Published in final edited form as:
Statistics in Medicine (2019), **38**(16): 3013–3025
DOI: 10.1002/sim.8155

Score tests based on a finite mixture model of Markov processes under intermittent observation

SHU JIANG

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: s64jiang@uwaterloo.ca

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

A mixture model is described, which accommodates different Markov processes governing disease progression in a finite set of latent classes. We give special attention to the setting in which individuals are examined intermittently and transition times are consequently interval censored. A score test is developed to identify genetic markers associated with class membership. Simulation studies are conducted to validate the algorithm, assess the finite sample properties of the estimators, and assess the frequency properties of the score tests. A permutation test is recommended for settings when there is concern that the asymptotic approximation to the score test is poor. An application involving progression in joint damage in psoriatic arthritis (PsA) provides illustration and identifies human leukocyte antigen markers associated with unilateral and bilateral sacroiliac damage in individuals with PsA.

Keywords: finite mixture model, intermittent observation, Markov process, multistate model, score test

1 INTRODUCTION

1.1 LITERATURE REVIEW

Understanding the determinants of disease progression in chronic conditions is key to advancing scientific understanding, for making prognoses, and in health policy decision making. Multistate models offer an appealing and powerful framework for modeling disease processes in settings where the degree of damage can be meaningfully characterized into a finite number of disjoint states. Among individuals with hepatitis C infection, for example, the extent of liver damage is quantified using a

five-point scale with state 1 representing no fibrosis, states 2 to 4 representing increasing degrees of fibrosis and state 5 representing cirrhosis (Sweeting et al., 2006). In diabetic retinopathy, the extent of damage is measured on an 11-point scale with state 1 representing no damage and state 11 severe damage; The Early Treatment Diabetic Retinopathy Study Research Group (Sweeting et al., 2006) reported on a clinical trial evaluating the effect of aggressive control of blood sugar on the rate of progression through states based on this. Multistate models have also proven useful in characterizing decline in cognitive function in dementia (Tyas et al., 2007), loss of functional ability in arthritic conditions (Husted et al., 2007), and progression of immunological disease (Gentleman et al., 1994), and the development of asymptotic vertebral fractures in patients with osteoporosis (Riggs et al., 1981) .

Despite careful modeling of available information on such processes, considerable unexplained variation in disease progression is often evident between individuals. While Markov models provide a natural and convenient starting point for modeling such processes, generalizations are warranted in such settings (Aalen, 1987). Satten (1999) considered a conditionally Markov model for a progressive multistate process where a single non-negative random effect was specified to act multiplicatively on each transition intensity to account for between-subject heterogeneity; extensions for clustered progressive processes with correlated random effects unique to each possible transition have also been developed (Sutradhar and Cook, 2008). Discrete random effect models are also useful with binary random effect models being among the most common. These accommodate zero inflation in the context of generalized linear models, with zero-inflated Poisson (Van den Broek, 1995) and zero-inflated negative binomial models (Yau et al., 2003) receiving the most attention. In the failure time setting, mixture models have been used to explain the presence of long-term survivors (Farewell, 1982) where interest may lie primarily in identifying covariates associated with membership in the sub-population of long-term survivors (Farewell, 1977; Kuk and Chen, 1992); while much of this work has been carried out to deal with right censoring (Sy and Taylor, 2000), more extreme forms of censoring have also been considered (Lam and Xue, 2005; Cook et al., 2008). Mixture models can be challenging to fit, so score tests have been developed to assess the need for them in the context of generalized linear models (Van den Broek, 1995; Deng and Paul, 2000) and failure time settings with right-censoring (Peng et al., 2001) or current status observation schemes (Jonas et al., 2017).

In the multistate setting, these are often called mover-stayer models in which a sub-population not at risk for disease progression are considered "stayers" while those who are at risk of progression are called "movers". Frydman (1984) developed maximum likelihood methods for this setting and Fuchs and Greenhouse (1988) outlined an expectation-maximization (EM) algorithm (Dempster et al., 1977) which accommodates censoring. A generalization of the mover-stayer formulation has also been developed which accommodates intermittent observation schemes (Cook et al., 2002). Cook et al. (2004) consider multivariate random effects that accommodate a point mass at zero and a continuous random effect for susceptible individuals. O'Keeffe et al. (2013) explore the use of random effect models with a mover-stayer inverse Gaussian and a compound Poisson distribution. Finite mixture models offer a useful generalization of the basic mover-stayer model but less has been developed in this setting. Here, the target population is envisioned as comprised of several distinct sub-populations, or classes, and the disease processes are allowed to differ in some ways between these classes. In general, it will not be known to which class an individual belongs, and so class membership is represented by a latent variable; in this case, the mixing distribution and the parameters governing the process dynamics in each class are estimated. The EM algorithm can again be useful in this setting (Dempster et al., 1977).

In many instances, it is not apparent when a disease process has progressed and so the precise times of transitions between states are not available. This will be the case in most of the examples given in the opening paragraph. When the precise state of a multistate process is only available at periodic assessment times, the number and times of transitions are unknown and resulting data are

referred to as panel data (Kalbfleisch and Lawless, 1985). Kalbfleisch and Lawless (1985) developed an efficient algorithm for maximum likelihood estimation under a Markov assumption that is implemented in the *msm* package by Jackson (2011). Grüger et al. (1991) described the conditions that need to be satisfied for the observation process to be ignorable and such analyses valid, which are in effect the sequentially missing at random assumption given by Hogan et al. (2004); see also Cook and Lawless (2014).

The purpose of this paper is to develop a model and algorithm for fitting a finite mixture of Markov processes under intermittent observation. To permit the use of this model as a basis for screening a large number of genetic markers, we develop score tests for marker effects on class membership. The remainder of this paper is organized as follows. In the next subsection, we introduce the University of Toronto Psoriatic Arthritis Registry and describe the data that motivates this work. In Section 2, we define notation and describe a model for a finite mixture of Markov processes. Specifically, we construct the likelihood for the setting where individuals are under intermittent observation and describe how to estimate the asymptotic covariance matrix for the estimates. Score tests are developed in Section 3 where their finite sample properties are studied by simulation. An application involving joint damage in patients with psoriatic arthritis is given in Section 4 and concluding remarks and topics for further research are given in Section 5.

1.2 Motivating Study: Sacroiliac Joint Involvement in Psoriatic Arthritis

The University of Toronto Psoriatic Arthritis Clinic is a tertiary referral center for individuals with PsA, an immunological condition that features both skin and joint involvement. Areas of skin affected have a characteristic red colour with silvery white plaques (Moll and Wright, 1973). Affected joints may exhibit pain, swelling and stiffness that can ultimately lead to joint damage and reduced functional ability (Gladman et al., 2005). A registry of patients was created in 1976 to study the disease course and it has been recruiting and following patients continuously since its inception. Upon entry to the clinic, patients provide serum samples for genetic testing and undergo a detailed clinical and radiological examination (Gladman and Chandran, 2010). Follow-up clinical and radiological assessments are scheduled annually and every 2 years, respectively, in order to track changes in joint damage (Rahman et al., 1998).

Spondylitis, one of the musculoskeletal manifestations of PsA, is characterized by inflammation of the sacroiliac (SI) joints, the spine and neck, and reduced lateral range of motion of the back. In an early study of spondylitis, Hanly et al. (1988) identified 52 of 220 (23.6%) patients with PsA as having this disease. Scientists are particularly interested in the involvement of SI joints in PsA since damaged of these joints can have a severe detrimental effect on functional ability and quality of life. A recent study by Harron et al. (2016) investigated the association between human leukocyte antigen (HLA) B and C loci and SI joint involvement in a cohort of patients with PsA. These authors used radiographic evidence of SI involvement to define axial disease defined as the presence of at least grade 2 radiographic damage (unilateral or bilateral) on a five-point grading scheme (Geijer et al., 2009). The cross-sectional analysis of Harron et al. (2016), however, did not account for the variable times patients with PsA may have been at risk for developing damage in the SI joints. The proposed analysis based on a finite mixture of multistate Markov processes is designed to address this limitation and thereby provide a valid basis for inferences regarding the effects of HLA markers on risk of SI joint involvement.

