Benchmarking Parametric Models of Disease Progression for Early Detection of Cognitive Decline

Carlos Platero[1,2] and Jorge Bengoa[2]

[1]Health Science Technology Group, Universidad Politécnica de Madrid, España
[2]Escuela Técnica Superior de Ingeniería y Diseño Industrial, Universidad Politécnica de Madrid, España

October 25, 2025

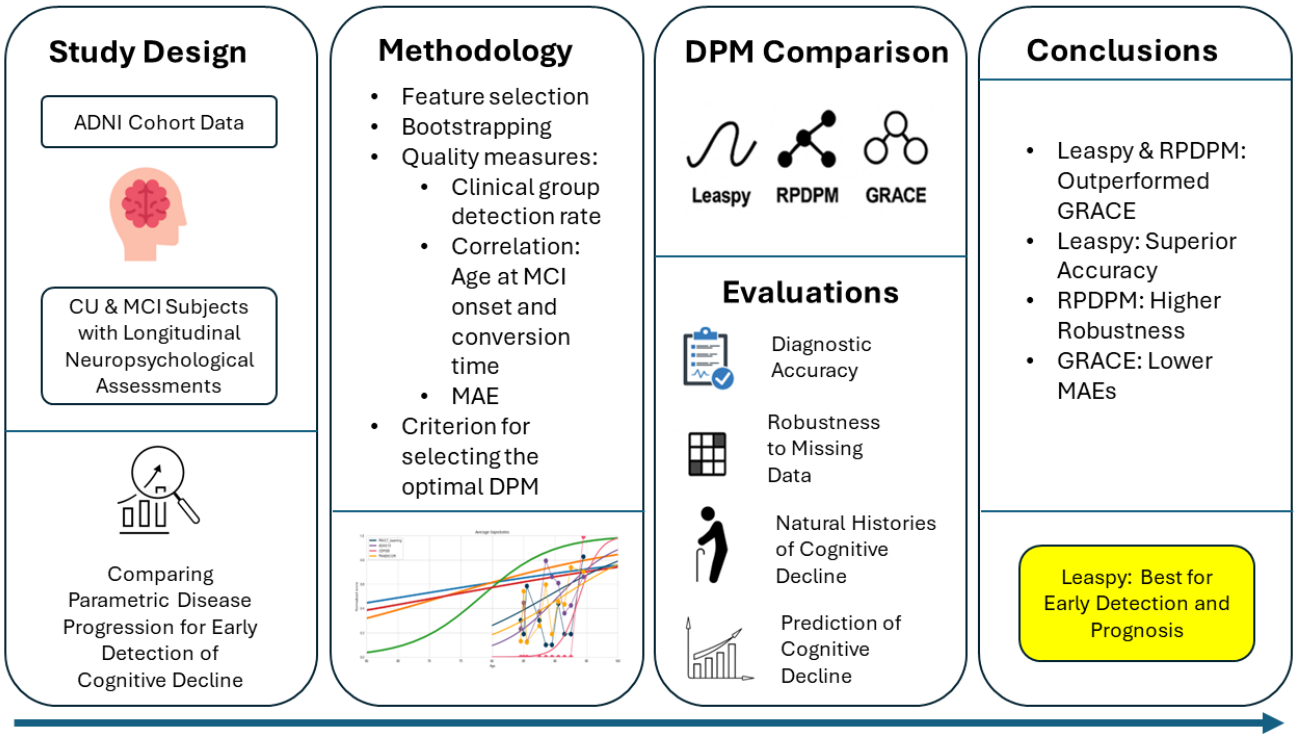

Figure 1: Graphical abstract summarizing the overall methodology. From left to right: (1) longitudinal neuropsychological data from cognitively unimpaired (CU) and mild cognitive impairment (MCI) subjects are collected; (2) multiple disease progression models (DPMs) are trained and evaluated; and (3) their performance is compared in terms of diagnostic accuracy, prediction of cognitive decline, and robustness to missing data.

Highlights

- We assess optimal neuropsychological subsets for disease progression modeling
- Evaluation criteria target both diagnostic accuracy and prognostic performance
- Marker subsets are selected based on accuracy in early diagnostic stage discrimination
- Leaspy excels overall; RPDPM is robust; GRACE performs well in reconstruction
- Publicly available code enables clinical adoption and external validation

Abstract

Background and Objective: Disease progression models (DPMs) are valuable tools for characterizing early cognitive decline in Alzheimer's Disease (AD) and supporting clinical decision-making. This study aimed to (1) evaluate the diagnostic and prognostic performance of parametric DPMs, (2) identify optimal subsets of neuropsychological markers for DPM construction, and (3) benchmark three parametric DPM frameworks in early detection tasks.
Methods: We analyzed longitudinal neuropsychological data from 1163 participants classified as cognitively unimpaired (CU) or with mild cognitive impairment (MCI). Three DPM approaches (Leaspy, RPDPM, and GRACE) were trained on selected marker subsets and evaluated using metrics related to diagnostic accuracy, time to conversion estimation, and robustness to missing data. Model performance was assessed via detection rates, area under the curve (AUC), mean absolute error (MAE), and Pearson correlation between estimated/observed onset ages.
Results: Leaspy achieved the highest diagnostic accuracy with an AUC of 0.96 and strong correlation with observed conversion time ($r = 0.78$). RPDPM showed superior robustness to missing data and maintained accurate predictions even with up to 40% data loss. GRACE offered the best trajectory fit (lowest error) but lower sensitivity to clinical transitions. A compact combination of neuropsychological tests, particularly CDRSB, ADAS13, and MMSE, was sufficient for reliable model training. Prognostic evaluation demonstrated that Leaspy provided the most consistent identification of individuals who converted to mild cognitive impairment within five years.
Conclusions: Parametric DPMs based solely on neuropsychological measures can effectively support early detection and prognosis of cognitive decline. Leaspy showed the best overall performance, while RPDPM proved more resilient to missing data. These models enable individualized disease timelines and can inform clinical decision-making and patient stratification. All code and data used are publicly available, facilitating reproducibility and clinical translation.

Keywords: Cognitively Unimpaired, Mild cognitive impairment, Alzheimer's disease, ADNI, Disease Progression Modeling, Parametric models

1 Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disorder in older adults. As a multifactorial condition with diverse risk factors, AD shows substantial heterogeneity, where markers are essential for diagnosis, prognosis, and early intervention. Revised criteria distinguish the clinical syndrome from its biological substrate [1, 2]. AD is defined by progressive cognitive decline and key neuropathological features, amyloid-β (Aβ) plaques, tau tangles, and neurodegeneration. Symptoms reflect disease progression but are not required for AD diagnosis. Neuropsychological tests assess cognitive impact, while imaging and fluid biomarkers reveal underlying pathology [3, 4, 5].

AD progresses non-linearly over decades, from normal cognition to dementia. Symptoms such as memory loss and functional decline often emerge long after initial brain changes [6]. Although its etiology remains unclear and no cure exists, recent models propose a non-deterministic progression shaped by interactions among amyloid, tau, genetics, and environment [7]. This complexity challenges the amyloid cascade hypothesis and underscores the need for personalized approaches. Early diagnosis and staging are critical for managing symptoms, identifying at-risk individuals, and advancing research.

Researchers use cross-sectional studies, predictive models [8], and increasingly, disease progression models (DPMs) to study AD. DPMs are particularly valuable when disease etiology is complex or poorly understood. These models leverage longitudinal data to quantitatively describe disease evolution and improve forecasting [9, 10, 11, 12, 13, 14, 15]. Inspired by theoretical frameworks describing marker trajectories across AD stages [16, 2], DPMs assume individuals follow a common trajectory with varying onset times and progression speeds. Their reliability, however, depends on proper calibration, validation, and sensitivity analyses to refine predictions and manage uncertainty [17].

DPMs can predict time-to-conversion to clinical stages, supporting personalized diagnosis and prognosis. While analogous to survival analysis, DPMs face challenges with longitudinal data, including high dimensionality, missing data, and temporal misalignment. Multivariate approaches address these by mapping outcomes onto a unified latent timeline, the Disease Progression Score (DPS), enabling direct comparison across markers with different sensitivities at various disease stages [18]. In summary, DPMs offer promising tools for: (a) validating disease models; (b) analyzing risk factors; (c) informing diagnosis and prognosis; (d) evaluating marker sensitivity to progression; and (e) identifying subgroups based on disease trajectories.

DPMs have been developed using a wide range of approaches. These can be broadly categorized into parametric models [9, 14, 15], which assume predefined (e.g., quasi-sigmoidal) biomarker trajectories, and non-parametric or data-driven models, including deep learning frameworks [19, 18, 20, 21, 22, 23]. Parametric methods, which are the focus of this study, offer key advantages in clinical contexts: they are generally simpler to implement, more computationally efficient, and their explicit mathematical formulation enhances interpretability. Furthermore, they perform robustly even with the sparse and irregularly sampled data typical of longitudinal clinical studies. However, their primary limitation is that the strong assumptions on trajectory shapes may not fully capture the complex, heterogeneous patterns of disease progression [14]. In contrast, non-parametric and deep learning approaches offer greater flexibility by learning trajectory shapes directly from the

data without prior assumptions. While powerful, these models often require larger datasets for effective training, are more susceptible to overfitting with sparse data, and typically function as "black boxes," posing significant challenges for clinical interpretability [24]. This study focuses on benchmarking parametric DPMs to establish their utility and performance boundaries in data-limited, early-stage clinical scenarios where model transparency is paramount.

Parametric DPMs often rely on longitudinal mixed-effects models [25], which capture population-level trends while accounting for individual variability. A major challenge is defining an appropriate time scale. Many models address this via temporal re-parameterization, aligning subjects using patient-specific time shifts [9, 12, 13]. More recent work incorporates not only onset timing but also variability in progression rates, improving model flexibility [15].

Developing DPMs faces two key challenges: missing data and irregular patient visit intervals [14, 24]. Parametric models address these by estimating trajectories directly from observed data, enabling reconstruction of a continuous disease timeline at the individual level. This allows for robust imputation and prediction of future time points.

The choice of markers used to train a DPM strongly influences the inferred disease trajectory. Traditionally, this has been explored through feature ablation, where the effect of removing markers is assessed [9, 19] or approaches that rely on feature preselection using predictive models [26]. In contrast, we propose a combinatorial approach, training DPMs on subsets of varying marker combinations. To evaluate performance, we introduce metrics that quantify both diagnostic and prognostic accuracy, enabling selection of the configuration that best predicts disease progression.

While numerous DPMs have been proposed, a critical gap remains in their standardized evaluation, particularly for the early detection of cognitive decline, an early and often subtle phase of AD, less detectable than the conversion to dementia [27]. The lack of a comprehensive benchmarking framework makes it difficult to compare the performance of different models and identify the most suitable approaches for clinical application. To address this limitation, this study has three primary objectives. First, we propose and implement a robust framework for evaluating both the diagnostic and prognostic performance of DPMs using clinically meaningful metrics. Second, we leverage this framework to identify optimal, low-dimensional subsets of neuropsychological markers that balance model accuracy with clinical feasibility. Third, we apply this methodology to benchmark three distinct parametric DPM frameworks: the established GRACE [9] and the more recent Leaspy [15] and RPDPM [14], trained and evaluated using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort [28]. Our contributions therefore focus on: (a) proposing procedures to evaluate DPMs; (b) selecting optimal marker sets for model construction; and (c) comparing DPM frameworks based on time-to-conversion estimates and trajectory reconstruction accuracy.

2 Methods

2.1 Study Participants and Data Source

This study utilized data from the ADNI database [28]. Clinical, demographic, and neuropsychological data for all participants were obtained through the R package ADNIMERGE [29]. To construct the DPMs, we selected two primary groups from the ADNI-1 to ADNI-3 phases: individuals classified as cognitively unimpaired (CU) at baseline and those diagnosed with Mild Cognitive Impairment (MCI) at baseline. All included participants had undergone longitudinal follow-up and completed a comprehensive battery of neuropsychological assessments, with data available from at least two time points.

Participants were followed longitudinally, and their clinical diagnoses at each visit were used as the reference standard. CU individuals who converted to MCI during follow-up were classified as progressive CU (pCU), while those who remained cognitively stable were designated as stable CU (sCU). The classification into stable or progressive groups was determined solely based on the initial and final diagnoses, without accounting for potential intermediate fluctuations in cognitive status. Additionally, MCI participants whose last visit indicated either reversion to CU or progression to dementia were excluded from the analysis to specifically characterize the transition from normal cognition to MCI.

Based on these criteria, the final cohort included 536 sCU, 114 pCU, and 513 stable MCI (sMCI) participants, contributing a total of 6,255 visits. The four-year conversion rate from CU to MCI was 10% (n = 66/650). Table 1 presents summary measures of baseline features for each diagnosis group; a complete version with additional neuropsychological tests and detailed statistical comparisons is available in the Supplementary Material (Table S.1). The identification of the selected participants within the ADNI cohort is available at: `https://github.com/cplatero/preAD_DPM`.

2.2 Neuropsychological assessment

This study focused exclusively on neuropsychological measures, as these are the most appropriate tools for tracking clinical symptoms from CU to MCI stage [30, 5]. A total of eleven neuropsychological tests were included in the analysis. These comprised the Mini-Mental State Examination (MMSE), the Clinical Dementia Rating Sum of Boxes (CDRSB), the Alzheimer's Disease Assessment Scale–Cognitive Subscale 13 (ADAS13), the Functional Assessment Questionnaire (FAQ), and the Logical Memory Delayed Recall Total Score (LDELTOTAL). In addition, performance on the Trail Making Test Part B (TRABSCOR), the Modified Preclinical Alzheimer Cognitive Composite (PACC), and several indices derived from the Rey Auditory Verbal Learning Test (RAVLT) were considered. These RAVLT-derived measures included Immediate Recall, Learning, Forgetting, and Percent Forgetting scores.

