



Longitudinal survival analysis and two-group comparison for predicting the progression of mild cognitive impairment to Alzheimer's disease

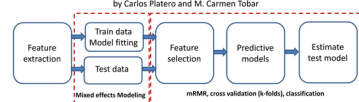


Carlos Platero*, M. Carmen Tobar

Health Science Technology Group, Technical University of Madrid, Ronda de Valencia 3, 28012 Madrid, Spain

GRAPHICAL ABSTRACT

*Longitudinal survival analysis and two-group comparison for predicting the progression of mild cognitive impairment (MCI) to Alzheimer's disease (AD)
by Carlos Platero and M. Carmen Tobar



Two approaches: a) two-group comparison and b) longitudinal survival analysis. The proposed approaches were developed and evaluated using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset with a total of 1330 visits from 321 subjects with a 3-year follow-up.

Visit	AUC	ACC	SEN	SPE
Baseline	0.860	78	79	76
Month 12	0.891	79	79	80
Month 24	0.925	84	85	82
Month 36	0.944	85	86	86

MRI markers: Hippocampal and pallidum volume and cortical thickness in the inferior parietal and medial temporal lobes. Neuropsychological measurements: ADAS13, FAQ and RAVLT Immediate.

ARTICLE INFO

Keywords:

Alzheimer's disease
Mild cognitive impairment
Longitudinal analysis
Magnetic resonance imaging

ABSTRACT

Background: Longitudinal studies using structural magnetic resonance imaging (MRI) and neuropsychological measurements (NMs) allow a noninvasive means of following the subtle anatomical changes occurring during the evolution of AD.

New Method: This paper compared two approaches for the construction of longitudinal predictive models: a) two-group comparison between converter and nonconverter MCI subjects and b) longitudinal survival analysis. Predictive models combined MRI-based markers with NMs and included demographic and clinical information as covariates. Both approaches employed linear mixed effects modeling to capture the longitudinal trajectories of the markers. The two-group comparison approaches used linear discriminant analysis and the survival analysis used risk ratios obtained from the extended Cox model and logistic regression.

Results: The proposed approaches were developed and evaluated using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset with a total of 1330 visits from 321 subjects. With both approaches, a very small number of features were selected. These markers are easily interpretable, generating robust, verifiable and reliable predictive models. Our best models predicted conversion with 78% accuracy at baseline (AUC = 0.860, 79% sensitivity, 76% specificity). As more visits were made, longitudinal predictive models improved their predictions with 85% accuracy (AUC = 0.944, 86% sensitivity, 85% specificity).

Comparison with Existing Method: Unlike the recently published models, there was also an improvement in the prediction accuracy of the conversion to AD when considering the longitudinal trajectory of the patients.

Conclusions: The survival-based predictive models showed a better balance between sensitivity and specificity with respect to the models based on the two-group comparison approach.

* Corresponding author at: Health Science Technology Group, Technical University of Madrid, Ronda de Valencia 3, 28012 Madrid, Spain.
E-mail address: carlos.platero@upm.es (C. Platero).

1. Introduction

Alzheimer's disease (AD), the leading cause of dementia in the aging population, is a degenerative brain disorder that causes a progressive decline in cognitive function (Holtzman et al., 2011; Jack, 2012). The aging global population causes the prevalence of AD to grow uncontrollably in the coming decades, creating a critical problem for health systems (Prince et al., 2013). The early identification of subjects who progress towards AD will allow the implementation of appropriate preventive treatments and interventions. Individual level clinical forecasting is complicated by the heterogeneous present in the prodromal phases of AD. Mild cognitive impairment (MCI) is often considered to be a transitional phase between healthy cognitive aging and dementia (Petersen et al., 2009). Thus, MCI represents a key prognostic and therapeutic target in the management of AD. However, MCI is a heterogeneous syndrome with varying clinical outcomes, and despite the high rate of conversion to AD (MCI progressing to AD (pMCI)), a significant number of MCI patients remain stable (sMCI) or regain normal cognitive function (Manly et al., 2008). Therefore, it is critical to understand AD and its progression and the successful identification of individuals with pMCI at early stages. A prognosis with high sensitivity and specificity in MCI-to-AD progression is extremely valuable in clinical practice and in medical research. In clinical practice, the reliability to accurately predict the diagnosis can help in clinical decisions on treatment strategies and early detection of at-risk subjects (Sperling et al., 2014). Clinical trial design could be improved, since it has been suggested that the low success of pharmaceutical trials in AD could be due to the inclusion in the study population of subjects too heterogeneous (Falahati et al., 2014).

Different modalities of disease indicators have been studied for AD progression, including neuroimaging biomarkers, blood tests, and neuropsychological assessments. Some studies have shown that baseline neuropsychological measures (NMs) have good power in predicting MCI progression to AD (Chapman et al., 2011). Other studies have demonstrated that neuroimaging and specifically structural magnetic resonance imaging (MRI)-based markers support an earlier and more precise MCI-to-AD diagnosis (Frisoni et al., 2010; Cuingnet et al., 2011; Eskildsen et al., 2015; Rathore et al., 2017). Some MRI-based markers complement each other, and markers applied in different parts of the brain are expected to be sensitive to different stages of the disease (Sørensen et al., 2017). Nevertheless, recent studies have suggested the combination of NMs with MRI-based markers to improve early diagnostic performance (Da et al., 2014; Moradi et al., 2015; Korolev et al., 2016; Gavidia-Bovadilla et al., 2017).

Most predictive models of MCI-to-AD conversion use only baseline data. Compared to cross-sectional analysis, longitudinal studies allow the observation of individual patterns of change, providing relevant information that can improve the differential diagnosis (Bernal-Rusiel et al., 2013b; Gavidia-Bovadilla et al., 2017; Minhas et al., 2017). The majority of studies in AD have reported longitudinal changes in structure only, such as the hippocampus or the temporal lobes. Nevertheless, longitudinal methods, such as linear mixed effects (LMEs) (Bernal-Rusiel et al., 2013a,b; Platero et al., 2019), allow the analysis of univariate or massive measures over time. The standard strategy for analyzing the association between longitudinal data and the conversion to AD is to perform a group comparison based on dichotomizing MCI subjects into converters and nonconverters (Chtelat et al., 2005; Moradi et al., 2015; Korolev et al., 2016; Gavidia-Bovadilla et al., 2017). However, the nature of the estimation problem with regard to MCI-to-AD conversion is not dichotomous. There is no homogeneity in the pMCI group (due to observed differences in conversion times) or in the sMCI group (due to uncertainties in the use of fixed cut-off periods in the follow-up). To overcome these drawbacks, survival models consider a unique clinical group that takes into account conversion times and finite follow-up or censoring (Kleinbaum and Klein, 2010; Sabuncu et al., 2014). Accordingly, predictive models of MCI-to-AD progression

were designed using Cox proportional hazards regressions. These models relate the biomarkers obtained from patients at baseline with conversion or censoring times and estimate the risk of AD (Devanand et al., 2007; Da et al., 2014; Pettigrew et al., 2016). However, the validity of the Cox models is based on the exploratory variables being time-independent or varying proportionally with time. Normally, the most discriminative markers between sMCI and pMCI patients are neither invariant nor proportional to time. For example, the classical hippocampal volume biomarker has annual atrophy ratios of 2% for the sMCI population and 3.5% in the pMCI group (Wolz et al., 2010; Iglesias et al., 2016; Platero et al., 2019). On the other hand, predictive models of MCI-to-AD progression with longitudinal data show better results than those based on cross-sectional studies. Following Sabuncu's proposal (Sabuncu et al., 2014), predictive models can be built by means of the temporal modeling of biomarker trajectories using LME combined with extended Cox survival analysis, which allows the use of exploratory variables that are time dependent (Kleinbaum and Klein, 2010). However, these authors only worked in a univariate study and were not applied as a predictive model of conversion from MCI to AD. Li et al. (2017) have also developed a predictive model of AD conversion estimating the risk of the univariate longitudinal trajectory of each marker, although recently they have built a multivariate approach (Li et al., 2018).

In the present study, we compared two different approaches of longitudinal multivariate predictive models to estimate the conversion of MCI to AD: a) comparison between two clinical groups, stable MCI and progressive MCI and b) survival analysis using the extended Cox model and logistic regression. Both approaches combined a very small number of MRI-based markers with standard cognitive measures, and both employed LME modeling to estimate the longitudinal trajectories of these measurements. The first approach was based on marginal longitudinal trajectory residues, while the second used risk ratios. Our experiments revealed that the survival-based predictive models with multisource data (i.e. MRI and NMs) present better prediction results starting from the second year of follow-up and showed a better balance between sensitivity and specificity over time than those of the predictive models based on the two-group comparison.

2. Materials

The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset was selected to evaluate the comparison among predictive models of MCI-to-AD conversion, where subjects have different numbers of visits to the clinic (Wyman et al., 2013). ADNI has been used by many publications that have focused on the early prediction of the conversion to AD (Moradi et al., 2015; Eskildsen et al., 2015; Korolev et al., 2016; Gavidia-Bovadilla et al., 2017). In the ADNI study, longitudinal brain T1-weighted MRI data were acquired at baseline and at regular intervals. MRI acquisition was performed according to the ADNI acquisition protocol (Jack et al., 2008). In our study, all images downloaded from ADNI were preprocessed using the N3 method (Sled et al., 1998) or grad warped, followed by B1 bias field correction and N3 intensity nonuniformity correction (Jack et al., 2008). The MRI-based markers and NMs used in this study correspond to measures of neurodegeneration at the 3-year follow-up. We included unbalanced longitudinal data from 321 subjects at multiple time points: baseline and 6, 12, 18, 24 and 36 months.

To predict MCI-to-AD conversion, we chose patients diagnosed with MCI at their baseline assessments and asked whether their diseases had converted to dementia within 3 years. At each visitation, a clinical diagnosis was made to identify MCI subjects whose diseases had converted to probable AD according to ADNI clinical assessments (The ADNI team, 2018). The conversion time was established between the baseline and the first visit where the patient was diagnosed with dementia, so long as the diagnosis was reconfirmed on subsequent visits. The sMCI group only included those MCI patients who were followed-

Table 1

Demographic and clinical details of the subset of the ADNI database used in this study. Data as represented as means and standard deviations (SDs) unless otherwise specified. ANOVA was used for baseline ages and neuropsychological scores. Gender and APOE $\epsilon 4+$ were analyzed with a chi-square test. A p -value < 0.01 was considered statistically significant. sMCI = Stable mild cognitive impairment; pMCI = Progressive mild cognitive impairment; MMSE = Mini-Mental State Examination; APOE $\epsilon 4+$ = APOE $\epsilon 4$ carriers or non-carriers.

Type (N. subjects)	sMCI (165)	pMCI (156)	F	p
Sex male (%)	110 (67%)	92 (59%)	1.71	0.19
Baseline age	74.9 (7.3)	74.6 (6.8)	0.13	0.72
Years of education	15.7 (3.1)	15.8 (2.8)	0.07	0.79
APOE $\epsilon 4+$ (%)	69 (42%)	108 (69%)	23.3	< 0.001
MMSE (Baseline)	27.6 (1.7)	26.6 (1.8)	23.53	< 0.001
MMSE (Month 6)	27.6 (2.2)	25.3 (2.9)	61.92	< 0.001
MMSE (Month 12)	27.9 (2.3)	24.9 (2.9)	102.53	< 0.001
MMSE (Month 24)	27.5 (3.0)	22.8 (3.9)	99.84	< 0.001
MMSE (Month 36)	27.3 (6.3)	21.3 (3.7)	63.71	< 0.001

up and whose diseases did not convert to probable AD. In addition, the last visits of the sMCI subjects defined the censorship times. The MCI subjects were divided into 165 sMCI subjects and 156 pMCI subjects according to the clinical follow-up for 36 months.