The particular formulation of our model is motivated by the possible sub-types of patients with SI joint involvement. We aim to identify factors associated with unilateral sacroilitis (i.e. only the left or right SI joint is involved) and bilateral sacroilitis (i.e. both the left and right SI joint are involved). The motivation comes from the fact that unilateral sacroilitis is considered to represent a distinct phenotype called psoriatic spondylitis whereas bilateral involvement is more likely representing ankylosing

spondylitis, an arthritic condition primarily affecting the spine. Individuals with ankylosing spondylitis will take time to develop evidence of bilateral involvement, and if a cross-sectional analysis is carried out based on a sample of individuals with a short disease duration, individuals may be classified as having unilateral involvement even though they have ankylosing spondylitis. We adopt a finite mixture model that accommodates a different course (unilateral or bilateral) of the disease in PsA with the aim of detecting HLA alleles associated with these different courses. We give the details of the model in the next section.

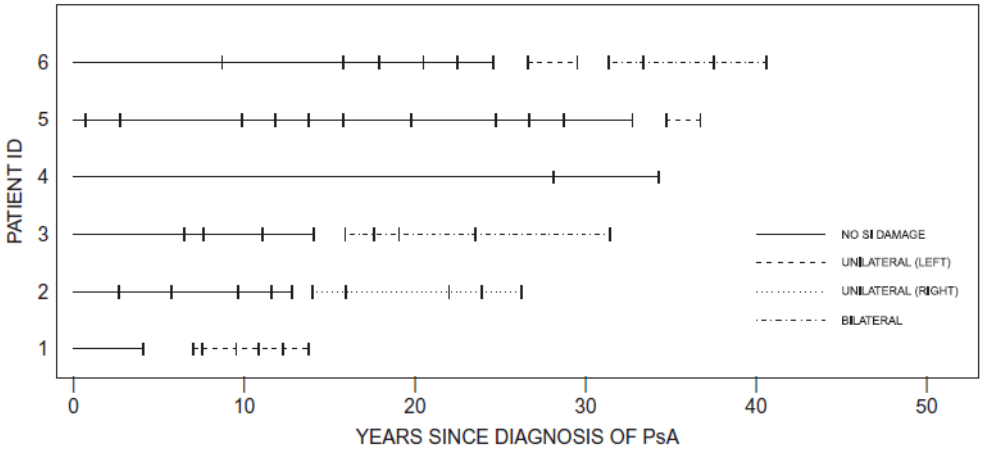


Figure 1: Plot of assessment times (hatch marks) and the type of joints damage (four types of line segments) between assessments from onset of PsA for a selected sample of patients from the University of Toronto Psoriatic Arthritis Clinic.

Figure 1 shows the time course of SI joint damage for a sample of six individuals in the available dataset. For each individual, the duration of follow-up since disease onset is represented by the length of the horizontal line and the vertical hatch marks denote the times joints are assessed at clinic visits. Four different types of line segments are used to convey the damage status of each individual at a given time with a solid line representing no SI joint involvement, a dashed line representing left side involvement, a dotted line representing right side involvement, and a dashed-dotted line representing bilateral involvement. The periods of time where no line segment is drawn are intervals in which the status is unknown because there was a different damage status for the visit at the left endpoint than at the right endpoint; since damage is assessed radiologically, the exact times at which damage occurs is unknown so the transition times are interval censored. We note from Figure 1 that some individuals develop SI damage shortly after diagnosis with PsA (e.g. individual 1) and some were not observed to develop damage despite long follow-up (e.g. individual 4). Moreover, some individuals that develop unilateral SI involvement progress quickly to the bilateral stage (see individuals 3 and 6), while some remain with unilateral involvement until the end of follow-up. The heterogeneity in the disease course motivates the formulation of a model that accommodates a class with no damage, two class with persistent unilateral (left or right) damage in alignment with the condition of psoriatic spondylitis, and a bilateral class corresponding to the condition of ankylosis spondylitis. Figure 2 displays the four multistate processes corresponding to the four latent classes; the definition of the states with a given number is the same in the four classes. State 0 represents no SI joint damage, state 1 corresponds to unilateral damage on the left side, state 2 corresponds to unilateral damage on the right side, and state 3 corresponds to bilateral damage. In class 0, individuals remain free of SI joint damage, in classes 1 and 2, they will ultimately experience unilateral SI joint damage consistent with the condition of psoriatic spondylitis, and in class 3, they will experience bilateral involvement consistent with ankylosing spondylitis. We denote $\lambda_{k\ell}$ as the intensity for transitions from state k to

state l that are denoted by λ_{kl} for $(k,l) \in \{(0,1),(0,2),(1,3),(2,3)\}$. We constrain λ_{01} to be the same for class 1 and 3 since, in both classes, it is the intensity for the onset of first damage in the left SI joint, and we likewise constrain λ_{02} to be the same in classes 2 and 3.

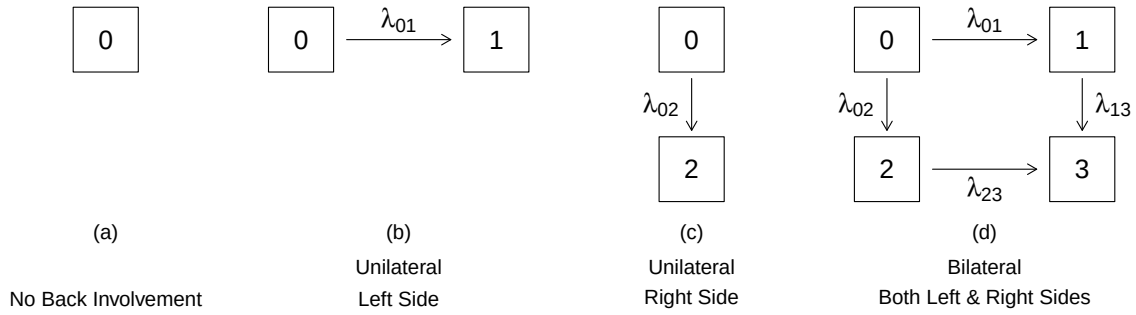


Figure 2: Multistate diagram for the processes of the four classes in the finite mixture model.

2 Model Formulation and Inference

2.1 Notation

We consider a $K+1$-state process with states labelled $k = 0, 1, \ldots, K$, with state 0 representing a healthy state and states $1, \ldots, K$ corresponding to varying degrees of damage. We let $Z(t)$ represent the state occupied at time t since disease onset, and $\{Z(s), 0 < s\}$ denote the associated stochastic process. Here, we develop the model and algorithm in a slightly more general context than the previous section by considering progressive processes in which transitions directly from k to l are possible for any $l > k$, $k = 0, 1, \ldots, K-1$. To accommodate heterogeneity in the disease course, we consider the setting where the population is comprised of $J+1$ distinct classes of individuals in which the processes for individuals in the same class are governed by a common transition probability matrix; the states with the same label in different classes are assumed to represent the same condition and so have the same interpretation. Let C be a discrete latent random variable representing the class label for an individual, $C = 0, 1, \ldots, J$, and let $X = (1, X_1, \ldots, X_{p-1})'$ be a $p \times 1$ covariate vector. We let $P(C = j|X;\beta) = \pi_j(X;\beta)$ denote the probability of belonging to class j given X, where $\sum_{j=0}^{J} \pi_j(X;\beta) = 1$. We formulate a multinomial logistic regression model (McCullagh and Nelder, 1989) using class $C = 0$ as the reference class so

$$P(C = j|X;\beta) = \frac{\exp(X'\beta_j)}{1 + \sum_{j=1}^{J} \exp(X'\beta_j)} \, , j = 1, \ldots, J \, , \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_J)'$ is a $p_0 \times 1$ parameter vector with $p_0 = p \times J$.

