Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A Bayesian Semiparametric Markov Regression Model for Juvenile Dermatomyositis

Maria De Iorio[a]*, Natacha Gallot[b], Beatriz Valcárcel[c] and Lucy R. Wedderburn[d]

Juvenile Dermatomyositis (JDM) is a rare autoimmune disease which may lead to serious complications, even to death. We develop a two state Markov regression model in a Bayesian framework to characterise disease progression in JDM over time and gain a better understanding of the factors influencing disease risk. The transition probabilities between disease and remission state (and vice-versa) are a function of time homogeneous and time-varying covariates. These latter type of covariates are introduced in the model through a latent health state function, which describes patient-specific health over time and accounts for variability among patients. We assume a nonparameteric prior based on the Dirichlet Process to model the health state function and the baseline transition intensities between disease and remission state and vice-versa. The Dirichlet Process induces a clustering of the patients in homogeneous risk groups. To highlight clinical variables that most affect the transition probabilities we perform variable selection using spike and slab prior distributions. Posterior inference is performed through Markov chain Monte Carlo methods. Data were made available from the UK JDM Cohort and Biomarker Study and Repository, hosted at the UCL Institute of Child Health. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: Dermatomyositis, Dirichlet process, Markov chain Monte Carlo, Random effects, Two-state Markov model

1. Introduction

Juvenile Dermatomyositis (JDM) is a rare autoimmune disease that occurs in childhood. An autoimmune disease is caused by dysfunctional antibodies or lymphocytes that attack self molecules present in the body. The incidence of JDM is about 2-3 per million children per year and it is more common in females than males (2:1 ratio) [1, 2]. JDM is characterised by chronic inflammation affecting muscle, skin and other organs, which can cause pain, suffering, long term disability, and even death. In severe cases of JDM, growth, bone health and development can all be negatively affected.

Examples of severe complications in JDM are [3]:

- calcinosis, lumps of calcium that form under the skin or in the muscle,
- ulcerative skin disease,
- interstitial lung disease that causes inflammation of lungs making it difficult to inhale enough oxygen,
- gastrointestinal complications including ulceration, perforation or haemorrhage which can be fatal.

Since Juvenile Dermatomyositis is a rare chronic disease, research effort has been limited compared to more common diseases such as cardiovascular diseases or cancer. In particular, there have been no studies modelling the progression of JDM from disease to remission (and vice-versa), trying to identify the epidemiological and medical features of JDM and to

[a] *Department of Statistical Science, University College London*
[b] *Veramed, UK*
[c] *Infection Control and Environmental Health, Norwegian Institute of Public Health*
[d] *UCL Institute for Child Health, Great Ormond Street Hospital for Children NHS Trust, Arthritis Research UK Centre for Adolescent Rheumatology at UCL and GOSH/ICH NIHR- Biomedical Research Centre*
* *Correspondence to: Department of Statistical Science, University College. Gower Street, London WC1E 6BT, UK. Email: m.deiorio@ucl.ac.uk*

unveil the biological mechanisms that underlie it. As a consequence, at present there are no reliable methods, either clinical features or biomarkers, with which to stratify patients according to the level of risk of severe complications and in order to direct medication choices appropriately. However, increasing efforts have been made in the last decades to collect data on patients affected by JDM in order to gain a better understanding of the disease [4]. The Juvenile Dermatomyositis Cohort Biomarker Study and Repository (UK JDM Cohort Study, http://www.juveniledermatomyositis.org.uk/, JDCBS), housed at UCL, is a large cohort of 530 JDM cases contributed from 17 centres across the UK, with detailed longitudinal clinical data, linked biobank (DNA, cells, serum, muscle biopsy), genome-wide genotyping data and detailed immunological and serological data on blood/tissue biomarkers. The centres involved in the Juvenile Dermatomyositis Research Group (JDRG) all have specialist Paediatric Rheumatology teams working in the field. Research work in this field has clearly shown that JDM is not in fact homogeneous but highly heterogeneous and that different clinical subtypes may have widely differing long term outcomes [5]. Even though no curative treatment of JDM has been discovered yet, there are treatments available to control the disease and prevent severe medical complications. The outcomes, especially the time to achieve remission, may be very different between patients whilst depending on the choice of the treatment.

In this work we develop statistical methods within a Bayesian framework to model disease progression in JDM, highlighting the major risk factors. Disease progression models present many advantages: a better understanding of JDM progression, resulting in a more accurate and earlier diagnosis for patients and in therapy choices specifically adapted to suit patient needs. The available data consists of longitudinal measurements of disease status (remission/disease) on paediatric patients, with few measurements over time for each subject. It is typical of longitudinal medical studies on disease progression that disease development is expressed in terms of distinct health stages, where patients are observed periodically and this status is recorded at the time of the visit, resulting in interval censored data. Often covariate information is collected at the time of visit. This type of data is usually referred to as panel data and are often modelled as observations at arbitrary times of a continuous-time process with multiple transient states. Multi-state models are generalizations of survival and competing risks models and have been successfully applied to model the complex evolution of chronic diseases, generally assuming a Markov structure in the evolution of the disease [6, 7]. The complexity of a multi-state model mainly depends on the number of states and the possible transitions from these states. In this framework we are interested in modelling the transition probabilities from one state to another. When we deal with homogeneous processes (i.e. such probabilities do not change over time) this task is relatively easy as the transition probabilities from a state to another can be expressed simply in terms of transition intensities, but this is not the case for inhomogeneous Markov processes. Moreover, in the presence of interval censored data, several paths are possible for transitioning from state h to state j between time s and time t, and this uncertainty needs to be accounted for in the model. Transition probabilities are usually modelled parametrically, e.g. [8, 9], but parametric assumptions are often too restrictive in applications. More flexible alternatives can be found, for example, in Fahrmeir *et al.* [10], who model the transition intensities in a Cox-type manner with smoothing splines for time-varying effects, and in Aalen *et al.* [11], who introduce dynamic versions of multi-state models based on an additive risk model. Kneib *et al.* [12] propose a Bayesian semiparametric multi-state model with flexible transition intensities based on penalised splines. The transition intensities are modelled as smooth functions of time and can be related to parametric as well as nonparametric covariate effects.

In this work we also adopt a Bayesian semiparametric approach to model the transition probabilities from disease to remission and inversely over a specified time interval. These are expressed as function of time-invariant and time-varying covariates, as well as baseline transition intensities and a subject-specific health function. These latter two components are modelled nonparametrically to account for inter-subject variability, using a Dirichlet process mixture prior. Dirichlet process mixture (DPM) models [13, 14] are arguably the most common nonparametric Bayesian prior and have proved successful in many applications due to their flexibility and ease of computation. DPM models are mixtures of a parametric kernel with a random mixing measure, in this case the Dirichlet process (DP) introduced by Ferguson [15], and they can accommodate for heterogeneity in the population, allow for outliers, clustering of individuals and over-dispersion. This higher level of flexibility is often difficult to achieve using a single parametric distribution. Moreover, we are able to account for missing covariate information and to identify the most important clinical features and therapy information affecting the rate of progression of JDM, by specifying a spike and slab prior on the parameters that govern the distribution of the time-varying covariates.

Posterior inference is performed through Markov chain Monte Carlo (MCMC) algorithms, in particular we employ the software `JAGS` [16] and the R package `R2jags` to implement the Gibbs sampling for our analysis. In Supplementary Material we provide the JAGS code.

In Section 2 we describe the data and in Section 3 the Bayesian semiparametric model for disease progression. In Section 4 we introduce our variable selection strategy. Finally, in Section 6 the results of the analysis are presented and in Section 7 we conclude with a discussion.

2. The Juvenile Dermatomyositis National (UK & Ireland) Cohort Biomarker Study and Repository for Idiopathic Inflammatory Myopathies

Data for the analysis were provided by the Juvenile Dermatomyositis Research group (JDRG, `/www.juvenile dermatomyositis.org.uk`), which is part of the Juvenile Dermatomyositis Cohort Biomarker Study and Repository (UK and Ireland) for Idiopathic Inflammatory Myopathies [4]. The latter is a multicentre cohort study in which children newly-diagnosed or previously diagnosed with idiopathic inflammatory myopathies in the UK have been followed up from 2000. One of its main objectives is to determine the clinical characteristics of JDM and to identify important biomarkers with the goal of improving diagnosis, therapeutic choices, treatment response and patient management. Patients are eligible to enter the study if they are diagnosed with an established or presumed myositis, including JDM, before their sixteenth birthday. Children participating in another registry are excluded. Demography data, medical history, examination findings comprising physical examinations, clinical features of JDM (skin, rash and joints) and muscle assessments, laboratory results and therapy information are prospectively collected in 14 centres across the UK. Each patient is examined at the study entry time, then every 3-4 months for the first two years and following that once a year. An initial inspection of the database has revealed a large quantity of missing data, data inconsistencies and data entry errors. As such extensive quality control has been carried out by our group.

To summarise, the analysis dataset includes 54 covariates that are observed on 157 individuals for a total of 766 time points, which implies 4.9 observation times per patient on average (range 2-15).

2.1. Patients' demographics

We focus our analysis on an information rich subset of the data. We consider 157 children. As fixed covariates (i.e. not changing over time) in our analysis we use data on three medical history covariates - myalgia, rash and weakness - appraised as potentially clinically relevant [17, 18], as well as sex (with female as baseline), ethnicity and age at diagnosis. The study population includes nearly three times as many female children (74.6%) as male ones (25.4%). It concurs with previous research [3, 18]: prevalence is higher in females. The mean age of patients at diagnosis is 8.0 years old and the median age is 7.7. Regarding ethnicity, most patients (76.3%) are white, 5.8% are black and the remainder are distributed among other ethnic groups.

2.2. Patients' physical examinations and laboratory results

Physical examinations and blood assessments give information to the clinicians about patients' general health over time, which allows them to assess the improvement in health for each patient. We have information over time on *height*, *weight*, *white blood cell (WBC) count*, *platelets*, *haemoglobin (Hb)* and *erythrocyte sedimentation rate (ESR)*. These covariates present deviations from normality and, therefore, it is reasonable to transform them prior to analysis. We use a log transformation for *WBC*, *haemoglobin* and *height* to reduce skewness. We employ a square root transformation for *weight* and *platelets*, as it is a variance stabilising transformation, while we choose a Box-Cox transformation for *ESR*. In summary, six continuous (transformed) covariates have been included in the final analysis, for which we assume a Normal distribution: *haemoglobin*, *platelets*, *weight*, *height*, *WBC* and *ESR*. Moreover, we include in the analysis a functional score, the CHAQ (Child Health Assessment Questionnaire) Score, which requires further modelling assumptions (see Subsection 3.4). Basic summaries of these variables in the original scale are given in Table 1in Supplementary Material.

2.3. Patients' symptoms and therapy data

Symptoms and therapy information were collected at the study entry and during follow-up. Since there are no reliable methods to accurately diagnose JDM, its diagnosis is made through a set of manifold examinations investigating the presence or absence of symptoms. These latter can be grouped into categories according to the type and/or location of symptoms: skin manifestations other than rash, rash distribution, joints, oedema, abdomen and other symptoms. Descriptive statistics for the symptoms are shown in Table 2 in Supplementary Material. The symptoms that seem to be more recurrent over time are *Gottron's papules* (40%), *nailfold changes* (30.4%), *rash* ($>$ 20%), *joints with limited range of motion* (17.5%) and *calcinosis* (15.7%). These statistics seem to be consistent with previous research [18, 1, 19, 17]. As mentioned previously, there are many missing data, especially for the symptoms, with the percentage of missing values varying from 2% to 24%.

Depending on different criteria such as disease severity, centre or physicians' personal judgements, a treatment (either a drug or a combination of drugs or one of these two associated with physiotherapy) is prescribed to the patient. The most often administered drugs are *methotrexate* (69.2%) and *oral steroids* (37.6%). See Table 3 in Supplementary Material. All covariates about patients' symptoms and therapy listed in Tables 2 and 3 in Supplementary Material were included in the model. Correlation between clinical covariates and the laboratory test results is investigated using logistic regression

and most symptoms seem correlated with at least one of the laboratory results. Also the treatments are correlated with laboratory results. For example, *oral steroids* seem to be strongly correlated with *WBC* and *hydroxychloroquine* with *weight*.