Current guidelines recommend the use of the same subjects as public cohorts to make studies more comparable (Wyman et al., 2013). Following this suggestion, the standardized two-year ADNI list for MCI subjects was used (Wyman et al., 2013). Eskildsen et al. (2013) published a list of MCI subjects from the ADNI database whose diseases had converted to AD or who had remained stable at the 3-year follow-up. In this study, Eskildsen's list was used and expanded with the MCI subjects of the standardized two-year ADNI list, who were followed for 36 months. An overview of the subject groups is given in Table 1. For each group, the total number of subjects, the number of males, and the Mini-Mental State Examination (MMSE) scores (Folstein et al., 1975) are shown. No significant differences in age and sex were observed among the clinical groups. There were significant differences in MMSE scores among the clinical groups at all time points. Table 2 reports the time points (Baseline, Month 6, Month 12, Month 24 and Month 36) that were available for the selected subjects.

2.1. Image processing

The FreeSurfer 5.3 software package was used for subcortical segmentation and surface-based cortical processing (Fischl et al., 2002; Fischl and Dale, 2000). All scans included in the study were pre-processed through the same longitudinal pipeline, including those participants with single time-point acquisitions (Reuter et al., 2012). The preprocessing was performed separately for each subject and scan. After cross-sectional preprocessing, FreeSurfer's longitudinal processing pipeline was used to achieve unbiased within-subject registration (Reuter et al., 2010), which is achieved by registering scans from each time point of a subject using a robust and inverse consistent registration

Table 2

Number and timing of scans per time point by clinical group. sMCI = stable mild cognitive impairment; pMCI = Progressive mild cognitive impairment.

Type (N. subjects)	sMCI	pMCI	Time from baseline (in month)
Baseline	165	156	0
Month 6	159	148	7.0(0.9)
Month 12	158	143	12.9(0.8)
Month 18	58	45	19.0(0.9)
Month 24	124	101	25.0(0.8)
Month 36	46	27	36.9(1.1)
Total	710	620	

algorithm. Several steps in the processing of the serial MRI scans, such as skull stripping, Talairach transformations and atlas registrations, were then initialized with common information from the subject-specific template. This implementation ensures that all time points are treated uniformly and with a very high degree of reproducibility (Reuter et al., 2012).

3. Method

A method with three fundamental stages was used to estimate the prediction of MCI-to-AD conversion: 1) Feature extraction: a pool of features was extracted from the MRI data and the cognitive measurements from the MCI population. 2) Feature selection: a feature ordering stage that uses the minimal-redundancy-maximal-relevance (mRMR) algorithm (Peng et al., 2005) to propose good subsets of markers for the prediction of AD conversion. 3) Classification: Two longitudinal classifier approaches were evaluated in terms of cross-validated classification accuracy.

A group of biomarkers was computed from structural MRI scans as potential predictors of MCI-to-dementia progression. The volumes of the hippocampus, amygdala, caudate nucleus, pallidum and putamen were selected (Sørensen et al., 2017) and normalized using the intracranial volume. The eight cortical thickness (CT) measures that were classified as *AD-vulnerable* cortical regions were also proposed as potential features (Pettigrew et al., 2016). These regions consisted of the entorhinal cortex, temporal pole, inferior temporal gyrus, middle temporal gyrus, inferior parietal cortex, superior parietal cortex, precuneus, and posterior cingulate cortex. All left and right hemisphere measures were independently considered and combined as bilateral features to be examined.

NMs are commonly used to classify dementia patients. These features test patients in multiple cognitive domains, such as episodic memory, learning and language (Ferreira et al., 2017). The cognitive features used in the models included total scores and subscores on five neuropsychological tests: Rey's Auditory Verbal Learning Test (RAVLT), Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-cog), MMSE, Clinical Dementia Rating (CDR), and Functional Activities Questionnaire (FAQ). These standard cognitive measurements are explained in the ADNI General Procedures Manual (The ADNI, 2018). See more information about extracted markers in supplementary material.

Next, longitudinal data were processed by means the mixed effects modeling. As the mRMR method does not provide any information on the optimal number of features, a combined feature ordering and classification stage was adopted to identify optimal feature subsets (Eskildsen et al., 2013; Korolev et al., 2016). Feature subsets of different dimensions were defined using the mRMR algorithm. In the classification stage, these feature subsets were evaluated in terms of the cross-validated classification accuracy. Fig. 1 shows an illustrated diagram of the main steps of the proposed framework.

3.1. Longitudinal classification

Cox proposed a proportional hazards model for survival data analysis (Cox, 1975). The Cox model expresses the relation between a hazard function at time t for a subject with an explanatory vector $X = \{X_1, X_2, \dots, X_p\}$, which is independent of time:

$$h(t, X) = h_0(t) \exp\left(\sum_{k=1}^p \alpha_k \cdot X_k\right),$$

where $h_0(t)$ is the baseline hazard function and $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_p\}$ is a vector of regression coefficients. It is a semi-parametric model that estimates that the risk of the clinical event (in our case, the conversion from MCI to AD) is given by the composition of the baseline hazard that depends on time and the linear combination of the exploratory variables, which were measured from a single baseline visit. The hazard

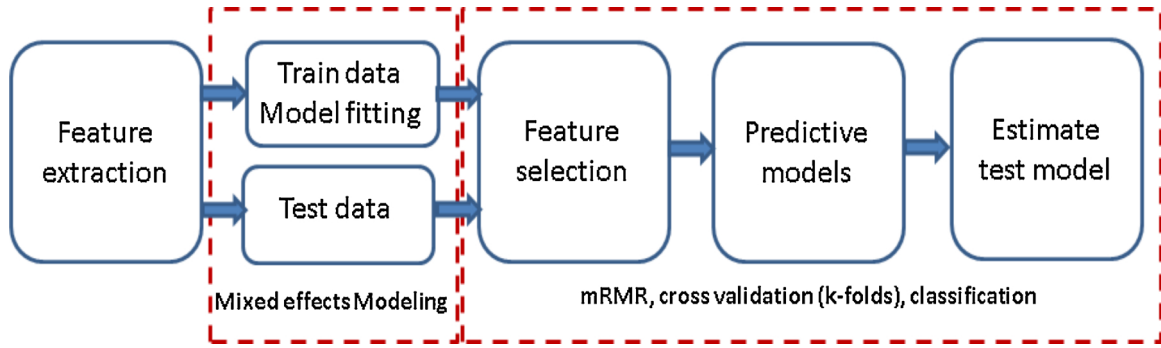


Fig. 1. Overview of the proposed comparative framework.

ratio (HR) quantifies the differential risk of a subject characterized by X_S in relation to a reference subject characterized by X_R :

$$HR(X_S, X_R) = \frac{h(t, X_S)}{h(t, X_R)} = \exp\left(\sum_{k=1}^p \alpha_k (X_{S,k} - X_{R,k})\right), \quad (1)$$

when $HR > 1$ indicates that the subject characterized by X_S has an increased risk of disease conversion with respect to the reference subject, X_R . Conversely, if $HR < 1$, the conversion risk is decreased. The main assumption of the Cox model is the ratio of the hazard functions of the two samples over time.

The partial likelihood maximization method is used to calculate the α coefficients (Cox, 1975). Let m denote the event of the conversion time from MCI to AD ($m \in \{1, 2, \dots, M\}$, where M is the number of converters such that M is less than the number of subjects in the population), and the partial likelihood is expressed as the product of the M terms. Each of these terms is calculated between the hazard functions of the subjects whose diseases converted at the m th event of time (t_m) in relation to all the subjects whose diseases have not yet converted:

$$L = \prod_{m=1}^M L_m(\alpha) \quad L_m(\alpha) = \frac{h(t_m, X_m)}{\sum_{r \in R_m} h(t_m, X_r)}, \quad (2)$$

where R_m is the so-called risk set at t_m , which indexes the subjects who remain as MCI patients at t_m . Subsequently, the partial likelihood function is maximized with respect to the unknown α parameters of the model. This set of equations will allow estimation of the α parameters using a numerical optimization method.

Variables such as gender or years of education can be considered independent of time. Age, meanwhile, is proportional to time. However, most of the biomarkers in a longitudinal study are neither constant nor proportional with time. The Cox model can be extended for independent and dependent variables over time:

$$h(t, X) = h_0(t) \exp\left(\sum_{k=1}^{p_1} \gamma_k \cdot X_k + \sum_{l=1}^{p_2} \delta_l \cdot Y_l(t)\right), \quad (3)$$

where the second term in the exponential includes the effects of p_2 time-varying variables $Y(t) = \{Y_1(t), Y_2(t), \dots, Y_{p_2}(t)\}$ with associated coefficients $\delta = \{\delta_1, \delta_2, \dots, \delta_{p_2}\}$. The estimation of the model parameters is also calculated by partial likelihood maximization. However, for each conversion event time, the risk set should take into account the values of the exploratory variables according to the conversion time, t_m . Therefore, to estimate the extended Cox model parameters, the values of the markers at each instant of conversion t_m for all the subjects of the risk set are required. However, in a longitudinal study, image data and neuropsychological tests are acquired at regular periods and unbalanced samples, i.e. the number of visits for each subject can be variable. Moreover, the acquisition times of the markers and the instants of conversion time do not necessarily coincide. To solve these drawbacks, the longitudinal trajectories of the measurements were modeled by LME models, allowing estimates of these measurements to

be known at any time.

LME models are extensions of linear models that allow both fixed and random effects, which consider the sources of variation between subjects and within subjects (Bernal-Rusiel et al., 2013b). The LME model is expressed as:

$$Y_i = Z_i \beta + W_i b_i + e_i,$$

where Y_i is the vector of a feature for the time points of subject i , Z_i is the design matrix for the fixed effects (including variables such as clinical group, age, sex, education, and scan time), and β are the fixed effects coefficients, which are identical for all subjects. In addition to the fixed effects, a mixed effects model is used for subject-specific random effects, where W_i is the design matrix for the random effects, b_i is a vector of the random effects, and e_i is a vector of measurement errors. The components of b_i reflect how the subset of regression parameters for the i th subject deviates from those of the population. The LME models were built with an intercept and slope as random effects to be included in the longitudinal trajectory (Bernal-Rusiel et al., 2013b; Platero et al., 2019).

The influence of sociodemographic characteristics was collected with the effect of age at baseline (Age_i), sex (Sex_i) and years of education ($Education_i$) (Jiang et al., 2014; Liu et al., 2012). Apolipoprotein E (APOE) genotype status is the most prevalent genetic risk factor for AD (Saykin et al., 2010). The APOE genotype was assessed, and patients were characterized as $\epsilon 4$ carriers or $\epsilon 4$ noncarriers. The interaction between APOE genotype status and time was also considered based on the evidence that $\epsilon 4$ accelerates atrophy during the prodromal phases of AD (Bernal-Rusiel et al., 2013b). The combinations of these fixed effects in the LME modeling did not show significant differences for the prediction of conversion from MCI to AD (Platero et al., 2019). The best predictive results were obtained when constructing LME models that added age, sex and years of education as covariates. The APOE genotype was not selected as a covariate in the LME modeling. This decision was consistent with APOE $\epsilon 4$ being a risk factor for AD; however, its value for individual patient predictions was limited (Da et al., 2014).