If $\mathcal{H}(t) = \{Z(s), 0 < s < t; X\}$ denotes the history at time t, the transition intensities are

$$\lim_{\triangle t \downarrow 0} \frac{P(Z(t + \triangle t^-) = l | Z(t^-) = k, \mathcal{H}(t), C = j)}{\triangle t} = \lambda_{kl}(t|\mathcal{H}(t), C = j; \alpha_j) = \lambda_{jkl}(t|\mathcal{H}(t); \alpha_j) \, , l > k \, ,$$

where α_j is the vector of parameters governing the transition intensities in class j, $j = 0, 1, \ldots, J$, and $\alpha = (\alpha_0', \alpha_1', \ldots, \alpha_J')'$ is a $q \times 1$ vector. We restrict attention to conditionally Markov processes for which $\lambda_{jkl}(t|\mathcal{H}(t);\alpha_j) = \lambda_{jkl}(t;\alpha_j)$. We do not model covariate effects on the transition intensities as the primary goal is model the latent class process. Moreover, the multistate models that define the classes are specified to account for the variable duration of follow-up and the fact that, for example,

individuals with only unilateral SI joint damage at one point in time may ultimately develop bilateral involvement. Moreover, estimability issues can arise when covariate effects are modeled in more than one part of a mixture model; this is particularly true when only panel data are available on the multistate processes. We let $\lambda_{jkl} = \lambda_{kl}$ for $\forall k \neq l$ so that the transition intensities between the same states in different classes are the same. Finally, we let $\theta = (\alpha', \beta')'$ denote the full $(q + p_0) \times 1$ parameter vector.

We now consider a sample of m independent individuals labeled $i = 1, \ldots, m$ and introduce a subscript i to write $\{Z_i(s), 0 < s\}$, C_i and X_i, $i = 1, \ldots, m$. We consider the panel data setting with inspection times for individual i that are denoted by $a_{ir},\ r = 0, \ldots, R_i$ where we assume $a_{i0} = 0$ and $Z_i(0) = 1$ with probability 1. With a fixed covariate X_i, the observed data for individual i are denoted by $\mathcal{D}_i = \{(Z_i(a_{ir}), a_{ir}), r = 0, 1, \ldots, R_i; X_i\}$. We define the likelihood contribution from a particular individual i for a finite mixture of Markov processes under panel observation as

$$L_i(\theta) = \sum_{j=0}^{J} \left\{ \prod_{r=1}^{R_i} P(Z_i(a_r)|Z_i(a_{r-1}), C_i = j, X_i; \alpha) \right\} P(C_i = j|X_i; \beta) \ . \tag{2}$$

While in the application, we assume $\{Z_i(s), 0 < s\} \perp X_i|C_i$, in what follows, we retain the process dependence on X_i given C_i for generality. To simplify the notation, we let

$$L_{ij}(\alpha) = \prod_{r=1}^{R_i} P(Z_i(a_r)|Z(a_{r-1}), C_i = j, X_i; \alpha) \ , \tag{3}$$

and we write the observed likelihood for individual i as

$$L_i(\theta) = \sum_{j=0}^{J} L_{ij}(\alpha)\pi_j(X_i; \beta) \ . \tag{4}$$

The model defined in (2) is based on the assumption that the latent classes are mutual exclusive and exhaustive; that is, each individual is a member of one and only one of the latent classes.

The maximum likelihood estimate (MLE) of θ is obtained by maximizing $L(\theta) = \prod_{i=1}^{m} L_i(\theta)$, or equivalently solving the $(q + p_0) \times 1$ observed data score equation $U(\theta) = 0$ where $U(\theta) = (U_1'(\theta), U_2'(\theta))' = \sum_{i=1}^{m} U_i(\theta)$ with $U_1(\theta) = \sum_{i=1}^{m} U_{i1}(\theta)$ a $q \times 1$ vector where $U_{i1}(\theta) = \partial \log L_i(\theta)/\partial\alpha$ and $U_2(\theta) = \sum_{i=1}^{m} U_{i2}(\theta))$ a $p_0 \times 1$ vector where $U_{i2}(\theta) = \partial \log L_i(\theta)/\partial\beta$. Also note that

$$\begin{aligned} U_{i1}(\theta) &= E_C\{S_{i1}(Z_i|C, X_i; \alpha)|\mathcal{D}_i\} = \sum_{j=0}^{J} S_{i1}(Z_i|C_i = j, X_i; \alpha)P(C_i = j|\mathcal{D}_i; \theta) \ , \\ U_{i2}(\theta) &= E_C\{S_{i2}(C|X_i; \beta)|\mathcal{D}_i\} = \sum_{j=0}^{J} S_{i2}(j|X_i; \beta)P(C_i = j|\mathcal{D}_i; \theta) \ , \end{aligned} \tag{5}$$

where $S_{i1}(Z_i|C_i = j, X_i; \alpha) = \partial \log L_{ij}(\alpha)/\partial\alpha$ and, $S_{i2}(C_i|X_i; \beta) = \partial \log P(C_i|X_i; \beta)/\partial\beta$. When we wish to write this more compactly, we let $S_i(\theta) = (S_{i1}'(\alpha), S_{i2}'(\beta))'$ denote the complete data score contributions from individual i, $i = 1, \ldots, m$.

2.2 Estimation and Inference via the EM Algorithm

Suppressing the subscript i and considering the contribution from a generic individual, the complete data likelihood is

$$\mathcal{L}(\theta) \propto \prod_{j=0}^{J} \{L_j(\alpha)\pi_j(X; \beta)\}^{I(C=j)} \ . \tag{6}$$

At the rth iteration of the EM algorithm, the E-step involves taking the conditional expectation of the log of (6) to obtain $Q(\theta;\theta^r) = E\{\log\ \mathcal{L}(\theta) \mid \mathcal{D};\theta^r\}$ where θ^r is the estimate of θ at the rth iteration with elements α^r and β^r, and $\mathcal{D}$ is the observed data. If we let

$$w_j^r = P(C=j|\mathcal{D};\theta^r) = \frac{L_j(\alpha^r)\pi_j(X;\beta^r)}{\sum_{j=0}^{J} L_j(\alpha^r)\pi_j(X;\beta^r)}\ , \tag{7}$$

we can write $Q(\theta;\theta^r) = Q_1(\alpha;\theta^r) + Q_2(\beta;\theta^r)$ w,here

$$\begin{aligned} Q_1(\alpha;\theta^r) &= \sum_{j=0}^{J} w_j^r \log L_j(\alpha)\ , \\ Q_2(\beta;\theta^r) &= \sum_{j=0}^{J} w_j^r \log \pi_j(X;\beta)\ . \end{aligned} \tag{8}$$

The M-step involves maximizing $Q(\theta;\theta^r)$ with respect to θ to obtain an updated estimate θ^{r+1}. Note that, if the α_j are functionally independent (i.e. there are no shared parameters among the Markov models for the different classes), then $Q_1(\alpha;\theta^r)$ can be maximized class by class by adapting the Fisher-scoring algorithm of Kalbfleisch and Lawless (1985) through the incorporation of weights. When different classes share transition intensities (e.g. as mentioned in Section 2.1 for common pairs of states in different classes), a slightly more involved adaptation can be used or one can simply use an all-purpose optimization function such as `nlm` or `optim` in R. The function $Q_2(\beta;\theta^r)$ has a similar form to the log-likelihood encountered in multinomial regression. We iterate between the E-step and M-step until the convergence criterion $\max|\theta^{r+1}-\theta^r| < \epsilon$ is satisfied, where ϵ is a specified tolerance.