Table 1: Baseline demographic and clinical characteristics of the studied population. Data are mean (SD) (minimum, maximum) or n (%). Group comparisons were conducted using the Wilcoxon rank-sum test with a Bonferroni correction or ANOVA with Bonferroni post hoc test is used for baseline age, education, neuropsychological score, except for gender and APOE4 convert where the chi-square test is used. Statistical significance is considered with $p - value \leq 0.01$. (a) Significant compared sCU to sMCI, (b) pCU vs. sMCI, (c) sCU vs. pCU.

Demographic and clinical characteristics				
Group	sCU	pCU	sMCI	Post-Hoc
Subjects	536 (46.1%)	114 (9.8%)	513 (44.1%)	
Visits	2677 (42.7%)	795 (12.7%)	2783 (44.6%)	
Female	298 (55.6%)	52 (45.6%)	203 (39.6%)	a
APOE4	155 (n=531 29.2%)	41 (n=114 36.0%)	219 (n= 506 42.7%)	a
Age	73.00 (6.16) (56.20 90.30)	75.50 (5.66) (63.20 90.10)	72.99 (7.58) (55.00 91.40)	b,c
Key Cognitive Outcomes				
MMSE	29.08 (1.10) (24.00 30.00)	28.96 (1.23) (24.00 30.00)	27.86 (1.82) (19.00 30.00)	a,b
CDRSB	0.03 (0.13) (0.00 1.00)	0.06 (0.17) (0.00 0.50)	1.30 (0.76) (0.00 5.50)	a,b
ADAS13	9.57 (4.32) (0.00 23.00)	11.48 (4.97) (1.00 29.33)	15.20 (5.92) (3.00 37.00)	a,b,c

2.3 Parametric DPM

We denote the marker measurement k for individual i at time j as y_{ijk}, where $i = 1, \ldots, n$, $k = 1, \ldots, p$ and $j = 1, \ldots, q_{ik}$. The model can be written as:

$$y_{ijk} = f(\boldsymbol{z}_i, t_{i,j}, \boldsymbol{\Theta}_k) + e_{ijk},$$

where f corresponds to the model, z_i are variables associated to the i-th subject, $t_{i,j}$ is the age of subject i in visit j, Θ_k are the average trajectory parameters, and e_{ijk} is an error term.

In general, the aim is to identify the parameters that most accurately represent the observations, such as those that optimize the likelihood [31, 14] or optimization the robust nonlinear regression [9].

2.3.1 Leaspy

It combines the concept of time-warp with translations of curves. The transformation of model to the subject's data consist in computing (1) a time-shift, δ_i, which translates the model along the time axis to accommodate for changes in age at disease onset, (2) an acceleration factor α_i, which scales the time interval to accommodate for differences in speed of progression, and (3) intermarker spacings $\boldsymbol{w}_{ik}$, which translate each biomarker differently to accommodate for differences in the timing and ordering among biomarkers. The first two parameters define a time-warp function that maps the chronological age of a subject to their Alzheimer Age, as defined by Koval et al. [15]. This personalized timescale represents a patient's position along a standardized disease progression timeline, independent of their actual age. The function therefore adjusts the dynamics of an individual's progression (i.e., accelerating or decelerating time) to align them with a common trajectory. In turn, $\boldsymbol{w}_{ik}$ shift the patient trajectories to account for phenotypic differences across subjects [15].

The long-term disease progression corresponds to the average population trajectory (γ_0), which is parameterized by a set of fixed-effect parameters: $\boldsymbol{p}_0, t_0$ and $\boldsymbol{v}_0$ [15]. Following the logistic model, t_0 represents the average age at the inflection point of the trajectory (i.e., the time of maximum progression velocity), serving as a temporal anchor. $\boldsymbol{p}_0$ defines the average normalized biomarker value (position) at time t_0, and $\boldsymbol{v}_0$ corresponds to the average speed (velocity) of progression at that same time point. We use normalized scores on marker values ensuring that geodesics take the form of a logistic curve for each coordinate, so with $p_k = \gamma_k(t_0)$ and $v_k = \dot{\gamma}_k(t_0)$ the multivariate model writes:

$$\gamma_k(t) = \left(1 + \left(\left(\frac{1}{p_k} - 1\right) \cdot exp\left(-\frac{u_k \cdot (t - t_0)}{p_k \cdot (1 - p_k)}\right)\right)\right)^{-1}$$

On the other hand, the function $f(\boldsymbol{z}_i, t_{i,j}, \boldsymbol{\Theta}_k)$ corresponds to the geometrical description of the individual trajectory on the Riemannian manifold, which produces the spatio-temporal transformation from the long-term into the i-short-term trajectory:

$$f(\boldsymbol{z}_i, t_{i,j}, \boldsymbol{\Theta}_k) = \left(1 + \left(\left(\frac{1}{p_k} - 1\right) \cdot exp\left(-\frac{w_{i,k} + u_k \cdot \alpha_i \cdot (t_{i,j} - \delta_i - t_0)}{p_k \cdot (1 - p_k)}\right)\right)\right)^{-1}$$

In order to estimate the diagnosis at each patient consultation, a linear transformation is also used that relates the patient's age to the time of disease progression, which is denoted as disease progression score (DPS):

$$s_{i,j} = \alpha_i \cdot (t_{i,j} - (\delta_i + t_0))$$

2.3.2 RPDPM

In the case of RPDPM [14], the geometric model for the trajectories is defined by:

$$f(s; \mathbf{\Theta}_k) = (a_k - d_k) g\left(s; \mathbf{\Theta}_k\right) + d_k$$

where a_k and d_k are normalization scalar parameters for each k-marker, s is the DPS and $g(s; \mathbf{\Theta}_k)$ represents a logistic function and $\mathbf{\Theta}$ is a vector parameter for each k-marker. DPS transforms, $s_{i,j}$, the patient's age $t_{i,j}$ into a timescale that defines the progression of the disease, similar to Leaspy, where α_i and β_i represent the temporal scalar parameters for the i-th patient:

$$s_{i,j} = \alpha_i \cdot t_{i,j} + \beta_i.$$

Finally, the multiobjective optimization for robust nonlinear regression is defined as:

$$\left\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{\Theta}}\right\} = \min_{i,j,k} \sum_{i,j,k} \omega_i \rho \left(\frac{y_{ijk} - f(z_i, t_{i,j}, \Theta_k)}{\sigma_k} \right)$$

where z_i is the individual parameter set, $\{\alpha_i, \beta_i\}$; $\rho(\cdot)$ is a maximum likelihood-type function and $\omega_i = 1/N_i$ is a weighting factor for normalizing the objective function with the number of available points per subject (N_i).

2.3.3 GRACE

The proposal of Donohue et al. [9] is expressed as:

$$f(\boldsymbol{z}_i, t^v_{i,j}, \mathbf{\Theta}_k) = g_k(t^c_{ijk} + \delta_i) + x'_{t^c_{ijk}} \beta_k + \alpha_{0ik} + \alpha_{1ik} t^c_{ijk}$$

where g_k is a continuously differentiable monotone function and δ_i is the unknown subject-specific time shift, which follows a normal distribution with mean zero and variance σ^2_δ. Short-term observation time is represented by t^c_{ijk}, which indicates the centered years in relation to the temporal evolution of the visits, $t^c_{i,j} = t^v_{ij} - (t^v_{i_1} + t^v_{i_{end}})/2$, where $t^v_{i,j}$ is the visit time and i_1 and i_{end} represented the first and last visit index of $i-$subject. In GRACE, DPS is defined as $s_{i,j} = t^c_{i,j} + \delta_i$. It is worth noting that both Leaspy and RPDPM use patients' age as the temporal variable, $t_{i,j}$; however, in GRACE, time refers to the interval between visits, $t^v_{i,j}$. $x'_{t^c_{ijk}}$ is the row vector for the fixed effects (including variables such as age and scan time), and β_k are the fixed effects coefficients. The parameters α_{0ik} and α_{1ik} are the subject- and outcome-specific random intercept and slope. These values reflect how the subset of regression parameters for the ith subject deviates from those of the population.

2.4 DPM's performances

Evaluating the performance of DPMs is essential and can be addressed using two complementary measures. First, by examining the model's ability to track diagnostic transitions over time, including estimating the conversion events such as the transition from CU to MCI or dementia. Second, by assessing the accuracy of the estimated temporal trajectories of patients' markers.

2.4.1 Clinical status classification

DPMs model marker trajectories using unsupervised learning. However, clinical status is inferred through a Bayesian classifier, which follows a supervised learning approach. Given the DPS at the j-th visit of subject i, denoted as $s_{i,j}$, the diagnosis is estimated using a Bayesian classifier based on class-conditional likelihoods fitted via kernel density estimation (KDE) according to Parzen's method [32]:

$$p(s|c) = \frac{1}{n \cdot h} \sum_{i,j} K\left(\frac{s - s_{i,j}}{h}\right)$$

where $K(\cdot)$ is the kernel function, $h \geq 0$ is the smoothing parameter, and c denotes a clinical group. The values $s_{i,j}$ correspond to training samples associated with class c, and n is the sample number. The predicted class for a test observation $s_{i,j}^{test}$ is the one that maximizes the posterior probability:

$$r = \arg\max_r \left(p(c_r) \cdot p(s_{i,j}^{test}|c_r)\right).$$

2.4.2 Evaluations of estimated trajectories

The accuracy of the estimated temporal trajectories for patient markers is evaluated using the Mean Absolute Error (MAE), a performance metric known for its robustness to outliers [33]. MAE is defined as the average absolute difference between observed and predicted values:

$$\text{MAE}_k = \frac{1}{n_k} \sum_{i,j} |y_{i,j,k} - \hat{y}_{i,j,k}| \tag{1}$$

where n_k represents the total number of measurements available for the k-th marker across all subjects and visits. The terms $y_{i,j,k}$ and $\hat{y}_{i,j,k}$ denote the observed and predicted values, respectively, for subject i at visit j.

2.5 Methodology for Constructing DPMs

The performance of any DPM, irrespective of its underlying framework, is intrinsically dependent on the selected markers for training. To enhance both diagnostic and prognostic capabilities, it is essential to identify the most informative marker combinations. In this study, we trained DPMs using multiple subsets of neuropsychological measures, varying both the combination and dimensionality, to determine the optimal configuration based on predefined progression-related quality metrics.

To evaluate the ability of DPMs to characterize the transition from CU to MCI, we defined the following performance metrics: 1. Percentage of stable individuals correctly identified (i.e., sCU and sMCI groups). 2. Percentage of pCU subjects detected during follow-up. 3. Area Under the Curve (AUC) of the Bayesian-KDE classifier at each visit, based on clinical diagnoses. 4. Pearson correlation between estimated and observed age at transition from CU to MCI in pCU individuals, $\rho\left(\tau_i, \hat{\tau}_i\right) = \frac{\sum_i (\tau_i - \bar{\tau})(\hat{\tau}_i - \bar{\hat{\tau}})}{\sqrt{\sum_i (\tau_i - \bar{\tau})^2}\sqrt{\sum_i (\hat{\tau}_i - \bar{\hat{\tau}})^2}}$, where τ_i and $\hat{\tau}_i$ are the observed and predicted age at cognitive decline onset. 5. Pearson correlation between estimated and time to conversion from CU to MCI in the pCU group, $\rho\left(\tau_i - t_{i,j=1}, \hat{\tau}_i - t_{i,j=1}\right)$.

For metrics 4 and 5, the clinical transition from CU to MCI in pCU subjects (τ_i) was operationalized as the midpoint between the last visit classified as CU and the first visit classified as MCI. DPM-estimated onset age ($\hat{\tau}_i$) was defined as the age at which the model's Bayesian-KDE classifier assigned a higher probability to MCI than to CU based on the DPS.

Conversion time was computed as the difference between this estimated age at MCI onset and the baseline age, $\tau_i - t_{i,1}$. Not all pCU individuals were included in this analysis —only those for whom the DPM predicted a transition to MCI. Accordingly, DPM performance captures both the accuracy of onset age estimation and the sensitivity to detect conversion, which are later used as key criteria for selecting the most suitable models.

In addition to diagnostic metrics, we also computed the MAE between observed and predicted marker values at each visit to assess the trajectory fitting quality.

Fig. 2 provides a schematic overview of the methodology for constructing the DPMs.

