Published in final edited form as:
Statistics in Medicine (2022), **41** (19): 3661–3678
DOI: 10.1002/sim.9441

The polytomous discrimination index for prediction involving multistate processes under intermittent observation

SHU JIANG

Division of Public Health Sciences,

Washington University School of Medicine, St. Louis, Missouri, USA

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: rjcook@uwaterloo.ca

Summary

With the increasing importance of predictive modeling in health research comes the need for methods to rigorously assess predictive accuracy. We consider the problem of evaluating the accuracy of predictive models for nominal outcomes when outcome data are coarsened at random. We first consider the problem in the context of a multinomial responsemodeled by polytomous logistic regression. Attention is then directed to the motivating setting in which class membership corresponds to the state occupied in a multistate disease process at a time horizon of interest. Here, class (state) membership may be unknown at the time horizon since disease processes are under intermittent observation. We propose a novel extension to the polytomous discrimination index to address this and evaluate the predictive accuracy of an intensity-based model in the context of a study involving patients with arthritis from a registry at the University of Toronto Centre for Prognosis Studies in Rheumatic Diseases.

Keywords: Classification, coarsening, discrimination, intermittent observation, multistate processes, predictive model, risk scores

1 Introduction

1.1 Introduction to Prediction

Predictive modeling is of increasing importance in the era of personalized and stratified medicine (Steyerberg, 2019). Much of the early work on methods for assessing prediction accuracy involved continuous outcomes where performance metrics include the proportion of explained variation (Schemper, 2003) and the "leave-one-out" analog called the PRESS statistic (Kutner et al., 2005), which better reflects out-of-sample performance. Any loss function can be specified of course, with the overall performance reflected by the expected loss. Harrell's concordance measure, called the C index, is another popular measure of performance which is geared towards assessing discriminatory power of a predictive model (Harrell et al., 1982). With dichotomous outcomes, point prediction yields a predicted response on the same scale as the response, while probabilistic prediction uses an estimated probability of the response. Point prediction has considerable appeal in medical research; discrimination measures based on misclassification rates and receiver operating characteristic curves are within this context (Hanley et al., 1989; Pepe, 2003), where the latter reflects of the utility of an underlying risk score for the classification of individuals. Since probabilistic prediction involves the use of estimated probabilities, the Brier score is natural to use when assessing predictive performance in this setting (Brier, 1950). Extensions of these measures have been proposed for right-censored data, where the goal is typically set to predicting the event status (failed/not failed) at some time horizon; due to censoring these extensions typically involve either imputation or use of inverse probability of censoring weights to address the fact that the failure status may be unknown for some individuals due to right-censoring (Uno et al., 2007, 2011).

Two general approaches are adopted for measuring predictive performance to deal with polytomous outcomes; the *hypervolume under the ROC manifold* (HUM) (Li et al., 2013) and the *polytomous discrimination index* (PDI) (Li et al., 2018; Van Calster et al., 2012). The former is a generalization of the area under the receiver operator characteristic curve. Consider a nominal outcome with K potential categories and a randomly selected K-tuple with one individual from each class. The HUM is the probability that the outcomes of all K individuals in this K-tuple are correctly classified. The term *volume under the surface* (VUS) is often used when the outcome has three categories (Mossman, 1999; Dreiseitl et al., 2000), where the term HUM is used when $K > 3$. The PDI for a particular category k is the probability that the subject in category k from a random K-tuple is correctly assigned to that category. This can be computed for each category k, and when these are averaged over all K categories an overall PDI is obtained.