To avoid computating the Hessian of the observed log-likelihood, we compute the observed information matrix $I(\theta) = -\partial U(\theta)/\partial\theta'$ based on the work of Louis (1982) who showed that

$$I(\theta) = E\{\mathcal{J}(\theta)|\mathcal{D}\} - E\{S(\theta)S'(\theta)|\mathcal{D}\} + U(\theta)U'(\theta)\,, \tag{9}$$

where $\mathcal{J}(\theta) = -\partial S(\theta)/\partial\theta'$ is the complete data information matrix. To compute $E\{S(\theta)S'(\theta)|\mathcal{D}\}$ note that it can be written as

$$E\{S(\theta)S'(\theta)|\mathcal{D}\} = \mathrm{var}\{S(\theta)|\mathcal{D}\} + E\{S(\theta)|\mathcal{D}\}E(S'(\theta)|\mathcal{D}\} = \mathrm{var}\{S(\theta)|\mathcal{D}\} + U(\theta)U'(\theta)\ , \tag{10}$$

since $U(\theta) = E\{S(\theta)|\mathcal{D}\}$. Substituting (10) into (9) gives $I(\theta) = E\{\mathcal{J}(\theta)|\mathcal{D}\} - \mathrm{var}\{S(\theta)|\mathcal{D}\}$. The term $\mathcal{J}(\theta)$ can be computed using the weights in (7) to take the expectation of the complete data observed information. To compute $\mathrm{var}\{S(\theta)|\mathcal{D}\}$, we express it as follows. First, we let In matrix notation we let $Y = (Y_0, Y_1, ..., Y_J)'$ where $Y_j = I(C=j)$, $j = 0, 1, \ldots, J$. Second, we note that, if $A = (\partial\log L_0(\alpha)/\partial\alpha, \ldots, \partial\log L_J(\alpha)/\partial\alpha)$ is a $q \times (J+1)$ matrix and $B = (\partial\log\pi_0(\beta)/\partial\beta, \ldots, \partial\log\pi_J(\beta)/\partial\beta)$ is a $p_0 \times (J+1)$ matrix, we can write

$$S(\theta) = H'Y\ , \tag{11}$$

where $H = (A', B')$ is a $(J+1)\times(q+p_0)$ matrix. Then, $\mathrm{var}\{S(\theta)|\mathcal{D}\} = H'\mathrm{cov}(Y|D)H$. Since C represents the class membership, $\mathrm{cov}(Y|D)$ is multinomial but with a conditional probability (7) giving $\mathrm{var}(Y_j|\mathcal{D}) = w_j(\theta)(1-w_j(\theta))$ for $j = 0, \ldots, J$ and $\mathrm{cov}(Y_{j_1}, Y_{j_2}|\mathcal{D}) = -w_{j_1}(\theta)w_{j_2}(\theta)$ for $j_1 \neq j_2$,j_1,$j_2 = 0, \ldots, J$. An estimate of $\mathrm{var}\{S(\theta)|\mathcal{D}\}$ can be obtained by inserting MLEs in place of the parameters. Summing contributions over all individuals for each term $I(\theta)$ and calculating the inverse of the observed information matrix yields standard errors.

3 SCORE TESTS FOR GENETIC EFFECTS

3.1 CONSTRUCTION OF THE TEST STATISTIC

Interest lies in the effect of genetic markers on class membership, so we extend the covariate vector in the class model; we let G denote a $p_1 \times 1$ genetic marker and set $W = (X', G')'$. The multinomial logistic regression model is then

$$P(C = j|W;\eta) = \frac{\exp(X'\beta_j + G'\gamma_j)}{1 + \sum_{j=1}^{J} \exp(X'\beta_j + G'\gamma_j)} = \frac{\exp(W'\eta_j)}{1 + \sum_{j=1}^{J} \exp(W'\eta_j)}\,, \tag{12}$$

where $\beta_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{j,p-1})'$ is the vector of coefficients of X in class j, $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jp_1})'$ are the coefficients of G, and $\eta_j = (\beta_j', \gamma_j')'$.

The null hypothesis of no genetic effects is H_0:$\gamma = 0$ where $\gamma = (\gamma_1', \ldots, \gamma_J')'$ and the alternative hypothesis is H_1:$\gamma \neq 0$. We could fit the model under H_1, but with a large number of candidate genetic markers, this can be computationally demanding with an EM algorithm required for each marker. We therefore opt to use score tests.

We let $\phi = (\theta', \gamma')'$ and let the complete data score vector be given by

$$\begin{aligned} U_{i1}(\phi) &= E_C\{S_{i1}(Z_i|C, W_i;\alpha)|D_i\} \\ U_{i2}(\phi) &= E_C\{S_{i2}(C|W_i;\eta)|D_i\} \\ U_{i3}(\phi) &= E_C\{S_{i3}(C|W_i;\eta)|D_i\}\,, \end{aligned}$$

where $S_{i1}(Z_i|C_i = j, W_i;\alpha) = \partial \log L_{ij}(\alpha)/\partial\alpha$ with $L_{ij}(\alpha)$ given by (3) with X_i replaced by $W_i = (X_i', G_i')'$, $S_{i2}(C_i|W_i;\eta) = \partial \log P(C_i|W_i;\eta)/\partial\beta$ and $S_{i3}(C_i|W_i;\eta) = \partial \log P(C_i|W_i;\eta)/\partial\gamma$.

The score test statistic (Boos, 1992) for testing H_0: $\gamma = 0$ is

$$T = U_3'(\widehat{\phi}_0) I^{\gamma\gamma}(\widehat{\phi}_0) U_3(\widehat{\phi}_0), \tag{13}$$

where $\phi_0 = (\alpha', \beta', 0')'$ and $\widehat{\phi}_0 = (\widehat{\alpha}', \widehat{\beta}', 0')'$ where $\widehat{\theta} = (\widehat{\alpha}', \widehat{\beta}')'$ is the MLE under the null, $U_3(\cdot)$ is a $p_1 \times 1$ score function for γ, and $I^{\gamma\gamma}(\cdot)$ is a lower diagonal $p_1 \times p_1$ submatrix of $I^{-1}(\widehat{\phi}_0)$. Under the null hypothesis, the score statistic (13) follows a $\chi^2_{p_1}$ distribution and a p-value for testing H_0:$\gamma = 0$ is obtained by computing $P(\chi^2_{p_1} > T_{\text{obs}})$ where T_{obs} is a realized value of (13).

3.2 SIMULATION STUDIES

The purpose of the simulation studies are to demonstrate the performance of a proposed finite mixture model and assess the empirical rejection rate of the score test. We consider 4 classes ($j = 0, 1, 2, 3$) and constrain α as in Figure 2. To model class membership for individual i, we first generate a Bernoulli covariate X_{i1} with $P(X_{i1} = 1) = 0.5$. Let $X_i = (1, X_{i1})'$, generate G_i as Bernoulli with probability of success 0.05 or 0.10 to explore performance in the setting of a relatively rare marker, and let $W_i = (X_i', G_i)'$, $i = 1, \ldots, m$. We set the coefficients of X_i in the multinomial regression model to $\beta_{11} = \beta_{21} = \log 1.1$ and $\beta_{31} = \log 1.2$. For generating the data under the null hypothesis, the coefficients for the genetic variable were set to $\gamma_1 = \gamma_2 = \gamma_3 = 0$. We then determined the intercepts β_{10}, β_{20} and β_{30} so that the marginal probabilities for the four classes are $P(C = 0) = 0.30$, $P(C = 1) = 0.25$, $P(C = 2) = 0.25$, and $P(C = 3) = 0.20$.