2.5.1 Data Splitting, Bootstrapping, and Normalization

To ensure a robust and unbiased evaluation, the dataset was divided into training, validation, and test sets. The split was stratified by clinical diagnosis (sCU, pCU, sMCI) to preserve the original group proportions across all subsets. Model training was performed on the training set, while the validation set was used to assess the candidate DPM. The test set, representing 20% of the data, was reserved for the final out-of-sample performance evaluation. Each model was subsequently individualized for every test subject to estimate biomarker trajectories and cognitive progression. To account for variability in training subsets and increase model robustness, we employed a bootstrapping strategy, generating multiple resampled training sets and training a DPM on each. Throughout this study, performance metrics are reported as the mean and standard deviation (SD) across these bootstrapped models. The SD should therefore be interpreted as the

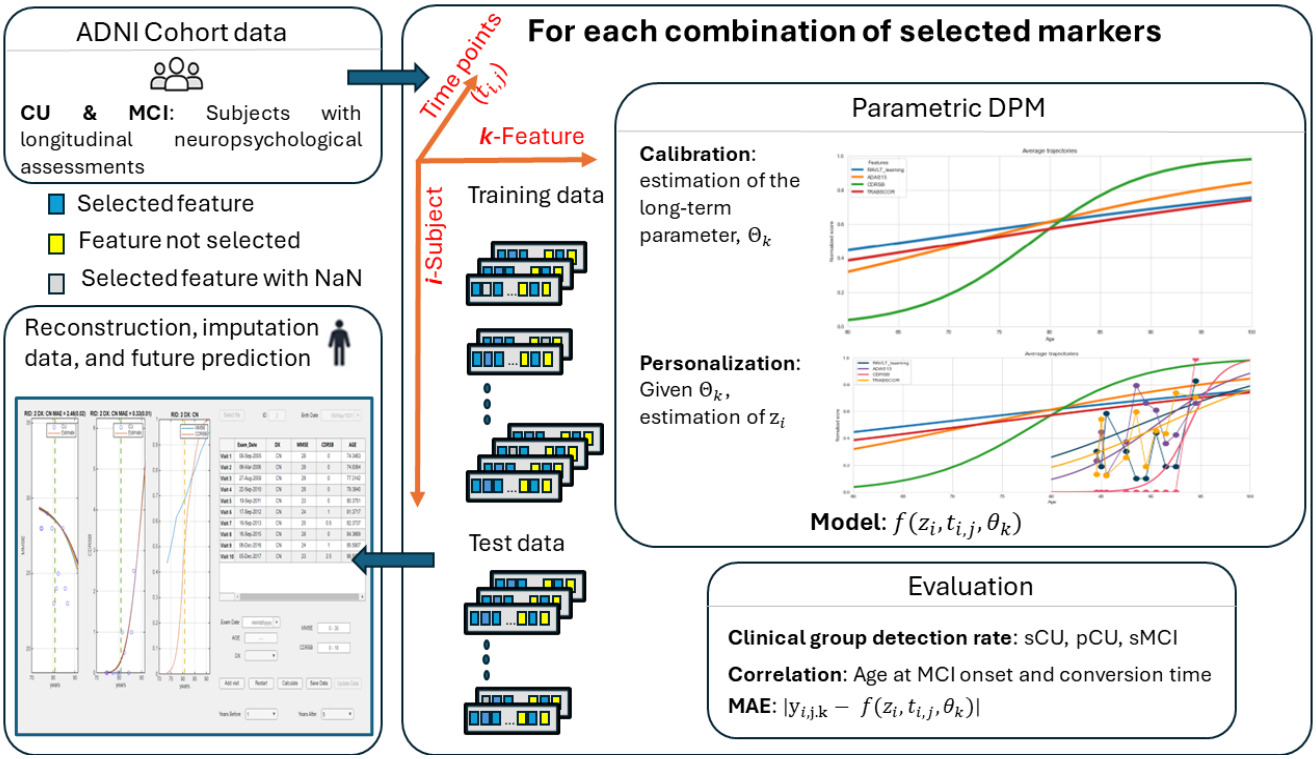

Figure 2: Schematic overview of the methodology for constructing Disease Progression Models (DPMs). The process starts with longitudinal neuropsychological assessments from cognitively unimpaired (CU) and mild cognitive impairment (MCI) subjects. Different DPM approaches (e.g., Leaspy, RPDPM, GRACE) are trained and evaluated based on their ability to capture disease dynamics, predict future cognitive decline, and handle missing data.

primary indicator of the stability and uncertainty of our reported metrics. While the number of bootstrap iterations is a configurable parameter in the implemented software, we decided to construct 10 DPMs for each marker combination. A fixed random seed was used in all experiments to ensure full reproducibility of our results.

In the case of Leaspy and RPDPM, personalization tools were available within their respective packages, allowing direct estimation of DPS and individual marker trajectories for unseen patients. Since the GRACE framework lacks this functionality, we developed a customization procedure to ensure comparability. Specifically, we estimated the subject-specific time shift based on observed markers and long-term model trajectories, and then used linear mixed-effects modeling to determine the intercept and slope for each marker trajectory.

Normalization was required for Leaspy and GRACE, which operate on standardized marker ranges. Each marker was scaled to the [0, 1] interval, where 0 represents the least severe value and 1 the most severe. Normalization parameters were derived exclusively from the training set, using percentile estimates weighted by clinical group proportions. These parameters were then applied to the test set and used for inverse transformation when necessary.

2.5.2 Criteria for Selecting Optimal Marker Combinations

To identify the optimal marker combinations for DPM construction, we developed a novel selection strategy that systematically builds and evaluates all possible models generated from different marker subsets and dimensional configurations. Unlike conventional approaches that assess feature importance post hoc through ablation [9, 19] or rely on a priori supervised preselection [26], our method performs a comprehensive a priori evaluation of entire DPM configurations. The novelty of our approach lies in optimizing model selection for a clinically meaningful objective: accurately predicting conversion to MCI. To this end, we introduced a composite criterion, which is designed to simultaneously balance two key aspects of prognostic performance: the temporal precision of the conversion time estimate and the model's sensitivity in identifying individuals who will progress. This ensures that the selected DPMs are not only statistically robust but are the most effective for the intended clinical application.

Specifically, models were ranked according to two stringent performance metrics: (1) the Pearson correlation between estimated and clinically observed conversion times in pCU subjects, and (2) the proportion of pCU individuals correctly identified. These two metrics were combined into a single composite criterion, defined as the product of the Pearson correlation coefficient and the detection rate of pCU subjects:

$$C = \rho\left(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1}\right) \cdot \text{Sensitivity}_{pCU}$$

Since the number of pCU subjects differed between the training and test sets, and, each DPM was trained on distinct bootstrapped partitions, we evaluated performance on both subsets. To ensure generalizability and avoid overfitting, we selected models with consistent quality scores across both sets, using the geometric mean of the composite criterion as the final selection metric:

$$\text{Criterion} = \sqrt{C_{\text{train}} \cdot C_{\text{test}}}$$

This strategy promotes the selection of robust and well-generalizing DPMs.

2.5.3 Natural Histories of Cognitive Decline as Inferred by DPMs

Each DPM inherently defines a natural history of cognitive decline by temporally aligning patients along a latent disease axis, independently of their clinical diagnosis. This alignment enables retrospective and prospective estimation of disease trajectories, offering a valuable framework to validate hypotheses regarding the temporal ordering of pathological events such as early amyloid or tau accumulation relative to the onset of dementia [34, 5].

To generate these timelines, we first estimated the age of cognitive decline onset, $\hat{\tau}_i$, for each individual. For CU individuals, the timeline was extrapolated into the future; for MCI individuals, it was inferred retrospectively. The onset age was defined as the point at which the probability of MCI exceeded that of CU, based on the Bayesian-KDE classifier applied to the DPS. This estimated onset served as a temporal anchor, $t^*_{i,j} = t_{i,j} - \hat{\tau}_i$, for aligning patients along a unified progression timeline.

Once aligned, we computed the long-term trajectory for each marker by averaging the personalized short-term trajectories across patients. Marker values at the estimated onset of cognitive decline were used to define population-level pathological scores. Additionally, we assessed each

marker's utility as a progression indicator by evaluating the degree of alignment between short-term and long-term trajectories. Markers exhibiting strong concordance were considered more sensitive and reliable for tracking disease evolution.

2.6 Robustness to Missing Data

To assess the robustness of DPMs under conditions of incomplete data, a common occurrence in real-world clinical settings, we simulated varying levels of missingness by randomly omitting values from selected markers during test set personalization. This was done progressively until predefined missing data thresholds were reached, ensuring that each visit retained at least one valid marker to maintain model applicability. This procedure enabled us to evaluate the capacity of each DPM to accurately personalize to new patients without requiring model retraining, thereby simulating the conditions under which these models might be deployed in clinical practice.

2.7 Impact of the Number of Patient Visits on Diagnostic Accuracy

Longitudinal studies are inherently affected by attrition and variability in follow-up duration, resulting in a heterogeneous number of visits across patients. To evaluate how the amount of longitudinal information influences model performance, we stratified the test cohort by the number of available visits used for trajectory personalization. We quantified diagnostic performance and trajectory accuracy using AUC, detection rates for clinical groups (sCU, pCU, sMCI), and MAE for marker predictions as a function of the number of visits available.

2.8 Prognostic Performance

While previous analyses focused on evaluating the diagnostic accuracy of DPMs, this subsection assesses their ability to predict future clinical trajectories. To this end, patient-specific marker trajectories were personalized using data from initial visits, and predictions were compared against observed marker values and clinical outcomes at later time points. For each individual, DPMs were used to forecast future visits by excluding the final visit from the input data, estimating marker values on that date, and comparing them to actual clinical observations. This procedure was extended to simulate forecasting windows of up to five years beyond the most recent visit used for personalization, provided that sufficient follow-up data existed for the patient.

The following methodology was applied to evaluate prognostic performance:

1. Trajectory Personalization: For each patient, the model was personalized using data from the first l visits. Prognostic accuracy was then assessed over subsequent visits, from $l+1$ to the last available time point. All predictions were generated using DPMs trained with previously selected optimal marker combinations.

2. Temporal Binning of Forecasts: Predictions were grouped by the time elapsed since the last visit used for personalization, in one-year intervals, up to a maximum of five years. Each bin

included a minimum of 30 patients to ensure statistical robustness and reduce sampling bias.

3. Prognostic Evaluation Metrics: Forecasted outcomes were evaluated using (a) the AUC of the cognitive status classifier, (b) detection rates for the clinical groups (sCU, sMCI, pCU), and (c) the MAE between predicted and observed marker values.

4. Definition of Clinical Groups: Prognostic outcomes were interpreted using the following classification: (a) sCU: CU diagnosis at both the first and l-th visits, and maintained at the forecasted visit. (b) pCU: CU diagnosis at the first and l-th visits, followed by conversion to MCI at the forecasted visit. (c) sMCI: Includes patients diagnosed with MCI at the first visit, and those who transitioned from CU to MCI by the l-th visit. Importantly, these individuals were not reclassified if their predicted visit diagnosis deviated from MCI (e.g., reversion to CU or progression to dementia).

5. Benchmarking with a Constant Predictor: To contextualize DPM performance, we compared each model to a "constant predictor" that assumes future marker values and clinical states remain unchanged from the last personalized visit. This naive approach achieves perfect detection for stable cases (sCU and sMCI) but fails entirely to detect progression (0% pCU sensitivity). Despite its simplicity, this method is known to perform competitively in short-term forecasts [19].

To specifically evaluate the capacity of DPMs to anticipate cognitive decline in pCU patients, we conducted a targeted analysis in which personalization was restricted to visits prior to the first recorded MCI diagnosis. Forecasts were made using various horizons, from the earliest visits up to five years before conversion. Three metrics were used to assess these predictions: (a) Probability of correctly identifying progression to MCI (pCU detection rate). (b) Correlation between the DPM-estimated and clinically determined age of cognitive decline onset. (c) Correlation between estimated and actual conversion time, defined as the interval between onset age and baseline age.

The estimated age of cognitive decline onset was determined as the point in time at which the Bayesian-KDE classifier assigned a higher probability to MCI than to CU. This analysis was extended to estimate progression up to 10 years beyond the last available visit.

3 Results

The DPMs described in the Methods section were trained and evaluated to assess their diagnostic and prognostic performance in detecting early cognitive decline. We used publicly available implementations of DPMs for this study. Specifically, we employed Leaspy (`https://leaspy.readthedocs.io/en/stable/`), RPDPM (`https://github.com/Mostafa-Ghazi/RPDPM`), and GRACE (`https://bitbucket.org/mdonohue/grace/src/master/`) to model longitudinal trajectories of neuropsychological measures. All code used for model training, feature selection, and evaluation (Leaspy in Python, RPDPM in MATLAB and GRACE in R) is openly available at `https://github.com/cplatero/preAD_DPM`.

3.1 Selection of DPMs and diagnostic performance

Table 2 summarizes the highest diagnostic quality scores achieved in the test set by the best-performing models from each framework. Although the reported metrics correspond to the test set, model selection was performed based on balanced performance across both the training and test sets, following the proposed criterion (see Section 2.5.2 for details). The results show that Leaspy achieved the highest overall performance, with a top AUC of 0.96 and a composite criterion score of 68.7. RPDPM also demonstrated strong performance, with a high pCU detection rate (81.4%) and the best correlation for conversion time ($\rho\left(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1}\right) = 0.88$). In contrast, GRACE consistently under-performed, showing lower pCU detection rates (around 54–64%) and weaker correlations. Notably, models with only two features (e.g., ADAS13 and CDRSB for Leaspy) performed nearly as well as those with four, supporting the use of low-dimensional configurations.

Framework	Features	% sCU	% sMCI	% pCU	AUC	$\rho(\tau_i, \hat{\tau}_i)$	$\rho(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1})$	Criterion
Leaspy	C, P, R, T	77.1 (1.3)	95.4 (0.5)	81.9 (2.5)	0.96 (0.006)	0.99 (0.001)	0.80 (0.011)	68.7
	A, C, M, L	77.1 (1.6)	94.8 (0.6)	80.4 (2.2)	0.96 (0.009)	0.99 (0.002)	0.78 (0.023)	68.2
	A, C	75.4 (1.3)	94.6 (0.7)	81.2 (5.6)	0.95 (0.009)	0.99 (0.001)	0.78 (0.027)	67.4
RPDPM	C, M, R, P	85.8 (1.2)	89.6 (2.1)	81.4 (1.4)	0.94 (0.011)	0.99 (0.002)	0.88 (0.052)	65.6
	C, L, M	90.8 (1.0)	94.1 (0.3)	77.2 (0.2)	0.94 (0.005)	0.99 (0.001)	0.89 (0.184)	63.8
	C, M	87.0 (0.9)	88.9 (1.6)	76.8 (1.4)	0.94 (0.006)	0.99 (0.001)	0.88 (0.031)	63.7
GRACE	C,F,M,R	77.4 (2.7)	74.8 (3.0)	64.0 (6.8)	0.93 (0.005)	0.98 (0.001)	0.62 (0.029)	36.9
	C,M,R	77.3 (1.8)	74.6 (3.4)	54.0 (3.9)	0.92 (0.007)	0.98 (0.001)	0.64 (0.033)	28.8
	C,M	78.6 (3.3)	74.2 (5.3)	54.8 (7.7)	0.92 (0.003)	0.98 (0.002)	0.54 (0.067)	25.2

Table 2: Diagnostic performance metrics of the DPMs on the test population using the best proposals with few markers, across the three evaluated approaches and ordered according to the proposed criterion. The diagnostic accuracy is reported for each clinic group (sCU, sMCI, and pMCI), along with Pearson correlations for the estimation of age at cognitive decline onset and time to conversion. All quality metrics are reported as the mean and standard deviation across DPMs trained via bootstrapping. A = ADAS13; C = CDRSB; F = FAQ; M= MMSE; L = LDELTOTAL; R= RAVLT forgetting; P = PACC; T=TRABSCORE

In terms of trajectory fitting accuracy, Table 3 shows that GRACE consistently yielded the lowest MAE for markers like MMSE (0.85), indicating superior curve-fitting. RPDPM achieved the lowest MAE for CDRSB (0.22). Conversely, Leaspy, despite its strong diagnostic performance, registered the highest MAE for several key markers. This highlights a key trade-off: GRACE excels at reconstruction, whereas Leaspy and RPDPM are better optimized for diagnostic classification.