In the case of the two-group comparison approach, the effect of the clinical group was considered; therefore, a Boolean variable for the clinical group ($Group_i$) and its interaction with time were added:

$$y_{ij} = (\beta_1 + \beta_2 \cdot Group_i + \beta_3 \cdot Age_i + \beta_4 \cdot Education_i + \beta_5 \cdot Sex_i + b_{i1}) + (\beta_6 + \beta_7 \cdot Group_i + b_{si}) t_{ij} + e_{ij}, \quad (4)$$

where $j = 1, \dots, n_i$ indexes the time points with n_i indicating the number of scans for subject i , y_{ij} is the j th measure of a feature for subject i , t_{ij} is the scan time from baseline (in years), and $\beta_r = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]^T$ and $\beta_s = [\beta_6, \beta_7]^T$ are the intercept and slope, respectively. The boolean variable $Group_i$ is true if the i th subject progresses to AD and false if the subject remains stable. However, in the survival analysis, the LME modeling did not consider the effect of the clinical group:

$$y_{ij} = (\beta_1 + \beta_2 \cdot \text{Age}_i + \beta_3 \cdot \text{Education}_i + \beta_4 \cdot \text{Sex}_i + b_{ri}) + (\beta_5 + b_{si})t_{ij} + e_{ij}. \quad (5)$$

3.2. Predictive models using the two-group comparison

Regarding the classification of subjects between the two clinical groups of MCI, we proposed to compare the longitudinal tendency of a feature using an LME model trained between these two clinical groups. For a feature, the difference between the longitudinal trajectory of the i th subject and the LME model is described by the random vectors b_i and e_i , which follow mean zero-Gaussian multivariate distributions, indicating a population-averaged mean of $E(Y_i) = Z_i\beta$ (Bernal-Rusiel et al., 2013b). Therefore, the longitudinal trajectory residue is defined as follows:

$$l_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - (Z_i)_j\beta), \quad (6)$$

where $(Z_i)_j$ is the j -row vector of the design matrix, which is activated by the boolean variable of the clinical group ($Group_i = 1$), i.e., the effects of AD progression compared to stable MCI. The l_i samples belonging to the pMCI group follow a normal distribution with zero mean and variance determined by b_i and e_i . In contrast, the l_i samples not belonging to the pMCI group will also follow a normal distribution with the same variance but with a bias in relation to zero-mean, which is related to the fixed effects of the clinical groups. The random variable l_i is used to train and classify the features by linear discriminant analysis (LDA). All longitudinal trajectory residues of the selected features were assumed to be independent. These marginal residues were used as data for training and testing of the LDA classifiers. Note that these classifiers do not require the adjustment of any parameter, as is traditionally done at this stage (Moradi et al., 2015; Korolev et al., 2016). Moreover, the marginal residues of the longitudinal trajectories of the markers as inputs to the LDA had been experimentally validated (Platero et al., 2019).

3.3. Predictive models using survival analysis

Given a population of MCI subjects, from which a series of longitudinal measurements of both MRI and neuropsychological tests have been extracted, the longitudinal measurements were modeled by LME using age, sex and education as covariates (without the clinical group effect, see Eq. (5)). Therefore, for each subject and each of the different conversion times, it was possible to estimate the value of each marker. In addition, it was also known whether the subject's disease had become AD in the follow-up period. For those subjects whose disease had converted to AD, the conversion time was calculated from the baseline. In the case of sMCI subjects, the censorship time was also known.

An extended Cox model was constructed for each significant discrete time. In our case, we had four models, i.e. one for the beginning of the study, and one each for the 12, 24 and up to 36 months follow-up. To build each of these four predictive models, the hazard ratios were calculated (see Eq. (1, 2),(3)) and converted into probabilistic terms of conversion from MCI to AD using the logistic regression model:

$$p(X_{S,v}) = \frac{1}{1 + \frac{1}{HR_v(X_{S,v}, X_{R,v})}}, \quad (7)$$

where HR_v is the hazard ratio in the visit v , ($v \in \{0, 12, 24, 36\}$) and $X_{S,v}$ and $X_{R,v}$ are the vectors of the exploratory variables of the subject and the reference in the visit v , respectively, and these vectors are formed with p_1 time-independent and p_2 time-varying variables, the latter being modeled by means LME. Note the bias due to the use of LME modeling in the time-dependent measurements as input to the extended Cox model (see supplementary materials). HR_v was built by means an extended Cox-LME model with $X_{R,v}$, which was calculated using a

random subset of the training population at visit v . This subset was sampled with the same number of subjects, both sMCI and pMCI patients. The components of $X_{R,v}$ were defined by the average values of this population and scaled by their standard deviations. Therefore, each exploratory variable of $(X_{S,v} - X_{R,v})$ was defined as a z-score. Therefore, if a subject with $X_{S,v}$ shows a $HR_v(X_{S,v}, X_{R,v}) > 1$, then $p(X_{S,v}) > 0.5$, and on the contrary, when $HR_v(X_{S,v}, X_{R,v}) < 1$, then $p(X_{S,v}) < 0.5$, where $p(X_{S,v})$ denotes the conversion probability into AD of the disease of subject S at visit v .

3.4. Feature selection and building the predictive models

For both approaches, a nested cross-validation (CV) procedure was used to avoid model overfitting and optimistically biased estimates of model performance (Korolev et al., 2016). The procedure consisted of two nested CV loops: an inner loop, designed to select the optimal feature subsets for the proposed models, and an outer loop, designed to obtain an unbiased estimate of model performance. Note that in this manner, double-dipping, i.e., using the same data for both feature selection and learning the classifier, was avoided. A nested k-fold CV procedure was applied. The value for k was fixed to 10, a value that was found through experimentation to generally result in a model skill estimate with low bias a modest variance (Kuhn and Johnson, 2013). Both the outer and inner CV loops used a 10-fold CV design. In the outer CV loop, the data were partitioned into model and test data (see Fig. 2). In the inner CV loop, the model data were again partitioned into training and validation data.

For each inner CV loop, a set of combinations of markers with different dimensions was proposed, which were subsequently evaluated in the outer CV loop. The feature ordering by means mRMR was combined with the evaluation of the predictive models constructed with the candidate feature subsets. These two steps of each inner CV loop were developed with the following activities: I) A resampling method searched for the first 10 subsets of each dimension that appeared most frequently in mRMR. The feature subsets with dimensions from 2 to 5 for single-source predictive models and from 4 to 10 for multi-source models were explored. For this purpose, random partitions of training data were subjected to the mRMR algorithm. We used the mutual information difference metric, and the features were normalized to zero mean and unit variance in the mRMR algorithm. This sequence (i.e., partitions of the random subsets of training, applying mRMR and proposing combination features) was repeated 100 times for each inner loop. For each dimension, the 10 combinations of features that most frequently appeared were selected. And II) Predictive models were built using only the training data with the above candidate feature subsets. Of these, the 3 top-performing combinations of markers in terms of classification accuracy by the evaluation of their corresponding models were selected. Therefore, for each outer iteration, 30 subsets for each dimension were evaluated. In the outer CV loop, for each candidate feature selection, a predictive model was built from the training data and its performances were evaluated with the withheld test data, which were not used during feature selection, model selection, or final model construction. For better replicability, the nested 10-fold CV procedure was repeated with different partitions of the data, generating multiple performance estimate values. In total, there were 30,000 evaluations of selected subsets for each dimension. Note that these 30,000 proposals for each dimension of the combinations of markers only use training information. Predictive models with more frequent appearances of the feature subsets (i.e. number of times that the combination of proposed features was evaluated by the CV procedure), higher AUC values and balanced sensitivity and specificity were selected. Fig. 2 shows the general procedure for the development of the predictive models and their subsequent evaluation.

A MATLAB implementation of our method is available at <https://www.nitrc.org/projects/twogrsurvana/>. As well as a demo with the ADNI data used in this study. The proposed algorithm may generate

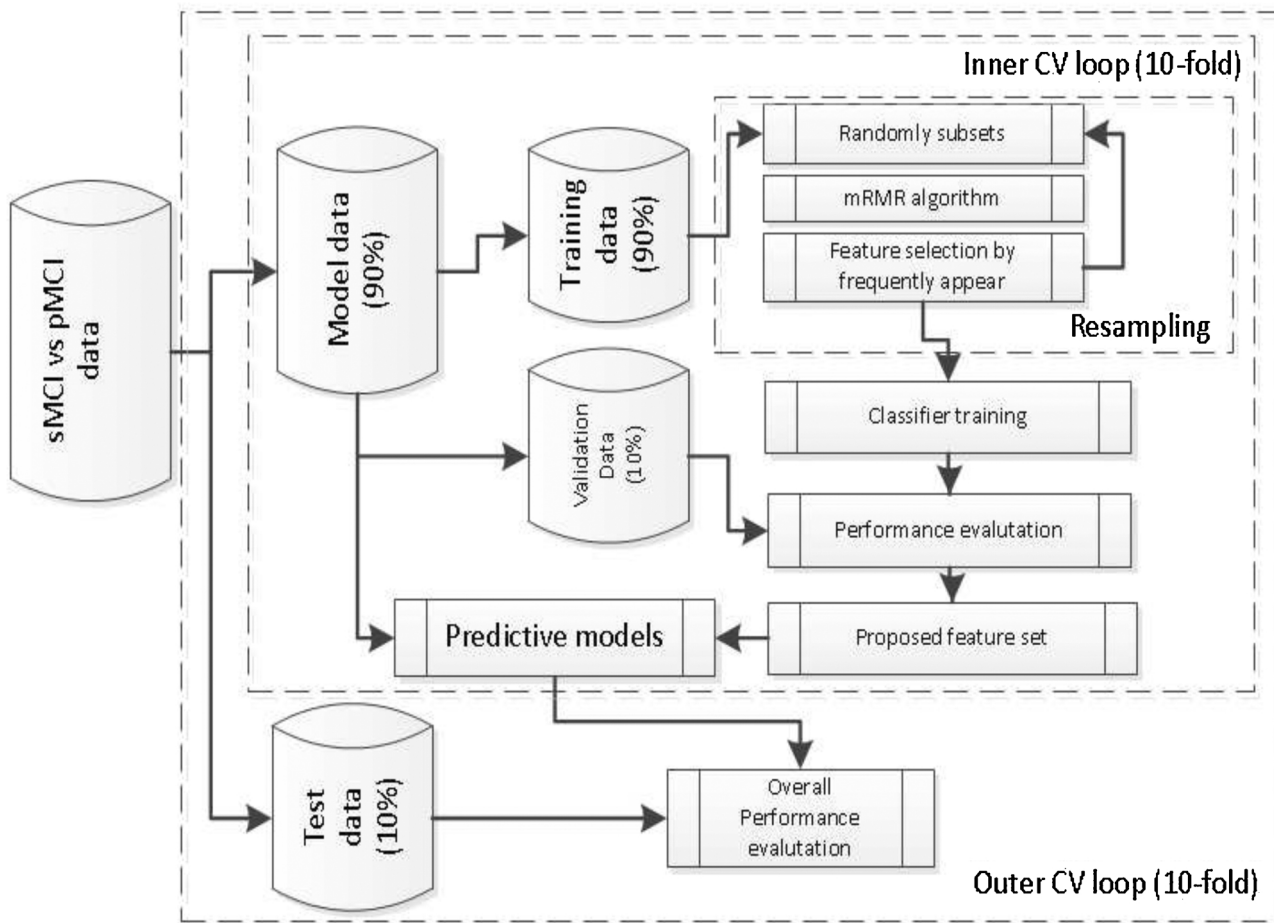


Fig. 2. Nested 10-fold cross-validation procedure for model development and evaluation.

new predictive models. Faced with a new problem, the subject identifiers, their visits, as well as a list of proposed markers to be explored must be provided. Note that both the feature selection, and the construction of the predictive models do not require tedious parameter adjustment, since mRMR does not require any parameter, nor the modeling of the trajectories with the LME approach, nor the classification tasks that also does not require any parameter.

