RESEARCH ARTICLE

Describing complex disease progression using joint latent class models for multivariate longitudinal markers and clinical endpoints

Cécile Proust-Lima[1,2] | Tiphaine Saulnier[1] | Viviane Philipps[1] | Anne Pavy-Le Traon[3] | Patrice Péran[4] | Olivier Rascol[3] | Wassilios G. Meissner[5,6] | Alexandra Foubert-Samier[1,2,5]

[1]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, Bordeaux, France

[2]Inserm, CIC1401-EC, Bordeaux, France

[3]MSA Reference Center and CIC-1436, Department of Clinical Pharmacology and Neurosciences, NeuroToul COEN Center, University of Toulouse 3, CHU of Toulouse, INSERM, Toulouse, France

[4]ToNIC, Toulouse NeuroImaging Center, Univ Toulouse, Inserm, UPS, Toulouse, France

[5]Univ. Bordeaux, CNRS, IMN, UMR5293, Bordeaux, France

[6]Dept. Medicine, University of Otago, Christchurch, and New Zealand Brain Research Institute, Christchurch, New Zealand

Correspondence
Cécile Proust-Lima, Univ. Bordeaux, Inserm, BPH, U1219, F-33000 Bordeaux, France.
Email: cecile.proust-lima@inserm.fr

Summary

Neurodegenerative diseases are characterized by numerous markers of progression and clinical endpoints. For instance, multiple system atrophy (MSA), a rare neurodegenerative synucleinopathy, is characterized by various combinations of progressive autonomic failure and motor dysfunction, and a very poor prognosis. Describing the progression of such complex and multi-dimensional diseases is particularly difficult. One has to simultaneously account for the assessment of multivariate markers over time, the occurrence of clinical endpoints, and a highly suspected heterogeneity between patients. Yet, such description is crucial for understanding the natural history of the disease, staging patients diagnosed with the disease, unravelling subphenotypes, and predicting the prognosis. Through the example of MSA progression, we show how a latent class approach modeling multiple repeated markers and clinical endpoints can help describe complex disease progression and identify subphenotypes for exploring new pathological hypotheses. The proposed joint latent class model includes class-specific multivariate mixed models to handle multivariate repeated biomarkers possibly summarized into latent dimensions and class-and-cause-specific proportional hazard models to handle time-to-event data. Maximum likelihood estimation procedure, validated through simulations is available in the lcmm R package. In the French MSA cohort comprising data of 598 patients during up to 13 years, five subphenotypes of MSA were identified that differ by the sequence and shape of biomarkers degradation, and the associated risk of death. In posterior analyses, the five subphenotypes were used to explore the association between clinical progression and external imaging and fluid biomarkers, while properly accounting for the uncertainty in the subphenotypes membership.

KEYWORDS
class-membership, clustering, heterogeneity, joint modeling, multiple system atrophy, multivariate longitudinal data, neurodegenerative diseases

Abbreviations: BP, blood pressure; MRI, magnetic resonance imaging; MSA, multi-system atrophy; UMSARS, unified MSA rating scale.

1 | INTRODUCTION

Some diseases are characterized by numerous markers of progression. Although not specific to, this is particularly the case in neurodegenerative diseases where pathological brain changes may induce multiple clinical signs on which the progression of a patient is assessed. Alzheimer's disease involves progressive impairment over decades of cerebral regions, multiple cognitive functions, functional dependency, and even depressive symptomatology or anxiety.[1,2] Parkinson's disease is a polymorph disease including progressive motor impairment, cognitive and behavioural disorders, and autonomic failure.[3,4] Multiple system atrophy (MSA), a rare neurodegenerative synucleinopathy with an annual incidence of 3/100,000 individuals,[5] is also characterized by the combinations of multiple dimensions, including autonomic failure, parkinsonism, and cerebellar ataxia.[6]

Describing the progression of such complex and multi-dimensional diseases is particularly difficult. One has to simultaneously account for the assessment of multivariate markers over time, the occurrence of clinical endpoints (eg, death, extreme dependency), and the suspected heterogeneity between patients. Yet, such description is crucial for understanding the natural history of the disease, staging patients diagnosed with the disease, unraveling subphenotypes, identifying novel therapeutic targets, and predicting the prognosis.

When interested in the change over time of markers along with the occurrence of endpoints, the dedicated statistical approach is the joint modeling methodology which simultaneously models the trajectory over time of a marker and the risk of an event when those two are correlated.[7-9] Traditionally based on the so-called "shared random effect" paradigm, joint models usually focus on how longitudinal markers impact the risk of an event by including some predictor of the marker trajectory in the time-to-event model. This is particularly useful to quantify the association between pre-determined characteristics of an endogenous marker and the clinical endpoints.[8,10] However, it might not be the best way to explore in a holistic way a complex disease progression that involves multiple markers along with clinical endpoints. In that perspective, joint latent class models (JLCMs),[9,11-13] another family of joint models, constitute a relevant alternative. JLCMs assume that the population of patients is heterogeneous, and that this heterogeneity explains why patients experiment different marker profiles and different event risks. This paradigm is much more descriptive than traditional joint models but apprehends the suspected heterogeneity present in many contexts and does not assume any particular nature of association between the marker and the event.[9] Over the years, several extensions of JLCM were proposed regarding the nature of the survival data with competing events,[14] recurrent events[15] or event history,[16] or regarding the nature of the longitudinal data by considering several markers measuring the same underlying phenomenon,[14,17] multiple Gaussian markers from high dimensional gene expression with a regularization step[18] and multiple Gaussian markers subject to limits of detection.[19]

In this work, we aimed to leverage the latent class approach to analyze multiple repeated progression markers over time and clinical endpoints, with the final goal of retrieving the subphenotypes of MSA progression and linking them with external biomarker information. Our contribution is fourfold. First, we developed a full methodology for the estimation of joint latent class models for multidimensional longitudinal data and survival time (possibly with competing causes). This model extends beyond the literature by handling multidimensional longitudinal data when Proust-Lima et al[14] considered multivariate longitudinal markers regrouped into a uni-dimensional latent process, and by considering multivariate Gaussian and non-Gaussian continuous markers possibly regrouped into distinct latent dimensions along with multi-cause (left-truncated) time-to-event when Sun et al[18] and Li et al[19] considered Gaussian markers possibly subject to detection limit, and classical survival data. Second, our methodology is made available to the community with a dedicated function in the user-friendly lcmm R software for latent process and latent class models estimation[20] along with documentation. Third, one critical but often ignored aspect of latent class models is the interpretation of the final latent classes, and their association with external information (covariates or outcomes). The uncertainty and miss-classification of any posterior class assignment has to be carefully accounted for to avoid spurious associations.[21-23] Following previous works in non-longitudinal mixture modeling,[24] our method includes two-stage posterior regressions for linking the latent class structure with external information while properly accounting for the uncertainty of the latent class structure. Fourth, we extensively describe a case study in MSA progression to show step-by-step how the JLCM methodology can help describe complex disease progression, identify disease subphenotypes and explore new research hypotheses.

The next section introduces the motivating MSA data. Section 3 details the multivariate JLCM methodology including the model, the maximum likelihood estimation procedure, and the strategy to associate the latent class structure with external information. Section 4 assesses the finite sample performances through simulations. Section 5 is dedicated to the MSA application. Finally, Section 6 concludes.

2 | THE FRENCH MULTI-SYSTEM ATROPHY COHORT

The French MSA cohort has been created in 2007 at the two sites (Toulouse and Bordeaux University hospitals) of the French Reference Center for MSA. After inclusion, patients are usually seen at least once a year by a movement disorder specialist with a clinical assessment that includes demographic information, medical history, neurological examination, diagnostic certainty and subtype, and a standardized clinical evaluation using the Unified MSA Rating Scale (UMSARS).[25] There is also a continuous search for death occurrence with a reporting of the exact date of death along with the cause of death.

The first objective of our study was to describe the clinical progression of MSA, and to uncover potential heterogeneous disease phenotypes using latent classes. We focused on the repeated measures of six UMSARS-derived markers regrouped in three different dimensions:

- Motor and function dimension assessed by a subscore of Activities of Daily Living (UMSARS I) and a subscore of Motor examination (UMSARS II);
- Supine blood pressure (BP) assessed by the systolic BP and the diastolic BP;
- The orthostatic change in BP assessed by the maximum change in systolic BP and in diastolic BP between supine position and standing position during 10 min (UMSARS III).

We leveraged the data of the 598 patients enrolled between 2007 and 2019 who had at least one measure of each marker during the follow-up and no missing information on the main known MSA characteristics: gender, age at inclusion, duration since first symptoms, subtype of MSA (Cerebellar or Parkinson), level of diagnosis certainty (possible or probable). See Table 1 for the sample description. Patients entered the study on average 4.5 years after the first symptoms onset (min–max = 0–24 years). During follow-up, 309 patients died with a median survival since first symptoms of 6.65 years (95% confidence interval [6.18, 7.13] years).

To describe the natural history of MSA, we considered the time since the first symptoms as a proxy of the time since disease onset. Figure 1 describes the individual observed trajectories for the six markers of progression under study of four randomly selected patients according to the time since the first symptoms.

The second objective of this application was to explore to what extent these phenotypes were associated with biological biomarkers which constitute potential therapeutic targets. Indeed, additional assessments were undertaken on small subsamples of the cohort to explore new research hypotheses. A MRI-subsample of 86 patients underwent a T1-weighted volumetric brain Magnetic Resonance Imaging (MRI) with a focus on the volumes of regions particularly involved in the neurodegenerative process such as the cerebellum (gray and white matter), the putamen and the pons. These volumes were measured using the FreeSurfer's image analysis pipeline (version 6). Another subsample underwent further cerebro-spinal fluid (CSF) and serum measurements of total α-synuclein concentration (for 23 patients) and neurofilament light chain (NfL) (for 52 patients). See Table S29 in supplementary materials for a description of the three subsamples.

