distribution of the training set (i.e. FAQ = 1, normalized Hippo = 5e-3, normalized Ventr = 1.7e-2). Figure 2 (left) shows the disease staging prediction obtained with formula (8) based on the value of each biomarker. We note that FAQ and hippocampal volume lead to similar posterior Gaussian distributions of disease staging, with expectation of respectively t_{FAQ}=-6 and t_{hippo}=-5.6 (indicated by the vertical lines in the figure), and standard deviation of sd_{FAQ}=6.3 and sd_{hippo}=5.9. The prediction associated with ventricles volume is wider and associated with higher uncertainty, with mean and standard deviation of respectively t_{ventr}=-3.8 and sd_{ventr}=6.1.

We now suppose that for the same patient we acquire a follow-up measurement for each biomarker at year 1, with values corresponding to the 50th percentiles of the distribution of the training set (i.e. FAQ = 5, normalized Hippo = 4.3e-3, normalized Ventr = 2.7e-2). The right hand side of Figure 2 shows the new prediction based on the joint baseline+follow-up information. For each biomarker the posterior distributions indicate an increase of the predicted disease stage with respect to the baseline scenario, while the prediction uncertainty is generally lower. Although the expectation for the 3 biomarkers is very similar (t_{FAQ}=-2.5, t_{hippo}=-3.5, and t_{Ventr}=-3.2), we notice that FAQ leads to the highest diagnostic confidence (sd_{FAQ}=2.6), followed by hippocampal volume (sd_{Hippo}=3.8), and finally by ventricles volume (sd_{Ventr}=5.7). Further assessment of the relationship between biomarker variability and model prediction is provided in supplementary Section AppendixA.2.3.

This illustrative example shows that the proposed probabilistic framework represents a valuable instrument for the assessment of the diagnostic value and uncertainty associated with different biomarkers, and can faithfully track the pathological progression of testing individuals along the modeled trajectories, from normal to pathological levels.

3.3. DPM for probabilistic diagnosis in ADNI.

We now assess the predictive results of the model when applied to the testing ADNI cohort. Figure 3 shows the individual posterior predictive distributions associated with the testing individuals, and the boxplot of the expected time-shift when using the model as statistical reference through formula (Figure 8). The figure reports the two different modeling scenarios based on baseline information only (Figure 3-1), and on the complete set of baseline and longitudinal measurements (Figure 3-2). We first note that the group-wise differences between the expected time-shifts are compatible for both scenarios, as shown by the similar boxplot distributions across groups (Figure 3-1b vs 3-2b). The con-

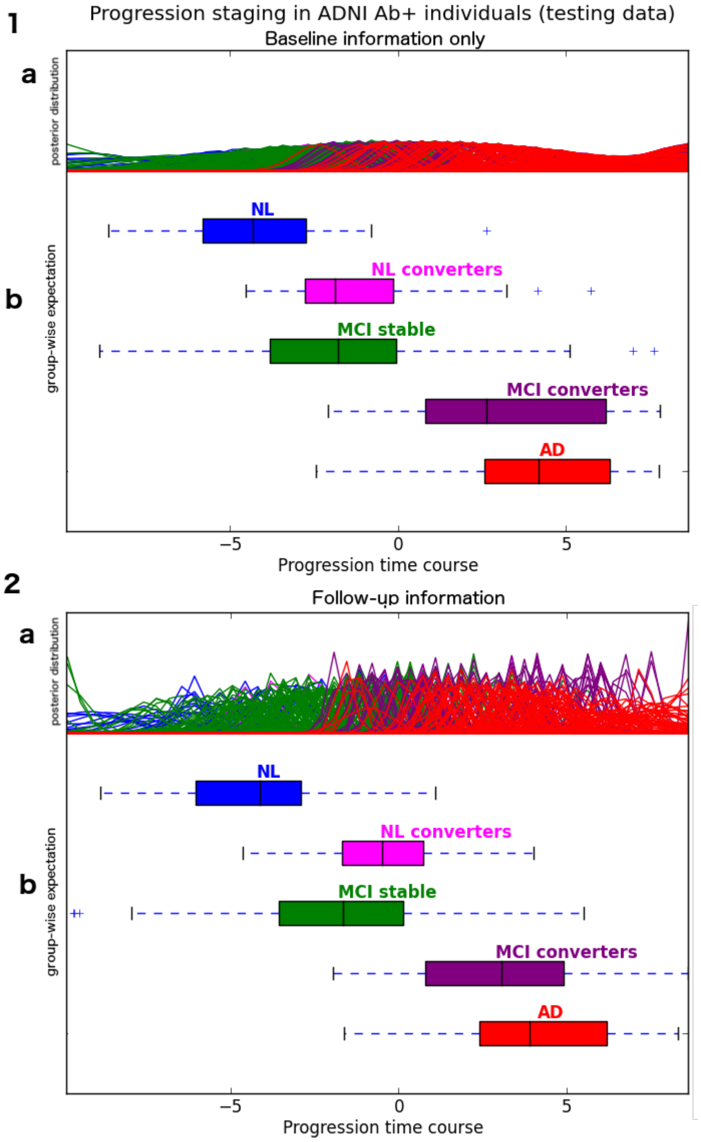


Figure 3: Posterior prediction for the individual time shift in testing data by using i) only the baseline information (1a-b), and ii) the baseline + follow-up information available for each testing subject (2a-b). Healthy individuals are generally displaced at the early stages of the pathology, while the predictions for MCI and AD patients are associated with respectively intermediate and late progression stages. The results are similar for both scenarios, although by adding the follow-up information we largely reduce the uncertainty in the prediction of the individual's pathological stage (subfigure 1a vs 2a). NL: normal individuals, MCI: mild cognitive impairment, AD: Alzheimer's patients.

sistency of the predictions is further illustrated in Figure A.9, where it is shown that the group-wise distribution and ordering of the predicted time-shifts in the testing data are compatible with those estimated in the training one.

However, the joint use of baseline and follow-up information largely reduces the uncertainty of the predictions (Figures 3-1a vs 3-2a). Indeed, the time distributions predicted when using baseline and follow-up information are narrower as compared to the wider confidence margins obtained by using the baseline information only. Therefore, adding follow-up measurements importantly improves the confidence of the model in determining the individual pathological stage.

As with the training case, for both scenarios the group-wise distribution of the expected time-shift shows a significant separation between the clinical groups according to the increase of the pathological stage (ANOVA, p <1e-4). Interestingly, the temporal positioning of the non converting MCI lies between controls and MCI converters, and is on average lower than the one of healthy individuals subsequently converted to cognitive impairment.

Figure 4 reports the classification results based on the baseline information only, and on increasing thresholds of the progression time course. Although the model is not optimized to explicitly classify the clinical groups, the simple thresholding based on the model predictions generally shows high face validity with respect to the clinical diagnosis. For all the considered scenarios, the highest accuracy is reached in a time window around the point $t = 0$, while the area under the ROC curve is .99, .88 and .83 for NL vs AD, MCI converters vs MCI stable, and NL converters vs NL stable, respectively.