4. Results

A total of 1330 scans from 321 subjects were processed using the proposed approaches. To validate the consistency of the applied longitudinal image processing pipeline, quality control was used between the scans and segmentations of each subject (see supplementary materials). For each visit by each subject, subcortical volumes and CT measurements were obtained using the longitudinal pipeline of Freesurfer (Reuter et al., 2012). The feature extraction stage resulted in 40 ROI-based MRI predictors of cortical and subcortical structures and

11 NMs. Considering the 51 features, first, a univariate analysis of each of the markers was performed, and then a multivariate analysis was carried out to generate the proposed predictive models.

4.1. Univariate analysis of the markers

Given the LME modeling of each longitudinal measurement on the study population, hypothesis tests were performed on the marker discrimination capacities according to the clinical group and time. In general, for a given contrast matrix C , the two competing hypotheses are as follows:

$$H_0: C\beta = 0 \text{ and } H_A: C\beta \neq 0,$$

where the β parameters were estimated from Eq. (4). Under the null hypothesis, it can be shown that the samples follow an F-distribution with different degrees of freedom depending on the contrast matrix (Bernal-Rusiel et al., 2013b). Table 3 shows the p-values for the top-most discriminating markers between sMCI and pMCI. For the first contrast matrix (C_1), the differences of each marker between the two

Table 3

The most significant p-values when comparing values between clinical groups at baseline (C_1) and over time (C_2). The ranking was determined based on the increasing order of the sum of the two p-values. The notation of the MRI-based features is given by an acronym of the brain structure, accompanied by a subindex, which indicates a measure of volumetry (V) or cortical thickness (T). A13 = ADAS13; ADV = AD-vulnerable (Pettigrew et al., 2016); MT = Middle temporal lobe; FAQ = Functional Activities Questionnaire; IT = Inferior temporal lobe; E = Entorhinal; RT_{im} = RAVLT Immediate; IP = Inferior parietal lobe; H = Hippocampus.

Features	A13	ADV_T	MT_T	FAQ	IT_T	E_T	RT_{im}	IP_T	H_V
p-values (C_1)	$3.1E - 18$	$1.2E - 12$	$7.8E - 12$	$3.6E - 11$	$4.8E - 9$	$9.0E - 9$	$7.9E - 15$	$6.0E - 10$	$4.0E - 11$
p-values (C_2)	$3.3E - 16$	$1.1E - 16$	$4.8E - 14$	$3.3E - 17$	$2.1E - 16$	$5.8E - 14$	$1.8E - 6$	$2.2E - 6$	$1.9E - 5$
Ranking	1	3	4	5	8	10	17	18	22

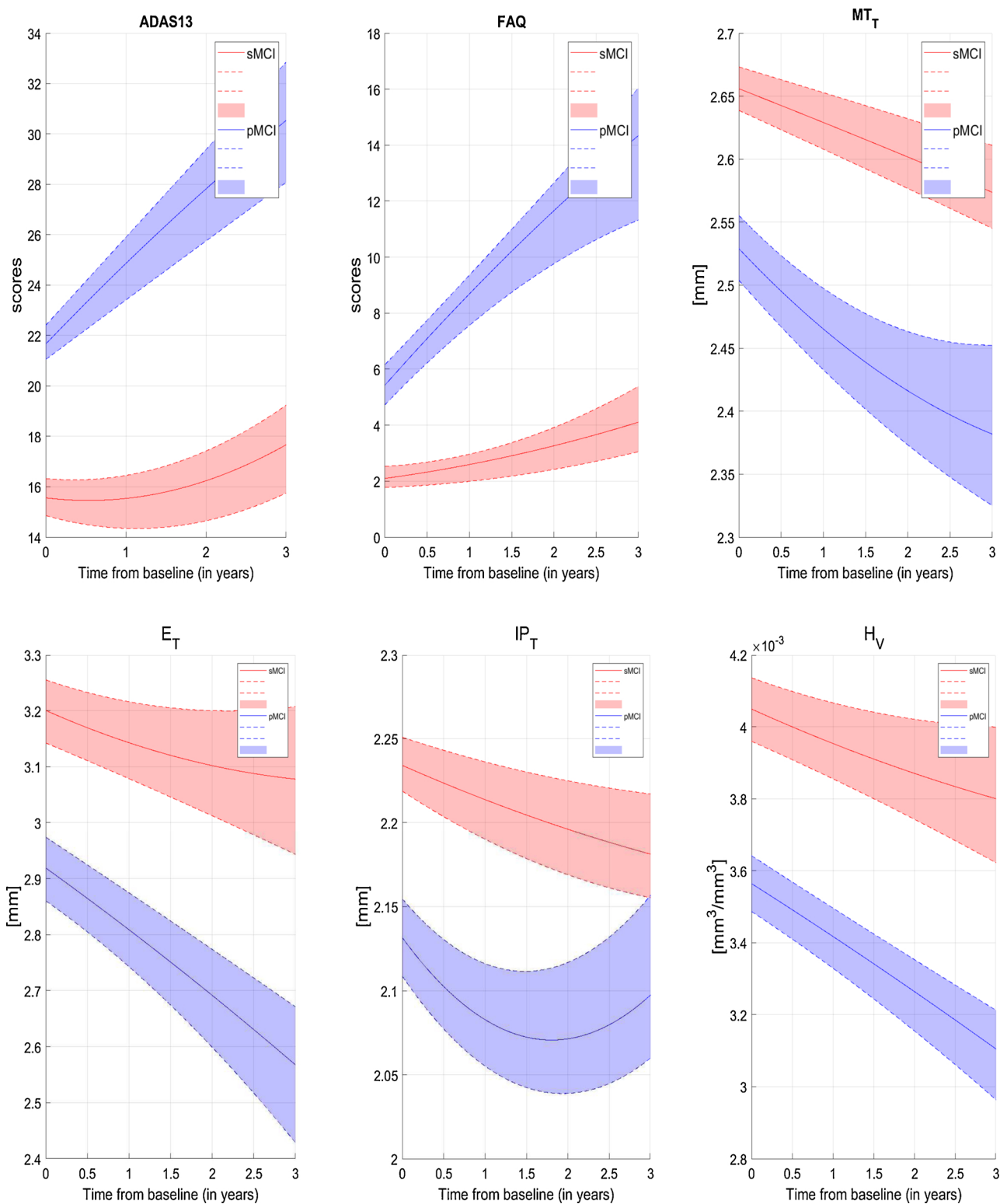


Fig. 3. The smoothed longitudinal trajectories of some of the best biomarkers under the univariate analysis. Dashed lines show 95% confidence intervals. The notation of the MRI-based features is given by an acronym of the brain structure, accompanied by a subindex that indicates a measure of volumetry (V) or cortical thickness (T). CTs are measured as the mean value of each ROI. The volume measurements are normalized by the intracranial volume (ICV). FAQ = Functional Activities Questionnaire; MT = Middle temporal lobe; E = Entorhinal; IP = Inferior parietal lobe; H = Hippocampus.

groups were measured at baseline. For the second contrast matrix (C_2), the discrepancies between sMCI and pMCI were evaluated over time. The markers were named without indication as to which hemisphere or subscore they corresponded. In addition, the ranking was determined based on the increasing order of the sum of the two p-values. The absent

markers in the sorting correspond well to marker subgroups (for example, the marker in the second position was ADAS11) or the same marker on the left, right or bilateral side (for example, the markers in the sixth and seventh positions were the cortical thickness of the left and right medial temporal lobe, respectively). It should be noted that

the first volume measure occupies position 22 in the ranking, corresponding to the normalized left hippocampal volume. Fig. 3 displays the smoothed longitudinal trajectories of some of the best biomarkers under the univariate analysis.

4.2. Performance of the predictive models of MCI-to-AD progression

However, adding the above mentioned first *n* markers did not guarantee that the prediction of AD conversion would be improved. It was proposed to apply a nested k-fold CV procedure, which selected the best subsets of features, and to build the predictive models with their subsequent evaluations (see Section 3.4). All training data was reused for the construction of predictive models in any of its visits and approaches used. Both proposed methods have as input to the construction of the predictive models the LME modeling of the longitudinal trajectories of the selected markers. LME modeling was calculated with all visits of training subjects (see Fig. 1). Another very different thing happens with the information of the test subjects. In the two-group comparison approach, the input to the linear discriminant analysis uses the longitudinal trajectory residue (see Eq. (6)). Therefore, all the measurements of a test subject, in the selected markers, from the baseline to the censorship visit were considered. Regarding the predictive models based on survival analysis, since the models were stratified by visits, the exploratory variables of the test subjects were compared with the reference population of that censored visit, then only the information of the censored visit was used, although in the construction of the extended Cox models, all the information of the training subjects was used, from the first visit to the last. Table 4 and 5 summarize the scores of the predictive models with the two types of approaches over time. Several predictive models instead of one for each

approach were presented due to these proposed models exhibiting combinations of markers with similar performance.

For predictive models with only MRI-derived markers (see Table 4), i.e. single-source models, the dimensions of the feature vector ranged from 2 to 5, and the selected markers were similar in both approaches. Hippocampal and pallidum volume and CT in the inferior parietal and medial temporal lobes were usually selected. Regarding the scores, as expected, we observed a sustained increase in AUC and accuracy as more visits of the subjects were considered in both approaches. The extended Cox models show a better balance between sensitivity and specificity throughout the longitudinal study. In addition, these survival models also reflected an increase in sensitivity and specificity scores over time. In contrast, the two-group comparison approach remained stable with a sensitivity of approximately 74% over time, and a specificity that increased with time. During the first year, the results of both approaches were comparable. However, from the second year onward, the survival analysis showed improved scores over those of the comparison method.

For multisource predictive models (i.e. MRI data and NMs), in either of the two approaches, the dimensions of the feature vectors range from 5 to 7, the sum of either two or three measurements from MRI data and of either three or four neuropsychological tests. There were many coincidences over the biomarkers chosen in the two approaches. In volumetry, one or two measures related to the hippocampus and pallidum appeared, especially on the left side. In cortical thickness, there was a unique measurement, one of either the inferior parietal lobe or the medial temporal lobe. Regarding NMs, the combined use of ADAS13, FAQ and RAVLT Immediate overwhelmingly coincided. In the same way as in the single-source models, the two approaches maintained similar scores during the first year. Starting in the second year, the

Table 4

Scores for predicting MCI-to-AD conversion using only MRI-based biomarkers. Each row shows the classification results of the predictive models according to the approach (Two-group comparison, 2G, and Survival analysis, SA) and visit (Baseline, bl, Month 12, m12, Month 24, m24, Month 36, m36). Numbers within parentheses are the 95% confidence interval except the frequency column. The notation of the MRI-based features is given by an acronym of the brain structure, accompanied by subindex and super-index. Subindex indicates a measure of volumetry (V) or cortical thickness (T). Super-index indicates whether the measurement is from the left (L), right (R) or bilateral (B) hemispheres. H = Hippocampus; P = Pallidum; MT = Medial Temporal; IP = Inferior Parietal; E = Entorhinal; AUC = Area under the curve; ACC = Accuracy; SEN = Sensitivity; SPE = Specificity; Frequency = Minimum and maximum range of the number of times that the combination of proposed features was evaluated by the cross-validation procedure.