2.4. Patients' remission

We now describe the clinical outcome. We consider a binary response representing disease or remission state in children affected by JDM over time. The clinically inactive disease for JDM has been defined by the Paediatric Rheumatology International Trials Organisation (PRINTO) criteria. Lazarevic *et al.* [20] classify patients as achieving remission if at least three out of the four following criteria are satisfied:

- Creatinin Kinase (CPK) ≤ 150 (CPK is a serum muscle enzyme),
- Childhood Myositis Assessment Scale (CMAS) ≥ 48 (CMAS determines the muscle strength/endurance),
- Manual Muscle Testing (MMT) ≥ 78, (MMT is manual testing of eight muscle groups),
- Physician Global Assessment Visual Analogue Scale (PhyGloVAS) ≤ 0.2 (PhyGloVAS is the physician's global assessment of the patient's overall well-being on a 10 cm Visual Analogue Scale).

FIGURE 1 ABOUT HERE

As JDM is a chronic disease, the children are expected to alternate between disease and remission over time. Figure 1 illustrates these transitions over vistis for all patients in the dataset. It is noteworthy that the health state, disease or remission, is known only at the time of observation and there is no available information on health status between two observation times. Furthermore, the time interval between two visits and the number of visits vary among patients. This situation is typical in medical applications. For ease of visualization, in Figure 1 the x-axis represents the visit number instead of the time on a continuous scale (the maximum number of follow-up visits is 15). On the y-axis, for each child, we report the the observed health status at the attendance visit.

Table 1. Summary of state transitions.

	To	
From	Disease	Remission
Disease	176	148
Remission	49	236

Overall, there are 609 observed transitions, of which 148 from disease state to remission. On only 49 occasions, remission is followed by a relapse (see Table 1). The transitions from one state to another will be our actual response variable in the analysis. These transitions are governed by transition intensities which represents the instantaneous risk of moving from one state to the other. An initial estimate of the baseline transition rates which does not account for covariate information and patient heterogeneity is reported in Table 2. The results have been obtained using the R package `msm` [6], which implements maximum likelihood estimation of Multi-state Markov models.

2.5. Exploratory Data Analysis

To gain an initial understanding of the relationship between the response variable and the time-varying covariates we perform a two sample t-test for continuous covariates and Fisher test for binary covariates. Most predictors show significant association with the response. Results are shown in Tables 4 and 5 in Supplementary Material. Moreover, we fit a generalized linear mixed models using the `lme4` package [21] in R, specifying a logit link function. Because of optimization problems due to collinearity between the predictors, we fit two separate models, one including only time-varying continuous covariates and the other only binary covariates, but always a subject specific random intercept and time from diagnosis as predictors. The results are presented in Tables 6 and 7 in Supplementary Material. Considering the fixed effect estimates, we conclude that the most significant predictors of health status are *Time from diagnosis, CHAQ score, haemoglobin, platelets* and *ESR* among the continuous covariates and *abnormal respiration, combined skin rash, periorbital rash, pain on motion, hydrotherapy* and *oral steroids* among the binary ones.

Table 2. Initial Transition Intensities: model with covariates. Estimates obtained using the R package `msm`. In brackets we report 95% confidence intervals.

	To	
From	Disease	Remission
Disease	-1.70 (-2.27,-1.28)	1.70 (1.28, 2.27)
Remission	0.79 (0.55, 1.13)	-0.79 (-1.13,-0.55)

Table 3. Hazard ratios: estimates obtained using the R package `msm`. In brackets we report 95% confidence intervals.

	To	
	Disease $\Rightarrow$ Remission	Remission $\Rightarrow$ Disease
Balck Ethnicity	1.725 (0.436, 6.823)	2.950 (0.579, 15.028)
Other Ethnicity	8.445 (0.409, 174.464)	24.926 (1.242, 500.345)
Sex	0.177 (0.057, 0.546)	0.333 (0.086, 1.2960)
Age at Diagnosis	1.175 (1.027, 1.345)	1.095 (0.929, 1.290)
Rash	21.946 (3.155, 152.683)	41.022 (4.536, 371.004)
Weakness	4.576 (0.026, 802.237)	9.248 (0.055, 1553.945)
Myalgia	0.009 (0.000, 2.877)	0.003 (0.000, 0.976)
CHAQ Score	0.244 (0.113, 0.530)	0.317 (0.123, 0.816)
Haemoglobin	21.377 (0.100, 4564.495)	17.186 (0.019, 14943.320)
Platelets	1.392 (1.099, 1.765)	1.558 (1.176, 2.065)
ESR	0.288 (0.087, 0.951)	0.234 (0.061, 0.899)

Using the R package `msm` we fit a Markov multi-state model which includes covariate information. In this case the effect of a vector of explanatory variables on the transition intensity for individual i at time j is modelled using proportional intensities, similarly to a Cox proportional hazard model. See [6] for details. We include all the time homogeneous covariates and, due to optimization problems, only *CHAQ Score, haemoglobin, platelets* and *ESR* among the time-varying continuous covariates. Inclusion of time-varying binary covariates is challenging in this set-up and will be discussed more extensively in Section 3.4. Estimates of baseline transition intensities are reported in Table 8 in Supplementary Material, while Table 3 displays hazard ratios for each covariate on each transition with 95% confidence intervals. Note that some of the confidence intervals are extremely wide, for example, the ones for *rash* and *haemoglobin*. This initial analysis highlights some of the challenges associated with this application, in particular, the inclusion of time-inhomogeneous binary covariates and the presence of strong collinearity among predictors. Moreover, this initial exploratory analysis has been conducted excluding individuals with missing data.

3. Bayesian Markov Regression Model

In this section we present our modelling strategy. As we record a binary clinical outcome over time, we employ a two-state Markov Model to model disease progression. We assume that the observed remission/disease state is determined by the overall health status of each patient, which is unobserved and modelled through a subject specific linear growth curve. The advantage of including a latent health curve is that it becomes straightforward to account for time-changing covariates.

Moreover, in a Bayesian framework, missing data are easily accounted for, at least in principle. Finally, we employ spike and slab priors for the parameters that link the covariates to the latent health status function to perform variable selection.

3.1. A two-state Markov Model

Each individual patient is assessed to be in one of two possible states: 'disease' ($Y_t = 0$) or 'remission' ($Y_t = 1$) at each time of observation t. To model disease progression over time we assume a continuous time two-state Markov Model [22]. The Markov assumption implies that future states are independent of the past states given the current state [23], i.e.:

$$P(Y_{t+1} \mid Y_t, \ldots, Y_1, Y_0) = P(Y_{t+1} \mid Y_t).$$

The movement on the discrete state space is determined by a set of transition intensities denoted by λ_{rs}, which corresponds to the instantaneous risk of moving from state r to s

$$\lambda_{rs} = \lim_{\delta t \to 0} \frac{P(Y(t+\delta t) = s \mid Y(t) = r)}{\delta t}$$

with $s = 1 - r$ and $r \in \{0, 1\}$. The matrix of transition intensities for the two-state model is then given by

$$Q = \begin{pmatrix} -\lambda_{01} & \lambda_{01} \\ \lambda_{10} & -\lambda_{10} \end{pmatrix}.$$

The probability, $p_{rs}(\varepsilon)$, of a transition from state r to state s during the interval ε can be easily derived using the Chapman-Kolmogorov equations [23].

$$\begin{aligned} p_{01}(\varepsilon) &= \frac{\lambda_{01}}{\lambda_{01} + \lambda_{10}} \left(1 - \exp\{-(\lambda_{01} + \lambda_{10})\, \varepsilon\}\right) \\ p_{10}(\varepsilon) &= \frac{\lambda_{10}}{\lambda_{01} + \lambda_{10}} \left(1 - \exp\{-(\lambda_{01} + \lambda_{10})\, \varepsilon\}\right) \end{aligned}$$

These probability determine the transition probability matrix $\mathbf{P}(\varepsilon)$ over the time interval ε:

$$\mathbf{P}(\varepsilon) = \begin{bmatrix} p_{00}(\varepsilon) & p_{01}(\varepsilon) \\ p_{10}(\varepsilon) & p_{11}(\varepsilon) \end{bmatrix} = \begin{bmatrix} 1 - p_{01}(\varepsilon) & p_{01}(\varepsilon) \\ p_{10}(\varepsilon) & 1 - p_{10}(\varepsilon) \end{bmatrix}. \tag{1}$$

Let N be the total number of patients in the study and let $\mathbf{Y}_i = Y_{i1}, \ldots, Y_{in_i}$ denote the interval-censored vector of binary responses over time for individual i, with $Y_{ij} \in \{0, 1\}$. Let $\mathbf{t}_i = t_{i1}, \ldots, t_{in_i}$ denote the observation times for patient i, where n_i is the total number of successive observations for individual i. Therefore, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ represents the trajectory over time of individual i. In absence of explanatory variables, the likelihood is calculated from the transition matrix $\mathbf{P}$ in (1). The contribution to the likelihood of individual i with observations times $t_{i1} < \ldots < t_{in_i}$, conditioning on the first state, is:

$$L_i(\lambda_i) = p(\mathbf{Y}_i \mid \lambda_i, Y_{i1}) = P(Y_{in_i}, \ldots Y_{i2} | Y_{i1}, \lambda_i) = P(Y_{in_i} | Y_{i,n_i-1}, \lambda_i) \times \ldots \times P(Y_{i2} | Y_{i1}, \lambda_i)$$

with the restriction $P(Y_{i1} \mid \lambda_i) = 1$. As the transitions from remission to disease and inversely may be subject to variation between and within individuals, we assume a subject specific vector of transition intensities $\lambda_i = (\lambda_{i01}, \lambda_{i10})$. The transitions between states are the actual observations, with probabilities given in (1). We denote with $\varepsilon_{ij} = t_{i,j+1} - t_{ij}$ the time interval between two successive observation times, t_{ij} and $t_{i,j+1}$, and with $m_i = n_i - 1$ the number of observed transitions for individual i. Therefore, the likelihood contribution for individual i becomes:

$$L_i(\lambda_i) = \prod_{j=1}^{m_i} p_{iY_i(t_{ij})Y_i(t_{i,j+1})}(\varepsilon_{ij}; \lambda_{iY_i(t_{ij})Y_i(t_{i,j+1})}); \quad Y_i(t_{ij}) \in \{0, 1\}$$

where $p_{irs}(\varepsilon)$ is the patient specific transition probability from state r to s over the interval ε. Finally, the full likelihood of the model is the product of probabilities of transition between observed states over all individuals i and transitions j

$$L(\lambda_1, \ldots, \lambda_N) = \prod_{i=1}^{N} L(\lambda_i) = \prod_{i=1}^{N} \prod_{j=1}^{m_i} \{p_{iY_i(t_{ij})Y_i(t_{i,j+1})}(\varepsilon_{ij}; \lambda_{iY_i(t_{ij})Y_i(t_{i,j+1})})\}$$

as individual trajectories are assumed independent given the parameters in the model.

It is easy to incorporate in the model time homogeneous covariates (which do not change with time), such as sex or age at diagnosis, by specifying a regression model on the transition intensities:

$$\lambda_{irs} = \tilde{\lambda}_{irs} \exp\{\mathbf{x}_i \beta_{rs}\} \tag{2}$$

where $\mathbf{X}_i = x_{i1}, \ldots, x_{ig}$ denotes the vector of time invariant covariates for individual i, $\beta_{rs} = (\beta_{rs1}, \ldots, \beta_{rsg})$ is the vector of regression parameters of length g for the transition from r to s and $\tilde{\lambda}_{irs}$ is the individual baseline intensity for the transition from state r to state s, with $r \in \{0, 1\}$ and $s = 1 - r$.