We consider the setting where interest lies in the disease course over the interval $(0, E]$ where 0 is the onset of the disease process and E is the end of the period. The transition intensities in the multistate framework are set so that $P(Z(E) = 1|C = 1) = 0.80$, $P(Z(E) = 2|C = 2) = 0.80$, and $P(Z(E) = 3|C = 3) = 0.70$, so that 80% of individuals with unilateral disease (classes 1 and 2) will experience damage by the end of the observation interval, and 70% of those with bilateral disease will

experience damage in both sides by this time. Let $A_i(t)$ count the cumulative number of assessments over $(0, t]$, $dA(t) = 1$ if an assessment occurs at time t and be zero otherwise, and let $\{A_i(s), s > 0\}$ denote the counting process which is taken to be a time homogeneous Poisson process. We set $E = 1$ without loss of generality and specify the rate ρ such that $E\{A_i(1)\} = \mu = 15$ or 30.

To assess the performance of estimators, we first fit the correct model under the constraint $\gamma = 0$ and examine the empirical performance of the estimators. We display these in a table reporting the empirical bias (EBIAS), the empirical standard error (ESE), and the empirical coverage probability (ECP) where the sample standard deviation is computed based on Louis' method as described in Section 2.2. To assess the empirical rejection rates, we simulate data under the null hypothesis, obtain the estimates under the null as described above, and compute the score statistic of the null hypothesis H_0: $\gamma = 0$. The empirical rejection rate is computed as the proportion of simulated samples for which the sample p-value is less than 0.05. We evaluate the empirical type I error rate when $\gamma = 0$ and the empirical power when $\gamma \neq 0$. For the latter, we consider several different values γ_1 and γ_3 including $\log 1 = 0$, $\log 1.25$, $\log 1.5$ and $\log 2$; for each combination of values, we set the intercepts of the multinomial models to give the same marginal probabilities of class membership as above.

Table 1: Empirical performance of estimators for β and α under the null model with $m = 2000$ individuals per simulation and and $nsim = 1000$ simulations; EBIAS is the empirical bias, ESE is the empirical standard error, ASE is the average model-based standard error, and ECP% is the percent empirical coverage probability.

	$E\{A_i(1)\} = 15$				$E\{A_i(1)\} = 30$			
	EBIAS	ESE	ASE	ECP%	EBIAS	ESE	ASE	ECP%
CLASS MODEL								
β_{10}	<0.001	0.147	0.149	96.3	0.008	0.140	0.141	95.7
β_{11}	-0.001	0.152	0.153	96.0	0.003	0.145	0.147	96.0
β_{20}	-0.029	0.148	0.149	95.1	0.007	0.141	0.141	95.9
β_{30}	0.001	0.157	0.159	96.1	0.022	0.150	0.153	96.3
β_{31}	<0.001	0.156	0.159	95.7	0.004	0.151	0.153	95.1
MULTISTATE MODEL								
α_{01}	-0.005	0.091	0.090	95.0	-0.012	0.088	0.086	95.3
α_{02}	-0.009	0.092	0.091	94.5	0.007	0.086	0.085	95.3
α_{13}	-0.027	0.214	0.216	94.4	-0.012	0.196	0.195	94.3
α_{23}	-0.017	0.214	0.215	94.0	-0.012	0.195	0.195	94.5

Table 1 reports on the finite sample properties of estimators of α and β for the setting when $\gamma = 0$ where there are $m = 2000$ individuals per sample and $nsim = 1000$ simulations carried out. Thee empirical biases are generally small, there is good agreement between the empirical standard error and the average model-based standard error from Louis' formula, and the empirical coverage probability of the 95% confidence intervals is compatible with the nominal level. Table 2 reports on the empirical rejection rates of the proposed score test under the null and alternative settings; we carry out $nsim = 1000$ simulations under the null (see the top row) and $nsim = 500$ simulations under the alternative when $\gamma \neq 0$. The type I error of the score test is compatible with the nominal level for both cases with $E\{A_i(1)\} = 30$ and for the case when $P(G = 1) = 0.10$ (and with higher

probabilities for the marker – results not shown) when $E\{A_i(1)\} = 15$, but it is slightly elevated when the marker is less common with $P(G = 1) = 0.05$ and $E\{A_i(1)\} = 15$. Figure 3 displays Q-Q plots of the empirical distribution of the score statistic against a χ^2_2 distribution under the null settings. Three of the Q-Q plots show good agreement between the sample quantiles and the quantiles of the χ^2_2 distribution. There are outliers evident when $E\{A_i(1)\} = 15$ with $P(G = 1) = 0.05$ suggesting the empirical distribution has a bigger right tail area; this is the setting where the empirical rejection rate is 6.8% in Table 2. To address this, we implement a permutation test in the application of the next section. In terms of power, there is generally an increase in the empirical power with increasing effect size under H_1; further remarks on power considerations are provided in Section 5.

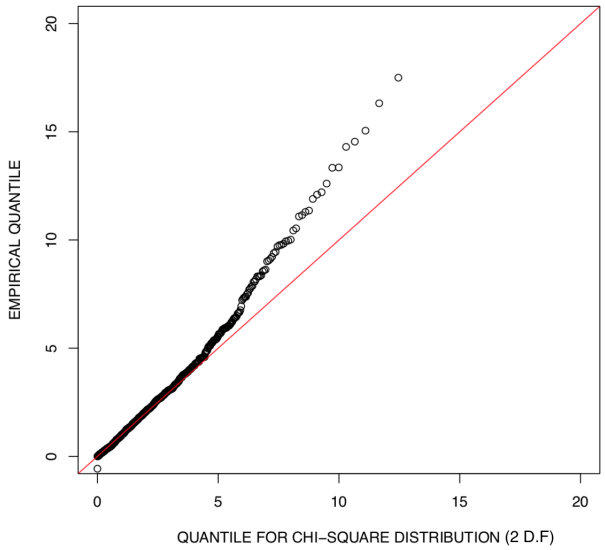
Table 2: Empirical rejection rates based on χ^2_2 approximation; $m = 2000$ individuals per sample with $nsim = 1000$ simulations when $(\gamma_1, \gamma_3) = (0, 0)$ and $nsim = 500$ for other settings.

		$E\{A_i(1)\} = 15$		$E\{A_i(1)\} = 30$	
γ_1	γ_2	$P(G = 1) = 0.05$	$P(G = 1) = 0.10$	$P(G = 1) = 0.05$	$P(G = 1) = 0.10$
0	0	6.8	6.0	6.2	5.3
0	$\log 1.25$	12.4	13.0	12.4	15.5
0	$\log 1.5$	23.6	35.9	26.0	42.4
0	$\log 2.0$	58.5	75.9	58.3	85.5
$\log 1.25$	$\log 1.25$	8.7	13.1	9.5	13.1
$\log 1.25$	$\log 1.5$	17.1	26.8	21.6	30.4
$\log 1.25$	$\log 2.0$	46.4	66.5	46.6	71.0
$\log 1.5$	$\log 1.5$	17.1	33.3	18.6	35.1
$\log 1.5$	$\log 2.0$	38.1	59.8	41.3	66.3
$\log 2.0$	$\log 2.0$	39.5	68.4	36.5	69.4