We further tested the model in presence of missing information, by computing the predictions when only one baseline biomarker is available (Figure 5). The predictive outcomes show important variations depending on the con-

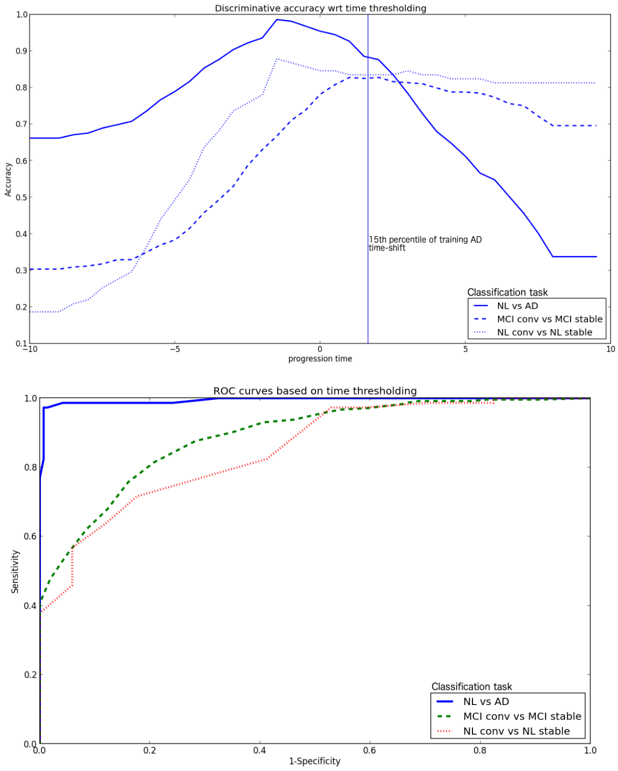


Figure 4: Predictive accuracy of the model when considering the joint set of available biomarkers measurements. The vertical bar indicates the reference threshold value of $t = 1.65$, corresponding to the 15th percentile of the time distribution of the training AD group. MCI: individuals with mild cognitive impairment, AD: Alzheimer's patients.

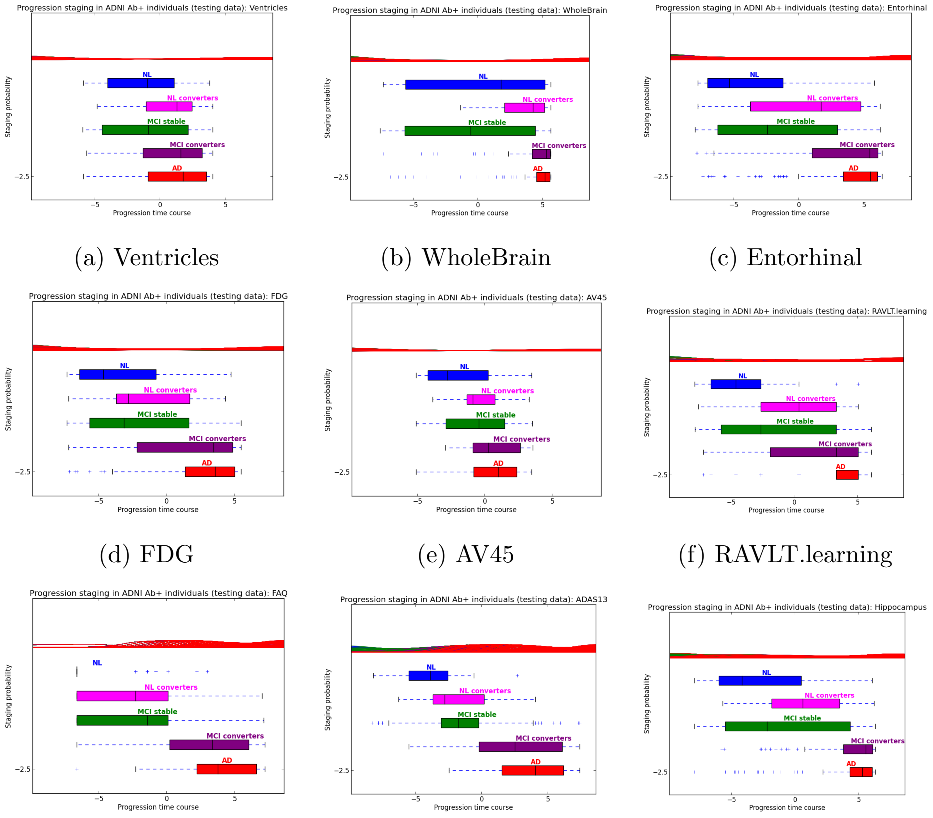
sidered biomarker, while the confidence bounds for the predictions are usually large, to denote increased uncertainty. We also note that FAQ, ADAS13, and hippocampal volume are the biomarkers leading to the largest group-wise separation, along with the lowest prediction uncertainty. This aspect is quantified in Table 3, reporting the discrimination results with respect to the nominal cut-off point of $t = 1.65$, corresponding to the 15th percentile of the distribution of the expected time-shift in the training AD group, as well as the area under the receiving operating characteristic curve (AUC). Although the highest discriminative results are consistently obtained when the biomarkers are used jointly, we note that the neuropsychological tests generally lead to the best predictive performance in identifying AD patients with respect to healthy individuals, followed by brain hypo-metabolism (FDG-PET), and temporal atrophy (Entorhinal and Hippocampal volume). This is related to the lower uncertainty of the modelled progressions, which leads to a more accurate identification of the individual staging along the pathological trajectory. The scenario sensibly changes in the other comparison scenarios (MCI conv vs stable and NL conv vs stable), where the sensitivity of the neuropsychological scores shows an important drop, while the other biomarkers (especially hippocampal and entorhinal volumes) provide comparable or even better discriminative performances.

These figures were similar when considering the single biomarkers within the longitudinal setting, where the neuropsychological tests still outperformed the other biomarkers in discriminating the clinical groups (Supplementary Figure A.11).

For the sake of comparison we finally benchmarked the predictive results provided by the disease progression model with respect those obtained by the classification analysis performed with standard statistical tools, such as a random forest classifier. We note that the comparison of the classification performance obtained on the heterogeneous data considered in this work is generally challenging, since the proposed DPM 1) accounts for missing data and non-fixed number of time points per individuals, and 2) is formulated in order to consistently handle both longitudinal and cross sectional measurements, either for training and prediction. To date there is not a consensus on the optimal approach to adopt to tackle these important modelling constraints, while the comparison between the classification performance obtained with complex machine learning methods is currently matter of scientific debate and investigation [26].

For this reason we restricted the random forest classification task to a standard statistical setting, in order to essentially provide a reliable benchmark for the classification performance of the proposed disease progression. To this end we trained the random forest on the classification between healthy individuals and AD patients based on the baseline measurements of the training group, while the missing entries in the testing data were imputed via nearest neighbour search, based on the available biomarkers. The classification results are reported in Supplementary Table A.5.

The performance of the random forest classifier is generally inferior to the one obtained with the proposed approach, as witnessed by the consistently lower AUC obtained for all the comparisons. The difference becomes more evident for the more challenging classification problems, such as the identification of conversion in MCI and healthy individuals. This result is indicative of the reliability of the classification results obtained by the proposed disease progression model, especially when considering that the random forest classifier is explicitly optimized to maximize the separation between groups, while the accuracy results reported in Table 3 are based on the empirical choice of a reference threshold in the training population.



(a) Ventricles

(b) WholeBrain

(c) Entorhinal

(d) FDG

(e) AV45

(f) RAVLT.learning

(g) FAQ

(h) ADAS13

(i) Hippocampus

Figure 5: Posterior prediction on testing data by using a single biomarker and the baseline information only.

	Biomarker									
	all	Hippo	Ventr	WholeBr	Entor	FDG	AV45	RAVLT	FAQ	ADAS13
NL vs AD (145 vs 74)										
Accuracy	.89	.81	.62	.76	.83	.80	.63	.82	.88	.83
Sensitivity	.83	.84	.52	.9	.82	.74	.82	.76	.84	.75
Specificity	.98	.76	.80	.46	.83	.89	.46	.94	.97	.98
AUC	.99	.87	.69	.7	.89	.87	.73	.91	.98	.98
MCI conv vs MCI stable (106 vs 243)										
Accuracy	.82	.67	.62	.69	.7	.71	.69	.67	.79	.79
Sensitivity	.65	.85	.5	.89	.74	.65	.37	.56	.63	.54
Specificity	.90	.59	.68	.60	.68	.73	.75	.71	.86	.9
AUC	.88	.79	.61	.78	.76	.74	.61	.66	.81	.82
NL conv vs NL stable (17 vs 74)										
Accuracy	.83	.70	.71	.54	.77	.76	.73	.83	.82	.83
Sensitivity	.18	.47	.41	.82	.52	.29	.27	.35	.17	.17
Specificity	.98	.77	.80	.47	.83	.89	.86	.94	.97	.98
AUC	.83	.71	.65	.63	.74	.65	.65	.7	.63	.68

Table 3: Classification results by using the reference time threshold of $t = 1.65$, corresponding to the 15th percentile of the training AD time distribution .