3.2. Latent health curve

As we have seen in the previous section, time invariant covariates are easily incorporated in the model through a regression model on the transition intensities. We now present our strategy to include time-varying covariates. When the time-varying covariates are continuous and are observed at the same time as the response, it is common practice to include them in the regression model for λ_i by taking as predictor the mean of the covariate over the interval of observation. Dealing with categorical covariates that change over time is more complex. A possible solution is to model jointly the categorical covariates and the response by specifying a Markov model on the state space given by the possible realisations of both Y and $\mathbf{X}$. This approach is most feasible when the number of categorical covariates and of the corresponding number of levels is small, as this would limit the dimension of the implied state space and the computational cost. As the JDM dataset contains a large number of binary symptoms that change over time, following [24] and [25], we introduce a subject-specific health function that varies over time to which we link the time-varying predictors. The health state is latent or unobservable, and represents the *true* well-being of an individual. Although it is not measurable, the covariate data can be used as a proxy for it. This strategy improves the utilization of data from diverse sources and also different patient populations potentially leading to better characterization of disease progression. We consider time-varying continuous measurements and binary variables representing the presence of symptoms and the administration of treatments. To model the health state function, we employ a growth curve which we assume linear in time:

$$\theta_{ij} = \eta_{0i} + \eta_{1i} t_{ij} + \rho_i; \qquad i = 1, \ldots, N; j = 1, \ldots, n_i \tag{3}$$

where θ_{ij} is the health state for the individual i at the time t_{ij}, η_{0i} is the baseline health state for the individual i across all observation times, η_{1i} characterises the subject-specific change over time of θ and ρ_i is the random component associated with the health state function for individual i. We assume the ρ_i to be independent and identically distributed $\mathrm{N}(0, 1)$ random variables. We fix the variance of ρ_i to 1 for identifiability reasons. The latent health curve is then linked to the transition probabilities by including θ_{ij} in the regression for the baseline intensities in (2). As the health state changes over time, the approximate effect of the latent curve for individual i can be estimated by assuming that its value is constant over each interval of observations ϵ_{ij} [6, 26], leading to a piecewise constant model for Q. We set θ_{ij} equal to its mean $\bar{\theta}_{i\epsilon}$ over the interval of observation ϵ, obtaining:

$$\lambda_{irs}(\epsilon) \quad = \quad \tilde{\lambda}_{irs} \exp\{\mathbf{x}_i \beta_{rs} + \gamma_{rs} \bar{\theta}_{i\epsilon}\} \tag{4}$$

where $r = 0, 1; s = 1 - r$, and the $\tilde{\lambda}_{irs}$ are independent given the remaining parameters in the model.

3.3. A Nonparametric Random Effects Distribution

In this section we describe the choice of prior distribution for the parameters that govern the latent health curve (η_{0i}, η_{1i}, see (3)) and the baseline transition intensities $\widetilde{\lambda}_{irs}$ (see (4)). An obvious parametric choice would be a Gamma distribution for the intensities and a multivariate Normal for the health curve parameters. In our analysis, we have found a need to move beyond the traditional parametric assumptions for the random effect distribution, as there is known inter-individual heterogeneity that cannot be described in a simple parametric model. Moreover, the random-effects distribution needs to accommodate the heterogeneity in the population and to allow for outliers, clustering and overdispersion. A second important element of the model proposed in this paper is the use of a semiparametric population model. This heterogeneity between patients and within patients is a common feature of many biomedical data. We employ the Dirichlet Process (DP; [15]) to define a flexible nonparametric model for an unknown random-effects distribution. A DP defines a probability model on the space of probability distributions so that if a random measure $G \sim \mathrm{DP}(\alpha, G_0)$, then G is almost surely discrete. G_0 is the base measure, a distribution around which the DP is centred, while $\alpha \in \mathrm{R}^+$ denotes the precision parameter. A constructive definition of the DP, extremely useful in applications, is the stick breaking representation [27].

Due to its discreteness, G can be represented as the infinite sum of point masses:

$$G = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}$$

where δ_{ϕ_k} is the Dirac measure taking value 1 in correspondence of ϕ_k and 0 otherwise. The weights w_k, conditional on $w_h, h < k$, are generated by rescaled beta distributions:

$$\frac{w_k}{\prod_{h<k} w_h} \overset{iid}{\sim} \text{Beta}(1, \alpha)$$

while the point masses ϕ_k are independent and identically distributed samples from the base measure, independent of the weights.

Let $\phi_i = (\widetilde{\lambda}_{i01}, \widetilde{\lambda}_{i10}, \eta_{0i}, \eta_{1i})$ denote the random effect vector. We can now rewrite the model for the health state function and the transition intensities as follows:

$$\begin{aligned}
\lambda_{irs}(\epsilon) &= \tilde{\lambda}_{irs} \ \exp\{\mathbf{x}_i \, \beta_{rs} + \gamma_{rs} \bar{\theta}_{i\epsilon}\} \\
\theta_{ij} \mid \eta_{0i}, \eta_{1i} &\sim \text{N}(\eta_{0i} + \eta_{1i} t_{ij}, 1) \\
\phi_i \mid G &\overset{iid}{\sim} G \\
G &\sim \text{DP}(\alpha, G_0) \\
G_0(\phi_i) &= \text{Gamma}(\widetilde{\lambda}_{i01}; \alpha_{01}, d_{01}) \times \text{Gamma}(\widetilde{\lambda}_{i10}; \alpha_{10}, d_{10}) \times \text{N}(\eta_{0i}; \mu_0, v_0) \times \text{N}(\eta_{1i}; \mu_1, v_1)
\end{aligned}$$

Setting a DP prior on the parameter vectors ϕ_i implies a non-zero probability that two or more vectors are equal. This, in turn, implies that the DP imposes a clustering structure on the data so that the observations will be grouped together in $K \leq n$ clusters, each characterised by a specific distribution. The parameters w_k include the prior probabilities of belonging to each cluster and ϕ_k denotes the cluster-specific parameter vector. Patients are clustered together according to their health trajectory over time and their baseline transition intensities. The advantage of this strategy is that the number of components K is also learned from the data through the posterior distribution. Thus, the vectors of individual-level parameters $\phi_1, \ldots, \phi_n$ reduce to the vectors of unique values $\phi_1^*, \ldots, \phi_K^*$ assigned to the n observations.

3.4. Time-varying covariates

The time-varying covariates are linked directly to the health state function. We start our discussion with the case of binary covariates. Suppose we have p binary covariates and let $\mathbf{H}_{ij} = H_{ij1}, \ldots, H_{ijp}$ be the vector of latent binary time-varying covariates for individual i at time t_{ij}. We assume that

$$\begin{aligned}
H_{ijh} &\sim \text{Bernoulli}\ (P(H_{ijh})) \\
P(H_{ijh}) &= \Phi\left(a_h \, \theta_{ij} - b_h\right)
\end{aligned} \tag{5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution and θ_{ij} is the $i-$th individual health state at time j. Hence, the prediction model for H_{ijh} is described as a probit model, which has the computational advantage of being easily implemented in a Gibbs sampler [28]. The parameter a_h indicates how much more likely the symptom is to be present given an individual health state. It can be interpreted as the effect on the probability that the symptom occurs for a one unit increment in the health state function θ. $\Phi(-b)$ is the probability that the symptom is present in the population when the health state function is equal to 0. van den Hout *et al.* [24] develop a similar approach using longitudinal data to model stroke with cognition as a latent time-dependent risk factor. Assuming that the symptoms are independent given the health state function, the likelihood contribution of the symptoms is then given by:

$$P(\mathbf{H}|\theta, \mathbf{a}, \mathbf{b}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \prod_{h=1}^{p} P(H_{ijh} = 1|\theta_{ij}, a_h, b_h)^{H_{ijh}} \ \ (1 - P(H_{ijh} = 1|\ \theta_{ij}, a_h, b_h))^{1-H_{ijh}}.$$

The health state function θ_{ij} is linked also to the time-varying continuous covariates through their expectations:

$$Z_{ijl} \ \sim \ \text{N}\left\{(c_{0l} + c_{1l} \, \theta_{ij}), \ \tau_l\right\}, \qquad l = 1, \ldots, q. \tag{6}$$

where Z_{ijl} denotes the continuous covariate l for individual i at time t_{ij}, q is the number of time-varying continuous covariates, c_{0l} is the baseline value of the continuous covariate l over all individuals and c_{1l} is the effect on the value of

continuous covariate l of an increase of one unit in the health state function. Moreover we assume that the continuous covariates are independent among themselves and from the binary symptoms given the state of health.

One of the continuous covariates, the CHAQ Score, presents an excess of zeros as it represents a score and a simple parametric density would not be appropriate to capture its distribution. Therefore we model the square root of such covariate with a mixture of a point mass at zero and a Normal distribution. The choice of the square root is due to the fact that it is a well known variance stabilising transformation. Let S denote the variable representing the CHAQ Score. Then, we assume

$$\begin{aligned} S_{ij} &\sim \pi_{ij}\delta_0 + (1-\pi_{ij})\mathrm{N}(m_0 + m_1\theta_{ij}, \tau_S) \\ \text{logit}\, \pi_{ij} &= \psi_0 + \psi_1\theta_{ij} \end{aligned}$$

where δ_0 denotes a point mass at 0. This type of strategy is known in the literature as a Zero Inflated model [29].

Finally, the full covariate model is given by

$$P(\mathbf{Z}, \mathbf{H}, S|\theta, \mathbf{a}, \mathbf{b}, \mathbf{c_0}, \mathbf{c_1}, \tau, \mathbf{m}, \psi, \pi, \tau_S) \propto P(\mathbf{H}|\theta, \mathbf{a}, \mathbf{b})\; P(\mathbf{Z}|\theta, \mathbf{c_0}, \mathbf{c_1}, \tau) P(S \mid \theta, \pi, \mathbf{m}, \psi, \tau_S)$$

with $\mathbf{Z} = \{Z_{ijl}; l = 1, \ldots, q, i = 1, \ldots, N, j = 1, \ldots, n_i\}, \mathbf{H} = \{H_{ijl}; h = 1, \ldots, p, i = 1, \ldots, N, j = 1, \ldots, n_i\}, S = \{S_{ij}; i = 1, \ldots, N, j = 1, \ldots, n_i\}$.

3.5. Missing values

The dataset contains a substantial amount of missing covariate data especially among the time-varying covariates (refer to Tables 2 and 3 in Supplementary Material). However, the objective is to include as much information as possible. Only completely observed baseline variables have been incorporated in the model. Therefore only 6 covariates were considered: sex, ethnicity, age at diagnosis and the presence or absence of rash, weakness and myalgia in the medical history of patients. These three last variables have been already reported to have a potential role in the prognosis of the disease [17, 18, 19]. Time-varying covariates present missing rates in the range $[0.018, 0.240]$. As we have specified a probability model for the time-inhomogeneous covariates, missing values are automatically accounted for in a Bayesian framework and they are imputed within the MCMC algorithm from the posterior predictive distribution of the missing observations given the observed data. The time-varying covariates are a function of the health state function θ. We implicitly make the assumption of Missing At Random (MAR) so that the missing data mechanism is ignorable. We also assume that the missing data mechanism is independent from θ. These two assumptions are the weakest ones in order to draw correct Bayesian inference according to Rubin [30].

3.6. Hyperprior specification

To complete the model, we specify hyperpriors on the remaining parameters. These hyperpriors are chosen mainly for computational reasons and are, in general, uninformative. We assume:

- Model for baseline intensities: β_{rsj} are independent normally distributed random variables with mean 0 and variance equal to 1000, $j = 1, \ldots, g; r = 0, 1, s = 1 - r$. Moreover, γ_{01} and γ_{10} are also independent normally distributed random variables with mean 0 and variance equal to 1000.
- Random effect distribution: in G_0 we set $\alpha_{01} = \alpha_{10} = d_{01} = d_{10} = 1$, $v_0 = v_1 = 100$, $\mu_0 \sim \mathrm{N}(0, 1000)$ and $\mu_1 \sim \mathrm{N}(0, 1000)$. For the precision parameter α of the DP we assume a Uniform distribution on the interval $(0.3, 5)$ as this prior choice leads to more stable computations in JAGS. Since the implementation in JAGS of the DP is based on approximating the infinite mixture with a finite one, the prior choice for α together with using a mixture of 50 components implies an approximation error smaller than 10^{-8} (see [31]). Furthermore, the prior specification on α affects the prior distribution of the number K of clusters in the mixture. The conditional mean and variance of K given the precision parameter of the DP and the sample size N are [32, 33]:

$$\begin{aligned} \mathbb{E}(K \mid \alpha) &= \sum_i^N \frac{\alpha}{\alpha + i - 1} \\ \text{var}(K \mid \alpha) &= \sum_i^N \frac{\alpha(i-1)}{\alpha + i - 1} \end{aligned}$$

From this we can derive that the our prior on α implies that marginally $E(K) = 11$ and $\text{var}(K) = 8$.