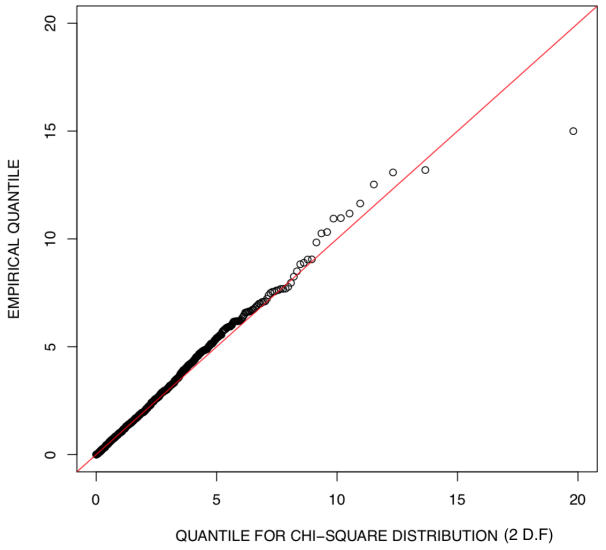
4 Application to Sacroiliac Damage in Psoriatic Arthritis

The methods developed in the previous sections were applied to data on joint damage in patients with PsA from the University of Toronto Psoriatic Arthritis Clinic. Here, interest lies in examining the effects of binary HLA markers on the nature of any back involvement in these patients. We examine the effects of HLA markers individually on SI joint damage while controlling for gender ($X_1 = 1$ for female, 0 for male) and early age of onset ($X_2 = 1$ for ≤ 40 old, 0 for > 40 years old). In the null model, we let $\beta_j = (\beta_{j0}, \beta_{j1}, \beta_{j2})'$ denote the parameters for class j membership, $j = 1, 2, 3$. Table 3 includes estimates, standard errors, and 95% confidence intervals for all parameters under the null model. From this fitted model, we conclude that the odds of females experiencing bilateral involvement (compared to no SI joint involvement) are lower than that of males ($OR = 0.42$, 95% CI: 0.29, 0.59, $p < 0.001$). There is no evidence of an effect of gender on unilateral involvement nor an effect of early onset on the nature of SI involvement.

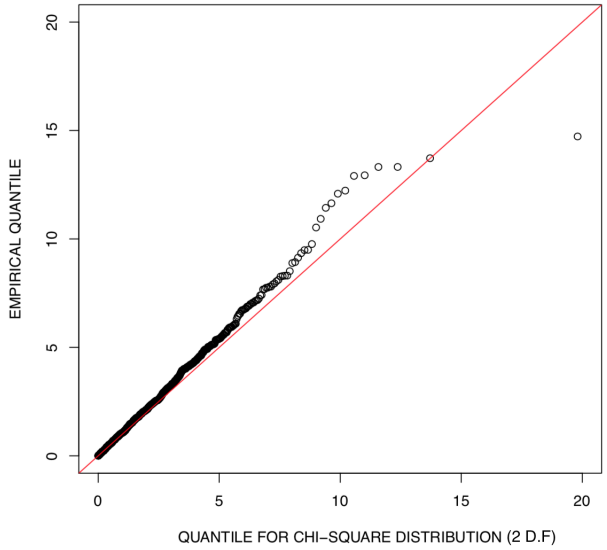
Under the alternative model, we again constrain $\gamma_1 = \gamma_2$ when deriving the test statistic since these may be interpreted as the effect of the HLA marker on unilateral SI joint damage. The null hypothesis of no genetic effect is given by H_0: $\gamma = 0$ where $\gamma = (\gamma_1, \gamma_1, \gamma_3)$ with γ_1 reflecting the effect of the marker on unilateral SI joint involvement compared to no involvement, and γ_3 reflecting the effect of the marker on bilateral SI joint involvement. The results of applying the score tests of



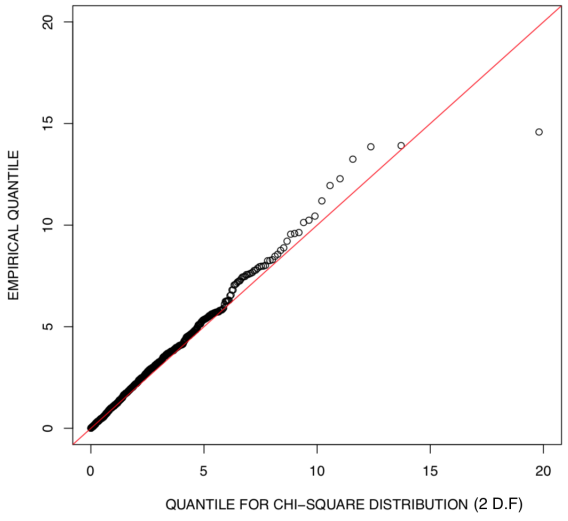
(a) $E\{A_i(1)\} = 15$ and $P(G = 1) = 0.05$



(b) $E\{A_i(1)\} = 15$ and $P(G = 1) = 0.10$



(c) $E\{A_i(1)\} = 30$ and $P(G = 1) = 0.05$



(d) $E\{A_i(1)\} = 30$ and $P(G = 1) = 0.10$

Figure 3: Q-Q plots for assessing the empirical distribution of the score statistic in relation to the two-degree-of-freedom (2 d.f.) chi-square distribution for different intensities for the visit process and different frequencies of the marker; two points are outside of the range for panel (a) with co-ordinates $(14.2, 40.1)$ and $(19.8, 40.2)$.

Table 3: Results of fitting the finite mixture model under the null hypothesis (omitting human leukocyte antigen markers) for the occurrence of sacroiliac joint damage.

		Parameter Estimates			Exponential Values		
MODEL		EST.	S.E.	95% CI	EXP	95% CI	p-value
CLASS							
1	β_{10}	-1.488	0.296	(-2.067, -0.908)	0.226	(0.127, 0.403)	
	β_{11}	-0.686	0.398	(-1.466, 0.095)	0.504	(0.231, 1.100)	0.085
	β_{12}	-0.139	0.414	(-0.951, 0.673)	0.870	(0.386, 1.960)	0.737
2	β_{20}	-2.268	0.638	(-3.520, -1.017)	0.104	(0.030, 0.362)	
	β_{21}	0.872	0.659	(-0.420, 2.164)	2.392	(0.657, 8.706)	0.186
	β_{22}	0.285	0.477	(-0.651, 1.220)	1.330	(0.522, 3.387)	0.550
3	β_{30}	0.914	0.150	(0.620, 1.209)	2.494	(1.859, 3.350)	
	β_{31}	-0.876	0.181	(-1.230, -0.521)	0.416	(0.292, 0.594)	<0.001
	β_{32}	-0.168	0.190	(-0.541, 0.205)	0.845	(0.582, 1.228)	0.377
MULTISTATE							
	α_{01}	-2.267	0.105	(-2.473, -2.060)	0.104	(0.084, 0.127)	
	α_{02}	-2.703	0.133	(-2.964, -2.442)	0.067	(0.052, 0.087)	
	α_{13}	0.281	0.162	(-0.037, 0.599)	1.324	(0.964, 1.820)	
	α_{23}	-1.762	0.190	(-2.135, -1.389)	0.172	(0.118, 0.249)	

Section 3.1 are given in Table 4 for a total of 96 HLA markers including HLA-A, HLA-B, HLA-C, HLA-DR, and HLA-DQB markers.

To address possible concerns about the adequacy of the χ^2_2 approximation for the score statistics in this setting, we computed p-values based on the permutation distribution of the score test statistic. Specifically, we consider $B = 10,000$ permutations of the genetic marker within the four strata defined by the binary gender and age of onset covariates. Permutation p-values are computed as the empirical probability of a score test being realized that is larger than the one observed for the given dataset. So, if T_{obs} is the observed test statistics given by (13) and $T^{(b)}$, $b = 1, \ldots, B$ denote the score statistics for the $B = 10,000$ permutation samples, then the permutation p-value is

$$p^{\ddagger} = \sum_{b=1}^{B} I(T^{(b)} \geq T_{\text{obs}})/B \tag{14}$$

Note the $p-$value based on the χ^2_2 approximation are labeled $p^{\dagger}$ in Table 4.