3 | METHODOLOGY

3.1 | The joint latent class model

3.1.1 | Latent class structure

The latent class methodology relies on a latent class/group structure underlying the variables under study. Let us consider a sample of N subjects ($i = 1, ..., N$), the latent class is defined by a latent discrete variable denoted c_i with value g if subject i belongs to latent class g ($g = 1, ..., G$). Its distribution is described by the latent class membership probability π_{ig} as follows:

$$P(c_i = g) = \pi_{ig} = \frac{e^{\xi_{0g} + X_{Ci}^\top \xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l} + X_{Ci}^\top \xi_{1l}}} \tag{1}$$

with constraints $\xi_{0G} = 0$ and $\boldsymbol{\xi_{1G}} = \boldsymbol{0}$ for reference class G. The probability π_{ig} can be either defined as a regression on time-independent covariates X_{Ci} or considered marginal by removing X_{Ci} from the equation.

TABLE 1 Description at baseline and over follow-up of the 598 MSA patients under study in the French MSA cohort.

| Characteristic	N (%)	Mean ± SD
At baseline		
Sex		
Male	297 (49.67%)	
Female	301 (50.33%)	
Hospital		
Bordeaux	309 (51.67%)	
Toulouse	289 (48.33%)	
Type of diagnosis		
MSA-C, with predominant cerebellar impairment	198 (33.11%)	
MSA-P, with predominant parkinsonism	400 (66.89%)	
Diagnosis certainty		
Possible	144 (24.08%)	
Probable	454 (75.92%)	
Age at inclusion		65.05 ± 8.09
Years since first symptoms		4.54 ± 2.58
Clinical markers		
Total UMSARS-I score		20.88 ± 7.35
Total UMSARS-II score		23.11 ± 8.25
Supine diastolic BP (in mmHg)		81.84 ± 13.37
Supine systolic BP (in mmHg)		140.8 ± 23.52
Maximum drop of diastolic BP (in mmHg)		−17.37 ± 14.88
Maximum drop of systolic BP (in mmHg)		−33.87 ± 25.16
During follow-up		
Repeated visits per patient		2.98 ± 2.08
Length of follow-up (years)		6.96 ± 3.33
Death	309 (51.67%)	

3.1.2 | Class-specific repeated outcome distribution

The latent class approach assumes that the outcomes under study have a class-specific distribution. In this work, we consider both quantitative repeated outcomes and time-to-event outcomes. Let Y_{kij} denote the repeated value of outcome k ($k = 1, ..., K$) for subject i at occasion j ($j = 1, ...n_{ki}$). The corresponding time of measurement is t_{kij}. Let T_i^* denote the time to an event of interest with L possible causes ($l = 1, ..., L$). The time-to-event can be right-censored by censoring time C_i, and left truncated (as in our motivating application) with delayed entry T_{0i}. The observed time is denoted $T_i = \min(T_i^*, C_i)$ with indicator $d_i = l$ when the event of cause l occurs before censoring and $d_i = 0$ otherwise.

Quantitative repeated outcomes
The class-specific trajectories of the repeated outcomes Y_{kij} over time is modelled using a mixed model. The most common case is a continuous Gaussian outcome which trajectory over time is modeled by a linear mixed model specific to latent class g:

$$
\begin{cases}
Y_{kij} = Y_{ki}^*(t_{kij}) + \epsilon_{kij} \\
Y_{ki}^*(t_{kij})|_{c_i=g} = Z_{ki}(t_{kij})^T b_{ki}|_{c_i=g} + X_{Lki}(t_{kij})^\top \beta_{kg}
\end{cases} \tag{2}
$$
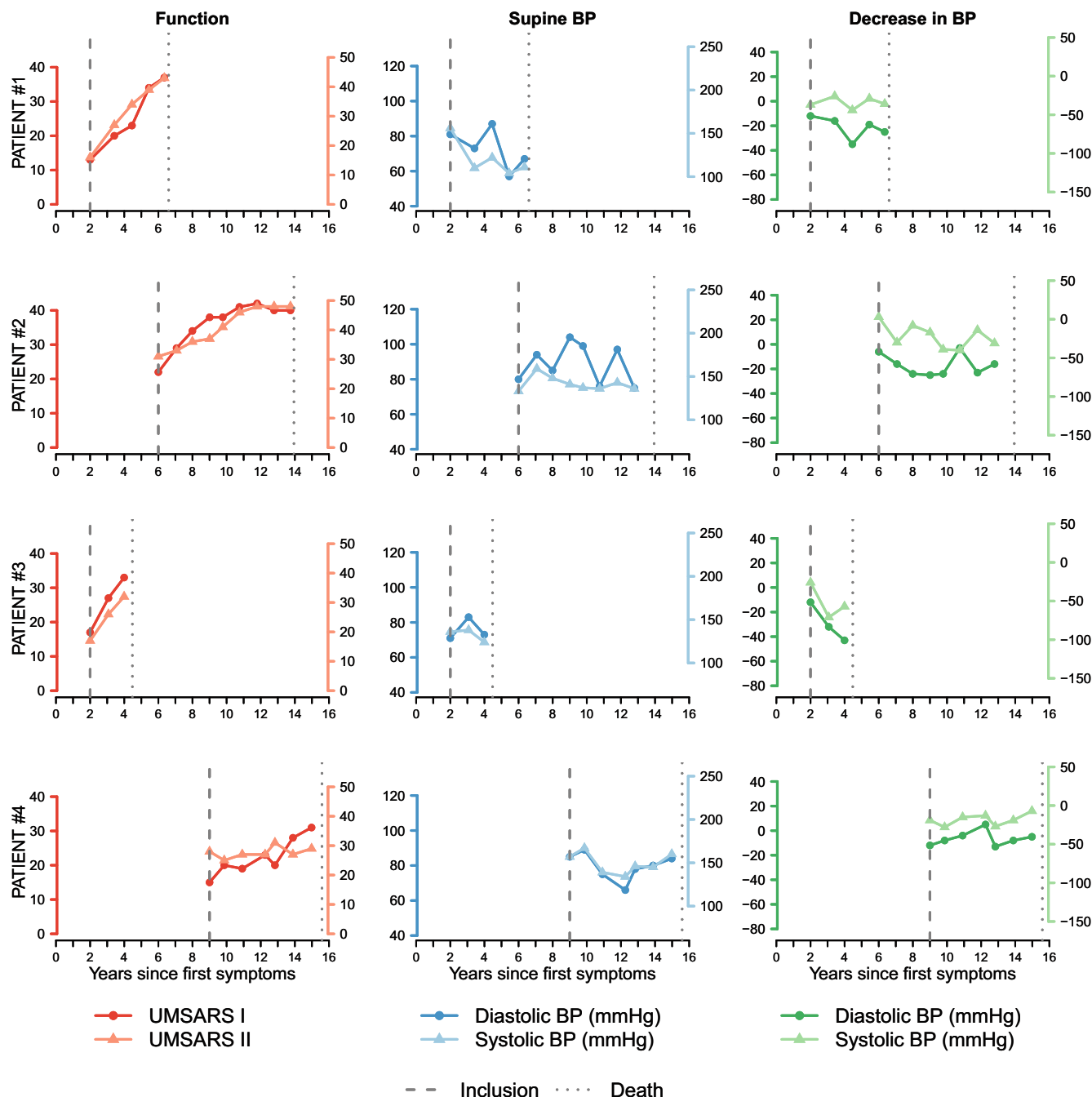
FIGURE 1 Observed marker trajectories for four randomly selected deceased patients according to the time since first symptoms. Are reported the time of inclusion in the cohort and the time of death. The six markers under study are the UMSARS I and II subscores both measuring functional dependency, the systolic and diastolic supine blood pressure (BP), and the maximum decrease in systolic and diastolic BP when standing up.

where $Y_{ki}^*(t_{kij})$ is the underlying level of the outcome without measurement error $\epsilon_{kij} \sim \mathcal{N}\left(0, \sigma_{\epsilon_k}^2\right)$; $X_{Lki}(t_{kij})$ is the vector of covariates associated with the fixed effects β_{kg}, and $Z_{ki}(t_{kij})$ is the vector of covariates (most of the time limited to functions of time) associated with the individual random effects b_{ki}, with a class-specific distribution: $b_{ki}|_{c_i=g} = b_{kig} \sim \mathcal{N}\left(\mu_{kg}, B_{kg}\right)$.

The mean trajectory over time t in each latent class is thus $\mathrm{E}(Y_{ki}^*(t)|_{c_i=g}) = Z_{ki}(t)^T \mu_{kg} + X_{Lki}(t)^\top \beta_{kg}$.

Quantitative repeated outcomes structured into latent dimensions

When some markers measure the same underlying construct as systolic and diastolic blood pressure for instance, the markers can be structured into a reduced number D of dimensions ($d = 1, \ldots D$), and the class-specific distribution applies to each dimension.

Following previous works,[26] we use a latent process approach to define the D latent processes $\Lambda_i^d(t)$ from the repeated measures of the K observed repeated outcomes. We consider that each outcome is a noisy measure of only one latent dimension using the following equation of observation:

$$H_k(Y_{kij}; \boldsymbol{\eta_k}) = \Lambda_i^{d(k)}(t_{kij}) + \epsilon_{kij} \tag{3}$$

where $\Lambda^{d(k)}$ is the latent dimension measured by outcome Y_k, and H_k is a bijective link function parameterized by $\boldsymbol{\eta_k}$ which puts each outcome k into the scale of the shared dimension $d(k)$.