Data	AUC	ACC (%)	SEN (%)	SPE (%)	Frequency	Optimal feature subsets
MRI ^{2G} _{bl}	0.769(0.763 0.775)	69.7(69.1 70.3)	74.1(73.2 74.9)	65.6(64.7 66.4)	376-981	H ^L _V , P ^L _V , MT ^R _T H ^L _V , P ^L _V , IP ^L _T , MT ^R _T E ^R _T , IP ^L _T
MRI ^{SA} _{bl}	0.769(0.763 0.775)	70.2(69.6 70.8)	68.8(68.0 69.7)	71.5(70.7 72.2)	756-1760	H ^L _V , P ^L _V , IP ^L _T , MT ^R _T H ^L _V , P ^L _V , E ^R _T , IP ^R _T H ^L _V , P ^L _V , IP ^L _T , E ^R _T
MRI ^{2G} _{m12}	0.797(0.791 0.803)	71.3(70.7 71.9)	73.9(73.0 74.7)	69.2(68.3 70.1)	473-552	H ^L _V , P ^L _V , IP ^L _T , MT ^R _T H ^L _V , P ^L _V , IP ^L _T , E ^R _T H ^L _V , P ^L _V , IP ^L _T
MRI ^{SA} _{m12}	0.789(0.783 0.795)	70.9(70.4 71.5)	71.4(70.5 72.3)	70.7(69.9 71.6)	636-1915	H ^L _V , P ^L _V , MT ^R _T H ^R _V , P ^L _V , IP ^L _T H ^R _V , P ^L _V , IP ^R _T
MRI ^{2G} _{m24}	0.810(0.804 0.816)	72.4(71.8 73.0)	73.9(73.0 74.8)	71.0(70.2 71.8)	363-573	H ^R _V , P ^L _V , IP ^L _T H ^L _V , P ^L _V , IP ^L _T , MT ^R _T H ^L _V , P ^L _V , MT ^R _T
MRI ^{SA} _{m24}	0.817(0.810 0.823)	73.7(73.0 74.3)	75.5(74.6 76.5)	71.9(70.9 72.8)	502-758	H ^R _V , P ^L _V , MT ^R _T H ^R _V , P ^L _V , IP ^R _T , E ^R _T H ^R _V , P ^L _V , IP ^L _T
MRI ^{2G} _{m36}	0.858(0.849 0.868)	74.2(73.3 75.1)	75.7(73.9 77.4)	72.9(71.7 74.2)	574-1471	H ^L _V , P ^L _V , IP ^L _T , E ^R _T , MT ^R _T H ^L _V , P ^L _V , IP ^L _T , MT ^R _T H ^R _V , P ^L _V , IP ^R _T , E ^R _T
MRI ^{SA} _{m36}	0.885(0.872 0.898)	78.3(76.9 79.7)	79.0(76.5 81.5)	78.1(76.2 80.0)	470-1165	H ^L _V , P ^L _V , IP ^L _T , E ^R _T , MT ^R _T P ^L _V , IP ^R _T , E ^R _T H ^R _V , P ^L _V , IP ^L _T

Table 5

Scores for predicting MCI-to-AD conversion using multisource biomarkers. Each row shows the classification results of the predictive models according to the approach (Two-group comparison, 2G, and Survival analysis, SA) and visit (Baseline, bl, Month 12, m12, Month 24, m24, Month 36, m36). Numbers within parentheses are the 95% confidence interval except the frequency column. The notation of the MRI-based features is given by an acronym of the brain structure, accompanied by subindex and super-index. Subindex indicates a measure of volumetry (V) or cortical thickness (T). Super-index shows the measurement is from the left (L), right (R) or bilateral (B) hemispheres. H = Hippocampus; P = Pallidum; MT = Medial Temporal; IP = Inferior Parietal; A13 = ADAS13; AQ4 = ADAS-Q4; FAQ = Functional Activities Questionnaire; RT_{im} = RAVLT Immediate; AUC = Area under the curve; ACC = Accuracy; SEN = Sensitivity; SPE = Specificity; Frequency = Minimum and maximum range of the number of times that the combination of proposed features was evaluated by the cross-validation procedure.

Data	AUC	ACC (%)	SEN (%)	SPE (%)	Frequency	Optimal feature subsets
(MRI + NM) $_{bl}^{2G}$	0.855(0.850 0.861)	75.9(75.3 76.5)	85.6(84.8 86.4)	67.1(66.2 68.0)	405-1480	$P_V^L, MT_T^R, A13, FAQ, RT_{im}$ $P_V^L, IP_T^R, A13, AQ4, FAQ, RT_{im}$ $H_V^L, MT_T^R, A13, FAQ, RT_{im}$
(MRI + NM) $_{bl}^{SA}$	0.860(0.856 0.864)	77.7(77.3 78.2)	79.2(78.6 79.9)	75.9(75.2 76.5)	266-461	$H_V^L, P_V^L, MT_T^R, A13, FAQ, RT_{im}$ $H_V^L, P_V^L, IP_T^L, A13, AQ4, FAQ, RT_{im}$ $P_V^L, IP_T^R, A13, FAQ, RT_{im}$
(MRI + NM) $_{m12}^{2G}$	0.894(0.890 0.897)	80.1(79.6 80.5)	82.5(81.9 83.1)	77.9(77.2 78.5)	709-1080	$P_V^L, IP_T^R, A13, FAQ, RT_{im}$ $P_V^L, MT_T^R, A13, FAQ, RT_{im}$ $H_V^B, P_V^L, IP_T^R, A13, AQ4, FAQ, RT_{im}$
(MRI + NM) $_{m12}^{SA}$	0.891(0.886 0.895)	79.3(78.7 79.9)	78.5(77.5 79.4)	80.2(79.3 81.1)	479-635	$P_V^L, IP_T^R, A13, FAQ, RT_{im}$ $P_V^L, MT_T^R, A13, AQ4, FAQ, RT_{im}$ $P_V^L, IP_T^R, A13, AQ4, FAQ, RT_{im}$
(MRI + NM) $_{m24}^{2G}$	0.908(0.905 0.911)	81.6(81.2 82.0)	77.8(77.1 78.5)	84.7(84.1 85.3)	646-1386	$P_V^L, IP_T^L, A13, FAQ, RT_{im}$ $P_V^L, E_T^B, IP_T^R, A13, AQ4, FAQ, RT_{im}$ $P_V^L, MT_T^R, A13, FAQ, RT_{im}$
(MRI + NM) $_{m24}^{SA}$	0.925(0.920 0.930)	83.6(82.9 84.2)	85.2(84.3 86.2)	82.4(81.4 83.3)	525-817	$P_V^L, IP_T^L, A13, FAQ, RT_{im}$ $P_V^L, MT_T^R, A13, FAQ, RT_{im}$ $P_V^L, IP_T^R, A13, FAQ, RT_{im}$
(MRI + NM) $_{m36}^{2G}$	0.917(0.908 0.926)	82.6(81.6 83.6)	74.2(72.1 76.2)	87.8(86.6 89.1)	672-1162	$P_V^L, MT_T^R, A13, AQ4, FAQ, RT_{im}$ $H_V^B, P_V^L, IP_T^R, A13, AQ4, FAQ, RT_{im}$ $P_V^L, MT_T^R, A13, FAQ, RT_{im}$
(MRI + NM) $_{m36}^{SA}$	0.944(0.949 0.949)	85.3(84.7 86.0)	86.3(85.2 87.4)	85.6(84.8 86.5)	692-1639	$P_V^L, MT_T^R, A13, FAQ, RT_{im}$ $H_V^L, P_V^L, MT_T^R, A13, FAQ, RT_{im}$ $P_V^L, MT_T^R, A13, AQ4, FAQ, RT_{im}$

prediction scores of the extended Cox models improved with respect to the two-group comparison approach.

In recent years, many studies have been published in the field of MCI-to-AD prediction. Reviews of these approaches can be found in Falahati et al. (2014), Rathore et al. (2017). Table 6 shows that our best prediction models, with either single- or multisource data, performed favorably compared with recently published models. For better compatibility with the present study, we limit this comparison to studies that used data from the ADNI dataset to predict MCI-to-AD progression

from 24 to 36 months of follow-up. In our study, we presented the baseline scores and the classification results from longitudinal trajectories belonging to the 36-month visit. This work stands out for obtaining baseline scores comparable with other cross-sectional approaches but with a smaller and identifiable feature vector. Unlike the other approaches, there was also an improvement in the prediction accuracy of the conversion to AD when considering the longitudinal trajectory of the patients. As more visits were made available from the patients, the predictive models exhibited not only improvement in the

Table 6

Comparison of approaches for predicting MCI-to-AD conversion. Our classification results are given at baseline as well as at the 36-month visit. Our predictive models were built using survival models with age, sex and years of education as covariates. AUC = Area under the curve; ACC = Accuracy; SEN = Sensitivity; SPE = Specificity.

Approach	Follow-up	Subject number sMCI/pMCI	Data	AUC	ACC (%)	SEN (%)	SPE (%)	Feature number
Gavidia-Bovadilla et al. (2017)	36		MRI		70.3	64.3	75.2	
Beheshti et al. (2017)	36	65/71	MRI	0.751	75.0	76.9	73.2	
Eskildsen et al. (2015)	36	227/161	MRI	0.763	71.9	69.6	73.6	5
Moradi et al. (2015)	36	164/100	MRI	0.766	74.7	88.8	51.6	309
Minhas et al. (2017)	24	65/54	MRI	0.811	77.5	53.9	89.2	38
Spasov et al. (2019)	36	228/181	MRI	0.79	72	63	81	
Present study	36	165/156	MRI	0.769	70.2	68.8	71.5	2-5
				0.885	78.3	79.0	78.1	
Korolev et al. (2016)	36	120/139	MRI,NM	0.87	79.9	83.4	76.4	10
Moradi et al. (2015)	36	164/100	MRI,NM	0.902	82.0	87.0	74.0	309
Gavidia-Bovadilla et al. (2017)	36		MRI,NM		76.7	70.8	81.6	
Minhas et al. (2017)	24	65/54	MRI,NM	0.889	84.3	70.4	92.3	45
Spasov et al. (2019)	36	228/181	MRI	0.925	86	87.5	84	
Present study	36	165/156	MRI,NM	0.860	77.7	79.2	75.9	5-7
				0.944	85.3	86.3	85.6	

Table 7

Scores for predicting MCI-to-AD conversion using ADAS13. For each visit (Baseline, bl, Month 12, m12, Month 24, m24, Month 36, m36), a predictive model was built with the extended Cox approach. Numbers within parentheses are the 95% confidence interval. AUC = Area under the curve; ACC = Accuracy; SEN = Sensitivity; SPE = Specificity.

Visit	AUC	ACC (%)	SEN (%)	SPE (%)
bl	0.773(0.766 0.780)	70.3(69.7 71.0)	69.5(68.5 70.5)	71.2(70.2 72.3)
m12	0.836(0.830 0.842)	73.6(72.9 74.3)	69.5(68.4 70.5)	77.3(76.3 78.2)
m24	0.848(0.841 0.855)	75.8(75.1 76.6)	72.0(70.7 73.2)	79.0(78.0 80.0)
m36	0.899(0.886 0.912)	76.7(75.3 78.2)	74.3(71.6 76.9)	78.2(76.1 80.2)

prediction but also a better balance between sensitivity and specificity. Finally, our study employed a greater number of subjects and visits of the ADNI database in the construction of the predictive models compared to other publications.

4.2.1. Correlations between the proposed predictive models and ADAS-Cog

The ADAS-Cog is the most widely used general cognitive measure in clinical trials of AD (Skinner et al., 2012). The ADAS-Cog assesses multiple cognitive domains including memory, language, praxis, and orientation (Skinner et al., 2012). The longitudinal ADAS-Cog score has been reported to be the strongest predictor of time from MCI to AD (Li et al., 2017). In addition, using the univariate analysis of the features, ADAS13 was the most discriminating measure between sMCI and pMCI patients. Table 7 shows the scores for predicting MCI-to-AD conversion using ADAS13 exclusively. These scores were obtained using the proposed survival analysis with ADAS13 and considering age, sex and education as covariables. Comparing these results with those presented in Tables 4 and 5, the scores using ADAS13 were similar to those obtained with the predictive models constructed with only MRI markers. However, the balance between sensitivity and specificity is better in the predictive models using MRI markers exclusively.