3.4. DPM staging and chronological age.

We finally compare the relationship between the predicted disease staging in training and testing set and the individual chronological age. We first note that both training and testing clinical groups were matched by age, with the exception of the 5 training healthy subjects converted to MCI (or AD) that were slightly older with respect to the reference training healthy population (p=0.02).

Nevertheless, when comparing the estimated time shift with respect to the chronological age of each individual we didn't report any significant correlation between these measures. Interestingly, the same lack of association is also quantifiable in the testing group (Figure 6). This result, in association with the strong relationship between time shift and clinical condition reported in Section 3.3, let us conclude that the model is describing the biomarker's variation essentially related to the pathological progression, which is orthogonal to the effect of healthy aging quantified by the chronological age. This result points to the effectiveness of the proposed approach in capturing significant effects related to the specific temporal progression of the disease.

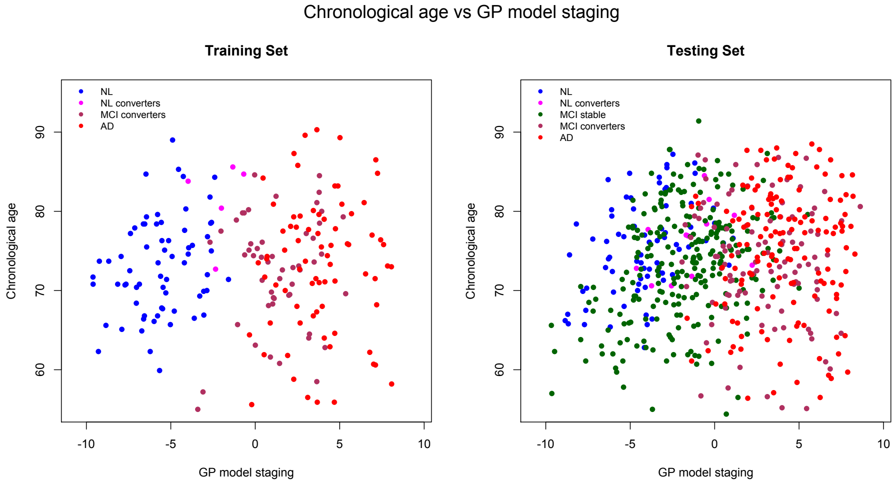


Figure 6: Chronological age (y-axis) vs model staging (x-axis). The estimated time-shift is decorrelated from the chronological age in both training and testing data ($p>0.05$).

4. Discussion

This study explores the use of DPM for probabilistic diagnosis and uncertainty quantification in an hypothetical clinical scenario. The proposed approach is based on the reformulation of DPM through a novel probabilistic approach aimed at leveraging on the longitudinal modeling of disease progression for prediction and quantification of the diagnostic uncertainty in neurodegeneration, by optimally combining the information provided by the several biomarkers into a biologically plausible and intelligible score quantified by the time shift. This work thus extends the previous contributions by proposing DPM as a probabilistic tool for diagnostic purposes, which can be used to quantify staging and predictive uncertainty of de-novo individuals in clinical trials. The disease progression model itself thus can be seen as a novel biomarker of pathological progression. We also note that the time shift is a *relative* measure of disease progression accounting for the biomarker variability observed in the training population. Thus, the point 0 is generally not associated with the conversion to

AD, as it is relative to the data initialization (in this case the study baseline).

We illustrated the use of DPM as benchmarking tool for the statistical comparison of biomarkers. The model allows the quantification of the variability associated with the single biomarkers, by identifying the related uncertainty in characterizing the progression from normal to pathological levels. The proposed model can be thus used as a reference for screening and enrichment purposes in clinical trials [27, 28, 29].

The modelled progression showed that neuropsychological tests generally lead to lower uncertainty for identifying the individual clinical stage, and to the higher separation power between healthy and AD groups. This finding is compatible with the results reported by previous disease progression models applied to ADNI, such as [7] and [15]. In this latter study ADAS13 consistently appeared among the first events distinguishing the normal disease stages from the pathological ones. Furthermore, our analysis further showed that volumetric measures such as hippocampal and entorhinal volumes provide equivalent if not superior diagnostic performances when tested on the more challenging problem of detecting conversion to dementia from healthy and MCI stages, especially in terms of improved AUC. Nevertheless, some care should be taken in drawing conclusions from the present analysis. Our model was based on the standard volumetric measures provided in the ADNI database, and we cannot exclude that a more precise quantification of morphological brain changes would lead to even better performance of volumetric biomarkers [30, 31]. Furthermore, the proposed model was not optimized in order to maximize the classification accuracy between clinical groups. For example, the results reported in Table 3 are based on the choice of the temporal threshold corresponding to the reference value of the 15th percentile of the AD distribution. This cut-off was not optimized to maximize the predictive outcome of the biomarkers, but was rather chosen based on heuristics aimed at illustrating the use of the model for predictive purposes. We thus cannot exclude that the optimization of the temporal threshold would lead to different figures for the classification task. The reported results are therefore indicative of the effectiveness of the model in faithfully representing the clinical spectrum of the disease. We note also that the reported figures are in line with those provided by state-of-art methods in AD classification, without requiring complex parameter optimization procedures, which would introduce additional levels of cross-validation and expose the results to selection bias and generalization issues [26].

Thanks to the probabilistic formulation we showed that the use of longitudinal information is important for reducing the uncertainty of the prediction, and thus allowing one to better identify the disease status associated to an individual. This important aspect is in agreement with the generally higher statistical power reported in previous Alzheimer's studies comparing longitudinal measurements to baselines ones [32, 25, 33].

In this work we focused on the modelling of the progression of amyloid positive individuals. This choice was motivated by the interest in assessing the model performance on an homogeneous clinical population likely to be representative of the Alzheimer's evolution. While the absence of pathological amylod levels seems indicative of non-Alzheimer's pathophysiology [34, 35], there is currently an active debate on the mechanisms of neurodegeneration not related to brain amyloidosis [36]. The investigation of these aspects goes beyond the scope of the present work, and future extensions of disease progression modeling will aim at identifying differential progressions underlying sub-pathologies, for example by reformulating the proposed random effect regression within the realm of Gaussian process mixture models [37, 38]. Analogously, the MCI population used for model training was composed exclusively by MCI individuals subsequently converted to AD, in order to train the model on a homogeneous data most likely to include the largest representation of individuals effectively affected by Alzheimer's disease. Although the inclusion of MCI stable could provide additional information on the intermediate pathological stages, this choice may probably lead to larger variability in the training set, as stable MCI are generally characterized by larger heterogeneity, either cross-sectionally and longitudinally, and higher diagnostic uncertainty. This modeling choice was also motivated by practical reasons since, thanks to the adopted data selection scheme, we were able to validate the model on a *large and independent* set of testing individuals including an important sample of MCI individuals across different clinical stages, thus providing a thorough and stringent assessment of the predictive qualities of the proposed approach.

4.1. Methodological considerations

From the methodological perspective, we proposed a novel probabilistic approach based on Gaussian process regression for disease progression modeling from time-series of biomarker measurements enabling novel applications beyond the state-of-art, such as the probabilistic prediction of disease staging in testing individuals. Furthermore, the model naturally accounts for missing data, and provides uncertainty quantification of the biomarker evolutions. Similarly to [8], in this work we focused on the modeling of disease staging represented by a time shift, although the proposed framework can naturally account for more complex time transformations, provided that a sufficient number of time points is available for each individual.

From the methodological point of view, the proposed model extends current approaches to GP-regression by consistently integrating time-reparameterization and monotonic constraints within a random effect regression framework. Monotonic GPs were introduced in [20] as a principled regularization solution to improve the plausibility of modeling results. For example, the strength of such a regularization approach in biomedical applications has been illustrated in survival analysis [39]. Our approach extends this framework by consistently integrating a latent time variable parameter within a random effect model formulation.