- Model for time-varying binary covariates: in (5) we assume a_h and b_h, $h = 1, \ldots, p$, a priori independent, normally distributed random variables with mean 0 and variance 1000.
- Model for time-varying continuous covariates: in (6), we specify independent $\mathrm{N}(0, 1000)$ priors for c_{0l}, c_{1l} and Gamma(1,1) prior for $1/\tau_l$, $l = 1, \ldots, q$.
- Zero Inflated model: we assume independent $\mathrm{N}(0, 1000)$ priors for m_0, m_1, ψ_0, ψ_1 and a Gamma(1,1) prior distribution for $1/\tau_S$.

4. Variable selection using spike and slab priors

The main objective of this analysis is to ascertain a subset of symptoms and continuous markers which are most associated with the latent health function and hence have a strong effect on disease progression. To this end, we want to identify those binary and continuous variables for which the posterior distribution of a_h and c_{1l} is concentrated away from zero. We specify a spike and slab prior for the regression coefficients $\mathbf{a}$ in (5) and $\mathbf{c}_1$ in (6) to implement Stochastic Search Variable Selection (SSVS, [34]).

$$a_h|\omega_h \sim (1-\omega_h)\,\delta_0 + \omega_h\, N(0,\ \tau_{ah}) \tag{7}$$
$$c_{1l}|\omega_l \sim (1-\omega_l)\,\delta_0 + \omega_h\, N(0,\ \tau_{cl}) \tag{8}$$

where δ_0 is a point mass at 0 and ω_h is a latent variable taking only values 0 or 1. The τ_{ah} and τ_{cl} are set equal to 1000. The use of ω_h in a mixture of Normals model directs the choice of the prior distribution for a_h and determines whether the symptoms h is a good proxy for the unobserved health state. If $\omega_h = 0$, then $a_h = 0$ else $a_h|\omega_h \sim \mathrm{N}\ (0,\ \tau_{ah})$. A similar interpretation holds for ω_l. The shrinkage properties of SSVS are sensitive to the shape of the spike and the slab (see e.g. [35, 36]). The precision of the Normal component is usually set to be small enough to identify the most relevant variables in the model. Larger values of the slab (Normal component) variance allow large effects to take on arbitrarily large values and encouraging stronger penalisation of small nonzero effects. Larger values of this variance are therefore more suited to sparse underlying models. On the other hand, small values of the slab variance reflect the belief that there are few close to zero effects and therefore are more suited to nonsparse models. In the implementation of the model, we have set the slab variances τ_{ah} and τ_{cl} equal to 1000. The random variables ω_h and ω_l are assigned Bernoulli priors:

$$\omega_h \sim \text{Bernoulli}(p_{\omega h}), \qquad h = 1, \ldots, p \tag{9}$$
$$p_{\omega h} \sim \text{Beta}(0.1, 0.1) \tag{10}$$
$$\omega_l \sim \text{Bernoulli}(p_{\omega l}), \qquad l = 1, \ldots, q \tag{11}$$
$$p_{\omega l} \sim \text{Beta}(0.1, 0.1) \tag{12}$$

The ω_h and the ω_l are assumed all a priori independent as well as the $p_{\omega h}$ and $p_{\omega l}$. The specification of the Beta hyper-prior has been proposed by [37] to induce extra sparsity on the regression coefficients, encouraging those associated with the covariates having no effect on the response variable to shrink toward zero. Note that in the original formulation of [34] the inclusion variables, ω_j, are assigned Bernoulli prior distributions with a common probability of success $\widetilde{p}$. Usually a Uniform$[0, 1]$ is chosen as hyper-prior for $\widetilde{p}$, but other options, e.g. a more informative Beta distribution, could also be used. In order to test the robustness of the Stochastic Search Variable Selection, we have conducted a sensitivity analysis using different values for $\tau_{ah}, \tau_{cl} \in \{\ 0.01,\ 0.1\}$ and different prior hyper-parameters for $p_{\omega l}$ and $p_{\omega h}$. An alternative and easy to implement strategy is the one proposed by Kuo and Mallick [38]. This involves introducing for each covariate an indicator variable $\omega_j, j = 1, 2, \ldots, p + q$ taking values in $\{0, 1\}$.

$$\widetilde{a}_h \mid \omega_h = \omega_h a_h, \qquad h = 1, \ldots, p$$
$$\widetilde{c}_{1l} \mid \omega_l = \omega_l c_{1l}, \qquad l = 1, \ldots, q$$
$$\omega_j \sim \text{Bernoulli}(\widetilde{p}), \qquad j = 1, \ldots, q + p$$

When $\omega_j = 1$, the j-th predictor influences the health function. When $\omega_j = 0$, then the distribution of the j-th predictor does not depend on θ_t. In the original formulation $\widetilde{p}$ is set equal to 0.5. We could extend this approach by specifying a Beta hyper-prior on $\widetilde{p}$ to induce sparsity. See [39] for a review of Bayesian variable selection methods. In what follows, covariates for which the mean of the posterior distribution of their coefficients ω_h and ω_l is higher than 0.5 are considered most relevant to explain disease progression.

We conclude this section noting that a different, but still effective alternative, to identify important predictors consists of specifying as prior for a_h and c_{1l} a *local scale mixture* of Normal distributions, such as the horseshoe [40] and the hyper-lasso [41] prior. This approach connects continuous prior distributions for regression parameters to models selection and involve intentional bias of the estimates to stabilise posterior inference. See [42] for a review.

5. Simulated Examples

We generate data on 100 subjects from a continuous time Markov Chain with two states, representing health (represented by 1) and disease status (represented by 0). We consider two groups of 50 individuals each. We fix $\eta_{0i} = 0$ in (3) and assume $\eta_{1i} = 1$ for the first group and $\eta_{1i} = -1$ for the second one. We generate observation times for each individual on the time interval $(0, 10)$ from a Normal distribution with mean 1, standard deviation 1, left truncated at 1/6 which corresponds to the minimum time distance between observations. The average number of observations per subject is equal to 8. We then generate the health function for each individual at the corresponding observation times. We simulate realisations from a continuous-time Markov process up to time $T = 10$ using the function `sim.msm` in the R package `msm` using the health function as time-inhomogeneous covariate. We fix the transition intensity matrix of the Markov process equal to

$$\mathbf{Q} = \begin{pmatrix} -0.13 & 0.13 \\ 0.25 & -0.25 \end{pmatrix}$$

The effect of time-dependent variables on the transition intensities can be modelled in the R package `msm` assuming that it is constant in between the observation times. In our simulations we set $\gamma_{01} = 0.08$ and $\gamma_{10} = 0.05$. The simulation parameters were based on a real data example described in [43]. Finally we generate two binary and two continuous time-varying predictor variables by setting in (5) and (6)

$$\begin{array}{lll} a_1 = 0.07, & b_1 = 0.2 & \\ a_2 = -0.08, & b_2 = 0 & \\ c_{01} = 2, & c_{11} = 3, & \tau_1 = 0.5 \\ c_{02} = 0, & c_{12} = 1 & \tau_2 = 1 \end{array}$$

We fit the Bayesian semiparametric model described in Section 3 to the simulated data. The posterior distribution of the number of cluster K has mode in 2 (see Figure 3 in Supplementary Material), which corresponds to the true simulation scenario. We summarise the clustering output by reporting the clustering that minimizes the posterior expectation of Binder's loss as described by [44] and implemented in the R package `mclust`. Our model is able to perfectly recover the original individual allocation. Moreover, in Figure 2 we display the predictive distribution of η_1 for a new hypothetical patient, which is bimodal with the two modes centred around -1 and 1, respectively. Moreover, the model is able to capture correctly the relationship between time-varying covariates and health function, as the posterior distribution of the regression coefficients linking the covariates to the health function are centred around the true values used to simulate the data. See Figure 3. Finally Figure 4 in Supplementary Material shows the posterior estimates of the transition probabilities $p_{rs}(\epsilon_{ij})$, with $\epsilon_{ij} = t_{i,j+1} - t_{ij}$, of the observed transitions for two randomly selected subjects.

To asses the ability of the model to select important covariates we generate data from a continuous time Markov chain, with eight states. Each state is denoted by (i, j, k) where $i, j, k \in \{0, 1\}$. The true model has the following transition intensity matrix:

$$\mathbf{Q} = \begin{pmatrix} -0.48 & 0.01 & 0.01 & 0.05 & 0.01 & 0.1 & 0.1 & 0.2 \\ 0.20 & -0.97 & 0.01 & 0.05 & 0.01 & 0.2 & 0.2 & 0.3 \\ 0.20 & 0.01 & -0.87 & 0.05 & 0.01 & 0.1 & 0.2 & 0.3 \\ 0.10 & 0.05 & 0.05 & -1.01 & 0.01 & 0.2 & 0.2 & 0.4 \\ 0.10 & 0.01 & 0.01 & 0.05 & -0.87 & 0.2 & 0.2 & 0.3 \\ 0.10 & 0.05 & 0.05 & 0.01 & 0.05 & -0.76 & 0.2 & 0.3 \\ 0.10 & 0.05 & 0.05 & 0.01 & 0.05 & 0.2 & -0.76 & 0.3 \\ 0.01 & 0.01 & 0.01 & 0.10 & 0.05 & 0.2 & 0.2 & -0.58 \end{pmatrix}$$

The columns (and rows) correspond to the following order of states:
$(0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1)$. The above $\mathbf{Q}$ implies that transition where two or three components are equal to one are favoured. We simulate 100 trajectories over the time interval (0,15) using the function `sim.msm` in the R package `msm`. We then introduce interval censoring by simulating, for each individual Markov chain, observation times from a Normal distribution with mean 2, standard deviation 1, truncated at 1/12 which corresponds to the minimum time distance between observations. To fit the Bayesian semiparametric model we use the first component of each state as response variable and the remaining components as time varying binary covariates. We

refer to these latter two variable as V_1 and V_2, which are obviously associated to the response variable given the data simulation process. Finally we generate three independent binary time-varying predictor variables from a Bernoulli with probability of success 0.5, which are also independent of the response. We perform variable selection using model (8) and specifying a Beta(1,1) prior on $p_{\omega h}$ and $p_{\omega l}$, since we have few predictors. We obtain a posterior probability of inclusion greater than 0.5 for V_1 and V_2 (approximately 0.7), while for the remaining covariates the posterior probability of inclusion is approximately 0.3.

6. Results

We ran the MCMC algorithm for 50000 iterations, discarding 20000 samples as burn-in and thinning every 5 iterations. We used a truncation of 50 clusters for the DP. In Figure 4 we plot the posterior distribution of the precision parameter α of the DP (posterior mean 1.34, posterior standard deviation 0.63) and the posterior distribution of the number of clusters for the general model. The analysis indicates that there are 6 main clusters among the patients, each characterised by a specific health function trajectory over time and associated transition intensities. In Figure 5 we show the health function $\theta(t)$ for six randomly chosen patients. It is important to notice the different patterns of the function, mainly captured by the slope.The analysis of the slope coefficients suggests that a negative slope corresponds to an improvement of the patients condition over time. Hence the slopes of θ_t for sicker patients are generally less steep, although they seem to ameliorate over time. On the other hand a sharp decline in the θ_t function reflects a faster recovery. The latent function θ_t seems to be able to satisfactorily characterise the underlying state of health of patients despite a few discrepancies when comparing to the observed states. This could be due to the linear assumption on the health curve, which can be too simplistic to represent some patients' progression.

FIGURE 2 ABOUT HERE

FIGURE 3 ABOUT HERE

FIGURE 4 ABOUT HERE

We now consider the effect of the time-homogeneous covariates on the transition intensities in (4), including the coefficient of the latent health function, γ_{rs}. Most of the fixed effect covariates appear to be important determinants of disease progression as their posterior distributions are centred away from zero. A weaker effect is noted for *ethnicity*. In Table 4 we report the posterior mean (posterior standard deviation) of the regression coefficients of the time-homogeneous covariates. A positive coefficient of a covariate implies a higher risk of transition. See also Figures 1 and 2 in Supplementary Material. Although patients health status tends to ameliorate with time, JDM is a chronic disease and children move from disease to remission and vice-versa over time. In general, transitions from remission to disease become less frequent as the child gets older. This is reflected in the estimates of the regression coefficients of the health function. Moreover, only some patients do really reach true remission, long lasting and off drugs. Current research shows ongoing disease well into adult life. For example, Sanner *et al.* [45] show that after 16.8 years after symptom onset 51-73% of JDM patients have active disease. Sanner *et al.* [46] report that majority of the patients in the study have cumulative organ damage during disease course (median 16.8 years) while Schwartz *et al.* [47] find signs of impaired cardiac systolic function with reduced long-axis strain in JDM patients seen 16.8 years after disease onset.