Next, we tried to verify whether there was a correlation between the ADAS13 scores and the predictions of the proposed models. Figs. 4 and 5 show the relationships between the ADAS13 scores of the subjects at the different visits with the predictions of the single- or multiple-sources models using the proposed survival analysis. It was observed that the linear correlation between ADAS13 and the results of the predictive models increased over time, more so in the multisource models. It is noteworthy how the single-source models, i.e. constructed exclusively with MRI data, correlated reasonably well with ADAS13. This was more obvious in the multisource models, as this measure was used in the construction of these models. It was also remarkable how the combination of ADAS13 with other markers, either MRI-based markers or NMs, improved the prediction results of the conversion of MCI to AD in relation to the ADAS13 scores.

5. Discussion

Many studies have focused on the ability to predict the conversion of MCI to AD with a single marker. For example, using two-group comparison approaches, greater rates of atrophy, as a biomarker, have been used in the hippocampus (Jack et al., 2004; Apostolova et al., 2010; Miller et al., 2013), entorhinal cortex (Jack et al., 2004; Miller et al., 2013) and the medial temporal lobe (Blasko et al., 2008). However, the complexity and heterogeneity of AD does not appear to be explained by the analysis of a single biomarker (Cuingnet et al., 2011; Falahati et al., 2014; Rathore et al., 2017). Advances in neuroimaging have provided the opportunity to study many variables simultaneously and observe inherent patterns in the data over time. Longitudinal studies of MCI patients basically have two objectives: (i) Maximize the accuracy of the prediction of conversion to AD using a combination of markers and (ii) Identify a small group of interpretable markers that can aid in the understanding of the evolution of AD. We developed

predictive models of MCI-to-AD progression that combine a very small subset of MRI-based markers with standard cognitive measures and included demographic and clinical information as covariates. The predictive models were built using longitudinal data. An extensive set of cortical and subcortical features from MRI data and neuropsychological tests were longitudinally extracted. Feature subsets of different dimensions were preselected using the mRMR algorithm, and a resampling method searched through each dimension for the feature subsets that appeared most frequently. Subsequently, the proposed feature subsets were evaluated in terms of the cross-validated classification accuracy. The optimal feature subsets to be used in the final models were determined according to the number of times that the combinations of proposed features were evaluated by the CV procedure, higher AUC values and the best balances between sensitivities and specificities (see Fig. 2, Tables 4, 5). Predictive models with more frequent appearances (i.e. number of times that the combination of proposed features was evaluated by the CV procedure), higher AUC values and balanced sensitivity and specificity were selected.

The main purpose of this paper was to compare two approaches to build longitudinal predictive models: a) Comparison between two clinical groups, i.e. sMCI vs. pMCI and b) Use of models based on survival analysis. Both approaches share LME modeling of the longitudinal trajectories of the markers. While the two-group comparison approach considered the differences between converter and nonconverter subjects, the survival analysis only explored the information of the subjects up until the conversion or censoring times. The analysis of the MMSE of the studied MCI population during the 36 months clearly shows that the problem is not dichotomous (see Table 1). The subjects labeled as sMCI maintained a MMSE value that practically remained constant throughout the study. However, the MMSE scores of the pMCI patients fell over time. The explanation for this is that the conversion times from MCI to AD were variable. There were subjects whose diseases convert at the beginning of the study and others at the middle or end. The conclusion was obvious: the pMCI group was not homogeneous with respect to conversion time; therefore, the main hypothesis in the two-group comparison approach was violated. There was no homogeneity in the pMCI group, as a result of different conversion times, or in the sMCI group, as a result of uncertainties due to a fixed cut-off period in the follow-up. To overcome these difficulties, statistical methods of survival analysis were used and the AD-prediction markers were reevaluated. The Cox proportional hazards model has previously been employed to predict the conversion from MCI to AD (Devanand et al., 2007; Desikan et al., 2009; Vemuri et al., 2011). However, this approach only analyzes baseline measurements, which restricts the associations among markers and the evolution of AD. An alternative approach was designed by combining an extended Cox model with LME modeling that examined the relationship between time-dependent markers and the timing of the conversion to AD or the censoring times of the samples.

With both approaches, a very small number of features were selected. These markers are easily interpretable, generating robust, verifiable and reliable predictive models. Moreover, age, sex and years of education of the patients were easily incorporated into the predictive models as covariates. Two-group comparison approaches used LDA-based classifiers, which were trained with marginal longitudinal trajectory residues. In contrast, survival analysis used risk ratios obtained from the extended Cox model, and these were transformed in probabilistic terms for the conversion from MCI to AD using the logistic regression model. CV strategies were used to avoid contamination of the training data with the testing data. The procedure consisted of two nested CV loops: an inner loop, designed to select the optimal feature set for the proposed models, and an outer loop, designed to obtain an unbiased estimate of model performance. The use of k-fold CV was recommended as a means of standardizing to evaluate the predictive models of MCI-to-AD progression (Wyman et al., 2013). Finally, a key feature of the dynamic prediction frameworks is that the predictive models can be updated as additional longitudinal measurements

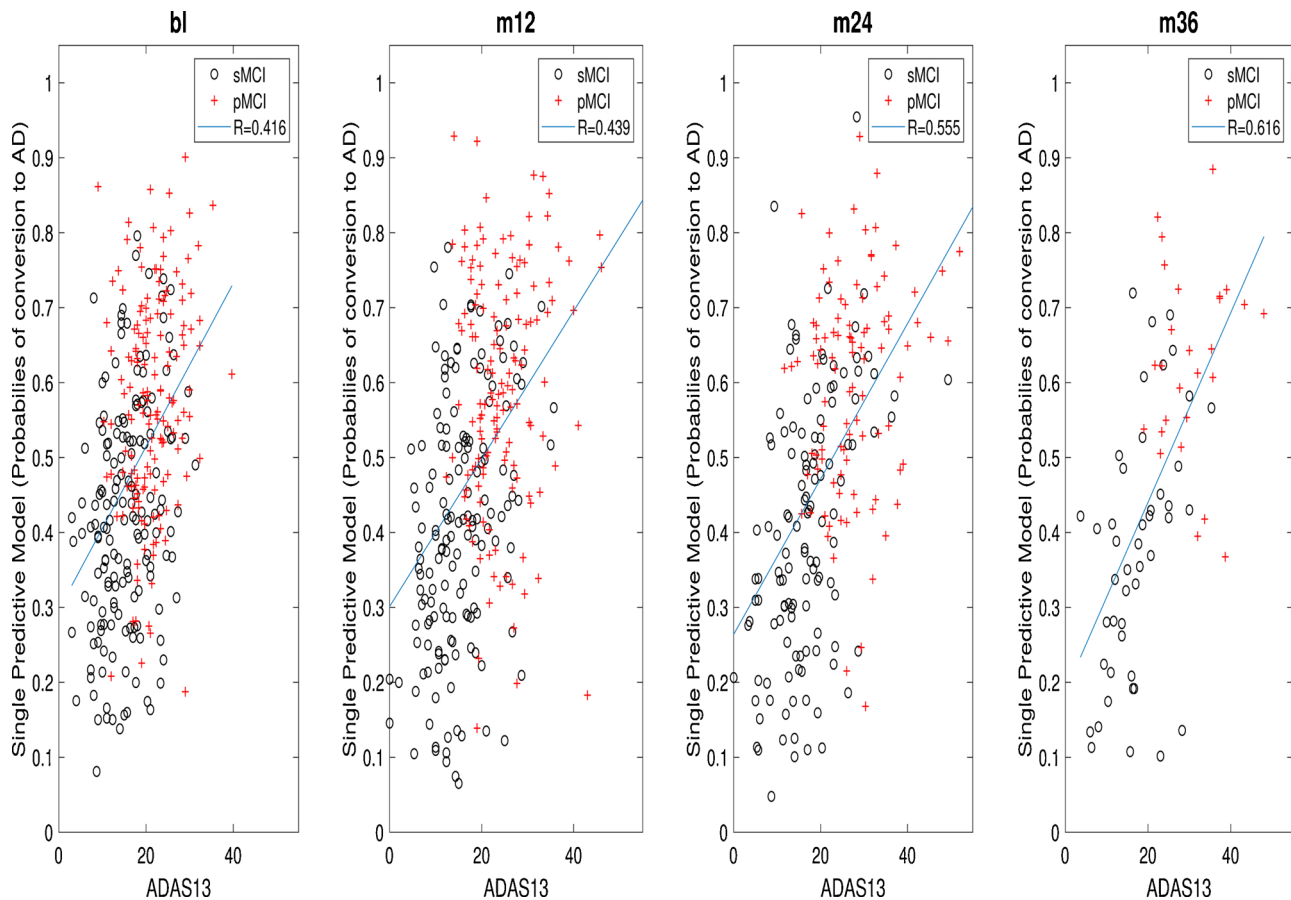


Fig. 4. Linear correlation between ADAS13 and single-source predictive models over time. The predictive models were built according to the proposed survival analysis and visit (Baseline, bl, Month 12, m12, Month 24, m24, Month 36, m36). The outputs of the predictive models are shown in terms of the probabilities of conversion to AD. The linear correlation coefficients (R) of each visit are presented in the legends. sMCI = Stable mild cognitive impairment; pMCI = Progressive mild cognitive impairment.

become available for the target subjects, providing instantaneous risk assessment.

Performing an analysis of 321 participants with 1330 visits provided a large number of subjects for training and testing the two proposed approaches and enabled us to establish a few main conclusions. Both strategies shared the following aspects: 1. Better scores were obtained when combining the MRI measurements with the neuropsychological tests. 2. The dimensions of the feature vectors were very low and the selected markers were usually always the same. Between two and five features in the predictive models used only MRI data, and between five to seven used multisource data. Among the selected MRI-based markers were the hippocampal and pallidum volumes and the cortical thicknesses of temporal and parietal regions. Regarding NM, ADAS13, FAQ and RAVLT Immediate were repeated. No differences were observed either in the dimensions of the feature vectors, nor the markers selected due to the approach used as well as over time. 3. As more visits were made available, the performance of the predictive models improved. 4. The selected features were similar over time for building predictive models at each visit. 5. Classifier scores were similar during the first year between the two approaches.

According to the approach, the predictive models differed in the following ways. 1. Survival models showed a better balance between sensitivity and specificity over time. 2. The prediction results from the survival models were better from the second year on. These discrepancies could be explained by the lack of homogeneity accelerated over time in the pMCI group. During the follow-up period, an increasing number of pMCI subjects progressed towards AD. Unlike the two-group comparison approach, survival analysis takes this effect into account,

given that the conversion time of each subject was considered in the estimations of these predictive models. Moreover, based on the classifier operations, in the two-group comparison approach only the residues between the trajectory and the available samples are weighted (see Eq. (6)), while in the survival analysis a logistic regression is applied depending on the patient's visit, i.e. it takes into account how the risk factors depend on the time of the visit (see Eq. (7)). This stratification of survival analysis models together with the consideration of conversion and censorship times are the two causes that produce an improvement in the predictions of diagnoses over time.

In both approaches, the number of feature vector dimensions was very low in comparison to other studies. The features most frequently selected in the NMs included ADAS13, FAQ and RAVLT Immediate. In the MRI data, the most frequently selected features included hippocampal and pallidum volumes and CT measures for several temporoparietal brain regions, with a preference for the left hemisphere. The selection of the hippocampus, middle temporal lobe, and inferior parietal cortex as predictors of MCI-to-dementia progression is consistent with the known pattern of grey matter atrophy associated with incipient AD (Thompson et al., 2003), and there is also evidence that AD-related atrophy occurs at a faster rate in the left hemisphere (Thompson et al., 2003). Similar findings have been reported in other studies (Belleville et al., 2014; Korolev et al., 2016; Sørensen et al., 2017), although most publications have used cross-sectional approaches, whereas this work is a longitudinal study.