The idea of estimating a time transformation in a GP regression framework has been previously used by [40] to account for uncertain measurement times to a microarray dataset of mRNA. However, in that work the estimation of the time uncertainty was subject to a strong prior constraint based on the assumption that the unknown biological time must be similar to the measured one. In the application proposed in our work such an assumption is no longer valid and would ultimately lead to implausible estimations. On the contrary, the proposed GP regression is able to recover the underlying time transformation thanks to the proposed monotonicity regularization.

Finally, thanks to the flexibility of the proposed Gaussian process framework, further extensions of the model will enable to consistently integrate a spatio-temporal covariance model, such as the efficient Kronecker form of [41], to provide a unified framework for jointly modelling time series of images and scalar biomarkers data in a coherent fully Bayesian setting.

5. Conclusions

This work proposes an extension of DPM for the accurate quantification of the diagnostic uncertainty in Alzheimer's disease. The proposed application shows that DPM provides at the same time a plausible description of the transition from normal to pathological stages along the natural history of the disease, as well as remarkable diagnostic performances when tested on de-novo individuals. The model used in this study can account for any missing data patterns (longitudinal or across biomarkers), and allows to directly quantify the uncertainty related to the missing information. It thus represents a novel and promising tool for the analysis of clinical trials data.

6. Further Information

The open-source code as well as the proposed predictive model trained on ADNI data will be available at the author's web-page: `https://team.inria.fr/asclepios/marco-lorenzi/`. The realization of this study required about 1.5kWh of computing power.

7. Acknowledgments

EPSRC grants EP/J020990/01 and EP/M020533/1 support DCA and SO's work on this topic. DCA and SO also received support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 666992 (EuroPOND) for this work. MF gratefully acknowledges support from the AXA Research Fund.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson

Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

[1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, H. M. Arrighi, Forecasting the global burden of Alzheimer's disease, Alzheimer's & Dementia 3 (3) (2007) 186–191.

[2] C. R. Jack, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, J. Q. Trojanowski, Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade, The Lancet Neurology 9 (1) (2010) 119–128.

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

[4] B. Mwangi, T. S. Tian, J. C. Soares, A review of feature reduction techniques in neuroimaging, Neuroinformatics 12 (2) (2014) 229–244.

[5] M. Lorenzi, I. J. Simpson, A. F. Mendelson, S. B. Vos, M. J. Cardoso, M. Modat, J. M. Schott, S. Ourselin, Multimodal image analysis in Alzheimer's disease via statistical modelling of non-local intensity correlations, Scientific reports 6 (2016) 22161.

[6] H. M. Fonteijn, M. J. Clarkson, M. Modat, et al., An event-based disease progression model and its application to familial Alzheimer's disease, in: IPMI, Springer, 2011, pp. 748–759.

[7] B. M. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, J. L. Prince, et al., A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort, Neuroimage 63 (3) (2012) 1478–1486.

[8] M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff, et al., Estimating long-term multivariate progression from short-term data, Alzheimer's & Dementia 10 (5) (2014) S400–S410.

[9] L. Younes, M. Albert, M. I. Miller, B. R. Team, et al., Inferring change-point times of medial temporal lobe morphometric change in preclinical Alzheimer's disease, NeuroImage: Clinical 5 (2014) 178–187.

[10] M. Bilgel, B. Jedynak, D. F. Wong, S. M. Resnick, J. L. Prince, Temporal trajectory and progression score estimation from voxelwise longitudinal imaging measures: Application to amyloid imaging, in: Proceedings of Information Processing in Medical Imaging, Springer, 2015, pp. 424–436.

[11] J.-B. Schiratti, S. Allassonniere, A. Routier, O. Colliot, S. Durrleman, A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data, in: IPMI, Springer, 2015, pp. 564–575.

[12] R. Guerrero, A. Schmidt-Richberg, C. Ledig, T. Tong, R. Wolz, D. Rueckert, Instantiated mixed effects modeling of Alzheimer's disease markers, NeuroImage 142 (2016) 113–125.

[13] R. V. Marinescu, A. Eshaghi, M. Lorenzi, et al., A vertex clustering model for disease progression: Application to cortical thickness images, in: IPMI, Springer, 2017, p. to appear.

[14] E. Yang, M. Farnum, V. Lobanov, et al., Quantifying the pathophysiological timeline of Alzheimer's disease, Journal of Alzheimer's Disease 26 (4) (2011) 745–753.

[15] A. L. Young, N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, D. C. Alexander, A data-driven model of biomarker changes in sporadic Alzheimer's disease, Brain 137 (9) (2014) 2564–2577.

[16] A. Kneip, T. Gasser, Convergence and consistency results for self-modeling nonlinear regression, The Annals of Statistics (1988) 82–112.

[17] A. Schmidt-Richberg, R. Guerrero, C. Ledig, H. Molina-Abril, A. F. Frangi, D. Rueckert, Multi-stage biomarker models for progression estimation in Alzheimer’s disease, in: International Conference on Information Processing in Medical Imaging, Springer, 2015, pp. 387–398.

[18] B. Shinkins, R. Perera, Diagnostic uncertainty: dichotomies are not the answer, British Journal of General Practice 63 (2013) 122–123.

[19] C. E. Rasmussen, Gaussian processes for machine learning, Springer, 2006.

[20] J. Riihimäki, A. Vehtari, Gaussian processes with monotonicity information., in: AISTATS, Vol. 9, 2010, pp. 645–652.

[21] H. Nickisch, C. E. Rasmussen, Approximations for binary Gaussian process classification, Journal of Machine Learning Research 9 (Oct) (2008) 2035–2078.

[22] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2015).
URL `https://www.R-project.org/`

[23] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, E. M. Stadlan, Clinical diagnosis of alzheimer’s disease report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer’s disease, Neurology 34 (7) (1984) 939–939.

[24] R. J. Bateman, C. Xiong, T. L. Benzinger, et al., Clinical and biomarker changes in dominantly inherited Alzheimer’s disease, New England Journal of Medicine 367 (9) (2012) 795–804.

[25] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, P. M. Thompson, The clinical use of structural mri in Alzheimer disease, Nature Reviews Neurology 6 (2) (2010) 67–77.

[26] A. F. Mendelson, M. A. Zuluaga, M. Lorenzi, B. F. Hutton, S. Ourselin, Selection bias in the reported performances of AD classification pipelines, NeuroImage: Clinical 14 (2016) 400–416.

[27] M. Lorenzi, M. Donohue, D. Paternico, C. Scarpazza, S. Ostrowitzki, O. Blin, E. Irving, G. Frisoni, Enrichment through biomarkers in clinical trials of Alzheimer's drugs in patients with mild cognitive impairment, Neurobiology of aging 31 (8) (2010) 1443–1451.

[28] P. Yu, J. Sun, R. Wolz, D. Stephenson, J. Brewer, N. C. Fox, P. E. Cole, C. R. Jack, D. L. Hill, A. J. Schwarz, et al., Operationalizing hippocampal volume as an enrichment biomarker for amnestic mild cognitive impairment trials: effect of algorithm, test-retest variability, and cut point on trial cost, duration, and sample size, Neurobiology of aging 35 (4) (2014) 808–818.

[29] D. L. Hill, A. J. Schwarz, M. Isaac, L. Pani, S. Vamvakas, R. Hemmings, M. C. Carrillo, P. Yu, J. Sun, L. Beckett, et al., Coalition against major diseases/european medicines agency biomarker qualification of hippocampal volume for enrichment of clinical trials in predementia stages of Alzheimer's disease, Alzheimer's & Dementia 10 (4) (2014) 421–429.

[30] R. Wolz, R. A. Heckemann, P. Aljabar, J. V. Hajnal, A. Hammers, J. Lötjönen, D. Rueckert, Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI, NeuroImage 52 (1) (2010) 109–118.

[31] D. M. Cash, C. Frost, L. O. Iheme, D. Ünay, M. Kandemir, J. Fripp, O. Salvado, P. Bourgeat, M. Reuter, B. Fischl, et al., Assessing atrophy

measurement techniques in dementia: Results from the MIRIAD atrophy challenge, NeuroImage 123 (2015) 149–164.