Figure 6 displays posterior inference for all parameters $\mathbf{a} = (a_1, \ldots, a_p)$ and highlights symptoms and therapies most associated with the latent health function and, therefore, with JDM. All the binary covariates, with exception of *Cyclophosphamide*, present a 95% credible interval which is not centred around 0 (see Figure 6), with most of them centred around positive values. This implies that the probability of symptom manifestation drops for a one unit decrease in the health state function or θ_t, thus suggesting that a decrease in the health state function (i.e. improvement in health) for an individual corresponds to a less likely manifestation of the symptoms. Moreover, only the coefficient for *Methotrexate* is centred around negative values. This can be interpreted as a better health condition being associated on being on Methotrexate therapy. In Figure 7 we plot the posterior distribution of the coefficients $\mathbf{c_1} = (c_{11}, \ldots, c_{1q})$, which represent

Table 4. Posterior mean (standard deviation) of the coefficients in the regression model for the transition intensities.

	Disease $\rightarrow$ **Remission**	**Remission** $\rightarrow$ Disease
Health Function	-24.64 (2.63)	-23.68 (2.62)
Black Ethnicity	14.92 (8.07)	15.79 (8.07)
Other Ethnicity	-12.14 (10.98)	-12.28 (11.09)
Sex	-26.56 (7.11)	-26.13 (7.15)
Age at Diagnosis	1.58 (0.85)	1.52 (0.85)
Rash	-25.84 (7.76)	-26.13 (7.72)
Weakness	10.23 (5.64)	11.66 (5.69)
Myalgia	26.44 (4.77)	26.25 (4.75)

the link between the continuous covariates and θ_t. They all present a posterior distribution centred away from zero. In particular, *Height* and *Weight* have negative coefficients. This is consistent with what is known, as these covariates can be thought as a proxy for growth and patients tend to get better as they grow older.

FIGURE 5 ABOUT HERE

Finally, in Figure 8 we show the regression coefficient of the health function in the Zero Inflated model for the CHAQ Score. In the left panel of Figure 8 we plot the posterior distribution of ψ_1 which corresponds to the logistic regression for the probability of the zero outcome, while in the right panel we show the posterior distribution of m_1 which links the mean of the continuous component to the health function. Both distributions are centred around positive values, implying that an improvement in health leads to a lower probability of a zero score.

To assess predictive performance, we fit the model leaving out 10 randomly selected patients among those with more than one transition. For this subjects we only include covariate information and only the initial state at time zero. As such we are left with 48 transitions to predict. We obtain posterior predictive probabilities of obtaining the observed outcome greater than 0.5 in 75% of the cases. Note that these estimates are obtained averaging over all the possible (simulated) trajectories for each patients.

FIGURE 6 ABOUT HERE

To highlight the most important determinants of disease progression, the Stochastic Search Variable Selection (SSVS) described in (8)-(12) is performed on the 36 potential binary predictors and the 6 continuous covariates. Using a Beta(0.1,0.1) prior for the probability of inclusion, we identify 22 binary covariates and five continuous markers with a posterior probability of inclusion greater than 0.5. The robustness of the variable selection procedure is tested by fitting the original version of the SSVS with a uniform prior on $\widetilde{p}$ and the Kuo-Mallick approach with $\widetilde{p}$ equal to 0.5. Results are reported in Table 5 for binary time-varying covariates and in Table 6 for continuous time-varying covariates. The results of the different methods are consistent with the SSVS based on a Beta$(0.1, 0.1)$ hyper-prior and the Kuo-Mallick method leading to more shrinkage of the inclusion probabilities. The results of the variable selection are consistent with existing clinical literature on JDM.

7. Conclusions

In this work, we have proposed a semiparametric model to describe the progression of JDM, a paediatric chronic disease, based on the Dirichlet process prior. The main modelling strategy includes a two-state Markov model and the specification of an underlying health state function which captures the patient's well being through a linear growth curve model. Time-varying covariates are linked directly to the health state function. This is an important feature of the model as it allows us to incorporate a large number of binary and continuous time-inhomogeneous covariates, which are often correlated. The

Table 5. Posterior probabilities of inclusion of the binary time-varying covariates. The first column corresponds to the SSVS described in (8)-(12) with a Beta(0.1,0.1) prior, the second column corresponds to the original version of the SSVS, while the third presents the results obtained employing the Kuo-Mallick approach.

	Beta(0.1,0.1)	Original SSVS	K-M
Abnomrmal Respisration	0.91	0.67	1.00
Gottron's Papules	0.92	0.67	1.00
Ulceration	0.92	0.67	1.00
Lipoatrophy	0.92	0.67	1.00
Oedema	0.92	0.67	1.00
Nailfold Changes	0.91	0.67	1.00
Calcinosis	0.09	0.36	0.09
Combined Skin Rash	0.91	0.67	1.00
Periorbital Rash	0.92	0.67	1.00
Periungual Rash	0.92	0.67	1.00
Trunk Rash	0.92	0.67	1.00
Small Joints Rash	0.92	0.66	1.00
Large Joints Rash	0.92	0.67	1.00
Arthritis	0.91	0.66	1.00
Pain On Motion	0.92	0.67	1.00
Joints with Limited ROM	0.92	0.67	1.00
Contractures	0.91	0.66	1.00
Periorbitial/Facial Oedema	0.92	0.66	1.00
Limb Oedema	0.91	0.67	1.00
Trunk Oedema	0.92	0.67	1.00
Abdominal Masses	0.09	0.34	0.00
Abdomen Tenderness	0.08	0.34	0.00
Hepatomegaly	0.21	0.63	0.22
Splenomegaly	0.10	0.34	0.01
Eyes	0.08	0.33	0.00
Other	0.08	0.33	0.01
Physiotherapy Dry Land	0.32	0.34	0.14
Physiotherapy Hydrotherapy	0.08	0.34	0.02
Oral Steroids	0.92	0.67	1.00
Intravenous Steroids	0.92	0.66	1.00
Methotrexate	0.08	0.33	0.00
Cyclosporin	0.09	0.33	0.01
Azathioprine	0.08	0.33	0.00
Cyclophosphamide	0.91	0.67	1.00
Hydroxychloroquine	0.09	0.34	0.00
Intravenous Immunoglobulin	0.10	0.34	0.01

Table 6. Posterior probabilities of inclusion of the continuous time-varying covariates. The first column corresponds to the SSVS described in (8)-(12) with a Beta(0.1,0.1) prior, the second column corresponds to the original version of the SSVS, while the third presents the results obtained employing the Kuo-Mallick approach.

	Beta(0.1,0.1)	Original SSVS	K-M
Hb	0.08	0.36	0.00
WBC	0.89	0.66	1.00
Platelets	0.71	0.50	0.18
ESR	0.91	0.46	0.80
Height	0.92	0.67	0.00
Weight	0.92	0.66	1.00

introduction of a latent health function is reminiscent of techniques used in Item Response Theory (IRT), which has been successufully applied in medical research. For example, Ueckert *et al.* [48] combine pharmacometric modelling and IRT to model cognition in clinical trials in Alzheimers Disease patients, while Gottpati *et al* [49] apply IRT methodology to integrate different versions of the main clinical endpoints used in Parkinsons disease studies into one unique framework, by mapping them to the same underlying latent variable(s). The same strategy can be extended to incorporate data from different sources. The model performs data-driven clustering of the patients according to their health profile over time and their baseline risk to move between health and disease states, in this way accounting for patients heterogeneity/similarity.

The data present a large amount of missing values. In absence of further information, we assume that the missing values for the covariates are Missing At Random (MAR) [30] and by specifying a model on the covariates we were able to deal with the missing observations in a Bayesian framework by imputing them in the MCMC algorithm. We were able to perform variable selection by employing a Stochastic Search Variable Selection approach which yields results consistent with current clinical understanding of JDM. Further advantages of the modelling strategy described in the paper include interpretability of the parameters, ability to cope with censoring, ease of computation and the potential to include more complex set-ups. Moreover, a more complex modelling of the health function could better capture relapses and intra-individual variability, leading to a more flexible strategy

The main drawback of our approach consists in defining the response variables as a binary outcome (remission/disease) by discretising the measurements of four clinically relevant continuous outcomes as described by the PRINTO criteria. For example, [50] compare different tools to assess skin disease in patients with Juvenile Dermatomyositis and highlight the need for a better one. However, PRINTO criteria are commonly used in clinical practice to define disease activity and as such it is clinically relevant to treat them as the main outcome. Current research involves modelling directly the longitudinal outcomes on which such criteria are based, to gain a better understanding of the JDM and of the clinical markers associated with disease activity and capable of predicting its evolution.

References

1. Symmons DPM, Sills JA, Davis SM. The incidence of juvenile dermatomyositis: Results from a nation-wide study. *Rheumatology* 1995; **34**(8):732–736, doi:10.1093/rheumatology/34.8.732. URL http://rheumatology.oxfordjournals.org/content/34/8/732.abstract.
2. Mendez EP, Lipton R, Ramsey-Goldman R, Roettcher P, Bowyer S, Dyer A, Pachman LM, the NIAMS Juvenile DM Registry Physician Referral Group F. US incidence of juvenile dermatomyositis, 19951998: Results from the national institute of arthritis and musculoskeletal and skin diseases registry. *Arthritis Care and Research* 2003; **49**(3):300–305, doi:10.1002/art.11122. URL http://dx.doi.org/10.1002/art.11122.
3. Batthish M, Feldman B. Juvenile dermatomyositis. *Current Rheumatology Reports* 2011; **13**(3):216–224, doi:10.1007/s11926-011-0167-9. URL http://dx.doi.org/10.1007/s11926-011-0167-9.
4. Martin N, Krol P, Smith S, Murray K, Pilkington CA, Davidson JE, Wedderburn LR. A national registry for juvenile dermatomyositis and other paediatric idiopathic inflammatory myopathies: 10 years' experience; the juvenile dermatomyositis national (UK and Ireland) cohort biomarker study and repository for idiopathic inflammatory myopathies. *Rheumatology* 2011; **50**(1):137–145, doi:10.1093/rheumatology/keq261. URL http://rheumatology.oxfordjournals.org/content/50/1/137.abstract.
5. Tansley S, McHugh N, Wedderburn L. Adult and juvenile dermatomyositis: are the distinct clinical features explained by our current understanding of serological subgroups and pathogenic mechanisms? *Arthritis Research and Therapy* 2013; **15**(2):211, doi:10.1186/ar4198. URL http://arthritis-research.com/content/15/2/211.
6. Jackson CH. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software* 2011; **38**(8):1–29. URL http://www.jstatsoft.org/v38/i08/.
7. Van Den Hout A. *Multi-State Survival Models for Interval-Censored Data*. Chapman & Hall Crc, 2014.