Then, the class-specific trajectory of dimension d is defined at any time t $(t \in \mathbb{R})$ by a class-specific linear mixed model similar to the one given in Equation (2):

$$\Lambda_i^d(t)|_{c_i=g} = Z_{di}(t)^T b_{di}|_{c_i=g} + X_{Ldi}(t)^\top \beta_{dg} \tag{4}$$

with $Z_{di}(t)$, $X_{Ldi}(t)$, b_{di} and β_{dg} having the same definition as in Equation (2) except that they apply to the latent dimension d instead of the outcome k.

Note that this more general formulation for the repeated outcomes modeling defined in Equations (3) and (4) includes different special cases:

- one Gaussian continuous marker by dimension, that is Equation (2), when H_k is the identity and each marker has its own underlying dimension (ie, $d(k) = k$ and $\Lambda_i^{d(k)}(t) = Y_{ki}^*(t)$).
- one non-Gaussian continuous marker by dimension when each marker has its own underlying dimension (ie, $d(k) = k$) but H_k is a nonlinear link function usually modelled using a basis of M I-splines functions $(I_m)_{m=1,\ldots,M}$, that is:

$$H_k(x; \boldsymbol{\eta_k}) = \eta_{0k} + \sum_{m=1}^{M} \eta_{mk} I_m(x) \quad \text{with } x \in \text{range}(Y_k) \tag{5}$$

In that case, some constraints need to be added so that Λ^k has a determined dimension; this is usually done with a 0 mean in the reference category (for the location constraint), and either $\sigma_{\epsilon_k}^2 = 1$ or first diagonal element of $B_{1k} = 1$ (for the dispersion constraint).
- multiple Gaussian and/or non-Gaussian continuous markers by dimension by appropriately define H_k either as $H_k(x; \boldsymbol{\eta_k}) = \eta_{0k} + \eta_{1k}x$ (with $x \in \mathbb{R}$) for a Gaussian marker or according to Equation (5) for a non-Gaussian marker. As above, in this multivariate case, each Λ^d needs to have a determined dimension with one constraint on the location, and one constraint on the dispersion.

3.1.3 | Class-specific times-to-event distribution

The time-to-event distribution in each latent class can be classically modelled within the cause-specific proportional hazard model framework where the class-specific instantaneous risk of event of cause l is defined as follows:

$$\alpha_{il}(t)|_{c_i=g} = \alpha_{0lg}(t; \boldsymbol{\zeta_g}) \exp\left(\boldsymbol{X_{Ti}}^\top \boldsymbol{\delta_{lg}}\right) \tag{6}$$

where $\alpha_{0lg}(t)$ is the instantaneous baseline hazard of cause l in latent class g and $\boldsymbol{X_{Ti}}$ is a vector of covariates associated with fixed effects $\boldsymbol{\delta_{lg}}$ (such fixed effects can also be considered as common over classes). Although any type of parametric hazards could be considered, we focus on Weibull hazards or approximate the baseline hazards by a small number of cubic M-splines. In addition, the baseline hazards can be either specific to each latent class or considered proportional across classes (ie, $\alpha_{0lg}(t; \boldsymbol{\zeta_g}) = \alpha_0(t; \boldsymbol{\zeta_0}) e^{\zeta_g}$ with $\zeta_G = 0$).

3.2 | Inference

3.2.1 | Maximum likelihood estimation

The joint latent class model defined in Section 3.1 can be estimated in the maximum likelihood framework for a given number of latent classes G. Let $\boldsymbol{\theta_G}$ denote the total vector of parameters for a G-class model. It includes all the parameters (subscripts k, g, d are not reported here for readability) for:

- the latent class structure ξ;
- the class-specific repeated outcomes distributions with fixed effects noted β and μ, variance-covariance of the random-effects vec(B) (parameterized using the Cholesky transformation), standard deviations of the errors σ_ϵ and parameters of the link functions when necessary η;
- the time-to-event outcomes distribution ζ and δ.

Thanks to the conditional independence between dimensions and time-to-event, the individual contribution l_i to the likelihood based on the joint distribution of the repeated outcomes $Y_i = \{Y_{kij} \text{ with } k = 1, ..., K, j = 1, ..., n_{ik}\}$ and the time to event (T_i, d_i) can be split as follows:

$$
\begin{aligned}
l_i(\theta_G) &= f(Y_i, (T_i, d_i); \theta_G) \\
&= \sum_{g=1}^{G} P(c_i = g; \theta_G) \times \quad f(Y_i | c_i = g; \theta_G) \quad \times \quad f((T_i, d_i) | c_i = g; \theta_G) \\
&= \sum_{g=1}^{G} \quad \pi_{ig} \quad \times \quad \prod_{d=1}^{D} f(Y_i^d | c_i = g; \theta_G) \quad \times \quad S_i(T_i | c_i = g, \theta_G) \prod_{l=1}^{L} \alpha_{il}(T_i | c_i = g; \theta_G)^{\mathbb{1}_{d_i = l}}
\end{aligned}
$$

with f the generic notation of a density function, and P a probability function.

The class-membership probability π_{ig} is given in (1), the instantaneous hazard $\alpha_i(T_i | c_i = g; \theta_G)$ is defined in (6) and the corresponding survival $S_i(T_i | c_i = g, \theta_G) = \exp\left(-\sum_{l=1}^{L} \int_0^{T_i} \alpha_{il}(u | c_i = g; \theta_G) du\right)$. Finally, the density function of the repeated outcomes is split into the product of the density functions of the subset of outcomes data, called Y_i^d, linked to each dimension d. Given the general formulation in Equations (3) and (4), the density function is:

$$
f(Y_i^d | c_i = g; \theta_G) = \phi(H(Y_i^d); m_i^d, V_i^d) \times \prod_{k=1}^{K(d)} \prod_{j=1}^{n_{ik}} J(H_k(Y_{kij})) \tag{8}
$$

where ϕ is the Gaussian density function with mean $m_i^d = X_{Ldi}\beta_{dg} + Z_{di}\mu_{dg}$ and variance $V_i^d = Z_{di}B_{dg}Z_{di}^\top + \Sigma_{di}$. $H(Y_i^d)$ denotes the vector $(H_k(Y_{ki}), k = 1, \dots, K(d))^\top$ and Σ_{di} is the diagonal matrix composed of values $\sigma_{\epsilon_k}^2$ with $k = 1, \dots, K(d)$. $J(H_k(x))$ is the Jacobian of the link function H_k. See Proust-Lima et al[26] for further details.

To take into account the delayed entry in T_{0i} when times-to-event are left-truncated, the final individual contribution is divided by the probability to still be at risk of the events at entry:

$$
l_i^{LT}(\theta_G) = \frac{l_i(\theta_G)}{\sum_{g=1}^{G} \pi_{ig} S_i(T_{0i} | c_i = g, \theta_G)} \tag{9}
$$

The final log-likelihood to be maximized is computed on all the subjects as $\mathcal{L}(\theta_G) = \sum_{i=1}^{N} \log(l_i^{LT}(\theta_G))$.

3.2.2 | Posterior classification

The posterior distribution of the latent classes can be derived from the observed data as:

$$
P(c_i = g | Y_i, (T_i, d_i); \theta_G) = \frac{P(c_i = g; \theta_G) \times f(Y_i | c_i = g; \theta_G) \times f((T_i, d_i) | c_i = g; \theta_G)}{f(Y_i, (T_i, d_i); \theta_G)} \tag{10}
$$

where each element is calculated similarly as in the individual contribution to the log-likelihood in (7).

This posterior distribution is classically computed at the point estimate $\hat{\theta}_G$, giving $\hat{\pi}_{ig} = P(c_i = g | Y_i, (T_i, d_i); \hat{\theta}_G)$, and the posterior classification is derived: each individual is assigned to the latent class that provides the maximum individual posterior probability, that is the most likely class $\hat{c}_i = \text{argmax}_{g=1,...,G} \hat{\pi}_{ig}$.

3.2.3 | Optimal number of latent classes

Maximum likelihood is obtained for a fixed number of latent classes G, and the optimal number of latent classes thus needs to be *a posteriori* determined. Selecting the optimal number of clusters in mixture problems is a wide area of statistical research.[27]

Among information criteria, the Bayesian Information Criterion (BIC) is usually favored. Defined as $\text{BIC}(G) = -2\mathcal{L}(\hat{\boldsymbol{\theta}}_G) + p\log(N)$ with the lower the better, it was repeatedly shown to correctly select the optimal number of latent classes in different mixture situations.[28,29] However, as it is a likelihood-based criterion, the BIC mainly focuses on the fit of the model to the data and may lead in some contexts to a poorly discriminant latent class structure.[27,30]

The discriminatory power of the latent class structure can be assessed by an entropy measure defined as $\text{EN}(G) = 1 + \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\hat{\pi}_{ig}\log(\hat{\pi}_{ig})}{N\log(G)}$ with values closer to one indicating higher discrimination of the classes.[31] However, as built only from the posterior probabilities, this entropy measure completely neglects the fit of the model.

When interested both in the fit and the clustering, the Integrated Classification Likelihood criterion (ICL)[31,32] has been considered. Defined as $\text{ICL}(G) = \text{BIC}(G) - 2 \times \sum_{i=1}^{N}\sum_{g=1}^{G}\mathbb{1}_{\hat{c}_i=g}\log(\hat{\pi}_{ig})$, this criterion penalizes the fit of the data by the discrimination power and thus can identify the latent class structure that provides the best balance between fit and discrimination. This is particularly useful in our exploratory context where we favor the identification of different relevant subphenotypes (ie, classes) rather than the best fit to the data.

3.2.4 | Multimodality

One critical issue with latent class models is the multimodality of the likelihood and the potential convergence toward local suboptimal maxima[33] with local optimizer. To ensure convergence toward the global likelihood maximum, we highly recommend the use of a gridsearch that replicates the estimation process for a large number of random initial values, and thus likely explores the entire parameter space and reaches the global likelihood maximum.