[32] W. Henneman, J. Sluimer, J. Barnes, W. Van Der Flier, I. Sluimer, N. Fox, P. Scheltens, H. Vrenken, F. Barkhof, Hippocampal atrophy rates in Alzheimer's disease. added value over whole brain volume measures, Neurology 72 (11) (2009) 999–1007.

[33] Z. Xu, X. Shen, W. Pan, Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes, PloS one 9 (8) (2014) e102312.

[34] B. A. Gordon, T. Blazey, Y. Su, A. M. Fagan, D. M. Holtzman, J. C. Morris, T. L. Benzinger, Longitudinal β-amyloid deposition and hippocampal volume in preclinical Alzheimer's disease and suspected non–Alzheimer disease pathophysiology, Jama neurology 73 (10) (2016) 1192–1200.

[35] E. C. Mormino, K. V. Papp, D. M. Rentz, A. P. Schultz, M. LaPoint, R. Amariglio, B. Hanseeuw, G. A. Marshall, T. Hedden, K. A. Johnson, et al., Heterogeneity in suspected non–Alzheimer disease pathophysiology among clinically normal older individuals, Jama neurology 73 (10) (2016) 1185–1191.

[36] C. R. Jack Jr, D. S. Knopman, G. Chételat, D. Dickson, A. M. Fagan, G. B. Frisoni, W. Jagust, E. C. Mormino, R. C. Petersen, R. A. Sperling, et al., Suspected non-Alzheimer disease pathophysiology – concept and controversy, Nature Reviews Neurology 12 (2016) 117–124.

[37] M. Lázaro-Gredilla, S. Van Vaerenbergh, N. D. Lawrence, Overlapping mixtures of Gaussian processes for the data association problem, Pattern Recognition 45 (4) (2012) 1386–1395.

[38] J. C. Ross, J. G. Dy, Nonparametric mixture of Gaussian processes with constraints., in: Proceedings of International Conference of Machine Learning (ICML), 2013, pp. 1346–1354.

[39] H. Joensuu, A. Vehtari, J. Riihimäki, T. Nishida, S. E. Steigen, P. Brabec, L. Plank, B. Nilsson, C. Cirilli, C. Braconi, et al., Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts, The lancet oncology 13 (3) (2012) 265–274.

[40] Q. Liu, K. K. Lin, B. Andersen, P. Smyth, A. Ihler, Estimating replicate time shifts using Gaussian process regression, Bioinformatics 26 (6) (2010) 770–776.

[41] M. Lorenzi, G. Ziegler, D. C. Alexander, S. Ourselin, Efficient Gaussian process-based modelling and prediction of image time series, in: IPMI, Springer, 2015, pp. 626–637.

AppendixA. Supplementary Information

AppendixA.1. Joint Model: marginal likelihood and inference

Given the sets of individual biomarker measurements $\mathbf{y} = \{(\mathbf{y}^j(t_i))_{i=1}^{k^j}\}_{j=1}^{N}$, and of D control derivatives $\mathbf{m} = \{m_{b_k}(t'_l)\}_{l=1}^{D}$ at points $t' = \{t'_l\}_{l=1}^{D}$ for the progression of each biomarker b_k, the random effect GP model posterior is:

$$\begin{aligned} p\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}, \mathbf{m}\right) &= \frac{1}{Z} p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu}|t) p(\mathbf{y}|\mathbf{f}, \boldsymbol{\nu}) p(\mathbf{m}|\dot{\mathbf{f}}) \\ &= p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu}|t) p(\mathbf{y}|\mathbf{f}, \boldsymbol{\nu}) \\ & \prod_k \prod_l \Phi\left(\frac{1}{\lambda} \dot{f}_{b_k}(t'_l)\right), \end{aligned} \tag{A.1}$$

where $\boldsymbol{\nu} = \{\nu^j\}_{j=1}^{N}$. Thanks to the linearity of GPs under derivation, we have that $Cov\left(\mathbf{f}(t), \dot{\mathbf{f}}(t')\right) = \frac{\mathrm{d}Cov(\mathbf{f}(t), \mathbf{f}(t'))}{\mathrm{d}t'}$, and that the joint distribution
775 $p\left(\mathbf{f}, \dot{\mathbf{f}} | t, t'\right)$ is again a GP

$$p\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | t, t'\right) \quad \sim \quad \mathcal{GP}\left(\mathbf{f}_{joint} | 0, \Sigma_{joint}\right),$$

with $\mathbf{f}_{joint} = \begin{pmatrix} \mathbf{f} \\ \dot{\mathbf{f}} \end{pmatrix}$ distributed as

$$\mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_G(\mathbf{f}(t), \mathbf{f}(t)) & \frac{\partial \Sigma_G(\mathbf{f}(t), \mathbf{f}(t'))}{\partial t'} \\ \frac{\mathrm{d}\Sigma_G(\mathbf{f}(t'), \mathbf{f}(t))}{\mathrm{d}t'} & \frac{\mathrm{d}^2 \Sigma_G(\mathbf{f}(t'), \mathbf{f}(t'))}{\mathrm{d}t'^2} \end{pmatrix}\right].$$

AppendixA.1.1. Approximated inference

Due to the non-Gaussianity of the derivative likelihood term, the direct inference on the posterior (A.1) is not possible due to its analytically intractable
780 form. For this reason, we employ an approximate inference scheme based on clas-

sical approaches to Gaussian process with binary activation functions [21]. Following [20], we compute an approximated posterior distribution $q\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}^j, \mathbf{m}\right)$ by replacing the derivative likelihood terms with local un-normalized Gaussian approximations:

$$\begin{aligned} q\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}^j, \mathbf{m}\right) =& \frac{1}{Z_{EP}} p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu} | t) p(\mathbf{y} | \mathbf{f}, \boldsymbol{\nu}) \\ & \prod_k \prod_l \tilde{Z}_{kl} \mathcal{N}(\dot{f}_{b_k}(t'_l) | \tilde{\mu}_{kl}, \tilde{\sigma}^2_{kl}), \end{aligned} \tag{A.2}$$

where

$$\prod_k \prod_l \tilde{Z}_{kl} \mathcal{N}(\dot{f}_{b_k}(t'_l) | \tilde{\mu}_{kl}, \tilde{\sigma}^2_{kl}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}) \prod_{k,l} \tilde{Z}_{kl},$$

with $\tilde{\boldsymbol{\mu}} = [\tilde{\mu}_{kl}]$, and $\tilde{\Sigma}$ is a diagonal matrix with elements $\tilde{\sigma}^2_{kl}$. It follows that the marginal posterior has a Gaussian form, $q\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}^j, \mathbf{m}\right) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with $\boldsymbol{\mu} = \Sigma \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}_{joint}$, and $\Sigma = (\Sigma^{-1}_{joint} + \tilde{\Sigma}^{-1}_{joint})^{-1}$, where

$$\tilde{\boldsymbol{\mu}}_{joint} = \begin{pmatrix} \mathbf{y} \\ \tilde{\boldsymbol{\mu}} \end{pmatrix}, \quad \text{and} \quad \tilde{\Sigma}_{joint} = \begin{pmatrix} \Sigma_\epsilon + \Sigma_S & 0 \\ 0 & \tilde{\Sigma} \end{pmatrix}.$$

AppendixA.1.2. Estimating the EP parameters.

The EP update of the local Gaussian approximation parameters is classically done by iterative moment matching with respect to the product between the cavity distributions $q_{-k'l'}\left(\dot{f}_{b_{k'}}(t'_{l'})\right)$ and the target likelihood term $\Phi\left(\frac{1}{\lambda} \dot{f}_{b_{k'}}(t'_{l'})\right)$.