Based on the χ^2_2 approximation, we find 11 markers are statistically significant at the 5% level. When controlling the false discovery rate using the Benjamini-Hochberg procedure (Benjamini and Y, 1995); however, only three of these are selected corresponding to HLA-B41, HLA-B47, and HLA-C17 all of which have frequencies between 0.5% and 1%. When computing the $p-$values based on the permutation distribution, we find an additional marker HLA-A26 is significantly associated with SI involvement having $p^{\ddagger} = 0.049$. Interestingly, when controlling for the false discovery rate based on the permutation $p-$values, no markers are selected.

For completeness, we estimate the effects of the identified HLA markers HLA-B41, HLA-B47 and HLA-C17 by fitting the three respective finite mixture models under the alternative. The contrasts that are parameterized naturally by the multinomial model are for unilateral versus no SI involvement and

Table 4: Results of applying the score test for each of the HLA-A, HLA-B, HLA-C, HLA-DR and HLA-DQB markers to the University of Toronto Psoriatic Arthritis Cohort; T_{obs} is the observed score statistics, $p^{\dagger}$ denotes p-value based on the χ^2_2 approximation, whereas $p^{\ddagger}$ denotes p-value based on the permutation distribution.

	T_{obs}	$p^{\dagger}$	$p^{\ddagger}$		T_{obs}	$p^{\dagger}$	$p^{\ddagger}$		T_{obs}	$p^{\dagger}$	$p^{\ddagger}$
HLA-A											
A1	1.762	0.414	0.371	A2	3.482	0.175	0.148	A3	7.152	0.028	0.033
A11	1.603	0.449	0.407	A24	0.823	0.663	0.626	A25	1.973	0.373	0.331
A26	5.881	0.053	0.049	A29	11.99	0.003	0.014	A30	1.784	0.410	0.367
A31	0.839	0.657	0.619	A32	3.199	0.202	0.170	A33	1.621	0.445	0.402
A68	1.278	0.528	0.482	A23	2.870	0.238	0.205	A28*	0.848	0.654	0.617
A34	0.550	0.760	0.728	A66	2.810	0.245	0.211	A69*	0.368	0.832	0.806
HLA-B											
B7	0.523	0.770	0.738	B8	1.015	0.602	0.556	B13	3.494	0.174	0.148
B14	3.182	0.204	0.171	B15	2.886	0.236	0.203	B62	1.916	0.384	0.343
B18	0.481	0.786	0.756	B27	8.086	0.018	0.026	B35	10.01	0.007	0.017
B37	0.801	0.670	0.634	B38	2.936	0.230	0.199	B39	3.314	0.191	0.161
B40	0.007	0.997	0.991	B44	0.669	0.716	0.682	B50	0.987	0.610	0.566
B51	0.221	0.895	0.878	B52	0.345	0.841	0.817	B55	0.668	0.716	0.683
B57	0.466	0.792	0.765	B58	0.291	0.865	0.843	B60	0.774	0.679	0.643
B61	0.319	0.853	0.829	B70	2.935	0.231	0.199	B41	23.53	$<$0.001	0.004
B45	2.741	0.254	0.219	B46	4.585	0.101	0.090	B47	22.53	$<$0.001	0.004
B48*	2.328	0.312	0.273	B49	3.893	0.143	0.124	B53	0.048	0.976	0.971
B56	1.768	0.413	0.370	B63	1.092	0.579	0.533	B67*	0.442	0.802	0.776
HLA-C											
C1	1.189	0.552	0.506	C2	9.095	0.011	0.021	C3	0.656	0.720	0.686
C4	7.316	0.026	0.031	C5	0.428	0.807	0.782	C6	2.562	0.278	0.243
C7	1.792	0.408	0.366	C8	1.754	0.416	0.373	C12	6.572	0.037	0.040
C14	0.235	0.889	0.871	C15	2.608	0.271	0.237	C16	0.822	0.663	0.627
C17	23.88	$<$0.001	0.004	C18*	1.609	0.558	0.404				
HLA-DR											
DR1	5.323	0.070	0.065	DR3	7.306	0.026	0.031	DR4	3.454	0.178	0.151
DR7	5.385	0.068	0.063	DR8	0.517	0.772	0.740	DR9	0.358	0.836	0.811
DR10	2.143	0.342	0.302	DR11	1.496	0.473	0.429	DR12	2.927	0.231	0.199
DR13	1.890	0.389	0.348	DR14	3.688	0.158	0.136	DR15	1.169	0.557	0.511
DR16	3.183	0.204	0.171								
HLA-DQB											
DQB201	5.665	0.059	0.055	DQB202	0.955	0.620	0.577	DQB301	5.252	0.072	0.067
DQB302	3.621	0.270	0.141	DQB303	1.828	0.410	0.359	DQB402	1.251	0.535	0.488
DQB501	4.099	0.129	0.112	DQB502	0.917	0.632	0.593	DQB503	1.438	0.487	0.442
DQB601	1.492	0.474	0.430	DQB602	0.543	0.762	0.731	DQB603	1.379	0.502	0.455
DQB604	2.448	0.294	0.259	DQB609	4.263	0.119	0.104	DQB299*	5.320	0.070	0.065
DQB305	1.028	0.598	0.552	DQB401	1.229	0.541	0.495	DQB605*	3.035	0.219	0.188

* markers of $<$0.5% in frequency

bilateral versus no SI involvement. None of these effects are significant, however, as the standard errors are larger for these comparisons. When assessing the effects on bilateral versus unilateral SI involvement based on the contrast $\gamma_3 - \gamma_1$, we find that, given some SI involvement, presence of HLA-B41, HLA-B27, and HLA-C17 each reduce the odds of bilateral (versus unilateral) involvement with OR=0.09 (95% CI: 0.02, 0.45; p=0.004), OR=0.09 (95% CI: 0.01, 1.00; p=0.050), and OR=0.09 (95% CI: 0.01, 1.00; p=0.050), respectively. These findings suggest these markers warrant further study with the ultimate goal of developing predictive models for SI joint involvement in PsA and particularly distinguishing between risk of psoriatic spondylitis and ankylosing spondylitis. Reliance on the permutation tests will not suggest any markers warrant further study.

The code for the computation of the score statistics and permutation p-values in the analyses is available from the first author upon request.

5 DISCUSSION

We have formulated a model for the finite mixture of Markov processes to accommodate latent classes of individuals who may experience different disease courses. We construct the observed data likelihood for the case when individuals are under intermittent observation and develop score tests for assessing the effect of the markers. This approach is especially convenient when markers are large in number since the model only needs to be fitted under the null hypothesis to assess the importance of the markers. We study the empirical performances of the proposed algorithm for model fitting and show that the coverage probabilities of confidence intervals are compatible with the nominal 95% level. Then, we study the type I error rate of the score test and show that the type I error is within the nominal 5% level for many settings. When the marker is rare and there may be concern about the adequacy of the χ^2 approximation, we describe how to conduct permutation tests of the marker effects that do not rely on asymptotic approximations. We also consider a multiple comparison procedure in the application to control the false discovery rate should that be of interest.

Harron et al. (2016) had a similar scientific objective to ours: to discover which HLA markers are associated with different types of SI involvement. Given the central role that disease duration plays in the development of joint damage, and under the assumption that HLA markers have a role in determining the nature of the disease manifestation, we feel that the model we developed provides a more natural basis for exploring these effects. Gaining insight into the power, or sample size requirements to meet power objectives, is important both in analyses based on observed disease status as in the work of Harron et al. (2016) and in the context of a finite mixture model that we develop. In practise, analyses are simply based on all of the data that is available, but if interest lies in a power or sample size calculation, the expected information matrix $\mathcal{I}(\theta) = E\{I(\theta)\}$ can be computed using Monte Carlo methods based on the observed information matrix in (9). Of course, this would require specification all parameters in the mixing distribution, the conditionally Markov intensities, the joint distribution of the covariates and the marker of interest, and the assessment time process. Having a sense of study power or sample size requirements to meet reasonable power objectives is important and we are exploring this issue both from the perspective of a standard design and a two-phase design (Lawless, 2018); the latter is appealing since the design could benefit from the available data in the first phase of such a study. This work is, however, beyond the scope of the current paper.