Framework	Leaspy			RPDPM			GRACE		
features	C,P,R,T	A,C,M,L	A,C	C,M,R,P	C,L,M	C,M	C,F,M,R	C,M,R	C,M
ADAS13	-	2.36(0.02)	2.32(0.01)	-	-	-	-	-	-
CDRSB	0.60(0.04)	0.59(0.04)	0.44(0.08)	0.23(<0.01)	0.23(<0.01)	0.22(<0.01)	0.25(<0.01)	0.25(<0.01)	0.25(<0.01)
MMSE	-	1.15(0.01)	-	1.13(<0.01)	1.16(<0.01)	1.09(<0.01)	0.85(<0.01)	0.85(<0.01)	0.85(<0.01)

Table 3: Mean and standard deviation values of the MAE, from the test population, for the three most frequently used markers selected by the DPM models, according to each approach used. A = ADAS13; C = CDRSB; F = FAQ; M= MMSE; L = LDELTOTAL; R= RAVLT forgetting; P = PACC; T=TRABSCORE

While models with three or more features were evaluated, our analysis revealed that lower-dimensional models offered the best balance of performance and clinical parsimony. Specifically, DPMs constructed with only two markers, typically CDRSB combined with either ADAS13 or MMSE, achieved composite criterion scores that were within 5% of the performance of more complex models across all frameworks. Given that the inclusion of additional markers did not yield a statistically significant improvement in prognostic or diagnostic accuracy, we focused our subsequent analyses on these more streamlined two-marker configurations. This decision was based on several factors: (a) low-dimensional models are more practical for clinical implementation; (b) these combinations yielded performance comparable to models using larger feature sets; (c) using the same pair of neuropsychological measures across frameworks allows for more direct and meaningful comparisons; and (d) these markers—CDRSB, MMSE, and ADAS13—are routinely employed in clinical settings for cognitive monitoring and dementia staging.

3.1.1 DPS and Conversion Times

Figure 3 illustrates key results related to the DPS and its ability to capture the onset and progression of cognitive decline for each approach. The first column shows the distribution of initial DPS values across clinical groups (sCU, pCU, sMCI) for each modeling approach. These DPS values represent the average across bootstrapped DPM runs. The DPS distributions show that Leaspy and RPDPM achieve better separation between sCU and sMCI clinical groups compared to GRACE, which exhibits significant overlap. The second column of the figure presents the correlation between the estimated age of cognitive decline onset in pCU and sMCI patients, as inferred from the DPMs, and the age derived from clinical diagnosis. The third column displays the estimated versus observed conversion times for pCU subjects. For predicting conversion times in pCU subjects, Leaspy and RPDPM demonstrated strong and comparable correlations with clinically observed times. GRACE was notably weaker in this regard, suggesting lower precision in prognostic estimates.

3.1.2 Natural Histories of Cognitive Decline Inferred by DPMs

The models' ability to estimate the age of conversion is shown in Figure 4 and Table 4. Fig. 4 depicts the distribution of estimated conversion times to cognitive decline, defined as the interval between the age at baseline ($t_{i,1}$) and the estimated age of conversion to MCI ($\hat{\tau}_i$), i.e., $t_{i,1} - \hat{\tau}_i$. Each DPM estimates $\hat{\tau}_i$ for all patients, including those classified as pCU. To obtain this estimate, a temporal window extending from 10 years prior to the first visit to 10 years after the last visit was considered. Not all CU and MCI patients yielded an estimated τ_i value. The histograms in Figure 4 reveal that Leaspy provides the clearest separation of clinical groups along the estimated

timeline, with most sCU subjects having a future conversion time (negative values) and sMCI subjects having a past conversion time (positive values).

Table 4 summarizes the characteristics of the conversion time distributions ($t_{i,1} - \hat{\tau}_i$) and evaluates the DPMs' capacity to estimate the age at cognitive decline onset ($\hat{\tau}_i$). Performance is first assessed by clinical group (%sCU, %sMCI, %pCU), as the models do not always identify the age of cognitive decline onset. For stable individuals, a correct estimate of τ_i corresponds to an age beyond the last visit for sCU cases, or prior to or equal to baseline for sMCI cases. These conditions are captured in the accuracy metrics (% sCU success, % sMCI success). In the case of pCU individuals, for whom clinical conversion to MCI is documented, the Pearson correlation between the clinically observed and DPM-estimated $\hat{\tau}_i$ is computed. As summarized in Table 4, Leaspy successfully estimated an onset for 100% of pCU individuals, whereas RPDPM was more conservative, identifying an onset in fewer stable individuals. While all models were effective, GRACE showed a lower correlation with observed conversion time ($\rho\left(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1}\right) = 0.59$) compared to Leaspy and RPDPM (both $\rho\left(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1}\right) = 0.72$).

Framework	**Features**	**% sCU**	**% sMCI**	**% pCU**	**% sCU success**	**% sMCI success**	$\rho\left(\tau_i, \hat{\tau}_i\right)$	$\rho\left(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1}\right)$
Leaspy	A, C	83.4 (2.7)	98.9 (0.7)	100 (0)	74.1 (2.4)	96.7 (1.2)	0.94 (0.02)	0.72 (0.06)
RPDPM	C, M	40.7 (0.7)	89.2 (1.1)	87.5 (0.7)	58.6 (3.6)	93.9 (0.8)	0.94 (0.01)	0.72 (0.05)
GRACE	C,M	79.6 (8.9)	100 (0)	99.4 (0.6)	69.3 (9.1)	81.6 (3.0)	0.77 (0.02)	0.59 (0.01)

Table 4: Percentage of correct detections of cognitive decline onset by clinical group (%sCU, %sMCI, %pCU), as well as the accuracy of onset estimation (%sCU success, %sMCI success, and Pearson correlations for pCU individuals). A = ADAS13; C = CDRSB; M= MMSE.

Fig. 5 displays individualized marker trajectories aligned according to the estimated age of decline onset, with time zero representing $\hat{\tau}_i$. Each visit's age is shifted accordingly ($t^*_{i,j} = t_{i,j} - \hat{\tau}_i$), producing a common timeline of cognitive decline. Despite being blind to clinical labels, the inferred natural histories display clear consistency: sCU trajectories predominantly lie to the left of time zero, sMCI to the right, and pCU in between. Leaspy demonstrates superior separation of clinical groups across neuropsychological markers, while GRACE exhibits greater overlap. In contrast, RPDPM produced fewer estimations, having failed to compute $\hat{\tau}_i$ for a subset of patients within the defined time window. The quality of $\hat{\tau}_i$, reflected in the alignment of clinical groups around marker trajectories, was further detailed in Table 4.

Table 5 compares marker values at the estimated transition from CU to MCI across the three DPMs. Neuropsychological scores (ADAS13, MMSE, CDRSB), CSF Aβ, and normalized hippocampal volume (NHV; computed as the sum of left and right hippocampal volumes normalized by intracranial volume) are presented, along with their baseline values stratified by clinical group. Reference cutoffs for cognitive impairment and biomarker positivity were taken from the literature [35, 36, 37, 38, 39]. Marker values at the estimated onset of decline are consistent across proposed DPMs, and for neuropsychological measures, they closely match clinical thresholds. In contrast, CSF Aβ and NHV values, although stable across methods, differ from clinical cutoffs, as these thresholds indicate amyloid pathology and neurodegeneration, respectively, within the AT(N) framework [1].

Fig. 6 (top row) shows pairwise comparisons of estimated transition from CU to MCI across

features	sCU	pCU	sMCI	clinical	Leaspy	RPDPM	GRACE
ADAS13	9.6	11.5	15.2	12 [35]	11.9	13.1	11.4
MMSE	29.1	29.0	27.9	27 [36]	28.6	28.5	28.7
CDRSB	0.03	0.06	1.3	0.5 [37]	0.6	0.6	0.5
CSF Aβ	1240	1098	1066	880 [38]	1092	1010	1082
NHV	5.0	4.7	4.6	3.8 [39]	4.7	4.5	4.7

Table 5: Comparison of baseline mean values of markers across clinical groups (sCU, pCU, and sMCI), the cut-off values used in clinical practice to identify cognitive decline (ADAS13, MMSE, CDRSB), amyloid pathology (CSF Aβ), or neurodegeneration (normalized hippocampal volume, NHV), and the estimated scores on the long-term trajectories of each approach at the time of cognitive decline onset for the different markers.

the three approaches. The bottom row displays estimates of conversion times ($t^*_{i,j=1} = \hat{\tau}_i - t_{i,1}$), encompassing the full patient sample. The range for estimating $\hat{\tau}_i$ remains fixed between 10 years before the first visit and 10 years after the last. The comparison reveals high concordance in $\hat{\tau}_i$ estimates between Leaspy and RPDPM ($\rho = 0.92$), with reduced agreement when including GRACE ($\rho = 0.81 - 0.85$).

3.2 Robustness to Missing Data

Tables 6 and 7 summarize the impact of missing data on DPM performance during patient-specific personalization, focusing on models constructed with two neuropsychological markers per approach. These tables report the quality metric scores as a function of the proportion of missing data (NaNs) in patient records. They also include scenarios in which valid marker values were artificially masked with NaNs to simulate missingness rates of 20% and 40% across the marker variables. The results clearly indicate that RPDPM is the most robust framework, maintaining highly stable diagnostic performance and low MAE even with 40% of data missing. Leaspy also showed good resilience, though its pCU detection rate declined more noticeably. In stark contrast, GRACE's performance degraded sharply with increasing missingness, highlighting its sensitivity to data sparsity.

3.3 Impact of the Number of Patient Visits on Diagnostic Accuracy

Figure 7 presents the performance metrics of the selected DPMs as a function of the number of visits used to personalize individual trajectories. The number of visits ranges from 2 to 16. Most individuals underwent between 2 and 7 visits, while fewer than 50 patients were followed for more than 12 visits. Consequently, the analysis focuses on patients with 2 to 10 visits, with the 10-visit group comprising only 45 individuals. As expected, diagnostic accuracy for progressive cases (pCU) and the correlation of conversion time estimates improved significantly with a higher number of visits for both Leaspy and RPDPM. This underscores the value of longer follow-up for accurate prognostication. In contrast, the MAE for trajectory reconstruction remained relatively stable and unaffected by the number of visits across all models.

Framework	NaN	% sCU	% sMCI	% pCU	AUC	$\rho(\tau_i, \hat{\tau}_i)$	$\rho(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1})$	Criterion
Leaspy (A,C)	-	75.3 (1.2)	94.6 (1.2)	81.2 (5.6)	0.95 (0.012)	0.99 (0.001)	0.78 (0.023)	67.0
	20	75.2 (2.9)	94.2 (0.9)	73.9 (10.1)	0.94 (0.010)	0.99 (0.001)	0.76 (0.015)	59.0
	40	74.7 (6.0)	92.3 (1.3)	68.1 (8.1)	0.93 (0.012)	0.99 (0.002)	0.71 (0.055)	50.4
RPDPM (C,M)	-	87.0 (0.9)	88.9 (1.6)	76.8 (1.4)	0.94 (0.006)	0.99 (0.001)	0.88 (0.031)	63.7
	20	84.6 (0.6)	88.1 (0.1)	72.7 (0.1)	0.93 (0.007)	0.99 (0.001)	0.86 (0.001)	61.6
	40	83.2 (0.8)	88.0 (0.7)	70.9 (2.4)	0.93 (0.003)	0.99 (0.001)	0.87 (0.016)	61.2
GRACE (C,M)	-	78.6 (3.3)	74.2 (5.3)	54.8 (7.7)	0.92 (0.003)	0.98 (0.002)	0.54 (0.067)	25.2
	20	78.7 (1.6)	72.9 (2.2)	38.4 (7.1)	0.90 (0.002)	0.98 (0.002)	0.50 (0.050)	16.6
	40	60.6 (3.1)	55.9 (0.9)	25.2 (6.6)	0.75 (0.001)	0.97 (0.002)	0.40 (0.061)	9.0

Table 6: Diagnostic performance metrics of the selected DPMs on the test patient set as a function of the percentage of missing data (original, 20% and 40%). The diagnostic accuracy is reported for each group at each visit (sCU, sMCI, and pMCI), along with Pearson correlations for the estimation of age at cognitive decline onset and time to conversion. All quality metrics are reported as the mean and standard deviation across DPMs trained via bootstrapping. A = ADAS13; C = CDRSB; M= MMSE.

Framework	Leaspy (A,C)			RPDPM (C,M)			GRACE (C,M)		
NaN	-	20	40	-	20	40	-	20	40
ADAS13	2.31(0.01)	2.34(0.03)	2.29(0.02)	-	-	-	-	-	-
CDRSB	0.47(0.07)	0.50(0.04)	0.48(0.06)	0.23(<0.01)	0.21(<0.01)	0.21(<0.01)	0.25(<0.01)	0.23(<0.01)	0.19(<0.01)
MMSE	-	-	-	1.09(<0.01)	1.09(<0.01)	1.05(<0.01)	0.85(<0.01)	0.79(<0.01)	0.64(<0.01)

Table 7: Mean and standard deviation of the MAE computed on the test set for the three markers by the selected DPMs, reported across different levels of missing data (original, 20% and 40%). A = ADAS13; C = CDRSB; M= MMSE.