In the GP case the cavity distribution has a straightforward Gaussian form:

$$\begin{aligned} q_{-k'l'}\left(\dot{f}_{b_{k'}}(t'_{l'})\right) &= \int \prod_{\substack{k \neq k', \\ l \neq l'}} \tilde{Z}_{kl} \mathcal{N}(\dot{f}_{b_k}(t'_l) | \tilde{\mu}_{kl}, \tilde{\sigma}^2_{kl}) d\dot{f}_{b_k}(t'_l) \\ &\sim \mathcal{N}(\dot{f}_{b_{k'}}(t'_{l'}) | \mu_{-k'l'}, \sigma_{-k'l'}). \end{aligned} \tag{A.3}$$

As shown in [20] for univariate monotonic regression, moments and updates of the approximation parameters can be computed in an analogous manner as in the classical GP classification problem [19].

AppendixA.1.3. Marginal Likelihood and hyper-parameter estimation

The model's log-marginal likelihood under the EP approximation is:

$$\begin{aligned}\log \mathcal{L} \quad = \quad & -\frac{1}{2}\log|\Sigma_{joint} + \tilde{\Sigma}_{joint}| \\ & -\frac{1}{2}\tilde{\boldsymbol{\mu}}_{joint}^{T}(\Sigma_{joint} + \tilde{\Sigma}_{joint})^{-1}\tilde{\boldsymbol{\mu}}_{joint} \\ & +\sum_{k}\sum_{l}\frac{(\mu_{-kl} - \tilde{\mu}_{kl})^2}{2(\sigma_{-kl}^2) + \tilde{\sigma}_{kl}^2)} \\ & +\sum_{k}\sum_{l}\log\Phi(\frac{\mu_{-kl}}{\sqrt{\lambda_k^2 + \sigma_{-kl}^2)}}) \\ & +\frac{1}{2}\sum_{k}\sum_{l}\log(\sigma_{-kl}^2 + \tilde{\sigma}_{kl}^2). \qquad \text{(A.4)}\end{aligned}$$

In what follows, the optimal parameters are obtained by maximising $\log \mathcal{L}$ through conjugate gradient descent, via alternate optimization between the hyper-parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_G^j$, and the individuals' time-shifts d^j. The position of the derivative points was updated at each iteration, according to the changes of the GP domain. Regularisation was also enforced by introducing Gaussian priors for the parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_G^j$. We note that the block structure of the GP covariance allows the computation of the gradients with respect to the biomarkers' and individual parameters by working on matrices of much smaller dimension than the one of the whole GP, thus considerably improving the numerical stability and the computational efficiency of the optimization procedure.

AppendixA.2. Model benchmarking on synthetic multivariate progressions

We benchmarked the model with respect to synthetic multivariate biomarker progressions. We generated random multivariate sigmoid functions for N_b biomarkers, $\mathbf{f}(\tau) = (f_{b_1}(\tau), f_{b_2}(\tau), \ldots, f_{b_{N_b}}(\tau))^\top$, with $f_{b_k}(\tau) = 1/(1+\exp(-\alpha_k \tau))$, $\tau \in [0, 15]$ and $\alpha_k \sim \mathcal{N}(0, .06)$, and we sampled N individual noisy trajectories at time points τ_k^j: $\mathbf{y}_k^j(\tau_k^j) = f_k(\tau_k^j) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For each individual we used the same initial sampling time point for every biomarker, while the number of samples per biomarker was allowed to independently vary between 1 and 4. The individual time points were subsequently centered by their mean μ_k^j to obtain shifted time-points $t_k^j = \tau_k^j - \mu_k^j$ defined in the interval $[-2, 2]$.

The model was applied to estimate biomarker progressions and individual time-shifts with respect to different combinations of trajectory noise σ, sample size N, and number of biomarkers N_b. The accuracy of the model in reconstructing the original time series was quantified by Pearson's correlation between the estimated time-shift d^j and the original individual time reference. The experiments were repeated 10 times for each configuration of parameters $\sigma \in \{0.1, 0.2, 0.3, 0.4\}$, $N_b \in \{4, 8\}$, and $N \in \{20, 100\}$.

AppendixA.2.1. Results.

Table A.4 reports summary correlations between time-shift estimation and the ground truth individual sampling time. The correlation values are generally high, and increase with lower noise levels. Interestingly, the increase in number of modelled biomarkers is associated with a better performance in recovering the underlying disease staging. We also observe that larger sample sizes are associated with higher correlation values, especially with increasing noise levels. We note however an exception for the case $\sigma = 0.1$ where, although the overall performance is still high, the correlation slightly decreases with $N = 100$.

		$N = 20$				$N = 100$			
		σ				σ			
		0.1	0.2	0.3	0.4	0.1	0.2	0.3	.4
N_b	4	.95 (.03)	.86 (.08)	.71 (.17)	.46 (.29)	.91 (.04)	.89(.04)	.76 (.17)	.75 (.12)
	8	.97 (.01)	.91 (.06)	.86 (.06)	.66 (.3)	.94 (.04)	.94 (.02)	.88 (.06)	.84 (.07)

Table A.4: Mean (sd) R^2 correlation coefficient across folds between estimated individual time-shifts and ground truth time reference.

AppendixA.2.2. Model benchmarking with respect to GRACE

The R package GRACE (v 1.0) was used to estimate the multivariate biomarker progression curves from the training set used in this study, by using default parameters and syntax:

```
grace.simulation.fits <- with(output_table,...
grace(Month, Y, Outcome, RID, group, plots = TRUE))
```

Figure A.7 shows the relationship between the estimated individual time-shift. Although the time range estimated by the GP model is roughly double with the respect to the GRACE one, there is a strong agreement between the *relative* positioning of the training individuals along the disease trajectory.

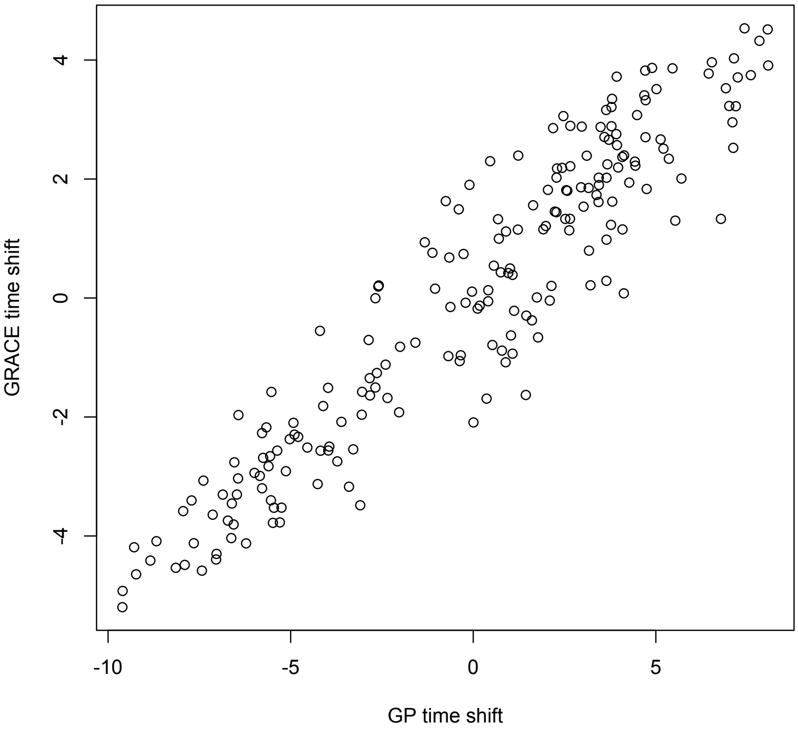


Figure A.7: Comparison between the shift estimated with our GP progression model (x-axis), and the one estimated by GRACE (y-axis). Although the time range estimated by the GP model is roughly double with the respect to the GRACE one, there is a strong agreement between the *relative* positioning of the training individuals along the disease trajectory.