We consider the challenge of measuring predictive performance based on multistate models for chronic disease processes which can be naturally characterized in terms of distinct stages (Cook and Lawless, 2018). Markov models are used routinely in such settings and considered here, where transition intensities are modulated by multiplicative covariate effects (Aalen et al., 2008; Andersen et al., 2012; Cook and Lawless, 2018). We suppose interest lies in predicting state occupancy at a particular time horizon based on a fitted multistate model. Examples of such problems are numerous, including prediction of nosocomial infections in the ICU and patient outcomes (Escolano et al., 2000), and prediction of outcomes following bone marrow transplantation (Keiding et al., 2001). Putter et al. (2006) report on a detailed analysis founded on a five-state model used to characterize disease course in breast cancer patients following surgery. States included on occupied when no events have occurred, and ones representing local recurrence, distant metastases, both local recurrence and distant metastases, and death; these authors also discuss the utility of multistate modeling for making predictions conditional on any observed history. More recently, Spitoni et al. (2018) discuss Brier score and Kullback-

Leibler type loss functions for predictions based on multistate process along with methods for estimating the corresponding expected loss functions under right-censoring – they then discuss extensions accommodating dynamic prediction.

The setting of interest involves a registry of patients attending a rheumatology clinic for periodic health assessments – at these clinic visits information is collected on the disease state; see Section 1.2 for full details of the motivating study. Specifically, we consider the problem of predicting state occupancy at a specified time horizon based on data arising from intermittent observation of a continuous-time multistate process. Intermittent observation of the processes makes it challenging to estimate the prediction accuracy of a model since it may not be known which state is occupied by some individuals at the time horizon of interest. We propose a novel extension to the PDI to accommodate such an intermittent observation scheme wherein the state occupied may be unknown for a subset of individuals in the validation sample. Our motivating application, described in detail under Section 1.2, involves the prediction of sacroiliac joint damage in patients in a psoriatic arthritis clinic, but there are many other clinical settings where this problem arises. In osteoporosis, for example, individuals are at risk of fractures (detected upon radiographic examination) due to weakened integrity of the bone. The development and evaluation of predictive models must deal with the fact that fracture status may be observed intermittently and so the event status at a particular time horizon may be unknown. Similarly in breast cancer prevention studies, individuals free of breast cancer are typically recruited but some may develop ductal carcinoma in situ, and ultimately progress to invasive breast cancer (Bergholtz et al., 2020); multistate processes can be used effectively to model this progression. Since individuals in prevention studies are screened periodically (i.e. annually or biannually) states are only known at the intermittent observation times of the disease process.

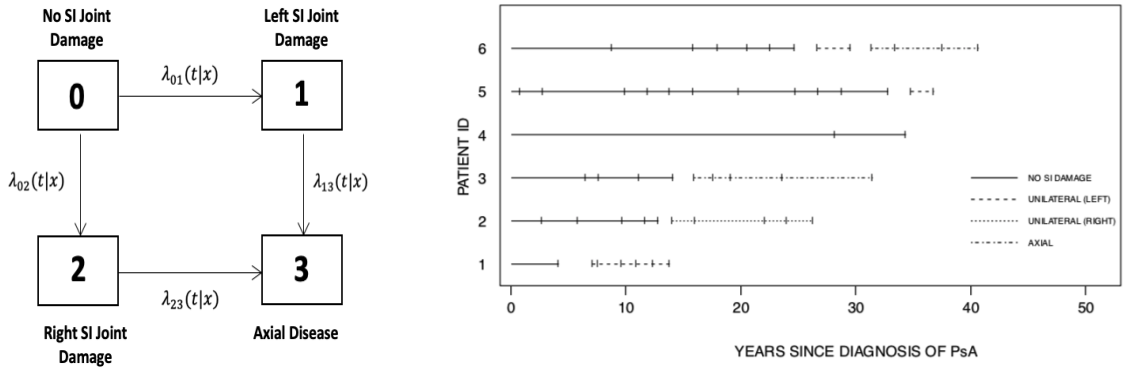
As a preliminary investigation, we considered a simplified version of the problem involving a simple categorical (polytomous) response with analyses based on a multinomial regression model. In this setting, we consider the problem where the validation sample does not report the categorical response for all individuals, but rather may simply indicate a set of possible categorical responses. We use the term coarsening to describe the general phenomenon whereby there is a loss of information about measurements; Heitjan and Rubin (1991) define coarsening to include "as special cases rounded, heaped, censored, partially categorized and missing data". Thus it is a slightly more general concept than what comes to mind from the term "missing data", and we use it here to encompass both the case where a set of possible categorical responses is reported rather than a single category, and the case where information is incomplete on state occupancy for a multistate process due to intermittent observation of a continuous-time process.