The proposed predictive models were also validated by analyzing their correlations with the ADAS13 scores, which is currently considered a strong predictor of the conversion from MCI to AD.

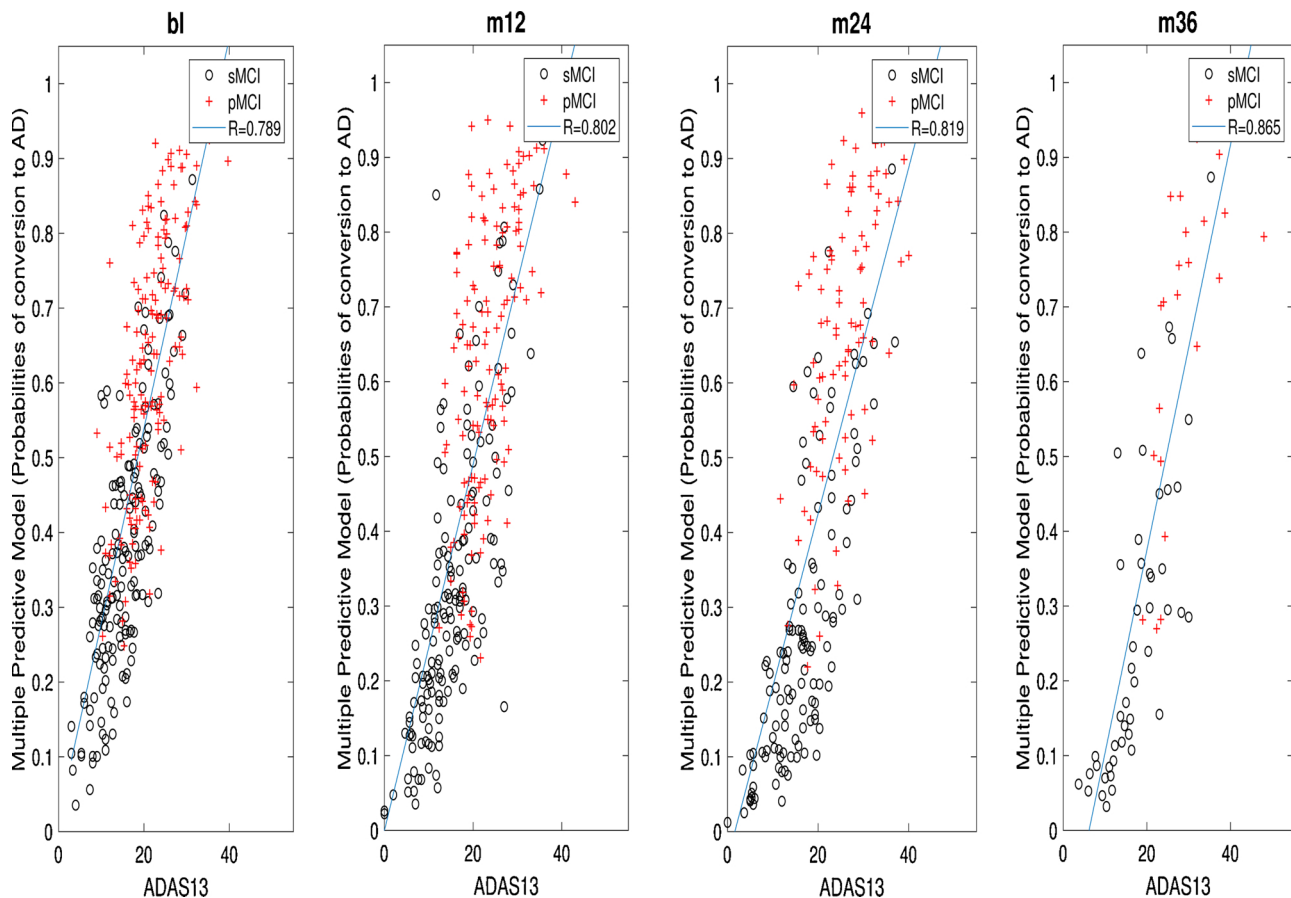


Fig. 5. Linear correlation between ADAS13 and multiple-source predictive models using the proposed survival analysis over time. The predictive models were built according to the proposed survival analysis and visit (Baseline, bl, Month 12, m12, Month 24, m24, Month 36, m36). The outputs of the predictive models are shown in terms of the probabilities of conversion to AD. The linear correlation coefficients (R) of each visit are presented in the legends. sMCI = Stable mild cognitive impairment; pMCI = Progressive mild cognitive impairment.

Verification was approached in two different aspects: a) A comparison between the ADAS13 scores and the predictive capacities of the proposed models and b) An analysis of the linear correlation between the ADAS13 scores and the probabilities of conversion to AD of the subjects' diseases at the different visits as estimated by the models. From the first proof, analyzing the tables of the prediction scores for AD conversion (see Tables 4, 5, 7), the classification results between sMCI and pMCI were similar between the ADAS13-only models and the models constructed with only MRI-based markers throughout the visits. Clearly, the multisource models yielded the best prediction results. Undoubtedly, the strongest validation was observed with the correlation analysis between the prediction probabilities of the models and the ADAS13 scores. Particularly relevant was the correlation between ADAS13 and the predictive models constructed exclusively with MRI-based markers, verifying the correlation between this cognitive measure and the selected structural T1-MRI measures, in addition to validating the proposed approaches. The highest correlations were found between ADAS13 and the multisource models where ADAS13 was usually present in their construction. Finally, it has been observed that the combination of ADAS13 with other markers proposed here, whether based on MRI data or other NMs, improved the prediction results of the conversion of MCI to AD.

5.1. Limitations

Despite the promising results, there are several limitations of our study. The data used here correspond to subjects who met the inclusion and exclusion criteria established by ADNI. The choice of cohort may

have affected the results of the predictive models. However, most of the studies that have investigated MCI-to-AD progression have used this database. In addition, we followed the recommendation of using standard patient lists, which allows a better comparison among the proposed approaches (Wyman et al., 2013). On the other hand, although 321 subjects with 1330 visits were processed, it would be desirable to increase the number of samples. To date, ADNI continues to recruit more subjects and visits. Therefore, future works should apply the proposed approaches for developing new predictive models using other public cohorts, like OASIS-3 (LaMontagne et al., 2018), and increase the number of samples to study.

The modeling of markers was implemented using the LME approach and the longitudinal trajectories were considered as linear. A linear function to model the dynamic changes of structural MRI-based markers is a well-accepted practice (Guerrero et al., 2016). Although other authors suggest the use of higher order polynomial models (Donohue et al., 2014; Schmidt-Richberg et al., 2016). On the other hand, experimental evidence suggests that linear models are not enough to describe cognitive decline measurements in AD progression (see supplementary materials). Neuro-psychological marker models have been hypothesized to follow sigmoid or quadratic shape (Jack and Holtzman, 2013; Buckley et al., 2018), with acceleration or deceleration of annual atrophy in the initial or final phases of the evolution of the disease. Therefore, more complex modeling of the evolution of the longitudinal trajectories of the markers (especially of the NMs) could improve the forecasts of the predictive models. A future line of work would be to improve the modeling of marker trajectories.

One of the main limitations in any AD study involving in vivo data is

the uncertainty of the diagnosis. This is the reason why the diagnosis will always be probable AD until an autopsy can confirm the diagnosis (Eskildsen et al., 2013). The clinical diagnosis of probable AD has an accuracy of 70–90% relative to pathological diagnosis (Beach et al., 2012). The predictive models were developed with the clinical diagnosis as the ground truth. This limitation could be overcome by improving the accuracy of the clinical assessments, which is beyond the scope of this study.

The criteria for the classification of MCI patients between stable and converters used in these studies are not unique (Moradi et al., 2015). We observed some discrepancy in the labeling of MCI patients in the list of Eskildsen et al. (2013), guided by the Wolz criteria (Wolz et al., 2011). In this study, the diagnosis from the ADNI was used at each visit. The conversion time and the corresponding labeling were based on the first visit with a diagnosis of dementia, staying with this prediction until the end of the follow-up time, i.e., up to 36 months.

Finally, because most studies established a 3-year cut-off period, we used this length of follow-up, which facilitates the comparisons among the proposed approaches (see Table 6). In addition, several studies claim that predictive models with follow-up beyond 3 years become less reliable in the prediction of conversion to AD (Pettigrew et al., 2016).

5.2. Future work

A new line of work would involve the optimization of the type of longitudinal classifier employed. The classifiers used here work well with low-dimensional feature vectors. Finding combinations of NMs and MRI markers that improve classification results may require searching for more powerful longitudinal classifiers.

Nothing prevents these approaches from being applied to other events in the progression of AD and even to other types of diseases with prodromal phases. In this sense, we are exploring the construction of predictive models in the preclinical phase of AD, i.e. a longitudinal study of control subjects towards MCI or AD. We are also developing predictive models for the case of Parkinson's disease (PD) between control subjects and PD subjects without medication, i.e. PD novo.

6. Conclusions

We compared two approaches for developing predictive models of MCI-to-AD progression that combine a very small subset of MRI-based markers with standard cognitive measures from a longitudinal study with a 3-year follow-up. The approaches were able to predict patients with converting diseases with an accuracy of 78%, a sensitivity of 79% and a specificity of 76% using only baseline patient data. In addition, the availability of new patient visits improves the performance of the proposed predictive models. For example, in the 36-month visit, the results were an accuracy of 85%, a sensitivity of 86% and a specificity of 85%.

Some of the conclusions of this study are as follows: 1) The proposed predictive models were built with only 2–7 highly stable features under cross-validation; these markers were consistent with Braak stages, were previously reported in MCI-to-AD conversion and are easy to transfer to new cohorts and clinical practice; 2) Multisource data for predicting MCI-to-AD conversion delivered a more accurate estimate than single-source data; 3) The survival-based predictive models showed a better balance between sensitivity and specificity, in addition to improving the scores starting from the 24-month visit, with respect to the models based on the two-group comparison approach; and 4) The proposed methods used relatively common clinical tests, such as MRI and NMs, as opposed to methods that rely on more expensive or invasive tests, such as PET-, CSF- and genotype-based markers, and can offer potential for monitoring treatment outcomes in future drug trials.

The scripts and list of the ADNI identifiers of the subjects used in this study are freely released for other users at <https://www.nitrc.org/projects/twogrsurvana/>. The MRI-based measurements and the

neuropsychological tests from each patient and each visit are also available.

Declarations of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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 through generous contributions from the following: Alzheimers Association; Alzheimers 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/how_to_apply/ADNI_Acknowledgement_List.pdf

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.jneumeth.2020.108698>.