AppendixA.2.3. Predictive performance under uncertain biomarker distribution

In this section we illustrate the relationship between the variability across an individual's biomarkers profile at a given time point and the subsequent time shift estimation. This point is tested in the following synthetic cases, where we considered three hypothetical baseline scenarios:

1. **Homogeneous, low severity**: all the biomarkers measurements for the individual correspond to the 10th percentile of the respective distribution
2. **Homogeneous, high severity**: all the biomarkers measurements for the individual correspond to the 90th percentile of the respective distribution
3. **Heterogeneous**: all the biomarkers measurements for the individual correspond to the 10th percentile of the respective distribution, while FAQ and ADAS are at the 90th percentile.

Figure A.8 illustrates the resulting log-likelihood of the prediction. We can see that for the homogeneous scenarios (blue and red curve), the log-likelihood is high and concentrated on the left and right extremities of the time axis, to indicate indeed greater confidence about low and high severity of the individual. On the contrary, the heterogeneous case (green curve) is characterized by (magnitude) lower log-likelihood values, and by an almost uniform profile across the time axis, to indicate higher uncertainty about the staging prediction.

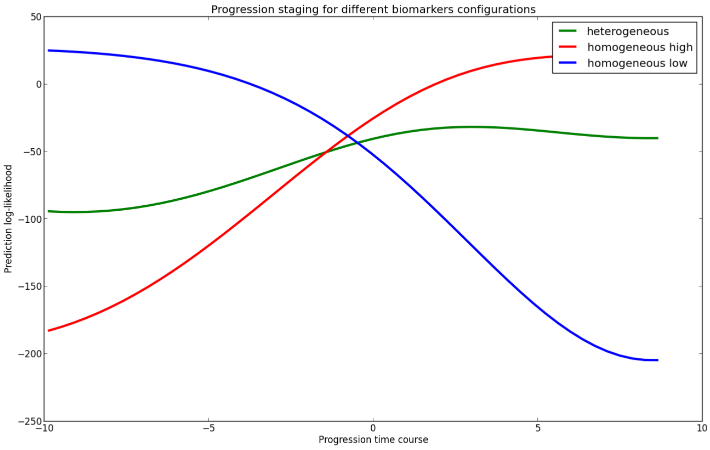


Figure A.8: Predictive uncertainty with respect to individual's biomarkers variability

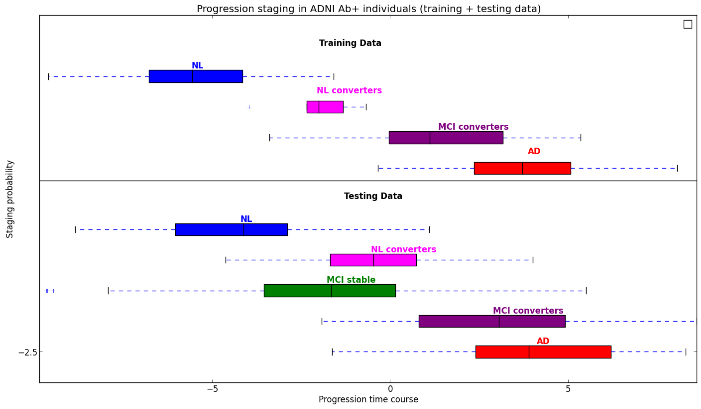


Figure A.9: Comparison between the time shift distribution in training and testing data. The group-wise distribution and ordering of the predicted time-shifts in the testing data are compatible with those estimated in the training one.

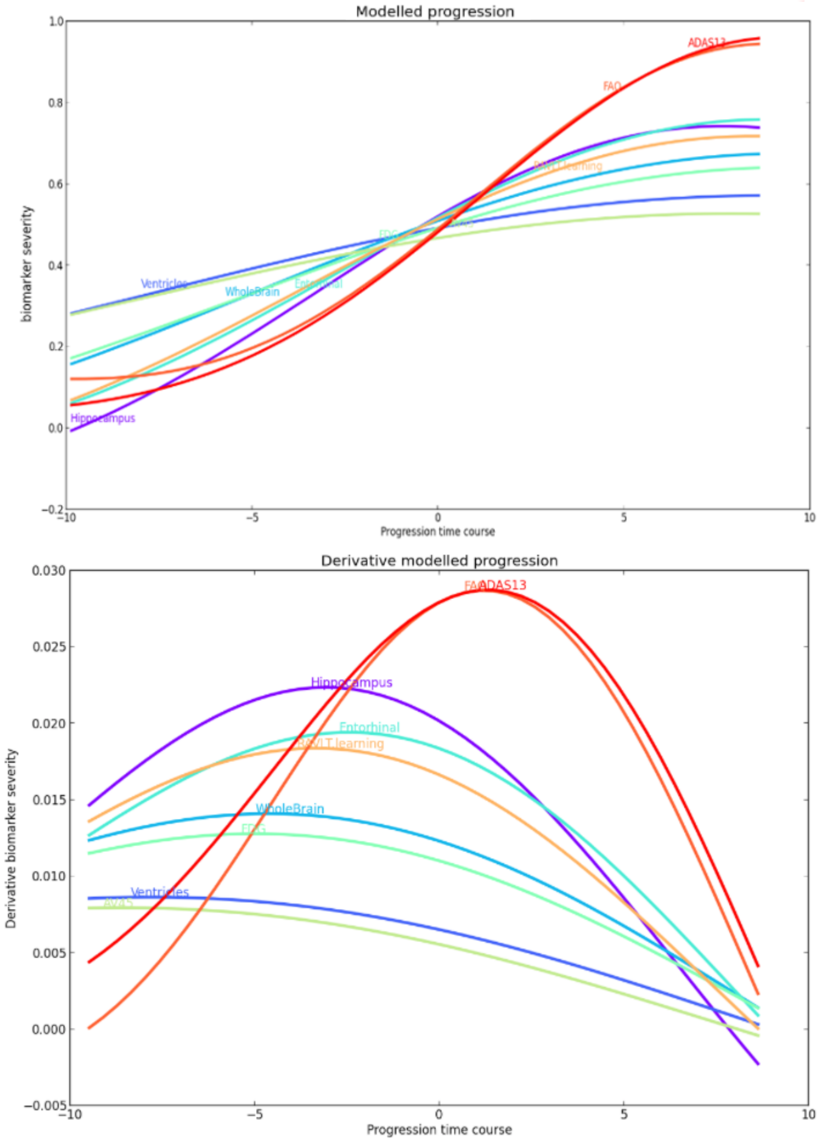
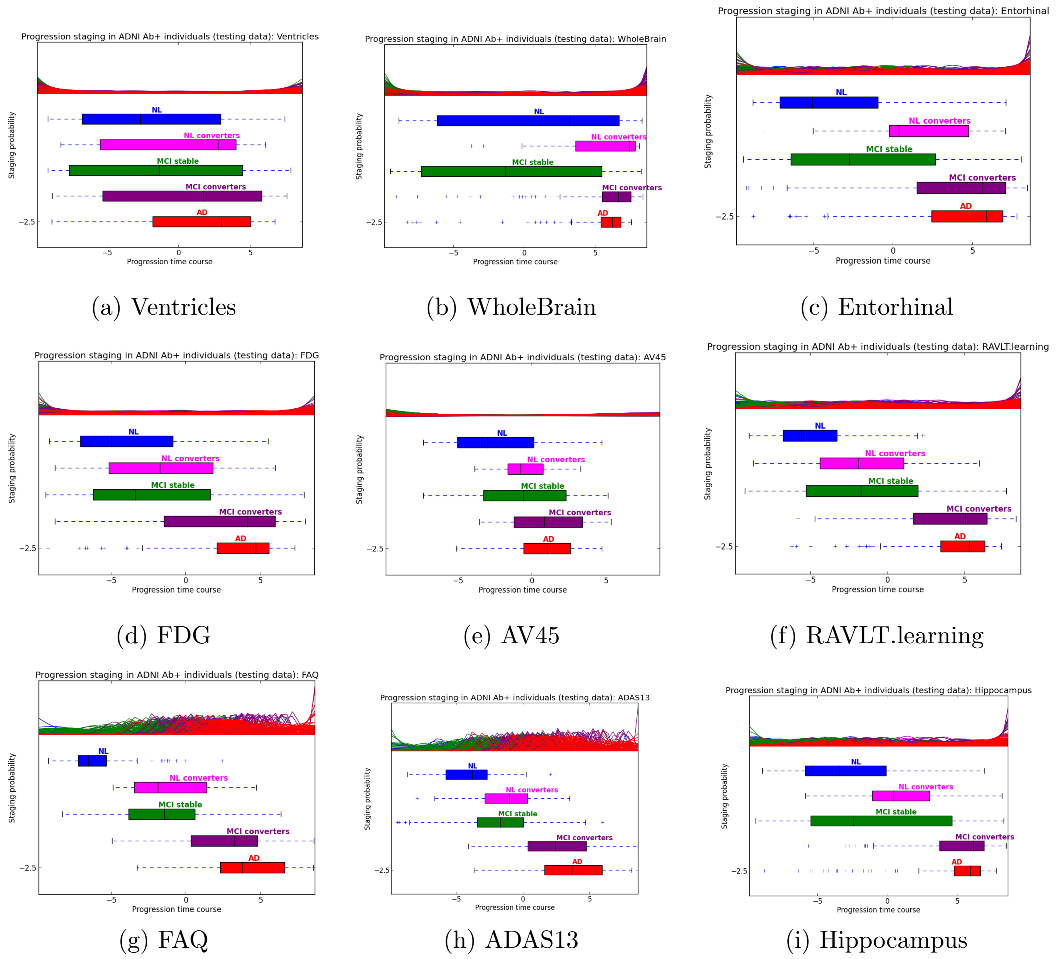