The remainder of the article is organized as follows. In Section 1.2, we describe the motivating problem involving data from the University of Toronto Psoriatic Arthritis Cohort where the goal is to predict different forms of sacroiliac joint damage in patients with psoriatic arthritis. In Section 2, we review the PDI for multinomial data and extend it to deal with coarsened data. The purpose of this section is to explore the impact of coarsening and methods for dealing with it in a simplified setting via multinomial representation. In Section 3, we introduce notation for the analysis of multistate processes under intermittent observation, and use the framework proposed in Section 2 to deal with prediction of the state occupied at a specified time horizon. Simulation studies are carried out to investigate the finite sample performance of the proposed estimation procedure under both the multinomial (Section 2) and the multistate (Section 3) settings. Section 4 involves an application to the motivating data where we use human leukocyte antigens to predict the presence of unilateral sacroiliac joint damage or axial disease in patients with psoriatic arthritis and estimate the PDI as a function of time. Concluding remarks and topics for future research are given in Section 5.

1.2 Prediction of Sacroiliac Involvement in Psoriatic Arthritis

The motivating problem arose in a collaboration with researchers at the Centre for Prognosis Studies in Rheumatic Disease in the University Health Network at the University of Toronto. These researchers maintain the Psoriatic Arthritis Cohort (UTPAC), founded in 1976 and now comprised of approximately 2000 patients. Upon recruitment, individuals provide biospecimens which are used for genetic testing and for proteomic analysis. Recruited patients are scheduled for annual clinic examination and biannual radiographic examination of joint damage. The broad aim of this cohort is to provide a platform for study of the clinical and radiological disease course, and to identify genetic and other types of risk factors for high disease activity and rapid progression of joint damage.

There has been considerable discussion in the rheumatology literature in recent years regarding the nature of spinal involvement in individuals with psoriatic arthritis. Ankylosing spondylitis is an arthritis condition with axial sacroiliac joint involvement, whereas other arthritic conditions tend to the development of unilateral sacroiliac joint damage. The aim of the current study is to predict spinal involvement in individual patients, and more specifically whether patients are likely to experience unilateral damage of the sacroiliac joints, or bilateral damage. The latter represents axial disease which is associated with greater pain and mobility impairment. Accurate prediction of axial disease is important as those at high risk may be given more intensive, potentially toxic and expensive, preventative therapy. Prediction of unilateral sacroiliac damage is also important; researchers speculate that this may represent a distinct disease process.



(a) State space diagram for onset of unilateral sacroiliac (SI) damage and axial disease.

(b) Timeline diagrams for visits and damage state for a sample of six patients.

Figure 1: Four state diagram for the onset of sacroiliac damage in psoriatic arthritis (panel (a)) and six sample timelines depicting data obtained from follow-up visits information on sacroiliac damage is acquired in patients from the University of Toronto Psoriatic Arthritis Cohort (panel (b)); gaps in timelines of panel (b) reflect periods during which the state is unknown.

We adopt a multistate model for the analysis. Figure 1(a) shows a simple four state model that can be used to characterize the onset of unilateral (left or right side) sacroiliac joint damage, as well as axial disease. The extent of joint damage is assessed using the New York Radiological Grading Criteria (Geijer et al., 2009) with damage defined here as grade 2 or higher. An individual free of sacroiliac damage makes a $0 \to k$ transition upon the onset of grade 2 or higher damage in the left ($k = 1$) or right ($k = 2$) sacroiliac joint, and enters state 3 upon the development of grade 2 or higher damage in their second sacroiliac joint. Figure 1(b) shows sample data for six individuals in the UTPAC with the length of the lines representing

how long they were under follow-up, and the vertical ticks representing visits at which x-rays are taken and damage can be assessed. We note that no transition times in Figure 1(a) are observed due to intermittent radiological assessments and hence we only know the state occupied at the times radiological assessments are made. Different line types in Figure 1(b) are used to depict the different states of sacroiliac damage; gaps are used to denote periods when the damage state is unknown.