3.4 Prognostic Performance

Fig. 8 summarizes the performance of the evaluated DPMs across prognostic horizons ranging from one to five years, reporting AUC scores for future cognitive state prediction and classification accuracy for clinical subgroups (%sCU, %sMCI and %pCU). As the prediction window extends, the performance of all DPMs declines. However, Leaspy consistently achieved the highest accuracy in predicting future converters to MCI (% pCU), clearly outperforming the constant predictor. It also depicts the evolution of the MAE across neuropsychological markers as a function of the prediction horizon for each modeling approach. The prediction error for all models increased with longer horizons.

Fig. 9 assesses the ability of the DPMs to estimate the age at which pCU patients convert to MCI. This includes both the detection of progression to MCI and the correlation between the estimated and clinically observed ages at MCI onset and conversion times. The analysis focuses exclusively on pCU individuals, using only the visits in which they were still clinically classified as CU for model personalization. It is important to note the distinction in the evaluation of pCU detection between Figs. 8 and 9. In the former, all visits were used for personalization except the last ones, which were withheld for evaluation. In contrast, the latter experiment limited personalization strictly to visits preceding clinical conversion, during which patients remained cognitively unimpaired. A key finding is that when using only pre-conversion data, Leaspy detected the highest percentage of future pCU converters, while RPDPM, when it made a prediction, yielded the strongest correlation with the clinical conversion time, highlighting a trade-off between sensitivity and temporal precision.

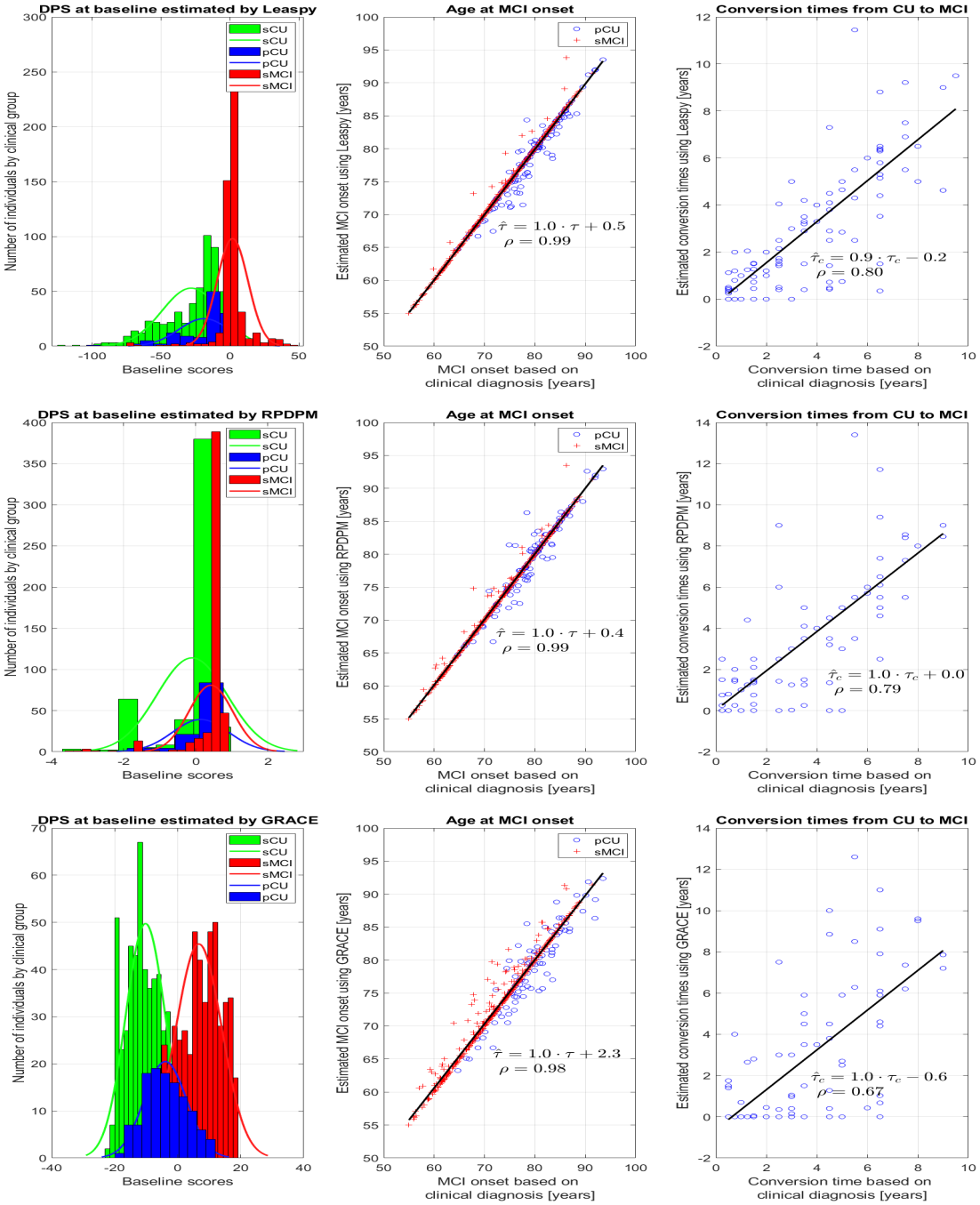

Figure 3: Diagnostic results of the DPMs for each approach, considering models built with two markers per approach: Leaspy, RPDPM, and GRACE. The first column shows histograms of the DPS at study baseline, separated by clinical group. The second column presents the correlation between the clinically determined age at cognitive decline onset and that estimated by each DPM. The third column shows the correlation between the time to conversion ($\tau_i - t_{i,1}$) according to clinical practice and the estimate provided by the DPM.

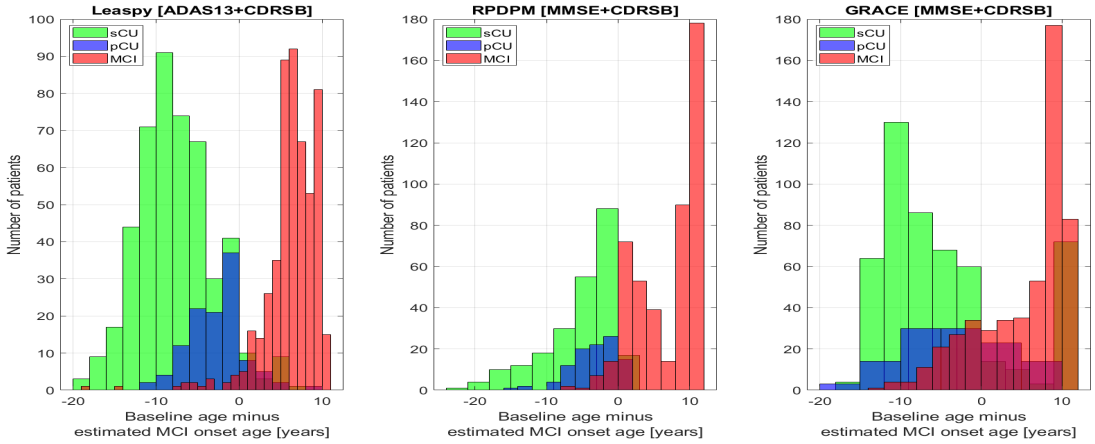

Figure 4: Histogram of the differences $t_{i,1} - \hat{\tau}_i$, where $t_{i,j=1}$ is the age at baseline and $\hat{\tau}_i$ is the age of MCI conversion according to the DPM's estimations, in the analyzed population, stratified by clinical group and approach. In this context, CU individuals are expected to exhibit negative values — indicating that $\hat{\tau}_i$ exceeds the age at the first visit — while MCI patients should tend to show positive values.

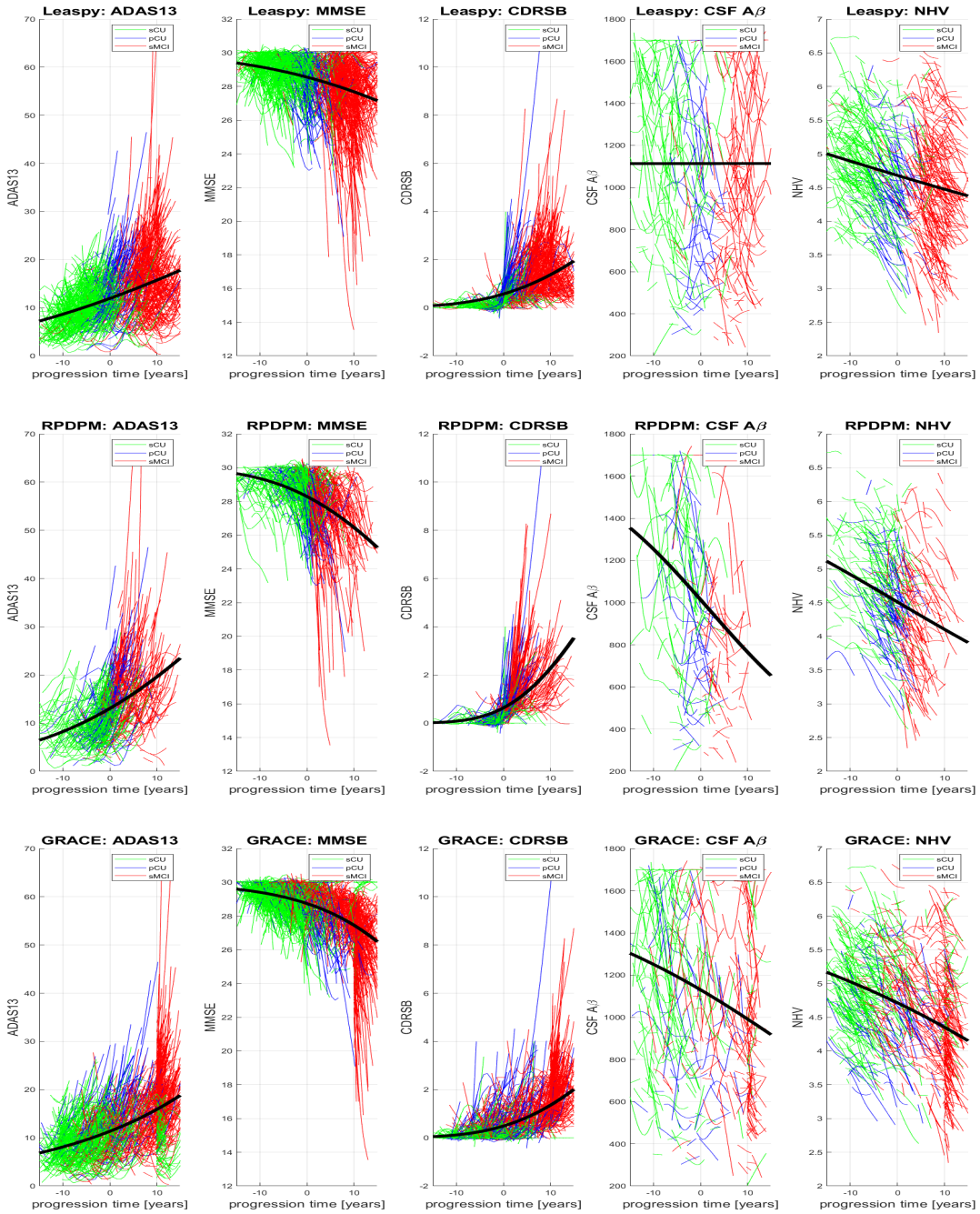

Figure 5: Estimated timelines generated by the Leaspy, RPDPM, and GRACE models. The independent variable corresponds to the time relative to the estimated transition from CU to MCI ($t^*_{i,j} = t_{i,j} - \hat{\tau}_i$), while the dependent variable represents the clinical marker scores. Individual trajectories of sCU subjects are shown in green, those of pCU patients in blue, and those of MCI subjects in red. Superimposed in black are the long-term trajectories.

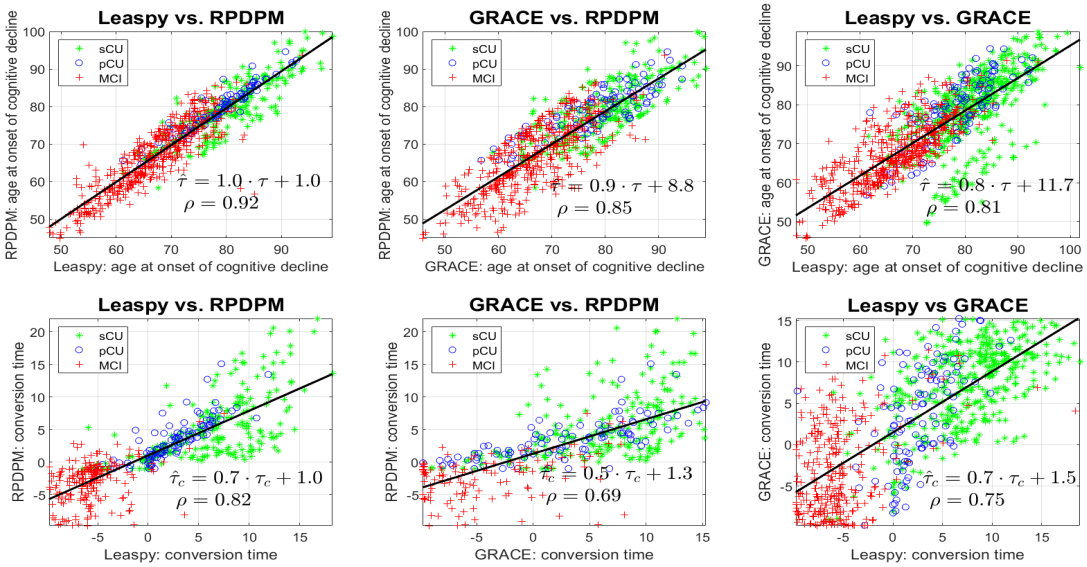

Figure 6: A pairwise comparison of Leaspy, RPDPM, and GRACE in estimating the age at transition from CU to MCI. The top row shows a comparison of the τ_i values across the three approaches. The bottom row presents the adjusted conversion times, defined as $t^*_{i,j=1} = \hat{\tau}_i - t_{i,j=1}$.