3.2.5 | Software

The maximum likelihood estimation of the joint latent class model is implemented in lcmm R package[20] with function mpjlcmm when considering $K \geq 1$ repeated markers and Jointlcmm when considering $K = 1$ marker. Log-likelihood optimization is carried out by a robust Marquardt-Levenberg algorithm combined with stringent convergent criteria on the log-likelihood, the parameter and the first and second derivatives. This optimizer has been demonstrated to provide correct inference even in the case of complex log-likelihoods and/or relatively flat regions.[34]

The package includes a gridsearch function for the parameter space exploration, and postfit functions for reporting the information criteria, posterior classification, goodness-of-fit, and predicted trajectories.[35]

3.3 | Association with external information

After a latent class model estimation, one may want to assess the external predictors $\boldsymbol{X}_i^{\text{extern}}$ of the latent class structure c_i, or one may want to determine how the latent class structure c_i relates with an external outcome Y_i^{extern} using regression techniques. In both cases, a naive approach consists in running the posterior regressions on an estimate of c_i (for instance the most likely class $\hat{c}_i$) instead of the true unknown c_i, and neglecting the uncertainty in the estimate of c_i. The inference quality of this naive method, usually called "2-stage"[21] or "3-step" method,[23] depends on the discrimination of the latent classes. While it can provide negligible bias in the case of highly separated classes (with high posterior probabilities, high entropy), it may become substantially biased in the case of rather poor discrimination.[21-23] Alternatives consist either in corrections in the multi-step analysis to account for the uncertainty,[23] or directly in the joint estimation of all the variables of interest including the external information $\boldsymbol{X}_i^{\text{extern}}$ and/or Y_i^{extern} to internally handle the measurement error in the latent class assignment.[21] While the joint estimation naturally handles the latent nature of the classes and the uncertainty, one drawback is that the external information becomes part of the latent class model estimation and may slightly change the latent class structure. In many situations, the statistician wants to determine the latent class structure using a set of outcomes (in our case Y_i and (T_i, d_i)), and relate this specific latent class structure with other external outcomes in posterior analyses (in our case $\boldsymbol{X}_i^{\text{extern}}$ and/or Y_i^{extern}).

To both separate the estimation process of the latent class structure from the posterior analyses (as done in multi-step techniques), and account for the latent nature of the class structure (as done in the joint estimation technique), we used an intermediate approach. We considered the joint likelihood including Y_i, (T_i, d_i) and the external information, either

X_i^{extern} or Y_i^{extern}. However, we did not re-estimate all the parameters involved, only those related to X_i^{extern} or Y_i^{extern} as explained below.

Case 1. The latent class structure as a predictor of an external outcome Y_i^{extern}

We consider the case of a continuous marker Y_i^{extern} either measured once or repeatedly over time. The same generic model as defined in (3) and (4) can be applied to Y_i^{extern} with its specific underlying level Λ^{D+1}. Note that the use of the identity in (3) and no random effects in (4) provides for instance a linear model for cross-sectional data. Let's define ψ_G^{extern} the total vector of parameters for this external outcome. By including the external outcome, the contribution to the joint log-likelihood previously defined in (7) now includes a fourth element:

$$
\begin{aligned}
l_i^{Y^{\text{extern}}}\left(\boldsymbol{\theta}_G, \boldsymbol{\psi}_G^{\text{extern}}\right) &= f\left(\boldsymbol{Y_i}, (T_i, d_i), Y_i^{\text{extern}}; \boldsymbol{\theta}_G, \boldsymbol{\psi}_G^{\text{extern}}\right) \\
&= \sum_{g=1}^{G} P(c_i = g; \boldsymbol{\theta}_G) \times f(\boldsymbol{Y_i}|c_i = g; \boldsymbol{\theta}_G) \times f((T_i, d_i)|c_i = g; \boldsymbol{\theta}_G) \times f(Y_i^{\text{extern}}|c_i = g; \boldsymbol{\psi}_G^{\text{extern}})
\end{aligned}
\tag{11}
$$

The posterior regression for Y_i^{extern} can be estimated by maximising $\mathcal{L}^{Y^{\text{extern}}}\left(\hat{\boldsymbol{\theta}}_G, \boldsymbol{\psi}_G^{\text{extern}}\right) = \sum_{i=1}^{N} \log\left(l_i^{Y^{\text{extern}}}\left(\hat{\boldsymbol{\theta}}_G, \boldsymbol{\psi}_G^{\text{extern}}\right)\right)$ according to $\boldsymbol{\psi}_G^{\text{extern}}$.

Case 2. External information X_i^{extern} as the predictor of the latent class structure

This case is usually sought when the joint latent class model does not already include predictors of the latent class structure in Equation (1). External predictors X_i^{extern} can be easily included a posteriori by updating Equation (1) with $X_{Ci} = X_i^{\text{extern}}$. The estimation technique is then very similar as for case 1. We define ξ_G^{extern} the total vector of parameters involved in the updated formula (1) according to X_i^{extern}, and we consider the following contribution to the joint log-likelihood where the component involving parameters ξ_G^{extern} is now re-estimated. Note that for clarity we mention here the condition on X_i^{extern}:

$$
\begin{aligned}
l_i^{X^{\text{extern}}}\left(\boldsymbol{\theta}_G, \boldsymbol{\xi}_G^{\text{extern}}\right) &= f\left(\boldsymbol{Y_i}, (T_i, d_i)|X_i^{\text{extern}}; \boldsymbol{\theta}_G, \boldsymbol{\xi}_G^{\text{extern}}\right) \\
&= \sum_{g=1}^{G} P(c_i = g|X_i^{\text{extern}}; \boldsymbol{\xi}_G^{\text{extern}}) \times f(\boldsymbol{Y_i}|c_i = g; \boldsymbol{\theta}_G) \times f((T_i, d_i)|c_i = g; \boldsymbol{\theta}_G)
\end{aligned}
\tag{12}
$$

The posterior regression of c_i according to X_i^{extern} can be estimated by maximising $\mathcal{L}^{X^{\text{extern}}}\left(\hat{\boldsymbol{\theta}}_G, \boldsymbol{\xi}_G^{\text{extern}}\right) = \sum_{i=1}^{N} \log\left(l_i^{X^{\text{extern}}}\left(\hat{\boldsymbol{\theta}}_G, \boldsymbol{\xi}_G^{\text{extern}}\right)\right)$ according to $\boldsymbol{\xi}_G^{\text{extern}}$.

4 | SIMULATION STUDY

We carried out a simulation study to explore the finite sample properties of the estimation procedure of the multivariate joint latent class model. The simulation study is fully detailed in Supplementary Materials, Section S1. This includes the generation algorithm, the description of the different scenarios, the results and their interpretation, as well as a replication script in Section 3. Briefly, we generated series of 300 samples of 250, 500, or 750 subjects constituted of three latent classes, $K = 2$ or 3 repeated outcomes with class-specific linear trajectories and one or two competing-cause survival times under class-specific Weibull risks or proportional piecewise exponential hazards with parameters chosen to achieve different levels of entropy (between 63% and 83%) and different proportions of events (between 18% and 77%). This lead to nine scenarios (see description in Table S1), six of them run for the three sample sizes. For each sample, the model was estimated using a grid of 100 random sets of initial values.

All the simulation results are reported in supplementary Tables S2 to S25. We also report the estimated parameters along with coverage rates of the 95% confidence interval in Figure 2 for scenario 4 that included two competing causes of event and an entropy of 0.71. Overall, the simulation results illustrate the correct estimation of the parameters in all the scenarios with negligible bias and good coverage rates of the 95% confidence intervals for samples of 500 and 750 subjects.
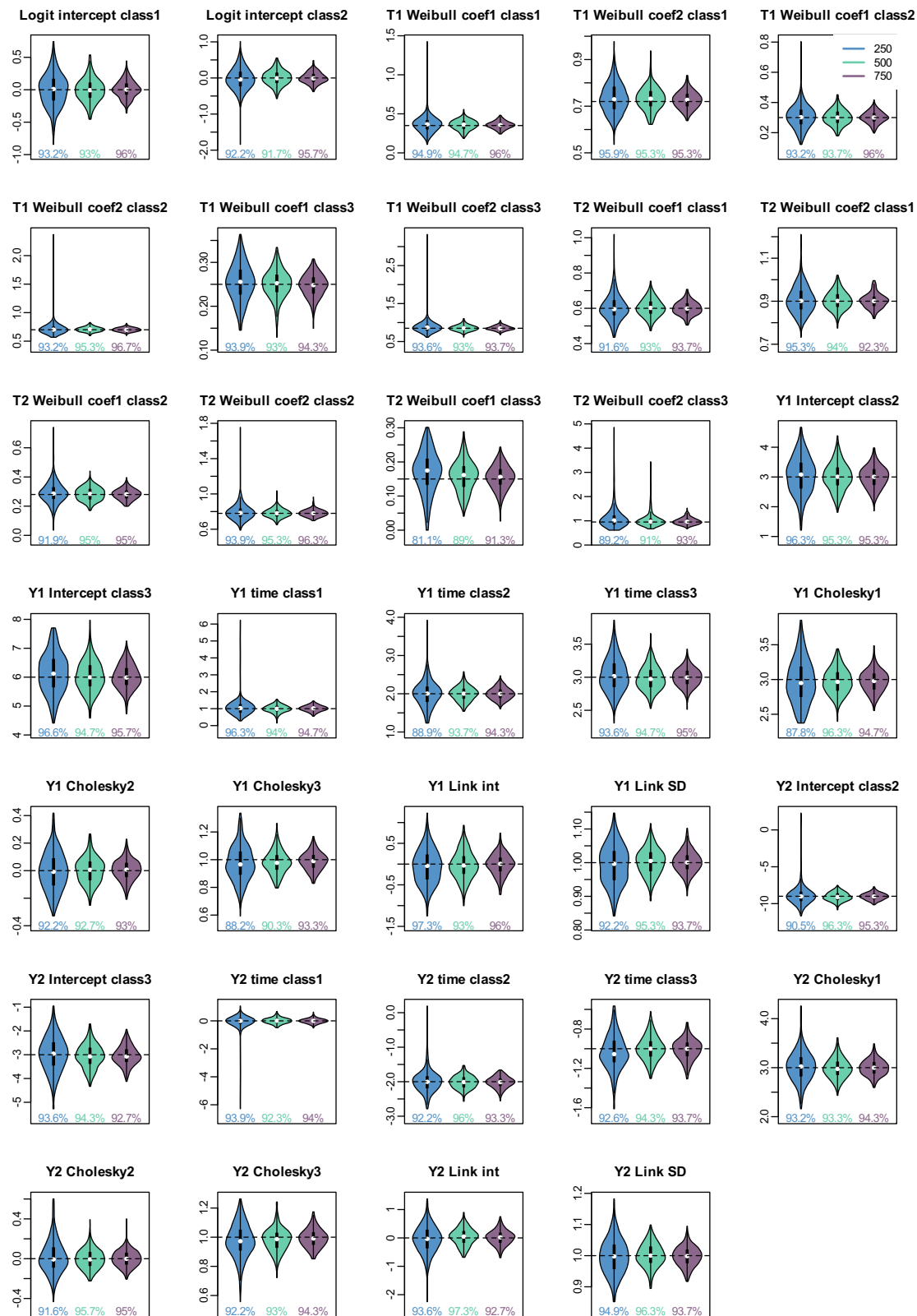
FIGURE 2 Violin plots of the parameters estimates in the 300 replicates of simulation Scenario 4 for 250, 500, and 750 subjects. The specified model included three latent classes, two markers with linear trajectory (Y1, Y2) and a class-specific and cause-specific Weibull risk of event for two causes of event (T1, T2).