2 Prediction with Multinomial Outcomes

2.1 The Polytomous Discrimination Index

Before considering the challenges involved in prediction and multistate processes under intermittent observation, we consider an analogous setting involving a coarsened multinomial random variable Y taking on the values $0, 1, 2, \ldots, K$. If $X = (X_1, \ldots, X_p)'$ is a $p \times 1$ covariate vector, let $P(Y = k|X) = \pi_k(X)$, $k = 1, \ldots, K$ and $P(Y = 0|X) = 1 - \sum_{k=1}^{K} \pi_k(X)$. Consider a multinomial regression model of the form

$$\log(\pi_k(X)/\pi_0(X)) = \bar{X}'\beta_k = \eta_k \ , \qquad k = 1, \ldots, K \ , \tag{1}$$

where $\bar{X} = (1, X')'$, $\beta_k = (\beta_{k0}, \ldots, \beta_{kp})'$ is $(p + 1) \times 1$ vector of regression coefficients, $k = 1, 2, \ldots, K$, and $\beta = (\beta_1', \ldots, \beta_K')'$ is $K(p+1) \times 1$ vector. We let $\eta_k = \bar{X}'\beta_k$ be a linear predictor associated with outcome k in (1), and $\eta = (\eta_1, \ldots, \eta_K)'$ denote the $K \times 1$ vector of linear predictors in which we suppress the notation for the dependence on X. For the purpose of predictive modeling, we refer to $\eta_1, \ldots, \eta_K$ as risk scores and η as the multivariate risk score with dimension K. We initially assume that β is known to focus on estimation of the PDI in an idealized setting, but investigate properties when β is estimated in Section 2.3. We consider the setting where the prediction for an individual with $X = x$ is $\hat{Y} = \underset{k}{\text{argmax}}\{\pi_k(x), k = 0, \ldots, K\}$. That is, the predicted outcome is the class with the highest probability of occurrence given $X = x$; alternative prediction rules can be adopted if additional costs or utilities are specified but in the absence of these we adopt this standard practice (Pepe, 2003).

Let $\mathcal{P}$ denote a population of interest and $\mathcal{P}_k = \{i : i \in \mathcal{P}, Y_i = k\}$ the sub-population of individuals in class k, $k = 0, 1, \ldots, K$. We then let $\boldsymbol{i} = \{i_0, i_1, \ldots, i_K\}$ denote a $(K + 1)$-tuple of individuals from $\mathcal{P}$ wherein $i_k \in \mathcal{P}_k$, $k = 0, 1, \ldots, K$, and let $\mathcal{P}^{(K+1)} = \{\boldsymbol{i} : i_j \in \mathcal{P}_j, \ j = 0, 1, \ldots, K\}$ be the set of all possible such $(K + 1)$-tuples. Next we let $\{X_{i_0}, X_{i_1}, \ldots, X_{i_K}\}$ denote the random set of covariate vectors associated with the $(K+1)$-tuple $\boldsymbol{i} \in \mathcal{P}^{(K+1)}$. Based on (1), let

$$\pi_k(x_{i_j}) = \frac{\exp(\bar{X}_{i_j}'\beta_k)}{1 + \sum_{l=1}^{K} \exp(\bar{X}_{i_j}'\beta_l)} \ , \tag{2}$$