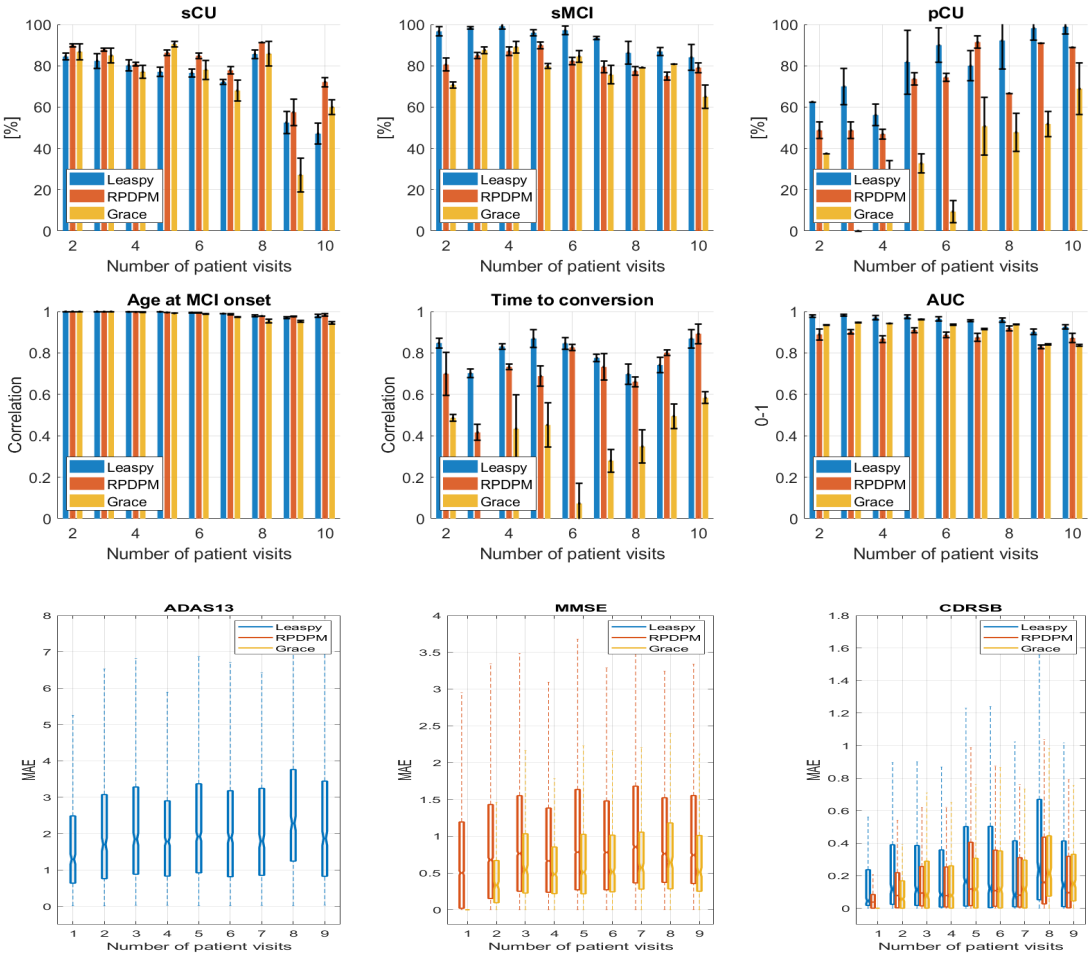

Figure 7: Impact of the Number of Patient Visits on DPM Performance. Diagnostic quality metrics of the DPMs as a function of the number of patient visits and the modeling approach used. Panels A–C: Diagnostic accuracy for identifying each clinical group (sCU, sMCI, and pMCI) is shown in the first row. Panels D–F: Pearson correlations for the estimated age at cognitive decline onset, time to conversion, and corresponding AUC values are presented in the second row. Panels G–I: Box-and-whisker plots depicting the distribution of the MAE for the ADAS13, MMSE, and CDRSB markers according to the number of visits per patient and the modeling approach, shown in the third row.

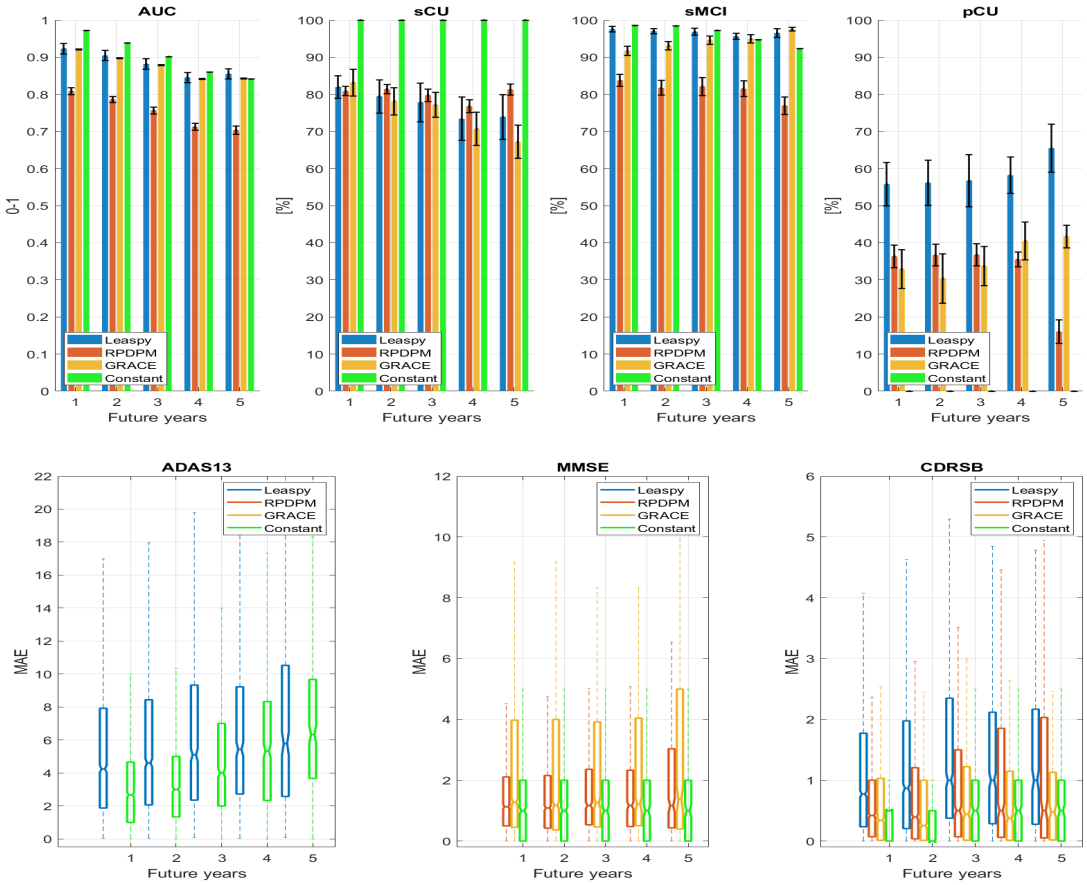

Figure 8: Prognostic Performance Across Prediction Horizons. Panels A–D: Quality metric scores of the DPMs as a function of the prediction horizon for each approach, including the classifier AUC at each future visit and the accuracy in identifying clinical groups (sCU, sMCI, and pMCI) in the first row. Panels E–G: Distribution of MAE values for the ADAS13, MMSE, and CDRSB markers as a function of the prediction horizon for each approach in the second row.

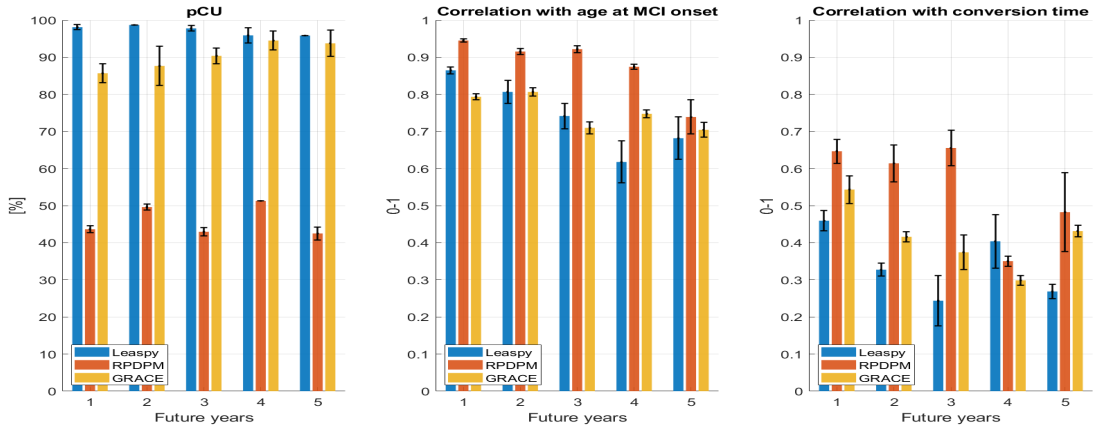

Figure 9: Prognostic performance metrics based on the estimated conversion time in pCU patients: a) probability of detecting progression to MCI, b) correlation with the age at onset of cognitive decline, and c) correlation with the conversion time.

4 Discussion

This study introduced a comprehensive framework for the development, evaluation, and benchmarking of parametric DPMs aimed at the early detection of cognitive decline. By leveraging this framework, we identified optimal, low-dimensional neuropsychological marker combinations and compared the performance of three distinct DPMs: Leaspy, RPDPM, and GRACE. Our findings demonstrate that parametric models trained exclusively on routinely collected clinical data can effectively characterize the transition from normal cognition to MCI, providing a valuable tool for both clinical research and practice.

4.1 Interpretation of the Findings

DPMs were trained exclusively on neuropsychological measures, which are sensitive to cognitive changes and commonly used in clinical follow-up. Furthermore, neuropsychological assessment remains essential in the diagnostic process of cognitive decline [40]. This choice enabled model training on a larger sample than would be feasible with imaging or fluid biomarkers and enhances real-world applicability, as these measures are routinely collected in clinical practice [5].

At baseline, several tests, such as ADAS13, LDELTOTAL, PACC, and RAVLT, showed statistically significant differences across clinical groups (sCU, pCU and sMCI), while others primarily distinguished MCI from CU group (see table 1 and the supplementary material table S.1). Around 10% of CU individuals progressed to MCI within four years, consistent with prior analyses of the ADNI cohort [41].

4.1.1 Selection of DPMs and diagnostic performance

In contrast to conventional ablation strategies, which remove features post hoc to assess their importance [9, 19], or approaches that rely on feature preselection using predictive models [26], we selected the best DPMs a priori based on diagnostic performance, prioritizing the detection of individuals progressing to the MCI stage and the estimation of their time to conversion. Notably, high-quality DPMs were obtained using only two or three neuropsychological measures, underscoring the feasibility of low-dimensional models for clinical application.

The most frequently selected combinations included CDRSB paired with either ADAS13 or MMSE, tools already established in clinical diagnostic criteria for AD [42, 37, 24]. Despite their unsupervised nature, the DPMs converged on clinically relevant markers, validating the models' alignment with existing diagnostic frameworks. In higher-dimensional models, LDELTOTAL, PACC, and RAVLT also emerged as discriminative features, further supporting their relevance in early-stage cognitive decline (see tables 1-2).

Leaspy and RPDPM demonstrated comparable performance, with consistently high detection rates for progressive and stable clinical profiles. In contrast, GRACE showed lower performance, particularly in identifying pCU subjects (detecting only around 60%), and exhibited a weaker correlation between estimated and clinical conversion times (Pearson ≈ 0.6). These limitations were also reflected across all diagnostic quality indicators. Although GRACE demonstrated lower

MAEs in trajectory fitting, this was accompanied by reduced sensitivity to clinical transitions. Thus, while GRACE provides better curve fitting, diagnostic accuracy remains the primary goal in clinical applications.

4.1.2 DPS and Conversion Times

The initial DPS distributions revealed greater overlap between CU and MCI groups in GRACE compared to Leaspy and RPDPM (Fig. 3), aligning with GRACE's lower AUC and reduced detection rates for sMCI and pCU (Table 2). This likely reflects GRACE's reliance on Gaussian priors for individual temporal shifts [9], which may fail to capture inter-subject variability. In contrast, Leaspy and RPDPM produced more dispersed and better-separated DPS distributions, enhancing conversion detection and improving correlations with observed onset and progression times in pCU subjects (Fig.3, columns 2–3; Table2).

Regarding age estimation at transition from CU to MCI, GRACE showed greater dispersion around the regression line, indicating reduced precision, although its Pearson correlation remained high and slopes approximated the identity line. Leaspy and RPDPM, however, showed tighter fits, suggesting more consistent alignment with clinical observations. For conversion time estimates, GRACE systematically underperformed, with a regression intercept shifted by one year, a meaningful bias given the time scale; and failed to detect several pCU cases, consistent with its lower detection rate (Table 2). Leaspy and RPDPM delivered more accurate and reliable conversion estimates.

4.1.3 Natural Histories of Cognitive Decline Inferred by DPMs

All evaluated DPMs enabled the construction of proposed timelines for cognitive decline. Among them, Leaspy yielded the most coherent representations, with estimated onset ages ($\hat{\tau}_i$) aligning closely with clinical progression (Fig. 4, Table 4). Pairwise comparisons also showed high agreement between Leaspy and RPDPM, while GRACE deviated more substantially (Fig. 6).

Once patients were temporally aligned using $t^*_{i,j} = t_{i,j} - \hat{\tau}_i$, Leaspy provided clearer clustering of sCU, pCU, and sMCI subjects along the progression axis (Fig. 5). This visualization supports its potential to reconstruct disease trajectories and validate biological hypotheses on the temporal ordering of symptoms [34, 5].