For samples of 250 subjects, the estimates were generally well estimated although scenarios 3, 4, 5 revealed a small bias and too low coverage rate for the class-specific survival parameters in cases where the number of events in the class was very low (<10 events). This was explained by rare extreme values and these estimations remained overall good over the replicates despite the very low proportion of events (see violin plots in Figures 2 and S3, for scenarios 4 and 5, respectively).

Scenario 8 also aimed to illustrate how a three-dimensional model could help identify further heterogeneity compared to uni-dimensional models (with 3 classes identified while uni-dimensional models could only retrieve 2 classes) (Supplementary section 1.3.8., Tables S23, S24).

5 | APPLICATION TO MSA PROGRESSION

We applied the joint latent class methodology to describe the progression of the six markers measured repeatedly over time and grouped into three dimensions: **function** with the sumscores UMSARS I and II, **supine blood pressure** with the diastolic and systolic measures, **orthostatic BP drop** with the maximum decreases in systolic and in diastolic blood pressure, along with the **time to death**. The specification of the model is summarized in Figure 3. Each parametric assumption was carefully assessed in separate preliminary analyses for $G = 1$ by comparing different candidates according to the Akaike Information Criterion (AIC) and BIC. For instance, better AIC were found when considering a Weibull baseline risk function compared to a basis of M-splines with three internal knots, when considering quadratic splines for the link functions compared to linear transformations, or when considering a linear trajectory of the dimensions over time compared to nonlinear trajectories approximated by natural cubic splines or polynomials. In addition, although the three underlying dimensions were defined based on clinical knowledge, we compared a latent process model (where the two constituting markers are assumed to measure the same underlying process) with a bivariate mixed model (where each marker has its own trajectory, and random effects are correlated between markers). The assumption of an underlying process was reasonable for the three dimensions under $G = 1$. The AIC and BIC concluded to the selection of the
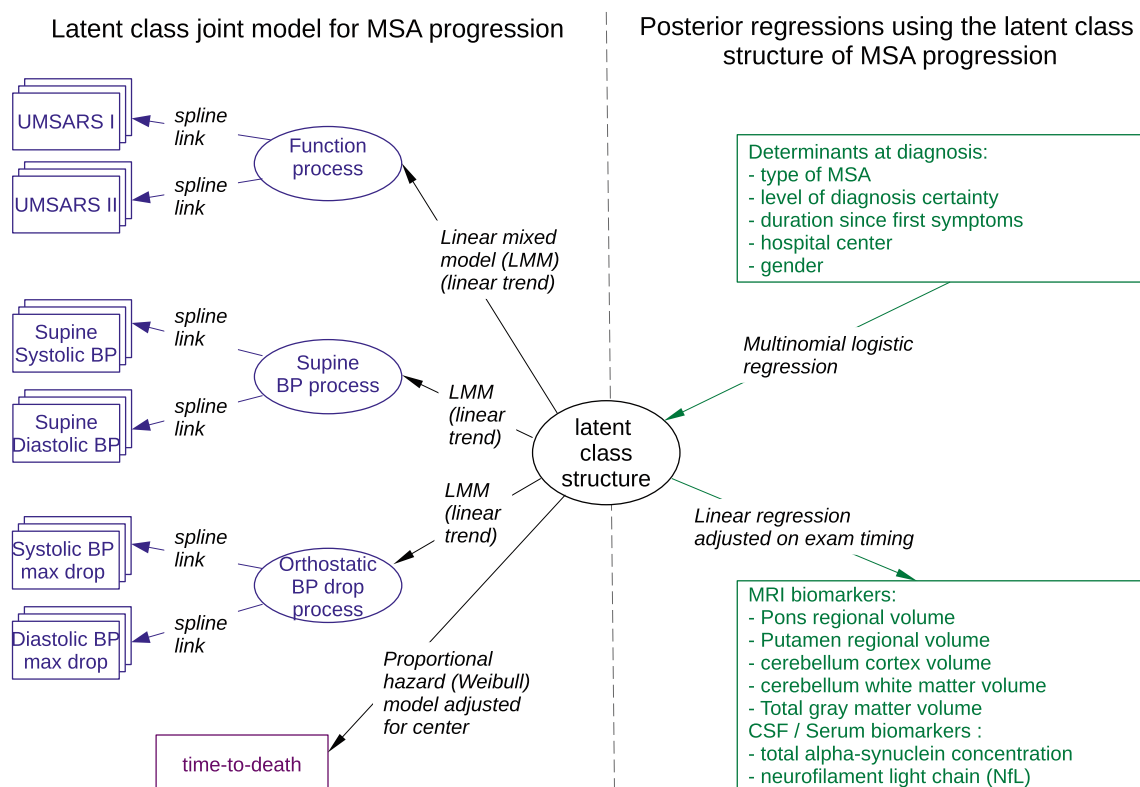
FIGURE 3 Diagram summarizing the definition of the joint latent class model for MSA progression (left part), and the associated posterior analyses (right part). Are reported the chosen specifications for each submodel (carefully determined in preliminary separated analyses after comparison with alternative candidates according to AIC).

latent process model for Supine BP (AIC/BIC = 26,170.01/26,275.45 and 26,175.04/26,245.33 for the bivariate model and for the latent process model, respectively) and Orthostatic BP (AIC/BIC = 25,483.22/25,588.67 and 25,471.33/25,541.63 for the bivariate model and for the latent process model, respectively). For the Function process, the bivariate mixed model provided a better fit (AIC/BIC = 21,731.81/21,837.25 and 21,909.83/21,980.13 for the bivariate model and for the latent process model, respectively) but the latent process model remained reasonable and was clinically justified as both UMSARS-I and UMSARS-II scores measure Function degradation.

5.1 | Selection of the number of latent classes

Joint latent class models assuming between 1 and 6 latent classes were repeatedly estimated using a grid of 100 random initial values. Figure 4 summarizes the three statistical criteria (BIC, Entropy, ICL) used for determining the optimal number of latent classes. While the goodness-of-fit was gradually improved when adding a new latent class, the entropy was clearly better for the five-class model (Entropy at 0.76) suggesting that although the six-class was even closer to the data, it did not provide a sufficiently high separation of the patients. As the ICL which accounts for both goodness-of-fit and discrimination also favored the five-class model, we retained these five subphenotypes of MSA clinical progression. The Sankey plot (displayed in Figure S5 in supplementary materials) describes the sequence of latent class splits with the increasing number of classes.