We have restricted attention to the case in which the transitional intensities for the unilateral classes (left or right) are the same as they are in the bilateral class. We also assume that the marker effects are the same for two unilateral classes. It is reasonable to assume these constraints on scientific grounds in this context, but in other settings, it may be desirable to relax these constraints to obtain more flexible models and test the plausibility of these assumptions. This may be feasible with larger datasets.

We have assumed that the assessment process satisfies a sequential missing at random assumption

of Hogan et al. (2004) If this assumption does not hold, joint models can be considered, or one can consider inverse intensity of visit weights (Lin et al., 2004) methods to correct for this since we are not exploring the effect of any time-varying covariates. An alternative approach would be to predict back involvement at a particular point in the disease process based on direct multinomial regression; in this case, weights are required to adjust for the selection bias arising from the need to restrict attention to individuals who can be definitively classified at the landmark time when assessing the predictive accuracy; see the work of Wu and Cook (2018).

ACKNOWLEDGEMENTS

This research was supported through grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849 and RGPIN 04027) and the Canadian Institutes for Health Research (FRN 13887) awarded to Richard Cook. Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research. The authors thank Dr. Dafna Gladman and Dr. Vinod Chandran for helpful discussions regarding the research at the Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

Aalen, O. (1987). Mixing distributions on a Markov chain. *Scandinavian Journal of Statistics*, 14(4):281–289.

Benjamini, Y. and Y, H. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300.

Boos, D. (1992). On generalized score tests. *Journal of the American Statistical Association*, 46:327 – 333.

Cook, R., Kalbfleisch, J., and Yi, G. (2002). A generalized mover-stayer model for panel data. *Biostatistics*, 3:407 – 420.

Cook, R. and Lawless, J. (2014). Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences*, 6(1):127–161.

Cook, R., White, B., Yi, G., Lee, K., and Warkentin, T. (2008). Analysis of a nonsusceptible fraction with current status data. *Statistics in Medicine*, 27(14):2715–2730.

Cook, R., Yi, G., and Lee, K. (2004). A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*, 60:436 – 443.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1 – 38.

Deng, D. and Paul, S. (2000). Score tests for zero inflation in generalized linear models. *Canadian Journal of Statistics*, 28:563 – 570.

Farewell, V. (1977). A model for a binary variable with time censored observations. *Biometrika*, 64:43 – 46.

Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041 – 1046.

Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of American Statistical Association*, 79:632 – 638.

Fuchs, C. and Greenhouse, J. (1988). The EM algorithm for maximum likelihood estimation in the mover stayer model. *Biometrics*, 44:605 – 613.

Geijer, M., Gadeholt Göthlin, G., and Göthlin, J. (2009). The validity of the New York radiological grading criteria in diagnosing sacroiliitis by computed tomography. *Acta Radiologica*, 50:664–673.

Gentleman, R., Lawless, J., Lindsey, J., and Yan, P. (1994). Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13:805 – 821.

Gladman, D., Antoni, C., Mease, P., Clegg, D., and Nash, P. (2005). Psoriatic arthritis: epidemiology, clinical features, course, and outcome. *Annals of the Rheumatic Diseases*, 64(suppl 2):ii14–ii17.

Gladman, D. and Chandran, V. (2010). Observational cohort studies: lessons learnt from the University of Toronto psoriatic arthritis program. *Rheumatology*, 50:25 – 31.

Grüger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47:595 – 605.

Hanly, J., Russell, M., and Gladman, D. (1988). Psoriatic spondyloarthropathy: a long term prospective study. *Ann. Rheumatol. Dis.*, 47:386 – 393.

Harron, M., Winchster, R., Giles, J., Heffernan, E., and FitzGerald, O. (2016). Certain class I HLA alleles and haplotypes implicated in susceptibility play a role in determining specific features of the psoriatic arthritis phenotype. *Clinical and Epidemiological Research*, 75:155 – 162.

Hogan, J., Roy, J., and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine*, 23:1455 – 1497.

Husted, J., Tom, B., Farewell, V., Schentag, C., and Gladman, D. (2007). A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: Does the effect change over time? *Arthritis and Rheumatology*, 56:840 – 849.

Jackson, C. (2011). Multi-state models for panel data: the *msm* package for R. *Journal of Statistical Software*, 38.

Jonas, S., Mbogning, C., Hässler, S., and Broët, P. (2017). A score test for comparing cross-sectional survival data with a fraction of non- susceptible patients and its application in clinical immunology. *PloS One*, 12:e0179896.

Kalbfleisch, J. and Lawless, J. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871.

Kuk, A. and Chen, C.-H. (1992). A mixed model combining logistic regression with proportional hazards regression. *Biometrika*, 79:531 – 541.

Lam, K. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika*, 92:573 – 586.

Lawless, J. (2018). Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis*, 24(1):28–44.

Lin, H., Scharfstein, D., and Rosenheck, R. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:791 – 813.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, 44:226–233.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London New York.

Moll, J. and Wright, V. (1973). Psoriatic arthritis. *Seminars in Arthritis and Rheumatism*, 3(1):55–78.

O'Keeffe, A., Tom, B., and Farewell, V. (2013). Mixture distributions in multi-state modelling: some considerations in a study of psoriatic arthritis. *Statistics in Medicine*, 32:600 – 619.

Peng, T., Dear, K., and Carrier, K. (2001). Testing for the presence of cured patients: a simulation study. *Statistics in Medicine*, 20:1783 – 1796.

Rahman, P., Gladman, D., Cook, R., Zhou, Y., Young, G., and Salonen, D. (1998). Radiological assessment in psoriatic arthritis. *British Journal of Rheumatology*, 37:760 – 765.

Riggs, B., Wahner, H., Dunn, W., Mazess, R., Offord, K., and Melton, L. (1981). Differential changes in bone mineral density of the appendicular and axial skeleton with ageing: relationship to spinal osteoporosis. *Journal of Clinical Investigation*, 67:328–335.

Satten, G. (1999). Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics*, 55:1228 – 1231.

Sutradhar, R. and Cook, R. (2008). Analysis of interval-censored data from clustered multistate processes: application to joint damage in psoriatic arthritis. *Journal of Royal Statistical Society (Series C)*, 57:553 – 566.

Sweeting, M., Angelis, D., Neal, K., Ramsay, M., Irving, W., Wright, M., Brant, L., Harris, H., and Trent HCV Study Group, and HCV National Register Steering Group (2006). Estimated progression rates in three united kingdom hepatitis c cohorts differed according to method of recruitment. *Journal of Clinical Epidemiology*, 59:144 – 152.

Sy, J. and Taylor, J. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, 56:227 – 236.

Tyas, S., Salazar, J., Snowdon, D., Desrosiers, M., Riley, K., Mendiondo, M., and Kryscio, R. J. (2007). Transitions to mild cognitive impairments, dementia, and death: findings from the nun study. *American Journal of Epidemiology*, 165:1231 – 1238.

Van den Broek, J. (1995). A score test for zero inflation in a poisson distribution. *Biometrics*, 51:738 – 743.

Wu, Y. and Cook, R. (2018). Variable selection and prediction in biased samples with censored outcomes. *Lifetime Data Analysis*, 24:72 – 93.

Yau, K., Wang, K., and Lee, A. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452.