Neuropsychological measures emerged as the most reliable indicators of disease progression across all DPMs. In contrast, CSF Aβ and normalized hippocampal volume exhibited greater within-group variability and limited alignment with long-term trajectories, likely due to floor/ceiling effects or nonlinear associations with clinical decline (see Fig. 5). These results further support the exclusive use of neuropsychological measures in constructing DPMs to characterize cognitive decline.

4.1.4 Robustness to Missing Data

Among the three frameworks, RPDPM demonstrated the highest robustness to missing data, maintaining stable performance across all evaluation metrics even with up to 40% of missing input. Both RPDPM and Leaspy preserved consistent scores for the detection of stable subjects (sCU, sMCI), as well as for AUC and the correlation between estimated and observed age of transition from CU to MCI, even at high levels of missingness. Furthermore, RPDPM retained high detection rates for pCU subjects and maintained stable correlations between clinical and estimated conversion times.

Regarding trajectory fitting, both Leaspy and RPDPM exhibited stable MAE values up to 40% missing data, indicating robust estimation of disease progression trajectories. In contrast, GRACE showed decreasing MAE with increasing data loss, likely due to its reliance on linear interpolation. With fewer data points, interpolation becomes more constrained, reducing variability and leading to closer alignment between adjacent values, potentially underestimating trajectory uncertainty.

4.1.5 Impact of the Number of Patient Visits on Diagnostic Accuracy

Leaspy demonstrates superior performance in detecting sMCI cases and consistently achieves slightly higher AUC values across varying numbers of visits compared to the other frameworks. In contrast, RPDPM excels in identifying sCU subjects, supporting previous findings regarding its diagnostic accuracy and robustness to missing data (see Tables 2–6). For both RPDPM and Leaspy, the accuracy of pCU detection improves with the number of available visits, as does the Pearson correlation between estimated and actual conversion time. GRACE shows a similar trend, though less pronounced. Notably, GRACE consistently achieves the lowest MAE values across all horizons, while Leaspy exhibits a slight increase in MAE with additional visits. In general, MAE appears relatively unaffected by the number of visits in all models.

The number of patient visits significantly influenced pCU detection and conversion time estimation in Leaspy and RPDPM. In contrast, performance metrics for stable groups (sCU, sMCI) remained relatively constant across visit counts. These findings highlight the importance of extended follow-up: longer observation periods improve model personalization and the detection of subtle longitudinal trends. DPM precision increases as follow-up duration extends, providing more reliable and individualized short-term trajectories. Conversely, limited data force the models to rely on population-level dynamics, reducing sensitivity to subject-specific changes.

4.1.6 Prognostic Performance

All DPMs were evaluated for their ability to forecast patient status up to five years after model personalization. As expected, AUC values progressively decline as the prognostic horizon increases. The constant predictor, which assumes that the future cognitive state remains unchanged from the last visit used for personalization, achieves 100% accuracy in identifying sCU patients but fails entirely in detecting pCU cases. Its performance for sMCI individuals is also suboptimal, as it does not account for transitions within the MCI category, such as progression to dementia or reversion to normal cognition. At shorter horizons, Leaspy performs similarly to the constant predictor in identifying stable individuals (sCU and sMCI) and in overall AUC. However, its advantage becomes more evident as the prediction window extends. Notably, Leaspy achieves the highest

accuracy in detecting future converters to MCI. In contrast, RPDPM and GRACE demonstrate lower prognostic performance across all horizons.

In line with the diagnostic findings, Leaspy exhibits the highest MAE in predicting marker trajectories, while RPDPM and GRACE display similarly lower error distributions. The constant predictor achieves the lowest MAE overall. For all models, MAE increases with longer prognostic horizons, reflecting the greater uncertainty associated with long-term forecasts.

When evaluating prognoses for pCU individuals based solely on pre-conversion visits, Leaspy, and to a lesser extent GRACE, were able to reliably estimate the age of conversion to MCI. In contrast, RPDPM generated fewer conversion estimates, following a more conservative approach. However, when such estimates were available, RPDPM achieved a stronger correlation with the clinically observed conversion times (Fig.9, Table4). These results highlight a trade-off between detection sensitivity (favored by Leaspy) and temporal precision in conversion time estimation (favored by RPDPM).

Overall, compared to the constant predictor, Leaspy achieved comparable performance in detecting stable patients and clearly outperformed it in identifying future converters. While RPDPM and GRACE also provided useful forecasts, their prospective performance was generally lower than that of Leaspy. However, it is important to note that Leaspy requires more training data to accurately fit the DPMs, making it more data-intensive than the other two approaches.

4.1.7 Interpretation of Model Performance and Clinical Implications

Our comparative analysis revealed distinct performance profiles for each DPM, confirming and extending findings from the original literature while highlighting a crucial trade-off between diagnostic sensitivity, trajectory fitting, and robustness. This has direct implications for their use in clinical and research settings. These key trade-offs are consolidated in Table 8.

Table 8: Summary of Key Performance Metrics Across DPM Frameworks. Values represent the best performance achieved by models with few markers.

Performance Metric	**Leaspy**	**RPDPM**	**GRACE**
Diagnostic Performance			
AUC (max)	**0.95**	0.94	0.92
pCU Detection (%)	**81.2**	76.8	54.8
Corr. with Conversion Time (ρ)	0.78	**0.88**	0.54
Prognostic Performance			
Future pCU Detection (1-year) (%)	**98.1**	42.4	85.7
Corr. with Conversion Time (ρ)	0.46	**0.65**	0.56
Trajectory Reconstruction			
MAE (Overall)	Higher	Lower	**Lowest**
Robustness to Missing Data			
Performance with 40% NaNs	Good	**Excellent**	Poor

In line with its design, Leaspy emerged as the superior model for both diagnostic and prognos-

tic tasks. Its flexible time-warping function, which models inter-subject variability in both disease onset and progression speed [15], likely explains its high accuracy in detecting progressive individuals ($AUC \approx 0.95$) and its strong correlation with clinical conversion times ($\rho\left(\tau_i - t_{i,1}, \hat{\tau}_i - t_{i,1}\right) \approx 0.78$). Clinically, this makes Leaspy a powerful tool for patient stratification in clinical trials or for identifying high-risk individuals who may benefit from early intervention.

RPDPM demonstrated exceptional robustness to missing data, maintaining stable performance even with up to 40% data loss. This result empirically validates the robust nonlinear regression formulation at its core [14]. This positions RPDPM as the most suitable model for deployment in real-world clinical settings, where patient data is often incomplete and irregularly collected. Its strong performance in these adverse conditions increases the feasibility of translating DPMs from controlled research cohorts to routine clinical practice.

Consistent with Donohue et al. [9], GRACE achieved the lowest Mean Absolute Error, excelling at trajectory reconstruction. However, our study highlights its primary limitation: a reduced sensitivity in detecting clinical transitions, particularly for progressive CU subjects. This suggests that while GRACE is valuable for modeling population-level biomarker trajectories in research, its utility for individualized clinical prognosis is limited compared to more flexible models.

Ultimately, this study underscores that no single DPM is universally optimal. The choice should be guided by the specific application: Leaspy for precise prognostic prediction in research, RPDPM for reliable use in messy clinical data, and GRACE for population-level trajectory analysis. A significant clinical implication from our work is that highly effective DPMs can be constructed using only two or three routinely collected neuropsychological measures, supporting the development of accessible and low-cost tools for monitoring cognitive decline.

To illustrate how these findings translate into a practical tool, we propose a workflow for a clinical decision support interface, depicted in Figure 10. This tool would allow a clinician to (a) input a patient's longitudinal visit data; (b) personalize the DPM by setting the time window for trajectory estimation; and (c) instantly visualize the comprehensive prognostic output. As shown in the figure, this output would include the patient's raw data alongside the estimated bootstrapped trajectories (in both original and normalized scales) and, most importantly, a clear visual marker of the DPM-estimated age of transition from CU to MCI. This provides a tangible pathway for integrating these DPMs into clinical practice to support individualized prognosis.

4.2 Contextualization within the Broader DPM Landscape

The DPMs evaluated here represent a specific class of parametric models. It is important to contextualize their performance relative to other modeling paradigms. While non-parametric and deep learning-based DPMs offer greater flexibility to model complex, non-linear trajectories, they often require larger datasets and function as "black boxes," hindering clinical interpretability [19, 20, 23]. Our findings confirm that simpler, interpretable parametric models can achieve high performance in early-stage disease detection, making them a more practical choice for many clinical applications where transparency is essential.

Furthermore, modeling the progression of AD presents unique challenges compared to other

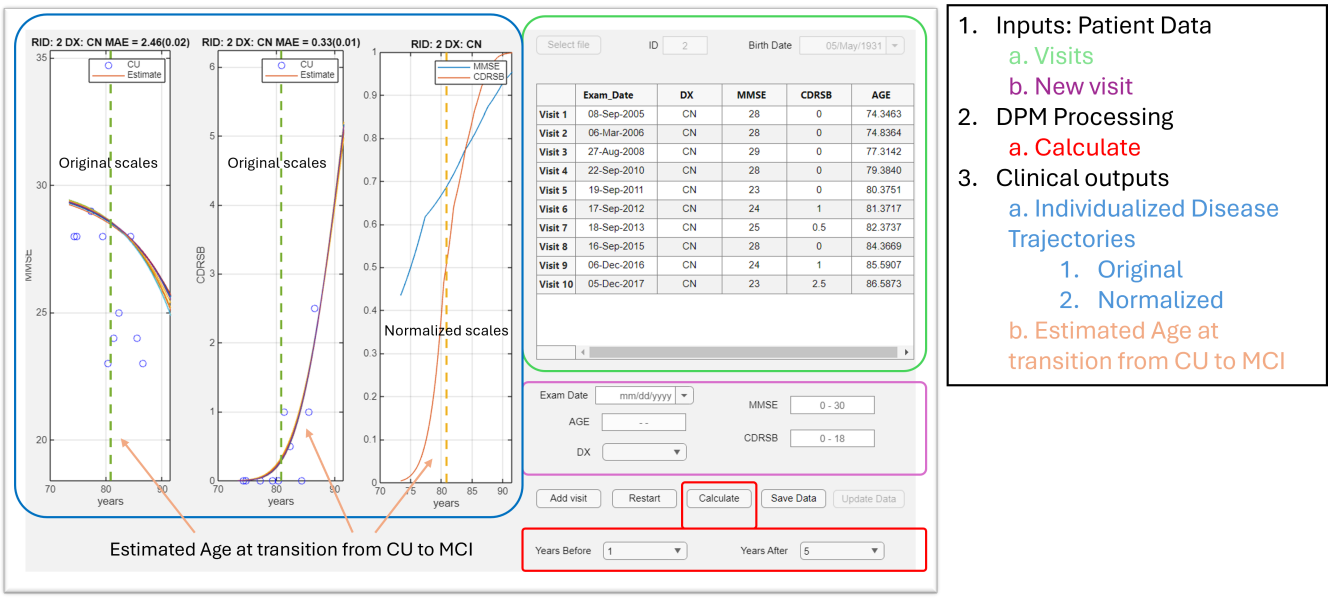

Figure 10: Conceptual workflow for a clinical decision support tool. (a) Patient data input. (b) Personalization settings for trajectory calculation. (c) Visualized output showing individual trajectories against bootstrapped DPMs and the estimated age of transition to MCI.

conditions where DPMs are applied, such as Huntington's disease [43, 44]. Unlike the more predictable, genetically determined trajectory in Huntington's, AD progression is multifactorial and highly heterogeneous. Moreover, the progression in early AD is characterized by a subtle and gradual cognitive decline that can be difficult to distinguish from normal aging. The transition from CU to MCI is characterized by subtle cognitive changes and significant etiological heterogeneity. Our results show that DPMs based on neuropsychological measures are highly sensitive to these early clinical manifestations, providing a complementary perspective to biomarker-based models that track underlying pathology [44, 5].

4.3 Strengths, Limitations, and Future Directions

The primary strength of this study is its novel, comprehensive evaluation framework, which balances diagnostic, prognostic, and robustness metrics to facilitate a holistic comparison of DPMs. The a priori selection of marker subsets based on a clinically relevant composite criterion is another key contribution. However, several limitations must be acknowledged.

First, the exclusive reliance on neuropsychological data. While this study's exclusive reliance on neuropsychological data is a limitation, it also highlights a key strength and provides a unique contribution to the field. Fluid and imaging biomarkers, such as CSF Aβ and hippocampal volume, are essential for tracking the underlying biological progression of Alzheimer's disease within the AT(N) framework [1, 2]. However, these biomarkers often show high variability and non-linear associations with clinical symptoms, especially in the early stages (see Fig. 5). Our findings demonstrate that DPMs built solely on neuropsychological measures are highly effective at modeling the clinical manifestation of the disease; the cognitive decline that directly impacts patients' lives. Therefore, these models do not compete with but rather complement biomarker-based research.

They can serve as a robust, low-cost, and clinically accessible framework for tracking symptomatic progression.

Second, our findings are based exclusively on data from the ADNI cohort. While unquestionably valuable for research, ADNI is a highly selected population that is predominantly white, highly educated, and has fewer comorbidities compared to the general population seen in routine clinical practice [45]. Real-world cohorts are far more heterogeneous, with greater diversity in socioeconomic status, ethnicity, and health profiles, including a higher prevalence of conditions like vascular disease, diabetes, or depression, all of which can influence cognitive performance and disease progression. This sampling bias may limit the generalizability of our models, as their performance could be attenuated in more complex, diverse populations. Therefore, while our results establish a strong proof-of-concept, further validation in larger, more representative "real-world" clinical datasets is an essential next step to confirm their external validity and clinical applicability.

Third, our classification of subjects as stable or progressive based solely on their baseline and final diagnoses represents a methodological simplification. This approach does not account for potential intra-individual diagnostic fluctuations during the follow-up period, which are common in early-stage cognitive decline. For instance, a patient might revert from MCI to CU before later progressing again, yet our method would classify them based only on their endpoint. This could introduce a misclassification bias, potentially affecting the training of the DPMs by including noisy labels. A more granular analysis, possibly using a state-transition model or a sensitivity analysis to assess the impact of these fluctuations, was beyond the scope of this work but is a key direction for future research to enhance model accuracy and clinical relevance.