5.2 | Five subphenotypes of MSA progression

The mean trajectories of the six markers and the predicted death probability characterizing the five subphenotypes of MSA progression are reported in Figure 5. The five subphenotypes differed by the shape and speed of progression of the three dimensions, and the risk of death. The largest class (Class 3) with 46.7% of the sample was characterized by a much slower deterioration of the function (UMSARS I and II) than others, and a relatively stable level of supine BP, and slight increase in orthostatic BP drop. The second largest class (Class 1) with 31.4% of the sample was characterized by a fast deterioration of the function but also a decrease in supine BP over time and a rather stable or slight decrease in orthostatic BP drop. Classes 2 and 5 comprised around 9% of the sample each and were both characterized by similar shapes of clinical progression: fast deterioration of the function, increase in supine BP and aggravation of the orthostatic BP drop. However, the timing was different. The patients from class 5 had a progression beginning right after the first symptoms while this progression was slightly delayed in class 2. Finally, the smallest class (Class 4), which included 3.7% of the sample, was characterized by a fast deterioration of the function, and at the same time, a
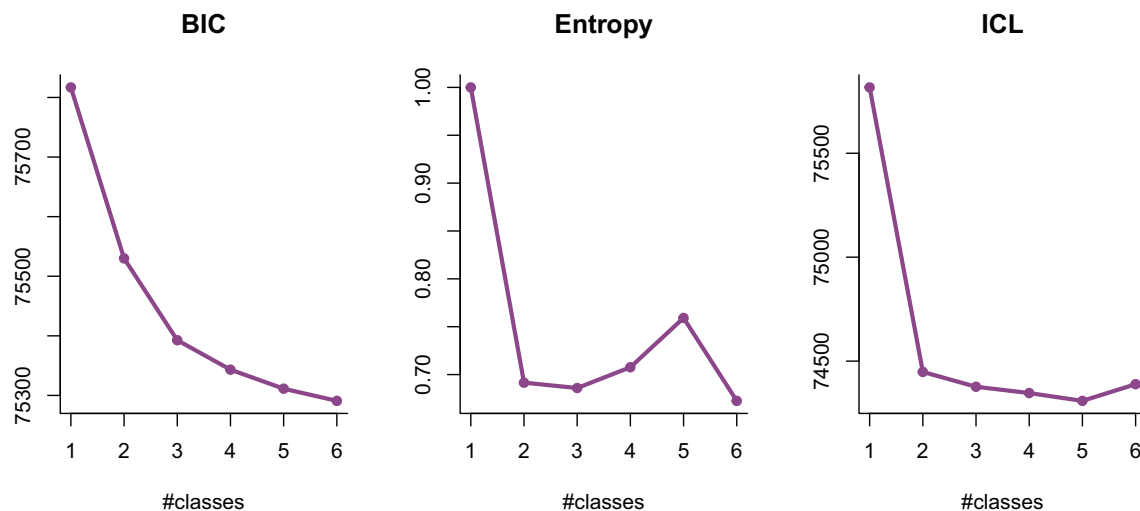
FIGURE 4 Comparison of models considering from 1 to 6 latent classes. Are reported the BIC for goodness-of-fit assessment, Entropy for discriminatory assessment, and ICL for an overall assessment.
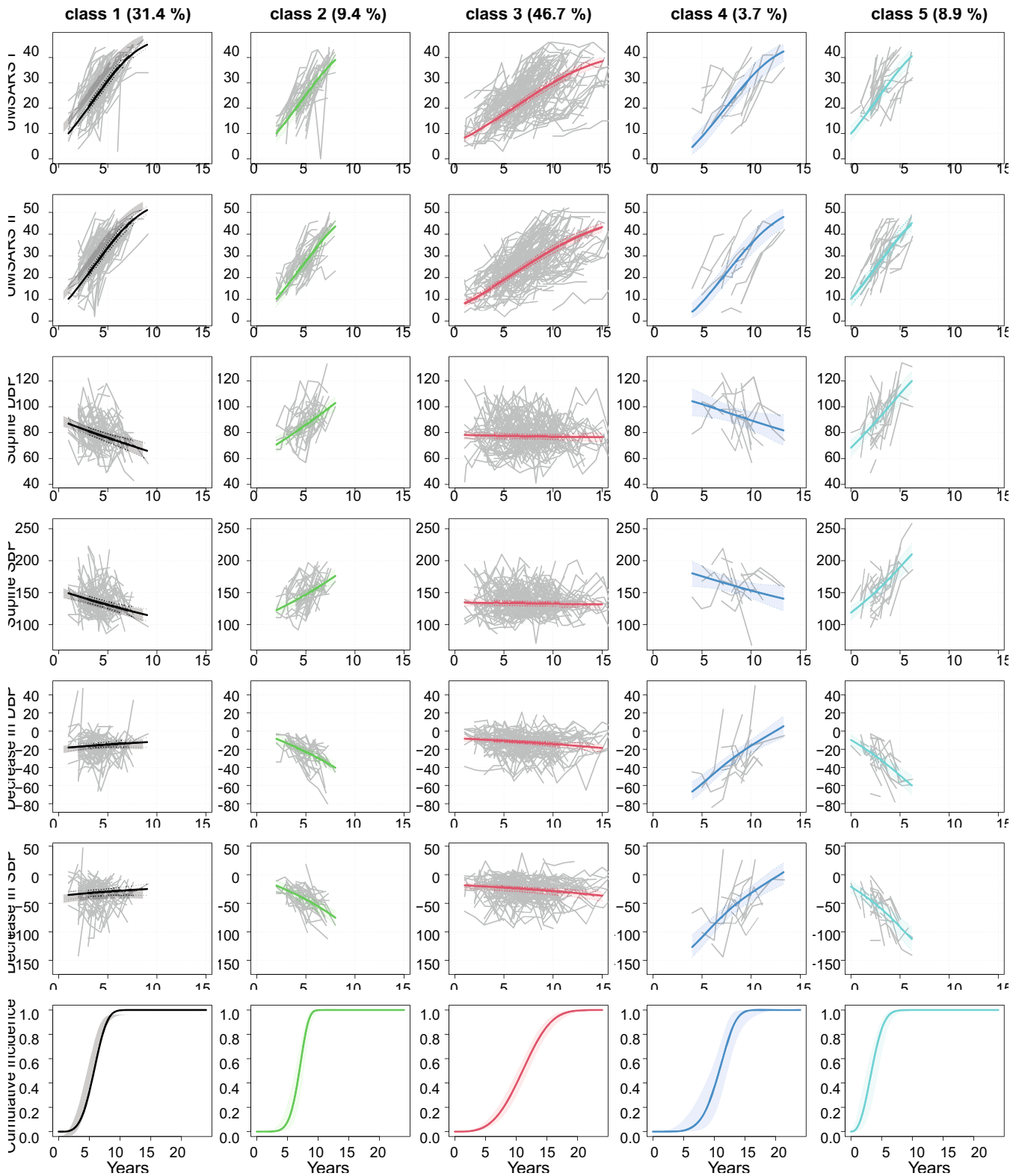
FIGURE 5 Predicted trajectories of the markers (and 95% in shades) and predicted death probability in the five latent classes. Are also reported in grey lines the observed trajectories of the patients *a posteriori* classified into the latent class.

decrease in supine BP and orthostatic BP drop which makes it very peculiar. As shown in the Sankey plot (Figure S4 in supplementary materials), the smallest latent class 4 was only identified when considering a fifth class. As this small class substantially differs from the others, this probably explains the gain in entropy observed between the 4- and 5-class models.

The risk of death was substantial in all the classes with a probability of death reaching 1 in all classes by 15 years after the first symptoms. It followed the functional degradation with a more progressive risk of death in class 3 compared to others and an earlier risk for class 5 and 1 compared to classes 3 and 4.

5.3 | Determinants of the latent classes

In posterior multinomial logistic regression, we assessed the determinants of the five subphenotypes (Figure S6 in supplementary materials). No difference was observed according to sex or MSA center. However, as previously identified, the duration between the first symptoms and the diagnosis substantially differed according to the latent class with later diagnoses for latent classes 3 and 4, and earlier diagnoses for latent classes 1 and 5. Patients with a cerebellar presentation of the disease were more likely classified in latent class 4 compared to others, and less likely classified in the fast progressors of class 1. Finally, patients with a probable diagnosis were much more likely classified in any other class than class 3 compared to patients with a possible diagnosis.

5.4 | Association with MRI and fluid biomarkers

Understanding the underlying biological mechanisms of MSA is particularly crucial for therapeutic development. Indeed, beyond MSA patients care, MSA constitutes a fast model for the group of α-synucleinopathies including Parkinson's disease. As such, identifying potential therapeutic targets, or differential biological mechanisms is of high importance. The classification may be useful to explore how new biomarkers differ according to this parsimonious summary of the MSA clinical progression.

We focused here on MRI biomarkers with five brain regions of interest (N = 86 patients), and serum and CSF measures of total α-synuclein (N = 23) and of NfL (N = 52). Due to the small sample sizes, we focused mainly on the differences between the two largest classes, class 1 of fast progressors and class 3 of slow progressors. The posterior linear regressions adjusted for the exam timing displayed in Figure 6 suggested a more preserved MRI structure for class 3 than class 1 with in particular a larger putamen volume and total gray matter. The concentration of NfL, a marker linked to the aggressiveness of axonal injury, also tended to be higher in the fast progression class 1 (classes 2 and 5 too) compared to the slow progression class 3, especially in the serum, while the CSF concentration of total α-synuclein was slightly lower for the fast progressors.

5.5 | Goodness-of-fit assessment

We followed the strategy described in Proust-Lima et al[9,14] to assess the goodness of fit of each part of the final joint latent class model:

- longitudinal data: we compared the trajectories of weighted mean predicted values in each class to the trajectories of weighted mean observations. Specifically, observation times were split into intervals. Then subject-and-class-specific marker predictions computed at each observation point of an interval were averaged with weights corresponding to the posterior individual probability to belong to each class. The same strategy was used for the observations. Applied to the selected five-class model, it showed that the weighted averaged predictions of each marker were very close to the weighted averaged observations (Figure S7 in supplementary materials).
- survival data: because of the delayed entry, we did not compare the predicted and observed survival functions or cumulative hazards in each class. Instead, we compared the weighted mean of class-and-subject-specific predicted instantaneous risk functions to the instantaneous risk function in each class estimated by a weighted piecewise hazard
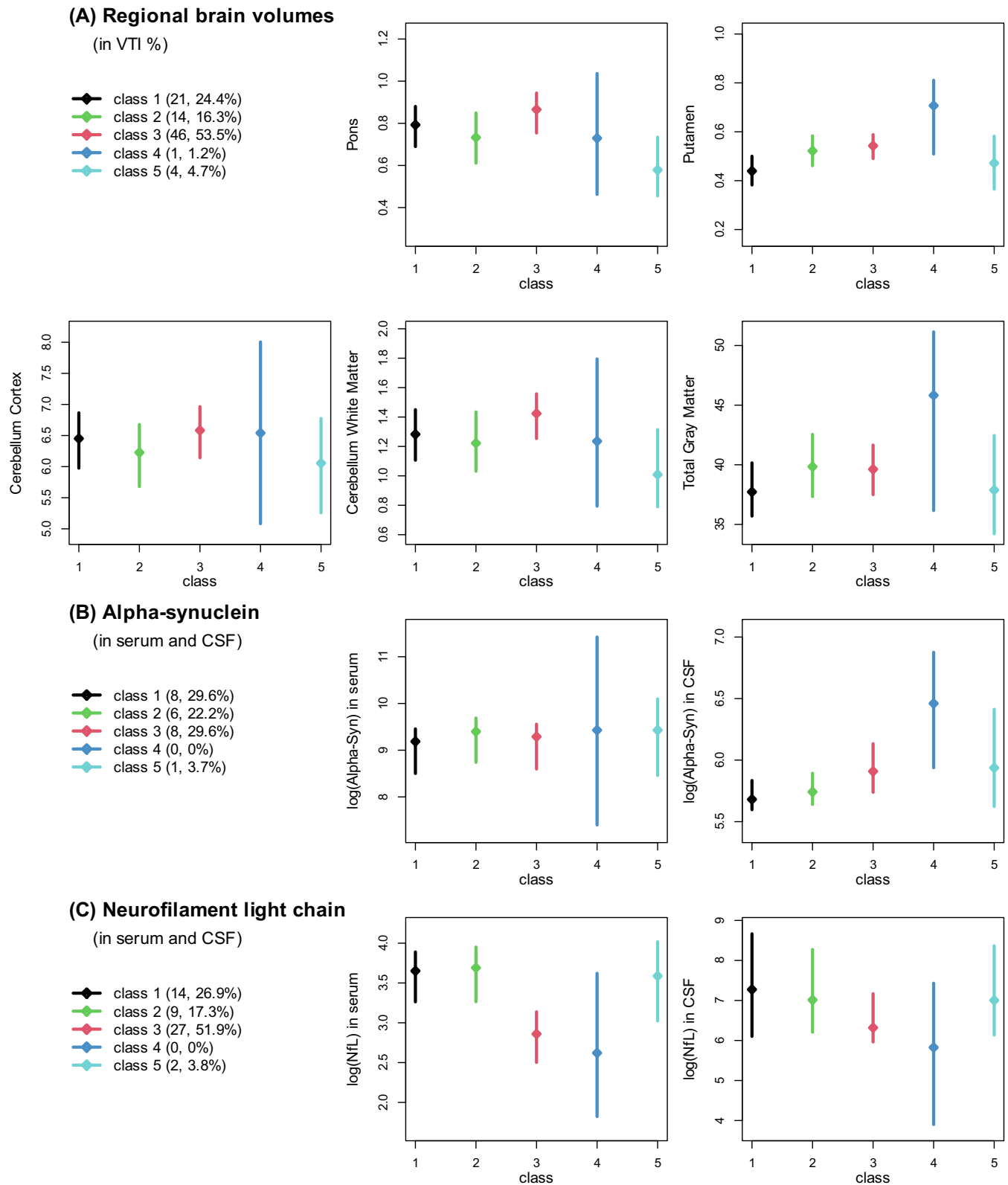
FIGURE 6 Posterior MRI and fluid biomarkers differences across classes predicted in separated linear regressions run on subsamples of 86 (A), 23 (B), and 52 (C) patients. Regressions are adjusted for the exam timing, and account for the uncertainty in the latent class assignments.

model with knots every two years. 95% confidence interval was obtained by non-parametric bootstrap with 200 samples. Applied to the selected five-class model, it showed that the weighted predicted risks of death were close to the observed ones (Figure S8 in supplementary materials).

- classification: the quality of the classification obtained from the five-class joint model was assessed by the posterior classification table (Table S30 in Supplementary Materials). In each class, the mean posterior probability of belonging to this class ranged from 77.7% in Class 1 to 88.6% for the slower progression class 3, and 89.4% for the small and peculiar class 4, indicating a clear discrimination between the latent classes.

5.6 | Comparison with unidimensional joint latent class models

In secondary analyses, we estimated a joint latent class model on each dimension taken separately. The model specification and the strategy of analysis remained the same. Four classes were identified for the function (12.5%, 29.8%, 41.8%, 15.9%–with a gradient of increasing slopes) and two small classes of rapid changes were distinguished for the supine BP (5.4%–with a substantial increase) and orthostatic BP (11.7%–with an amplification of the orthostatic hypotension), the rest of the sample having a rather stable progression of both BPs (Figure S9). When comparing the posterior classifications with the one of the five-class multidimensional joint model, the four latent classes of Function progression discriminated mainly classes 1, 3, and 4 while the two small classes of Supine and Orthostatic BP rather corresponded to latent classes 2 and 5.

6 | DISCUSSION

Given the complexity of some diseases and the richness of the data collected in cohort studies, methods to summarize multivariate longitudinal information, and capture heterogeneity become real assets in biostatistics. With this work and the associated implementation in the R package lcmm (function mpjlcmm), we provide a relevant and effective solution validated in simulations for summarizing information from multivariate markers measured repeatedly over time and clinical endpoints. In the MSA example, the approach summarized 6 marker trajectories and risk of death into five subgroups of patients with different profiles of progression that suggest distinct subphenotypes of the disease.

In addition to unraveling a heterogeneous clinical progression, this method provides a simple parsimonious summary of complex disease progression that can then be used to explore new research directions, and markers of interest. For instance, in MSA, although based on a very small subset of patients, posterior analyses of the classification suggested a preserved MRI structure in the slow progression class 3 compared to other classes, and higher NfL for the rapid progression classes confirming the higher aggressiveness of these profiles. A slightly lower concentration in CSF total α-synuclein was also observed for the fast progressors which suggests a higher pathological sequestration of α-synuclein in the brain for these patients.[36] The differences in total α-synuclein were small across classes compared to the differences in NfL. This is probably due to the fact that α-synuclein is a marker of the pathophysiological process rather than a marker of progression. Although further research is needed to confirm these observations, they illustrate how this statistical methodology opens up perspectives in a complex disease such as MSA to improve the understanding of pathophysiological processes. The statistical model relies on the assumption of conditional independence between processes, that is, all the correlations between such rich data can be captured by a few latent classes. We are aware that this assumption is likely violated. However, our objective was not to properly assess the nature of the association between the markers and clinical endpoints but to explore and identify heterogeneous profiles of progression that could be used as a parsimonious summary to be considered for external analyses. In this context, we did not further test the independence assumption. We relied instead on criteria such as the entropy or the ICL (which gives a balance between discrimination ability and goodness-of-fit) to assess the quality of the classification, and its discrimination ability. With a final entropy of 0.76, and mean posterior class-membership probabilities between 0.78 and 0.89, the five latent classes showed a very convincing split of the population into distinct profiles of progression that can be referred to as subphenotypes.

We carried out additional simulations to explore the behavior of the method under a residual correlation of 0.2 and 0.3 between the markers (See subsection 1.3.9 in supplementary materials). Overall, the parameters' distributions under misspecification showed a small bias but were not too much impacted (Figure S4, Tables S25–S27). Moreover, the percentage of individuals correctly classified remained similar to the one under true conditional independence (about 85% overall).

Previous works had formally addressed the issue of conditional independence in joint latent class models by focusing on simpler settings (single repeated marker and event time). Jacqmin-Gadda et al[37] and Proust-Lima et al[20] developed score tests to evaluate whether there is residual dependence between a repeated marker and clinical endpoints. The same strategy could be undertaken to provide a score test for conditional independence adapted to the multivariate context. Andrinopoulou et al[38] and Liu et al[39] also developed joint models that included both marker-event dependence on latent classes and on shared random-effects. However, these hybrid models already showed substantial numerical problems in the univariate situation, and their applications were limited to a too small number of classes to be useful in our context of progression summary. We thus leave these directions of development for future research.

Beyond the multivariate nature of the processes in play, the study of chronic disease progression usually induces additional complexities that our methodology and the associated software can handle: (i) Gaussian and non-Gaussian distributions of markers managed by defining parameterized link functions following previous work of the authors,[26] (ii) markers measuring the same underlying dimension handled by shared latent processes, (iii) delayed entry taken into account in the estimation procedure, (iv) competing risk setting with cause-and-class specific proportional hazard models (although not detailed here as not relevant in the MSA context, it is included in the software solution). Still, some issues are left for future improvements. First, although the theory could apply to other natures of repeated markers, especially within the exponential family with generalized linear mixed models applied to the latent dimensions, we only focused on continuous markers. Second, we described the trajectory according to the time since the first symptoms under the assumption that, at their inclusion in the cohort, the patients were able to accurately determine the time since their first symptoms. Dealing with this type of uncertainty calls for methodologies based on latent disease time[40] that could be combined with the latent class approach in the future. It is important to recall that our solution, although flexible, remains fully parametric. As such, each part of the model (eg, existence of underlying latent dimensions, link functions, shape of trajectory, baseline risk functions, selection of the number of latent classes) has to be carefully determined in preliminary analyses and posterior evaluations. In the application, we postulated notably the existence of underlying processes, each one measured by two markers. This was clinically justified and seemed reasonable given the data both in preliminary analyses and in the posterior comparison of predictions versus observations (Figure S7). Another essential caution with the use of joint latent class models and mixture models in general is that they constitute flexible approaches to model asymmetric distributions or heavy-tailed distributions even in the absence of a real latent class structure (see for instance Bauer and Curran[30] and discussants). This is why in this work where the latent class structure was central, we did not rely only on the goodness-of-fit but also on the discrimination performances with the entropy (Figure 4) and posterior class-membership probabilities (Table S30).