8. Shen S, Han SX, Petousis P, Weiss RE, Meng F, Bui AA, Hsu W. A Bayesian model for estimating multi-state disease progression. *Computers in Biology and Medicine* 2017; **81**:111–120.
9. van den Hout A, Matthews FE. Multi-state analysis of cognitive ability data: A piecewise-constant model and a weibull model. *Statistics in medicine* 2008; **27**(26):5440–5455.
10. Fahrmeir L, Klinger A. A nonparametric multiplicative hazard model for event history analysis. *Biometrika* 1998; :581–592.
11. Aalen OO, Fosen J, Weedon-Fekjær H, Borgan Ø, Husebye E. Dynamic analysis of multivariate failure time data. *Biometrics* 2004; **60**(3):764–773.
12. Kneib T, Hennerfeind A. Bayesian semi parametric multi-state models. *Statistical Modelling* 2008; **8**(2):169–198.
13. Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics* 1974; :1152–1174.
14. Lo A. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics* 1984; **12**:351–357.
15. Ferguson TS. A Bayesian analysis of some nonparametric problems. *The annals of statistics* 1973; :209–230.
16. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.
17. Feldman BM, Rider LG, Reed AM, Pachman LM. Juvenile dermatomyositis and other idiopathic inflammatory myopathies of childhood. *Lancet* Jun 2008; **371**(9631):2201–2212, doi:10.1016/s0140-6736(08)60955-1. URL http://dx.doi.org/10.1016/s0140-6736(08)60955-1.
18. McCann LJ, Juggins AD, Maillard SM, Wedderburn LR, Davidson JE, Murray KJ, Pilkington CA, on behalf of the Juvenile Dermatomyositis Research Group. The juvenile dermatomyositis national registry and repository (UK and Ireland) – clinical characteristics of children recruited within the first 5yr. *Rheumatology* 2006; **45**(10):1255–1260, doi:10.1093/rheumatology/kel099. URL http://rheumatology.oxfordjournals.org/content/45/10/1255.abstract.
19. Gowdie PJ, Allen RC, Kornberg AJ, Akikusa JD. Clinical features and disease course of patients with juvenile dermatomyositis. *International Journal of Rheumatic Diseases* 2013; **16**(5):561–567, doi:10.1111/1756-185X.12107. URL http://dx.doi.org/10.1111/1756-185X.12107.
20. Lazarevic D, Pistorio A, Palmisani E, Miettunen P, Ravelli A, Pilkington C, Wulffraat NM, Malattia C, Garay SM, Hofer M, *et al.*. The PRINTO criteria for clinically inactive disease in juvenile dermatomyositis. *Annals of the Rheumatic Diseases* 2013; **72**(5):686–693, doi:10.1136/annrheumdis-2012-201483. URL http://ard.bmj.com/content/72/5/686.abstract.
21. Bates D, Sarkar D, Bates MD, Matrix L. The lme4 package. *R package version* 2007; **2**(1):74.
22. Cook RJ. A mixed model for two-state Markov processes under panel observation. *Biometrics* 1999; **55**(3):915–920, doi:10.1111/j.0006-341X.1999.00915.x. URL http://dx.doi.org/10.1111/j.0006-341X.1999.00915.x.
23. Cox D, Miller H. *The theory of stochastic processes*. Wiley publications in statistics, Wiley, 1965. URL https://books.google.co.uk/books?id=5rpEAAAAIAAJ.
24. van den Hout A, Fox JP, Klein Entink RH. Bayesian inference for an illness-death model for stroke with cognition as a latent time-dependent risk factor. *Statistical Methods in Medical Research* 2011; doi:10.1177/0962280211426359. URL http://smm.sagepub.com/content/early/2014/03/19/0962280211426359.abstract.
25. Dunson DB. Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association* 2003; **98**(463):pp. 555–563. URL http://www.jstor.org/stable/30045281.
26. Kapetanakis V, Matthews FE, van den Hout A. A semi-Markov model for stroke with piecewise-constant hazards in the presence of left, right and intervalcensoring. *Statist. Med.* 2012; :n/adoi:10.1002/sim.5534. URL http://dx.doi.org/10.1002/sim.5534.
27. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.
28. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 1993; **88**(422):669–679. URL http://www.jstor.org/stable/2290350.
29. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**(1):1–14. URL http://www.jstor.org/stable/1269547.
30. Rubin DB. Inference and missing data. *Biometrika* Dec 1976; **63**(3):581–592, doi:10.1093/biomet/63.3.581. URL http://dx.doi.org/10.1093/biomet/63.3.581.
31. Ohlssen DI, Sharples LD, Spiegelhalter DJ. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine* 2007; **26**(9):2088–2112, doi:10.1002/sim.2666. URL http://dx.doi.org/10.1002/sim.2666.
32. Liu JS. Nonparametric hierarchical bayes via sequential imputations. *The Annals of Statistics* 1996; :911–930.
33. Jara A, García-Zattera MJ, Lesaffre E. A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics & Data Analysis* 2007; **51**(11):5402–5415.
34. George EI, McCulloch RE. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 1993; **88**(423):881–889, doi:10.1080/01621459.1993.10476353. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476353.
35. Rockova V, Lesaffre E, Luime J, Löwenberg B. Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in medicine* 2012; **31**(11-12):1221–1237.
36. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in medicine* 2016; **35**(7):1159–1177.
37. Lucas J, Carvalho C, Wang Q, Bild A, Nevins J, West M. Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics* 2006; **1**:0–1.
38. Kuo L, Mallick B. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B* 1998; :65–81.
39. O'Hara RB, Sillanpää MJ, *et al.*. A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis* 2009; **4**(1):85–117.
40. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika* 2010; **97**(2):465–480.
41. Griffin JE, Brown PJ, *et al.*. Structuring shrinkage: some correlated priors for regression. *Biometrika* 2012; **99**(2):481.
42. Hahn PR, Carvalho CM. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* 2015; **110**(509):435–448.
43. Van Den Hout A. *Multi-state Survival Models for Interval-censored Data*. CRC Press, 2016.
44. Fritsch A, Ickstadt K, *et al.*. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian analysis* 2009; **4**(2):367–391.
45. Sanner H, Sjaastad I, Flatø B. Disease activity and prognostic factors in juvenile dermatomyositis: a long-term follow-up study applying the paediatric rheumatology international trials organization criteria for inactive disease and the myositis disease activity assessment tool. *Rheumatology* 2014; :keu146.
46. Sanner H, Gran JT, Sjaastad I, Flatø B. Cumulative organ damage and prognostic factors in juvenile dermatomyositis: a cross-sectional study median 16.8 years after symptom onset. *Rheumatology* 2009; **48**(12):1541–1547.

47. Schwartz T, Sanner H, Gjesdal O, Flatø B, Sjaastad I. In juvenile dermatomyositis, cardiac systolic dysfunction is present after long-term follow-up and is predicted by sustained early skin activity. *Annals of the rheumatic diseases* 2013; :annrheumdis–2013.
48. Ueckert S, Plan EL, Ito K, Karlsson MO, Corrigan B, Hooker AC, Initiative ADN, *et al.*. Improved utilization of adas-cog assessment data through item response theory based pharmacometric modeling. *Pharmaceutical research* 2014; **31**(8):2152–2165.
49. Gottipati G, Berges AC, Yang S, Chen C, Karlsson MO, Plan EL. Item response theory modelling to leverage data from historical parkinsons disease trials while integrating data from a newer version of the clinical endpoint ; .
50. Campanilho-Marques R, Almeida B, Deakin C, Arnold K, Gallot N, de Iorio M, Nistala K, Pilkington CA, Wedderburn LR, on behalf of the Juvenile Dermatomyositis Research Group (JDRG). Comparison of the utility and validity of three scoring tools to measure skin involvement in patients with juvenile dermatomyositis. *Arthritis Care & Research* 2016; :n/a–n/adoi:10.1002/acr.22867. URL http://dx.doi.org/10.1002/acr.22867.

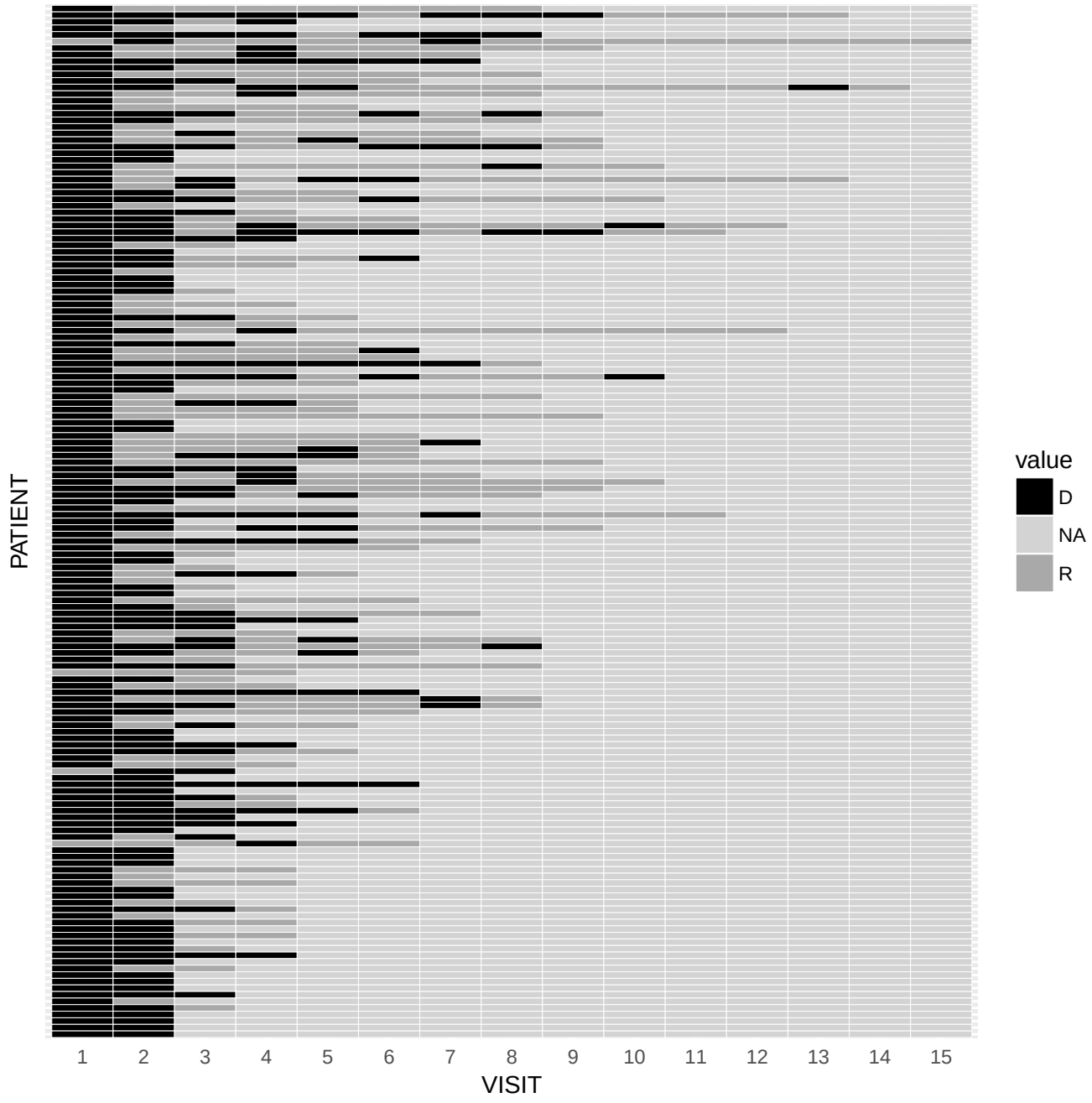

Figure 1. Transitions over visits for all patients affected by Juvenile Dermatomyositis.The x-axis represents the number of the follow-up visit, while the y-axis shows the health status observed at the attendance visit. Black denotes disease state (D), dark grey indicates remission (R), while light grey indicates that the patient has not been observed at that visit (NA).