Fourth, the parametric nature of the models introduces constraints in their ability to fully capture the heterogeneity of disease trajectories. These models are not well suited to represent non-monotonic patterns such as transient cognitive improvements or recovery. Although the models demonstrated robustness, further studies in independent and more diverse cohorts are needed to confirm their generalizability and clinical applicability.

Finally, estimating the time to clinical events—such as the transition from CU to MCI, or from MCI to dementia—will enable the evaluation of novel AD biomarkers. For instance, it is well established that patients with amyloid pathology progress more rapidly to dementia than those without it. Therefore, it would be possible to compare the proposed natural disease trajectories between groups with and without amyloid pathology, according to amyloid PET criteria or hybrid CSF biomarker ratios such as ptau/Aβ_{42}, with the classification of amyloid pathology based on the proposed plasma AD biomarkers.

5 Conclusions

This study addressed three main goals related to the development and evaluation of parametric Disease Progression Models (DPMs) for the early detection of cognitive decline. First, we proposed and implemented a framework for evaluating both diagnostic and prognostic performance of DPMs. This included multiple clinically meaningful metrics, such as the detection rate of progressive CU cases, the correlation between estimated and observed onset ages, and the reconstruction

accuracy of marker trajectories. These measures allow for a comprehensive assessment of each model's capacity to detect early disease changes and forecast future progression.

Second, we introduced a systematic procedure to identify optimal subsets of neuropsychological markers for model construction. Instead of relying on feature ablation or supervised feature selection, we employed a combinatorial approach coupled with a composite criterion that balances diagnostic sensitivity and prognostic accuracy. This enabled the identification of low-dimensional marker combinations—primarily including CDRSB, ADAS13, and MMSE—that support clinically feasible and robust DPMs.

Third, we benchmarked three parametric DPM frameworks—Leaspy, RPDPM, and GRACE—on a large ADNI cohort. Leaspy emerged as the most accurate and clinically useful model, showing superior detection of progressive subjects and more reliable estimation of onset age. RPDPM demonstrated greater robustness to missing data, while GRACE yielded lower trajectory errors but reduced diagnostic sensitivity. These comparative results were consistent across diagnostic, prognostic, and robustness analyses.

Overall, this work supports the integration of parametric DPMs into early-stage AD monitoring and digital prognosis tools, reinforcing their value for personalized disease modeling based on routinely collected clinical data.

Ethics Statement

This study is based on de-identified, publicly available data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI), accessible at: `https://adni.loni.usc.edu/`. All original data collection procedures were approved by the ADNI Institutional Review Boards, and written informed consent was obtained from all participants. The present analysis involves no new data collection or direct interaction with human subjects.

Funding

This work was supported by the Comunidad de Madrid, grant number MINA CM P2022 BMD 7236.

CRediT authorship contribution statement

Carlos Platero: Writing – original draft, Writing – review & editing, Methodology, Conceptualization, Investigation, Visualization, Software, Methodology, Validation.

Jorge Bengoa: Software, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Data collection for this study was funded by the ADNI (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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.ucla.edu/wp-content/uploads/howtoapply/ADNIAcknowledgementList.pdf`

We are grateful to the authors of Leaspy, GRACE, and RPDPM for making their code publicly available, which facilitated the comparative evaluation presented in this study.

We acknowledge the use of ChatGPT (OpenAI) for support with language and technical editing.

References

[1] Jack Jr, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., et al.: NIA-AA research framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia **14** (2018) 535–562

[2] Jack Jr, C.R., Andrews, J.S., Beach, T.G., Buracchio, T., Dunn, B., Graf, A., Hansson, O., Ho, C., Jagust, W., McDade, E., et al.: Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. Alzheimer's & Dementia **20** (2024) 5143–5169

[3] Ebenau, J.L., Timmers, T., Wesselman, L.M., Verberk, I.M., Verfaillie, S.C., Slot, R.E., Van Harten, A.C., Teunissen, C.E., Barkhof, F., Van Den Bosch, K.A., et al.: ATN classification and clinical progression in subjective cognitive decline: The Science project. Neurology **95** (2020) e46–e58

[4] Bucci, M., Chiotis, K., Nordberg, A.: Alzheimer's disease profiled by fluid and imaging markers: tau PET best predicts cognitive decline. Molecular Psychiatry (2021) 1–11

[5] Platero, C.: Temporal modeling and AT profiles in the early phase of Alzheimer's disease. Journal of Alzheimer's Disease Reports **9** (2025) 25424823241306097

[6] Braak, H., Braak, E.: Neuropathological staging of Alzheimer-related changes. Acta Neuropathologica **82** (1991) 239–259

[7] Frisoni, G.B., Altomare, D., Thal, D.R., Ribaldi, F., van der Kant, R., Ossenkoppele, R., Blennow, K., Cummings, J., van Duijn, C., Nilsson, P.M., et al.: The probabilistic model of Alzheimer disease: the amyloid hypothesis revised. Nature Reviews Neuroscience **23** (2022) 53–66

[8] Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronné, R., Faouzi, J., Koval, I., Louis, M., et al.: Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. Medical Image Analysis **67** (2021) 101848

[9] Donohue, M.C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R.G., Raman, R., Gamst, A.C., Beckett, L.A., Jack Jr, C.R., Weiner, M.W., Dartigues, J.F., et al.: Estimating long-term multivariate progression from short-term data. Alzheimer's & Dementia **10** (2014) S400–S410

[10] Guerrero, R., Schmidt-Richberg, A., Ledig, C., Tong, T., Wolz, R., Rueckert, D., Alzheimer's Disease Neuroimaging Initiative (ADNI), et al.: Instantiated mixed effects modeling of Alzheimer's disease markers. NeuroImage **142** (2016) 113–125

[11] Schmidt-Richberg, A., Ledig, C., Guerrero, R., Molina-Abril, H., Frangi, A., Rueckert, D., Alzheimer's Disease Neuroimaging Initiative, et al.: Learning biomarker models for progression estimation of Alzheimer's disease. PloS one **11** (2016)

[12] Li, D., Iddi, S., Thompson, W.K., Donohue, M.C., Alzheimer's Disease Neuroimaging Initiative: Bayesian latent time joint mixed effect models for multicohort longitudinal data. Statistical methods in medical research **28** (2019) 835–845

[13] Lorenzi, M., Filippone, M., Frisoni, G.B., Alexander, D.C., Ourselin, S., Alzheimer's Disease Neuroimaging Initiative, et al.: Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. NeuroImage **190** (2019) 56–68

[14] Ghazi, M.M., Nielsen, M., Pai, A., Modat, M., Cardoso, M.J., Ourselin, S., Sørensen, L.: Robust parametric modeling of Alzheimer's disease progression. NeuroImage **225** (2021) 117460

[15] Koval, I., Bône, A., Louis, M., Lartigue, T., Bottani, S., Marcoux, A., Samper-Gonzalez, J., Burgos, N., Charlier, B., Bertrand, A., et al.: AD course map charts Alzheimer's disease progression. Scientific Reports **11** (2021) 8020

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

[17] Moravveji, S., Doyon, N., Mashreghi, J., Duchesne, S.: A scoping review of mathematical models covering Alzheimer's disease progression. Frontiers in Neuroinformatics **18** (2024) 1281656

[18] Jung, W., Jun, E., Suk, H.I., Alzheimer's Disease Neuroimaging Initiative, et al.: Deep recurrent model for individualized prediction of Alzheimer's disease progression. NeuroImage **237** (2021) 118143

[19] Nguyen, M., He, T., An, L., Alexander, D.C., Feng, J., Yeo, B.T., Alzheimer's Disease Neuroimaging Initiative, et al.: Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage **222** (2020) 117203

[20] Ghazi, M.M., Sørensen, L., Ourselin, S., Nielsen, M.: Carrnn: A continuous autoregressive recurrent neural network for deep representation learning from sporadic temporal data. IEEE Transactions on Neural Networks and Learning Systems (2022)

[21] Al Olaimat, M., Martinez, J., Saeed, F., Bozdag, S., Initiative, A.D.N.: Ppad: a deep learning architecture to predict progression of alzheimer's disease. Bioinformatics **39** (2023) i149–i157

[22] Cheng, H., Yuan, S., Li, W., Yu, X., Liu, F., Liu, X., Bezabih, T.T.: De-accumulated error collaborative learning framework for predicting alzheimer's disease progression. Biomedical Signal Processing and Control **89** (2024) 105767

[23] Jia, N., Jia, T., Zhang, Z.: A residual gru method with deep cross fusion for alzheimer's disease progression prediction using missing variable-length time series data. Biomedical Signal Processing and Control **102** (2025) 107253

[24] Xu, L., Wu, H., He, C., Wang, J., Zhang, C., Nie, F., Chen, L.: Multi-modal sequence learning for Alzheimer's disease progression prediction with incomplete variable-length longitudinal data. Medical Image Analysis **82** (2022) 102643

[25] Saint-Jalmes, M., Fedyashov, V., Beck, D., Baldwin, T., Faux, N.G., Bourgeat, P., Fripp, J., Masters, C.L., Goudey, B., Alzheimer's Disease Neuroimaging Initiative, et al.: Disease progression modelling of Alzheimer's disease using probabilistic principal components analysis. Neuroimage **278** (2023) 120279

[26] Platero, C.: Categorical predictive and disease progression modeling in the early stage of Alzheimer's disease. Journal of Neuroscience Methods **374** (2022) 109581

[27] Karaman, B.K., Mormino, E.C., Sabuncu, M.R., Alzheimer's Disease Neuroimaging Initiative: Machine learning based multi-modal prediction of future decline toward Alzheimer's disease: an empirical study. PLoS One **17** (2022) e0277322

[28] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.S., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L.M., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q.: The Alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. Alzheimer's & Dementia **13** (2017) 561–571

[29] the ADNI team: ADNIMERGE: Alzheimer's Disease Neuroimaging Initiative. (2021) R package version 0.0.1.

[30] Aisen, P.S., Petersen, R.C., Donohue, M.C., Gamst, A., Raman, R., Thomas, R.G., Walter, S., Trojanowski, J.Q., Shaw, L.M., Beckett, L.A., et al.: Clinical core of the Alzheimer's disease neuroimaging initiative: progress and plans. Alzheimer's & Dementia **6** (2010) 239–246

[31] Schiratti, J.B., Allassonnière, S., Colliot, O., Durrleman, S.: A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. Journal of Machine Learning Research **18** (2017) 1–33

[32] Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics **33** (1962) 1065–1076

[33] Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. Geoscientific Model Development **7** (2014) 1247–1250

[34] Platero, C., Tohka, J., Strange, B.: Estimating dementia onset: AT(N) profiles and predictive modeling in mild cognitive impairment patients. Current Alzheimer Research **20** (2023) 778–790

[35] Monllau, A., Pena-Casanova, J., Blesa, R., Aguilar, M., Bohm, P., Sol, J., Hernandez, G.: Diagnostic value and functional correlations of the ADAS-Cog scale in Alzheimer's disease: data on NORMACODEM project. Neurologia (Barcelona, Spain) **22** (2007) 493–501

[36] O'Bryant, S.E., Humphreys, J.D., Smith, G.E., Ivnik, R.J., Graff-Radford, N.R., Petersen, R.C., Lucas, J.A.: Detecting dementia with the mini-mental state examination in highly educated individuals. Archives of neurology **65** (2008) 963–967

[37] Petersen, R.C., Aisen, P., Beckett, L.A., Donohue, M., Gamst, A., Harvey, D.J., Jack, C., Jagust, W., Shaw, L., Toga, A., et al.: Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. Neurology **74** (2010) 201–209

[38] Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J.Q., Bittner, T., Lifke, V., Corradini, V., Eichenlaub, U., Batrla, R., et al.: CSF biomarkers of Alzheimer's disease concord with amyloid-β PET and predict clinical progression: a study of fully automated immunoassays in BioFINDER and ADNI cohorts. Alzheimer's & Dementia **14** (2018) 1470–1481

[39] De Francesco, S., Galluzzi, S., Vanacore, N., Festari, C., Rossini, P.M., Cappa, S.F., Frisoni, G.B., Redolfi, A.: Norms for automatic estimation of hippocampal atrophy and a step forward for applicability to the italian population. Frontiers in Neuroscience **15** (2021) 656808

[40] Alzola, P., Carnero, C., Bermejo-Pareja, F., Sanchez-Benavides, G., Pena-Casanova, J., Puertas-Martin, V., Fernandez-Calvo, B., Contador, I.: Neuropsychological assessment for early detection and diagnosis of dementia: current knowledge and new insights. Journal of Clinical Medicine **13** (2024) 3442

[41] Kim, Y.J., Kim, S.E., Hahn, A., Jang, H., Kim, J.P., Kim, H.J., Na, D.L., Chin, J., Seo, S.W.: Classification and prediction of cognitive trajectories of cognitively unimpaired individuals. Frontiers in Aging Neuroscience **15** (2023) 1122927

[42] Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.: The Alzheimer's Disease Neuroimaging Initiative. Neuroimaging Clinics **15** (2005) 869–877

[43] Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., et al.: An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. NeuroImage **60** (2012) 1880–1889

[44] Wijeratne, P.A., Eshaghi, A., Scotton, W.J., Kohli, M., Aksman, L., Oxtoby, N.P., Pustina, D., Warner, J.H., Paulsen, J.S., Scahill, R.I., et al.: The temporal event-based model: Learning event timelines in progressive diseases. Imaging Neuroscience **1** (2023) 1–19

[45] Veitch, D.P., Weiner, M.W., Miller, M., Aisen, P.S., Ashford, M.A., Beckett, L.A., Green, R.C., Harvey, D., Jack Jr, C.R., Jagust, W., et al.: The Alzheimer's Disease Neuroimaging Initiative in the era of Alzheimer's disease treatment: a review of ADNI studies from 2021 to 2022. Alzheimer's & Dementia **20** (2024) 652–694