Figure A.10: Joint temporal progression of the biomarkers (top) and derivative of the modelled average trajectory (bottom). ADAS13 and FAQ are characterised by very similar longitudinal profiles, and show the largest changes in the latest stages of the pathology (peak of the derivative at t>0). On the contrary, the change in hippocampal volume is more strongly associated with the earlier stages of the pathology. AV45 and ventricles volumes are the least informative and are associated with the lowest changes (lowest derivative values).



(a) Ventricles

(b) WholeBrain

(c) Entorhinal

(d) FDG

(e) AV45

(f) RAVLT.learning

(g) FAQ

(h) ADAS13

(i) Hippocampus

Figure A.11: Posterior prediction on testing data by using a single biomarker and the follow-up information.

	NL vs AD (145 vs 74)	MCI conv vs MCI stable (106 vs 243)	NL conv vs NL stable (17 vs 74)
Accuracy	.95	.69	.86
Sensitivity	.94	.89	.28
Specificity	.97	.61	.97
AUC	.96	.75	.62

Table A.5: Classification results by using a random forest classifier trained on the whole set of biomarkers for the baseline training data. The missing values in the testing data were imputed by nearest neighbour search.

AppendixA.2.6. Data preparation

```
#Loading ADNIMERGE library

library("ADNIMERGE")

#Identifying clinical subgroups

ridAD = unique(subset(adnimerge,DX=="Dementia")$RID)
ridNL = unique(subset(adnimerge,DX=="NL")$RID)
ridMCI = unique(subset(adnimerge,DX=="MCI")$RID)

#In the next steps converted/reverted individuals are manually identified
# and clinical groups are defined accordingly

ADreverted = c(167, 1226, 4641)
ridAD = ridAD[!ridAD%in%ADreverted]

NLconverted = c(15, 22, 35, 55, 61, 106, 112, 127, 156, 171, 210, 223, 232,
259, 420, 454, 459, 467, 520, 545, 548, 555, 558, 602, 605, 622, 680, 722,
778, 779, 842, 843, 883, 899, 920, 972, 985, 1063, 1123, 1169, 1190, 1194,
1200, 1202, 1203, 2150, 4041, 4071, 4092, 4218, 4262, 4385, 4474, 4506, 4566,
4577, 4579, 4652, 4855, 5096, 5121, 5207, 5273)
ridNL = ridNL[!ridNL%in%NLconverted]


ridConv = subset(adnimerge,RID%in%ridMCI&DX=="MCI to Dementia")$RID
```

```
ridReverter = c(429, 4706)
ridConv = ridConv[!ridConv%in%ridReverter]

ridMCI = c(ridMCI,ADreverted)
ridNConv = ridMCI[which(!ridMCI%in%ridConv)]
ridNConv =         ridNConv[!ridNConv%in%c(ridConv,ridAD,ridNL,NLconverted)]

ridAD = ridAD[-which(ridAD%in%ridConv)]


#Amyloid positive individuals are retained for subsequent analysis

Abpos = read.csv("AbposADNI.csv",skip=1)
ridABpos = Abpos$RID

Set = subset(adnimerge,RID%in%c(ridNConv,ridConv,ridAD,ridNL, NLconverted),
select=c("RID","Month","DX","Hippocampus",
"Ventricles","WholeBrain","Entorhinal","FDG","AV45",
"RAVLT.learning","FAQ", "ADAS13","ICV.bl"))

#Brain volumes are scaled for ICV

Set$Hippocampus = Set$Hippocampus/Set$ICV.bl
Set$WholeBrain = Set$WholeBrain/Set$ICV.bl
Set$Entorhinal = Set$Entorhinal/Set$ICV.bl
Set$Ventricles = Set$Ventricles/Set$ICV.bl

```

```
Set = subset(Set,select=c("RID","Month","DX","Hippocampus","Ventricles",
"WholeBrain","Entorhinal","FDG","AV45", "RAVLT.learning","FAQ", "ADAS13"))

#Identifying individuals with at least one measurements for each biomarker
# (training set)

RIDnoNA = subset(Set,Month==0)$RID[which(apply(is.na(subset(Set,Month==0)),
1,any)==FALSE)]

SetnoNA = subset(Set,RID%in%ridABpos&RID%in%RIDnoNA&RID%in%
c(ridConv,ridAD,ridNL,NLconverted))

#Sampling training set composed by 200 individuals, and testing set composed by
#remaining ones

trainRID = sample(unique(SetnoNA$RID),200)
trainSet = subset(Set,RID%in%trainRID)
testSet = subset(Set,!RID%in%trainRID&RID%in%Abpos$RID)

#Ranking of biomarkers values according to
# training set distribution



newSet = trainSet

for (i in seq(4,length(names(newSet))))
    {
        newSet[,i] = rank(newSet[,i],na.last='keep')/
        length(newSet[,i][which(!is.na(newSet[,i]))])
        }



newSet_test = testSet

for (i in seq(4,length(names(newSet)))){
        for (j in seq(1,length(newSet_test[,i])))
        {
        if (!is.na(testSet[j,i]))
                {
                newSet_test[j,i] = rank(c(testSet[j,i],trainSet[,i]),
                na.last='keep')[1]/
                (length(trainSet[,i][which(!is.na(trainSet[,i]))])+1)
                }

        }
}


# Scaling the biomarkers to increasing abnormality order

newSet$FDG = 1-newSet$FDG
newSet$Hippocampus = 1-newSet$Hippocampus
newSet$WholeBrain = 1-newSet$WholeBrain
```

```
newSet$Entorhinal = 1-newSet$Entorhinal
newSet$RAVLT.learning = 1 - newSet$RAVLT.learning

newSet_test$FDG = 1-newSet_test$FDG
newSet_test$Hippocampus = 1-newSet_test$Hippocampus
newSet_test$WholeBrain = 1-newSet_test$WholeBrain
newSet_test$Entorhinal = 1-newSet_test$Entorhinal
newSet_test$RAVLT.learning = 1 - newSet_test$RAVLT.learning


# Output

write.csv(newSet,"ADNIDataTrain.csv")
write.csv(newSet_test,"ADNIDataTest.csv")

write.csv(ridAD,"ridAD.csv",row.names=FALSE)
write.csv(NLconverted,"ridNLconverted.csv",row.names=FALSE)
write.csv(ridConv,"ridConv.csv",row.names=FALSE)
write.csv(ridNL,"ridNL.csv",row.names=FALSE)
write.csv(ridNConv[ridNConv%in%ridABpos],"ridNConv.csv",row.names=FALSE)
```