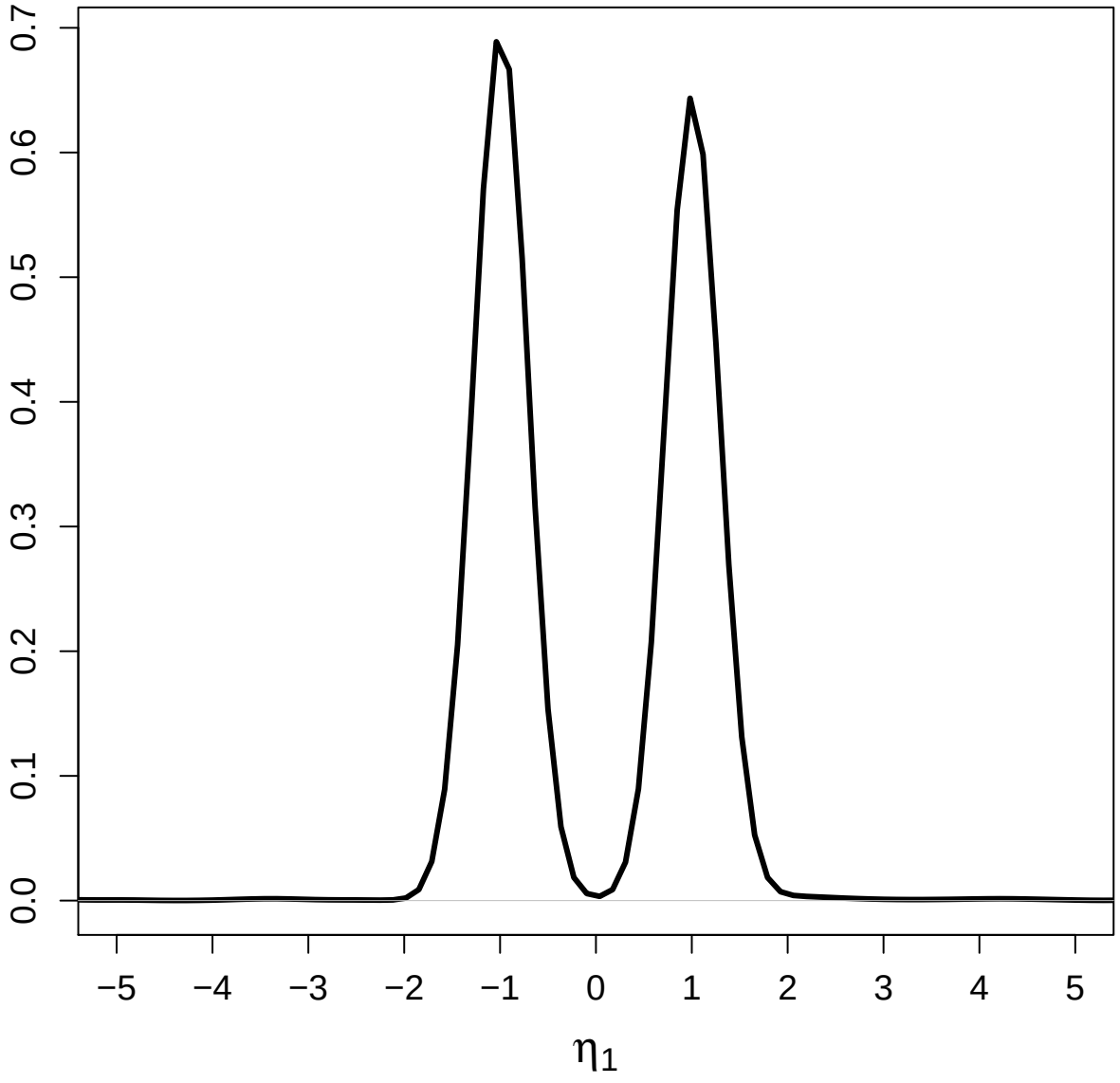

Figure 2. Simulated data: marginal predictive distribution of the parameter η_1 of the health function.

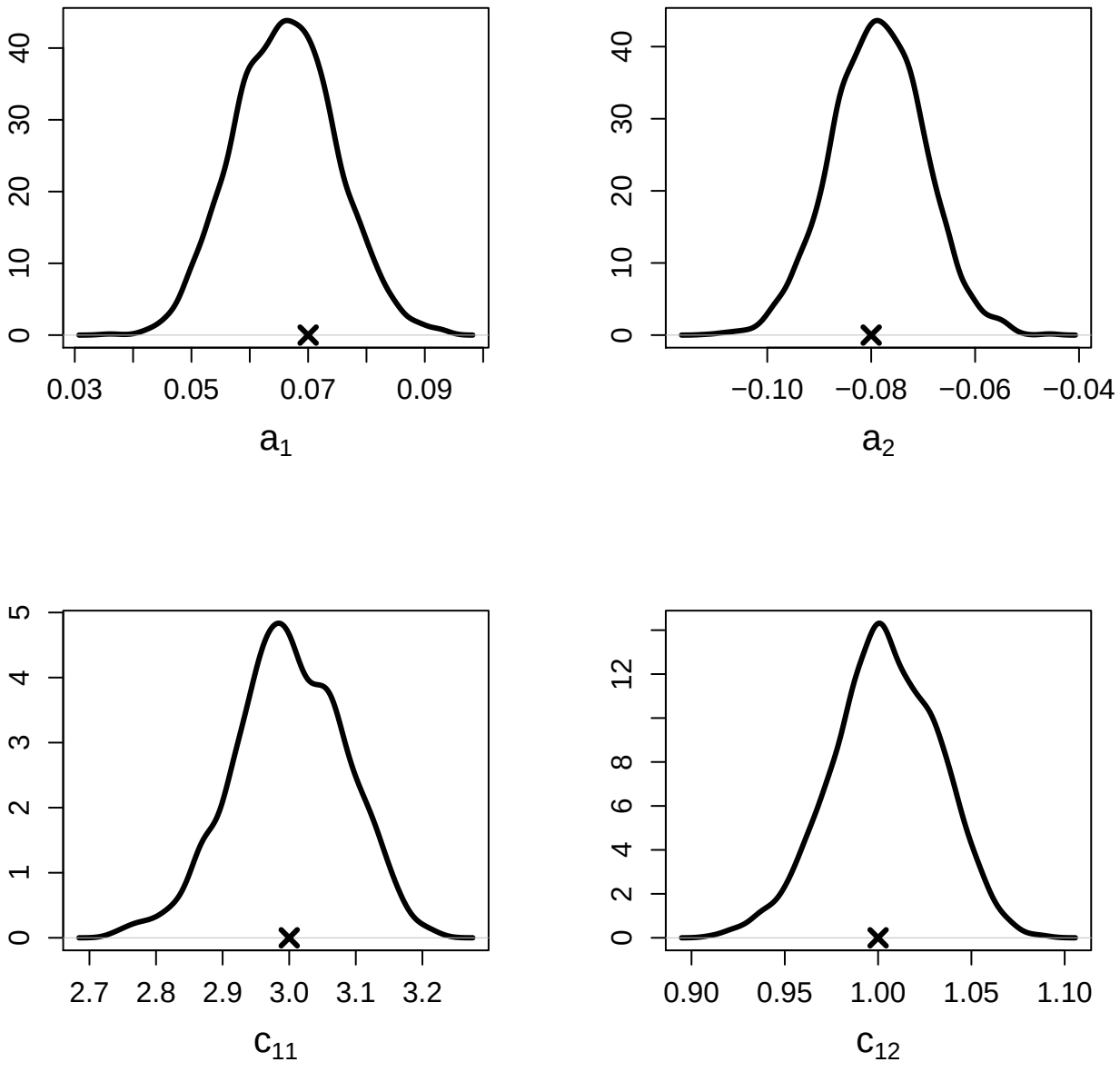

Figure 3. Simulated data: posterior distribution of the regression coefficients a_1, a_2, c_{11}, c_{12}, linking the time-varying covariates to the health function. The crosses indicate the true value used in the simulations.

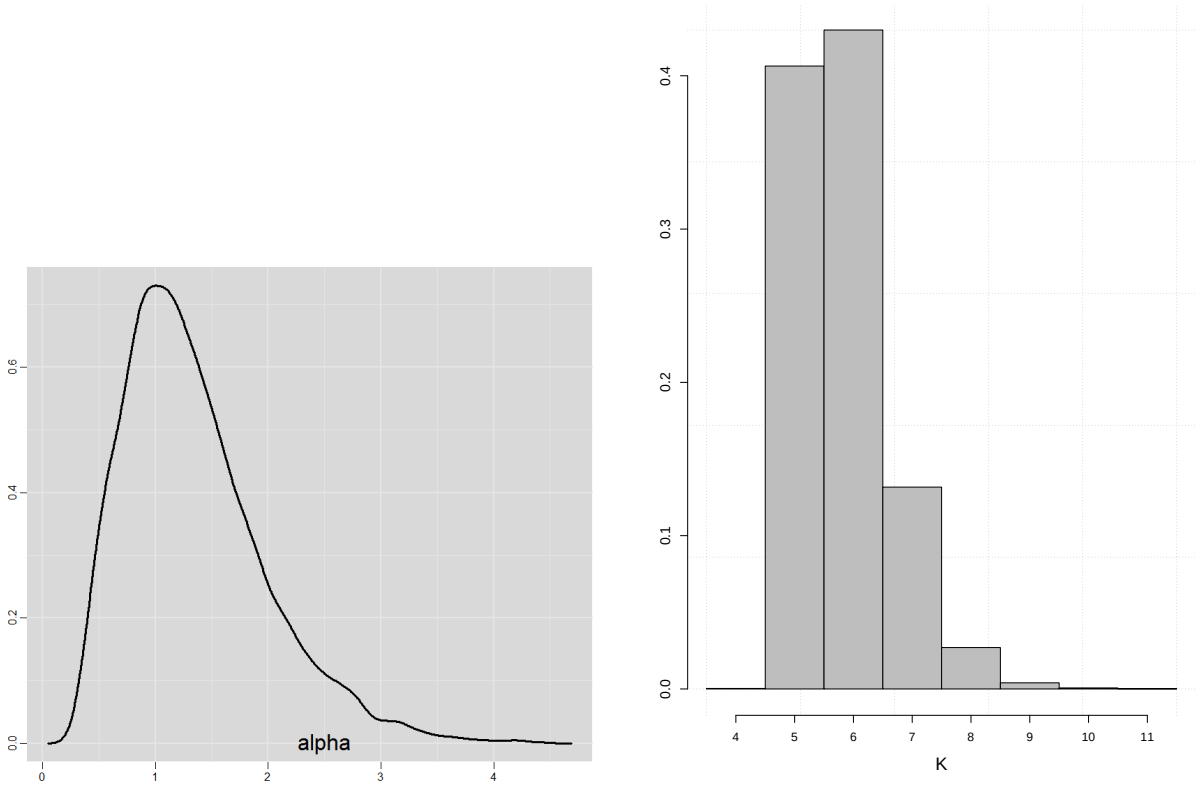

Figure 4. Posterior density of the precision parameter α (left panel) and posterior distribution of the number of clusters K for the general model (right panel).

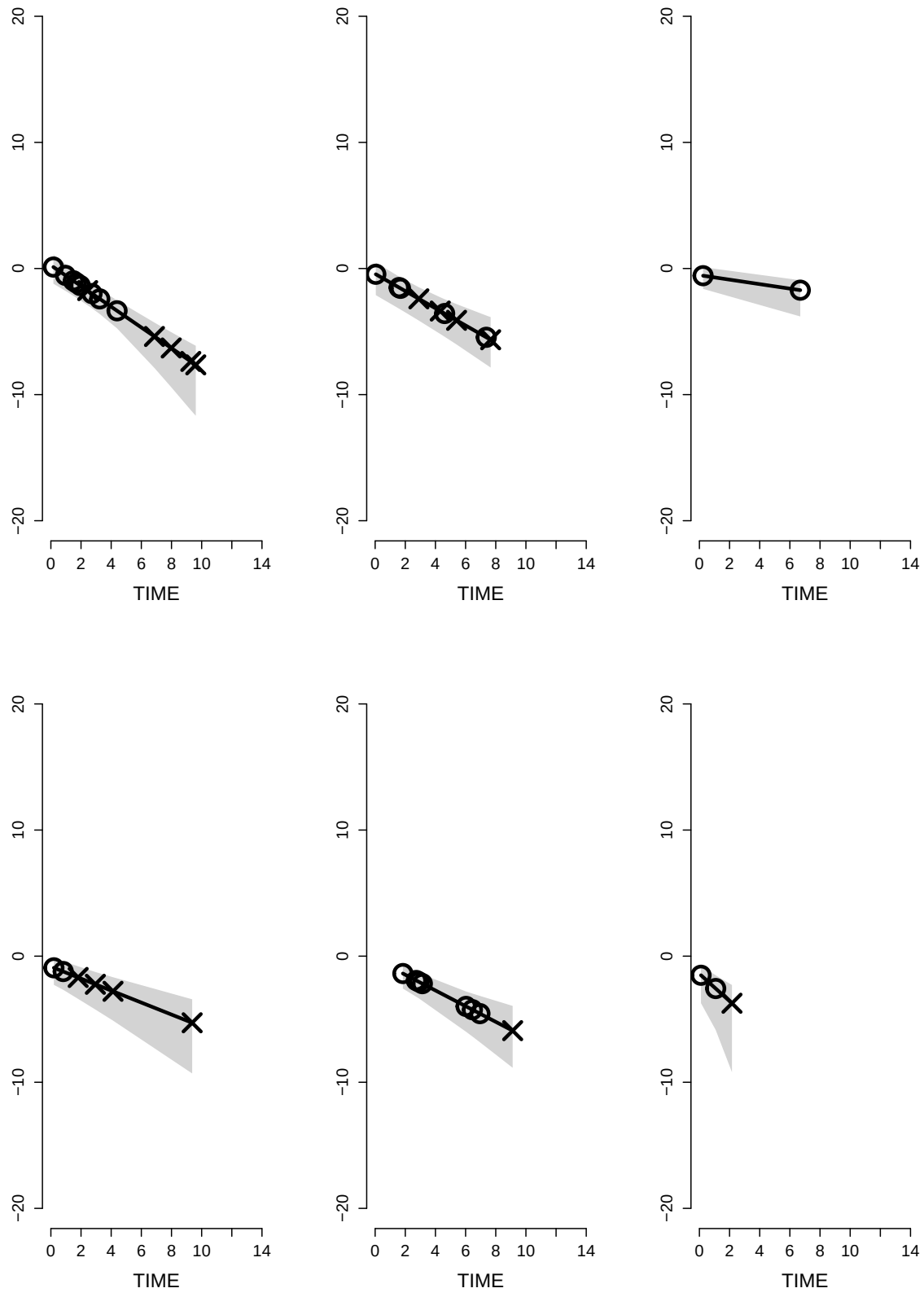

Figure 5. Health state function for six randomly selected patients over time. Circles represents disease state, while crosses denote remission.