Linking latent classes to external outcomes, as done in MSA with MRI, CSF and plasma markers, constitutes one direction of research of its own due to the difficulty to account for the uncertainty in the estimated latent class structure.[21-24] In our work, we chose to directly integrate the external outcomes into the joint model program to correctly handle the uncertainty on the latent class membership, as suggested by others in a different latent class framework.[21,24]

In conclusion, the multi-dimensional latent class methodology described here is a powerful, flexible and effective tool for exploring disease progression, especially in complex settings as encountered in MSA with different markers of different dimensions and no clear biological assumption behind. It opens to a deeper understanding of the disease progression, and exploration for phenotypes differences. Although limited in our motivating example to several MRI, CSF and plasma markers, posterior analyses based on latent classes can also apply in high dimensional contexts with omics information for instance.

AUTHOR CONTRIBUTIONS

Cécile Proust-Lima and Viviane Philipps developed the statistical methodology and carried out the validation by simulations. Tiphaine Saulnier, Cécile Proust-Lima, and Alexandra Foubert-Samier conceived, designed and carried out the data analysis. Alexandra Foubert-Samier, Wassilios Meissner, Anne Pavy-Le Traon, Patrice Péran and Olivier Rascol provided the data. Cécile Proust-Lima drafted the manuscript. Viviane Philipps, Tiphaine Saulnier, Alexandra Foubert-Samier substantially revised the manuscript. All authors read and approved the final manuscript. All authors take responsibility for the integrity of the data and the accuracy of the data analysis.

ACKNOWLEDGMENTS

for the MRI acquisition and Centre d'Investigation Clinique (CIC) for the global coordination of the clinical study within INSERM-DGOS 2013–2014 grant.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT

Programs to simulate data and validate the inference method are available in the supplementary material of this article. MSA data are not shared.

ORCID

Cécile Proust-Lima ⬥ https://orcid.org/0000-0002-9884-955X

REFERENCES

1. Amieva H, Le Goff M, Millet X, et al. Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. *Ann Neurol.* 2008;64(5):492-498. doi:10.1002/ana.21509
2. Jack CR, Knopman DS, Jagust WJ, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* 2013;12(2):207-216.
3. Weintraub D, Tröster AI, Marras C, Stebbins G. Initial cognitive changes in Parkinson's disease. *Mov Disord.* 2018;33(4):511-519. doi:10.1002/mds.27330
4. Swick TJ, Friedman JH, Chaudhuri KR, et al. Associations between severity of motor function and nonmotor symptoms in Parkinson's disease: a post hoc analysis of the RECOVER study. *Eur Neurol.* 2014;71(3-4):140-147. doi:10.1159/000355019
5. Wenning GK, Colosimo C, Geser F, Poewe W. Multiple system atrophy. *Lancet Neurol.* 2004;3(2):93-103. doi:10.1016/s1474-4422(03)00662-8
6. Meissner WG, Fernagut PO, Dehay B, et al. Multiple system atrophy: recent developments and future perspectives. *Mov Disord.* 2019;34(11):1629-1642. doi:10.1002/mds.27894
7. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics (Oxford, England).* 2000;1(4):465-480.
8. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R.* Biostatistics Series. Boca Raton: Chapman & Hall/CRC; 2012. doi:10.2333/bhmk.29.81
9. Proust-Lima C, Séne M, Taylor JMG, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: a review. *Stat Methods Med Res.* 2014;23(1):74-90. doi:10.1177/0962280212445839
10. Sène M, Bellera CA, Proust-Lima C. Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. *J Soc Franç Stat.* 2014;155(1):134-155.
11. Lin H, Turnbull BW, McCulloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *J Am Stat Assoc.* 2002;97(457):53-65.
12. Liu Y, Lin Y, Zhou J, Liu L. A semi-parametric joint latent class model with longitudinal and survival data. *Stat Interface.* 2020;13(3):411-422. doi:10.4310/SII.2020.v13.n3.a10
13. Huang Y, Lu X, Chen J, Liang J, Zangmeister M. Joint model-based clustering of nonlinear longitudinal trajectories and associated time-to-event data analysis, linked by latent class membership: with application to AIDS clinical studies. *Lifetime Data Anal.* 2018;24(4):699-718. doi:10.1007/s10985-017-9409-0
14. Proust-Lima C, Dartigues JF, Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Stat Med.* 2016;35(3):382-398. doi:10.1002/sim.6731
15. Han J, Slate EH, Peña EA. Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Stat Med.* 2007;26(29):5285-5302. doi:10.1002/sim.2915
16. Rouanet A, Joly P, Dartigues JF, Proust-Lima C, Jacqmin-Gadda H. Joint latent class model for longitudinal data and interval-censored semi-competing events: application to dementia. *Biometrics.* 2016;72(4):1123-1135. doi:10.1111/biom.12530
17. Proust-Lima C, Joly P, Dartigues JF, Jacqmin-Gadda H. Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Comput Stat Data Anal.* 2009;53(4):1142-1154. doi:10.1016/j.csda.2008.10.017
18. Sun J, Herazo-Maya JD, Molyneaux PL, Maher TM, Kaminski N, Zhao H. Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics.* 2019;75(1):69-77. doi:10.1111/biom.12964

19. Li M, Lee CW, Kong L. A latent class approach for joint modeling of a time-to-event outcome and multiple longitudinal biomarkers subject to limits of detection. *Stat Methods Med Res.* 2020;29(6):1624-1638. doi:10.1177/0962280219871679

20. Proust-Lima C, Philipps V, Liquet B. Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *J Stat Softw, Articles.* 2017;78(2):1-56. doi:10.18637/jss.v078.i02

21. Elliott MR, Zhao Z, Mukherjee B, Kanaya A, Needham BL. Methods to account for uncertainty in latent class assignments when using latent classes as predictors in regression models, with application to acculturation strategy measures. *Epidemiology.* 2020;31(2):194-204. doi:10.1097/EDE.0000000000001139

22. Clark S, Muthén B. *Relating Latent Class Analysis Results to Variables Not Included in the Analysis*; 2009. Accessed November 27, 2020. http://www.statmodel.com/download/relatinglca.pdf

23. Bakk Z, Oberski DL, Vermunt JK. Relating latent class assignments to external variables: standard errors for correct inference. *Politic Anal.* 2014;22(4):520-540.

24. Bakk Z, Kuha J. Two-step estimation of models between latent classes and external variables. *Psychometrika.* 2018;83(4):871-892. doi:10.1007/s11336-017-9592-7

25. Wenning GK, Tison F, Seppi K, et al. Development and validation of the unified multiple system atrophy rating scale (UMSARS). *Mov Disord: J Mov Disord Soc.* 2004;19(12):1391-1402. doi:10.1002/mds.20255

26. Proust-Lima C, Amieva H, Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *Br J Math Stat Psychol.* 2013;66(3):470-487. doi:10.1111/bmsp.12000

27. Celeux G, Fruewirth-Schnatter S, Robert CP. *Model Selection for Mixture Models–Perspectives and Strategies.* first ed. Boca Raton: CRC Press; 2018.

28. Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture Modeling: a Monte Carlo simulation study. *Struct Equ Model Multidiscip J.* 2007;14(4):535-569. doi:10.1080/10705510701575396

29. Morgan GB. Mixed mode latent class analysis: an examination of fit index performance for classification. *Struct Equ Model Multidiscip J.* 2015;22(1):76-86. doi:10.1080/10705511.2014.935751

30. Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychol Methods.* 2003;8(3):338-363. doi:10.1037/1082-989x.8.3.338

31. Morgan GB, Hodge KJ, Baggett AR. Latent profile analysis with nonnormal mixtures: a Monte Carlo examination of model selection using fit indices. *Comput Stat Data Anal.* 2016;93:146-161. doi:10.1016/j.csda.2015.02.019

32. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Anal Mach Intell.* 2000;22(7):719-725. doi:10.1109/34.865189

33. Hipp JR, Bauer DJ. Local solutions in the estimation of growth mixture models. *Psychol Methods.* 2006;11(1):36-53. doi:10.1037/1082-989X.11.1.36

34. Philipps V, Hejblum BP, Prague M, Commenges D, Proust-Lima C. Robust and efficient optimization using a Marquardt-Levenberg algorithm with R package marqLevAlg. *R J.* 2021;13(2):365-379.

35. Proust-Lima C, Philipps V, ADaB L. Lcmm: extended mixed models using latent classes and latent processes. 2022.

36. Mollenhauer B, Locascio JJ, Schulz-Schaeffer W, Sixel-Döring F, Trenkwalder C, Schlossmacher MG. Synuclein and tau concentrations in cerebrospinal fluid of patients presenting with parkinsonism: a cohort study. *Lancet Neurol.* 2011;10(3):230-240. doi:10.1016/S1474-4422(11)70014-X

37. Jacqmin-Gadda H, Proust-Lima C, Taylor JM, Commenges D. Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics.* 2010;66(1):11-19. doi:10.1111/j.1541-0420.2009.01234.x

38. Andrinopoulou ER, Nasserinejad K, Szczesniak R, Rizopoulos D. Integrating latent classes in the Bayesian shared parameter joint model of longitudinal and survival outcomes. *Stat Methods Med Res.* 2020;29(11):3294-3307. doi:10.1177/0962280220924680

39. Liu Y, Liu L, Zhou J. Joint latent class model of survival and longitudinal data: an application to CPCRA study. *Comput Stat Data Anal.* 2015;91:40-50. doi:10.1016/j.csda.2015.05.007

40. Li D, Iddi S, Thompson WK, Donohue MC. Alzheimer's Disease Neuroimaging Initiative. Bayesian latent time joint mixed effect models for multicohort longitudinal data. *Stat Methods Med Res.* 2019;28(3):835-845. doi:10.1177/0962280217737566

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.