References

- Apostolova, L.G., Mosconi, L., Thompson, P.M., Green, A.E., Hwang, K.S., Ramirez, A., Mistur, R., Tsui, W.H., de Leon, M.J., 2010. Subregional hippocampal atrophy predicts Alzheimer's dementia in the cognitively normal. *Neurobiol. Aging* 31, 1077–1088.
- Beach, T.G., Monsell, S.E., Phillips, L.E., Kukull, W., 2012. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *J. Neuropathol. Exp. Neurol.* 71, 266–273.
- Beheshti, I., Demirel, H., Matsuda, H.A.D.N.I., et al., 2017. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resonance imaging using feature ranking and a genetic algorithm. *Comput. Biol. Med.* 83, 109–119.
- Belleville, S., Fouquet, C., Duchesne, S., Collins, D.L., Hudon, C., et al., 2014. Detecting early preclinical Alzheimer's disease via cognition, neuropsychiatry, and neuroimaging: qualitative review and recommendations for testing. *J. Alzheimer's Dis.* 42, S375–S382.
- Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., Sabuncu, M.R.A.D.N.I., et al., 2013a. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. *Neuroimage* 81, 358–370.
- Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R.A.D.N.I., et al., 2013b. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage* 66, 249–260.
- Blasko, I., Jellinger, K., Kemmler, G., Krampla, W., Jungwirth, S., Wichart, I., Tragl, K.H., Fischer, P., 2008. Conversion from cognitive health to mild cognitive impairment and Alzheimer's disease: prediction by plasma amyloid beta 42. medial temporal lobe atrophy and homocysteine. *Neurobiol. Aging* 29, 1–11.
- Buckley, R.F., Mormino, E.C., Amariglio, R.E., Properzi, M.J., Rabin, J.S., Lim, Y.Y., Papp, K.V., Jacobs, H.I., Burnham, S., Hanseeuw, B.J., et al., 2018. Sex, amyloid, and apoe ϵ 4 and risk of cognitive decline in preclinical Alzheimer's disease: findings from three well-characterized cohorts. *Alzheimer's Dementia* 14, 1193–1203.
- Chapman, R.M., Mapstone, M., McCrary, J.W., Gardner, M.N., Porsteinsson, A., Sandoval, T.C., Guillily, M.D., DeGrush, E., Reilly, L.A., 2011. Predicting conversion from mild cognitive impairment to Alzheimer's disease using neuropsychological tests and multivariate methods. *J. Clin. Exp. Neuropsychol.* 33, 187–199.
- Chetelat, G., Landeau, B., Eustache, F., Mzenge, F., Viader, F., de la Sayette, V., Desgranges, B., Baron, J.C., 2005. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: A longitudinal MRI

- study. *NeuroImage* 27, 934–946.
- Cox, D.R., 1975. Partial likelihood. *Biometrika* 62, 269–276.
- Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., et al., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage* 56, 766–781.
- Da, X., Toledo, J.B., Zee, J., Wolk, D.A., Xie, S.X., Ou, Y., Shacklett, A., Parmpi, P., Shaw, L., Trojanowski, J.Q., et al., 2014. Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. *NeuroImage: Clin.* 4, 164–173.
- Desikan, R.S., Cabral, H.J., Fischl, B., Guttman, C.R., Blacker, D., Hyman, B.T., Albert, M.S., Killiany, R.J., 2009. Temporoparietal MR imaging measures of atrophy in subjects with mild cognitive impairment that predict subsequent diagnosis of Alzheimer disease. *Am. J. Neuroradiol.* 30, 532–538.
- Devanand, D., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G., Honig, L., Mayeux, R., et al., 2007. Hippocampal and entorhinal atrophy in mild cognitive impairment prediction of Alzheimer disease. *Neurology* 68, 828–836.
- 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., 2014. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia* 10, S400–S410.
- Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L.A.D.N.I. et al., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage* 65, 511–521.
- Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Collins, D.L., 2015. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. *Neurobiol. Aging* 36, S23–S31.
- Falahati, F., Westman, E., Simmons, A., 2014. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. *J. Alzheimer's Dis.* 41, 685–708.
- Ferreira, F.L., Cardoso, S., Silva, D., Guerreiro, M., de Mendonça, A., Madeira, S.C., 2017. Improving prognostic prediction from mild cognitive impairment to Alzheimer's disease using genetic algorithms. In: *International Conference on Practical Applications of Computational Biology & Bioinformatics* 180–188.
- Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl. Acad. Sci.* 97, 11050–11055.
- Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355.
- Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatric Res.* 12, 189–198.
- Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77.
- Gaviddia-Bovadilla, G., Kanaan-Izquierdo, S., Mataró-Serrat, M., Perera-Lluna, A.A.D.N.I. et al., 2017. Early prediction of Alzheimer's disease using null longitudinal model-based classifiers. *PLOS ONE* 12, e0168011.
- Guerrero, R., Schmidt-Richberg, A., Ledig, C., Tong, T., Wolz, R., Rueckert, D., Adni, A.D.N.I. et al., 2016. Instantiated mixed effects modeling of Alzheimer's disease markers. *NeuroImage* 142, 113–125.
- Holtzman, D.M., Morris, J.C., Goate, A.M., 2011. Alzheimers disease: the challenge of the second century. *Sci. Transl. Med.* 3 775r1–775r1.
- Iglesias, J.E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., Reuter, M.A.D.N.I. et al., 2016. Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. *NeuroImage* 141, 542–555.
- Jack Jr., C.R., Holtzman, D.M., 2013. Biomarker modeling of Alzheimers disease. *Neuron* 80, 1347–1358.
- Jack, C., Shiung, M., Gunter, J., O'Brien, P., Weigand, S., Knopman, D., Boeve, B., Ivnik, R., Smith, G., Cha, R., et al., 2004. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. *Neurology* 62, 591–600.
- Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, L., Ward, J.C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magnet. Reson. Imaging* 27, 685–691.
- Jack Jr., C.R., 2012. Alzheimer disease: new concepts on its neurobiology and the clinical role imaging will play. *Radiology* 263, 344–361.
- Jiang, J., Sachdev, P., Lipnicki, D.M., Zhang, H., Liu, T., Zhu, W., Suo, C., Zhuang, L., Crawford, J., Reppermund, S., et al., 2014. A longitudinal study of brain atrophy over two years in community-dwelling older individuals. *NeuroImage* 86, 203–211.
- Kleinbaum, D.G., Klein, M., 2010. *Survival analysis*. Volume 3. Springer.
- Korolev, I.O., Symonds, L.L., Bozoki, A.C.A.D.N.I. et al., 2016. Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLOS ONE* 11, e0138866.
- Kuhn, M., Johnson, K., 2013. *Applied predictive modeling*. Volume 26. Springer.
- LaMontagne, P.J., Keefe, S., Lauren, W., Xiong, C., Grant, E.A., Moulder, K.L., Morris, J.C., Benzinger, T.L., Marcus, D.S., 2018. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimers disease. *Alzheimer's Dementia* 14, P1097.
- Li, K., Chan, W., Doody, R.S., Quinn, J., Luo, S., 2017. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *J. Alzheimer's Disease* 58, 361–371.
- Li, K., O'Brien, R., Lutz, M., Luo, S., Initiative, A.D.N.I., et al., 2018. A prognostic model of Alzheimer's disease relying on multiple longitudinal measures and time-to-event data. *Alzheimer's Dementia* 14, 644–651.
- Liu, Y., Julkunen, V., Paajanen, T., Westman, E., Wahlund, L.O., Aitken, A., Sobow, T., Mecocci, P., Tsolaki, M., Vellas, B., et al., 2012. Education increases reserve against Alzheimer's disease evidence from structural MRI analysis. *Neuroradiology* 54, 929–938.
- Manly, J.J., Tang, M.X., Schupf, N., Stern, Y., Vonsattel, J.P.G., Mayeux, R., 2008. Frequency and course of mild cognitive impairment in a multiethnic community. *Ann. Neurol.* 63, 494–506.
- Miller, M.I., Younes, L., Ratnanather, J.T., Brown, T., Trinh, H., Postell, E., Lee, D.S., Wang, M.C., Mori, S., O'Brien, R., et al., 2013. The diffeomorphometry of temporal lobe structures in preclinical Alzheimer's disease. *NeuroImage: Clin.* 3, 352–360.
- Minhas, S., Khanum, A., Riaz, F., Khan, S., Alvi, A., 2017. Predicting progression from mild cognitive impairment to Alzheimer's disease using autoregressive modelling of longitudinal and multimodal biomarkers. *IEEE J. Biomed. Health Inform.*
- Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J.A.D.N.I., et al., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage* 104, 398–412.
- Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intel.* 27, 1226–1238.
- Petersen, R.C., Roberts, R.O., Knopman, D.S., Boeve, B.F., Geda, Y.E., Ivnik, R.J., Smith, G.E., Jack, C.R., 2009. Mild cognitive impairment: ten years later. *Archiv. Neurol.* 66, 1447–1455.
- Pettigrew, C., Soldan, A., Zhu, Y., Wang, M.C., Moghekar, A., Brown, T., Miller, M., Albert, M., Team, B.R., et al., 2016. Cortical thickness in relation to clinical symptom onset in preclinical AD. *NeuroImage: Clin.* 12, 116–122.
- Platero, C., Lin, L., Tobar, M.C., 2019. Longitudinal neuroimaging hippocampal markers for diagnosing Alzheimers disease. *Neuroinformatics* 17, 43–61.
- Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P., 2013. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's Dementia* 9, 63–75.
- Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155, 530–548.
- Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: a robust approach. *NeuroImage* 53, 1181–1196.
- Reuter, M., Schmahs, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage* 61, 1402–1418.
- Sørensen, L., Igel, C., Pai, A., Balas, I., Anker, C., Lillholm, M., Nielsen, M.A.D.N.I., et al., 2017. Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry. *NeuroImage: Clinical* 13, 470–482.
- Sabuncu, M.R., Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B.A.D.N.I., et al., 2014. Event time analysis of longitudinal neuroimaging data. *NeuroImage* 97, 9–18.
- Saykin, A.J., Shen, L., Foroud, T.M., Potkin, S.G., Swaminathan, S., Kim, S., Risacher, S.L., Nho, K., Huentelman, M.J., Craig, D.W., et al., 2010. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimer's Dementia* 6, 265–273.
- Schmidt-Richberg, A., Ledig, C., Guerrero, R., Molina-Abril, H., Frangi, A., Rueckert, D., Initiative, A.D.N.I., et al., 2016. Learning biomarker models for progression estimation of alzheimers disease. *PLOS ONE* 11, e0153040.
- Skinner, J., Carvalho, J.O., Potter, G.G., Thames, A., Zelinski, E., Crane, P.K., Gibbons, L.E., Initiative, A.D.N.I., et al., 2012. The alzheimers disease assessment scale-cognitive-plus (adas-cog-plus): an expansion of the adas-cog to improve responsiveness in mci. *Brain Imaging Behav.* 6, 489–501.
- Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* 17, 87–97.
- Spasov, S., Passamonti, L., Duggento, A., Liò, P., Toschi, N., Initiative, A.D.N.I., et al., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage* 189, 276–287.
- Sperling, R.A., Rentz, D.M., Johnson, K.A., Karlawish, J., Donohue, M., Salmon, D.P., Aisen, P., 2014. The a4 study: stopping ad before symptoms begin? *Sci. Transl. Med.* 6 228fs13-228fs13.
- the ADNI team: ADNIMERGE: Alzheimer's Disease Neuroimaging Initiative. (2018) R package version 0.0.1.
- The ADNI General Procedure Manual. (https://adni.loni.usc.edu/wp-content/uploads/2010/09/ADNI_GeneralProceduresManual.pdf) Accessed: 2018-05-05.
- Thompson, P.M., Hayashi, K.M., De Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman, D., Hong, M.S., Dittmer, S.S., Doddrell, D.M., et al., 2003. Dynamics of gray matter loss in Alzheimer's disease. *J. Neurosci.* 23, 994–1005.
- Vemuri, P., Weigand, S.D., Knopman, D.S., Kantarci, K., Boeve, B.F., Petersen, R.C., Jack, C.R., 2011. Time-to-event voxel-based techniques to assess regional atrophy associated with MCI risk of progression to AD. *NeuroImage* 54, 985–991.
- Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., 2010. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. *NeuroImage* 52, 109–118.
- Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J.A.D.N.I., et al., 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PLOS ONE* 6, e25446.
- Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., et al., 2013. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's Dementia* 9, 332–337.