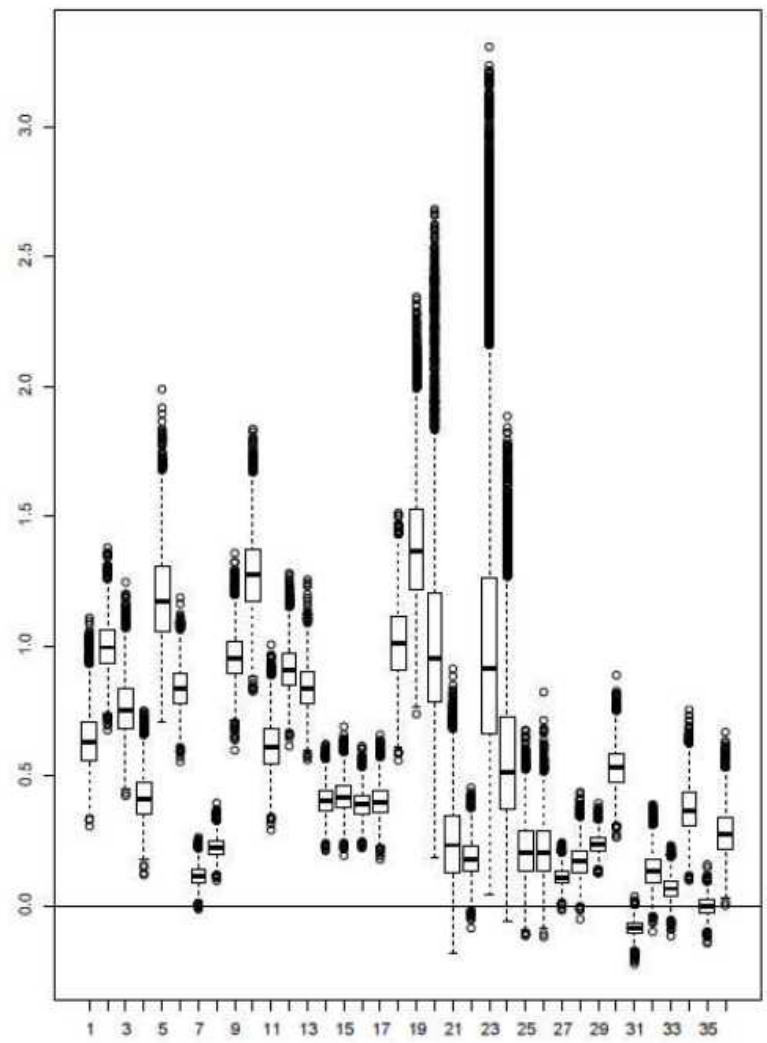

Figure 6. Boxplots of the posterior distribution of parameters $a_h, h = 1, \ldots, 36$ of the linear growth curve. The symptoms are ordered according to the list on the right.

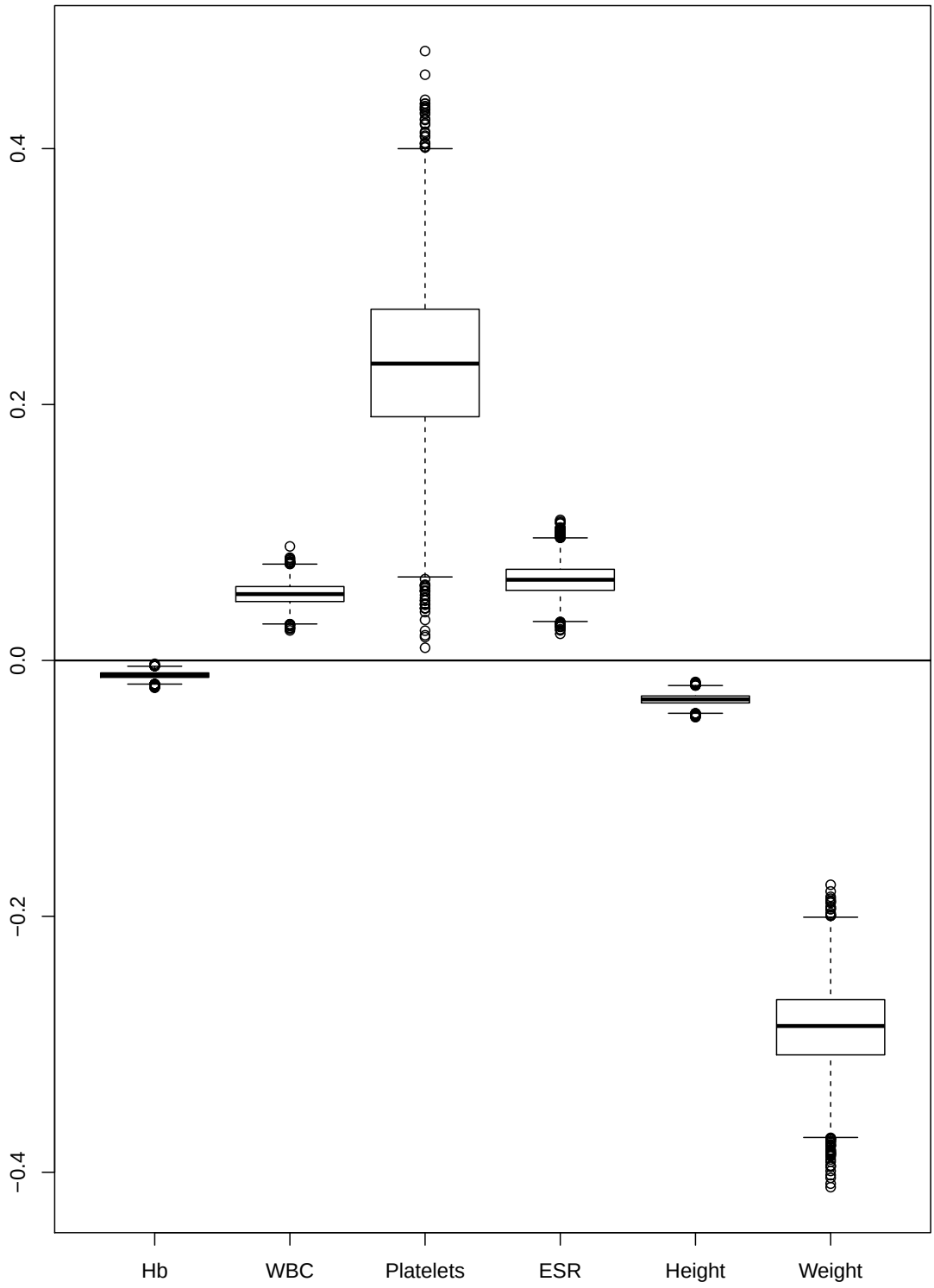

Figure 7. Boxplots of the posterior distribution of the parameters $\mathbf{c}_1$ in the continuous covariate model.

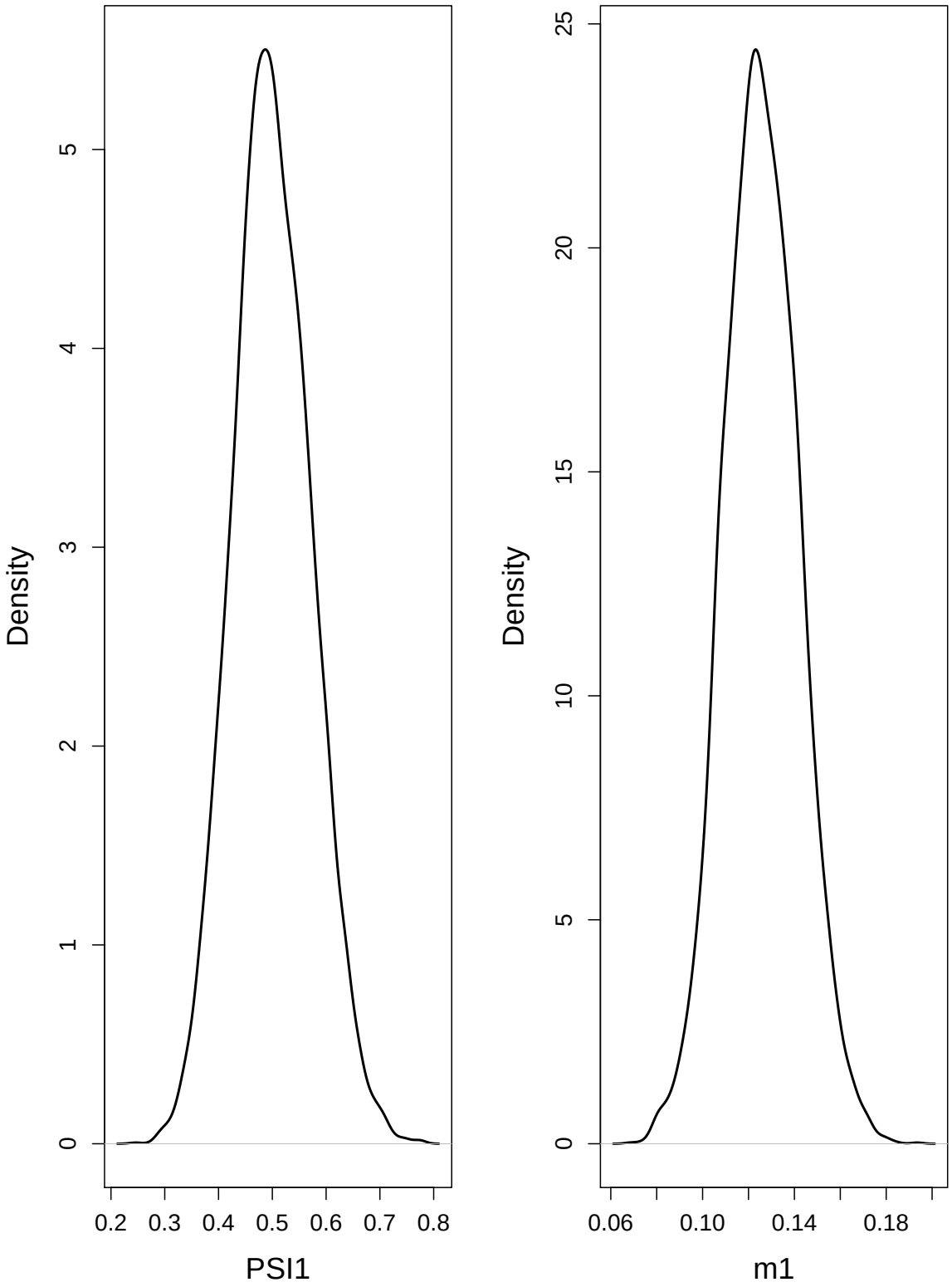

Figure 8. Posterior distribution of the regression coefficients of the health function in the Zero Inflated model for the CHAQ Score: posterior distribution of ψ_1 in the left panel and posterior distribution of m_1 in the right panel.