be the conditional probability of a response in class k given covariate x_{i_j} which we denote more compactly as $\pi_{i_j k}$ in what follows. The PDI for category k, denoted by Δ_k, is defined as the probability that, among the individuals in a randomly selected $(K+1)$-tuple, the individual in class k is assigned to class k. To define this we first let

$$A_k(\boldsymbol{i}) = I(\pi_{i_k k} > \pi_{i_j k} \ , \ j \neq k, j = 0, \ldots, K) \tag{3}$$

indicate such an assignment for $\boldsymbol{i} \in \mathcal{P}^{(K+1)}$. Then

$$\Delta_k = E\{A_k(\boldsymbol{i})\} \ , \tag{4}$$

where the expectation is over $\{X_{i_0}, X_{i_1}, \ldots, X_{i_K}\}$. The dimension of the integration necessary to compute (4) can be reduced by working with the multidimensional risk scores since η represents a sufficient dimension reduction from X if $K < p$. Strong assumptions for the multivariate covariate distribution are required to compute Δ_k, so it is more commonly estimated empirically as we next describe.

Here we consider a validation sample $\mathcal{S}$ comprised of n independent individuals drawn at random from a target population for which the prediction model is to be applied. Let $\mathcal{S}_k = \{i : i \in \mathcal{S}, Y_i = k\}$ be the subset of individuals in the validation sample who are known to be in class k, where $|S_k| = n_k$, $k = 0, 1, \ldots, K$. In what follows $\boldsymbol{i} = (i_0, \ldots, i_K)$ is a vector of labels for a $(K+1)$-tuple of individuals constructed from the validation sample with $i_k \in \mathcal{S}_k$, $k = 0, 1, \ldots, K$; we let $\mathcal{T} = \{\boldsymbol{i} : i_k \in \mathcal{S}_k, k = 0, 1, \ldots, K\}$ denote the set of all $\prod_{k=0}^{K} n_k$ possible $(K+1)$-tuples based on the validation sample. An estimating function for Δ_k can then be defined as,

$$U_k(\Delta_k) = \sum_{\boldsymbol{i} \in \mathcal{T}} (A_k(\boldsymbol{i}) - \Delta_k) \ , \tag{5}$$

with solution

$$\widehat{\Delta}_k = \frac{1}{\prod_{k=0}^{K} n_k} \sum_{\boldsymbol{i} \in \mathcal{T}} A_k(\boldsymbol{i}) \ , \tag{6}$$

$k = 0, 1, \ldots, K$. The overall polytomous discrimination index Δ is defined as the simple average of the category-specific measures:

$$\widehat{\Delta} = \frac{1}{K} \sum_{k=0}^{K} \widehat{\Delta}_k \ . \tag{7}$$

2.2 Estimation with Coarsened Validation Data

We now consider the case in which the true class membership is unknown for some individuals in the validation samples due to coarsening. Methods for dealing with coarsened data are well developed, and we do not consider the formation of a predictive model but rather how to evaluate predictive accuracy in terms of the PDI when responses are only known to be in one of a set of classes. Let $\mathcal{C}_i$ be the coarsened response for individual i where, if $K = 2$ for example, $\mathcal{C}_i \in \{0, 1, 2, (0,1), (0,2), (1,2), (0,1,2)\}$ and the first three elements 0, 1 and 2 are realized when there is no coarsening. We omit the noninformative outcomes defined as those that have probability one (e.g. $\mathcal{C}_i = (0,1,2)$ when $K = 2$).

As in Section 2.1, we use a subscript on individual labels to denote the class they are in, but here we introduce a superscript p to indicate that these may be pseudo-individuals who are conceptualized to represent the possible class membership of an individual whose response is coarsened. For example if $\mathcal{C}_i = (j, k)$, then there are two pseudo-individuals associated with individual i, with one pseudo-individual assigned to class j and another to class k; we label these pseudo-individuals i_j^p and i_k^p, respectively, but note that the values of i_j^p and i_k^p are equal to i —- the subscript represents the class considered for a particular allocation. We further let $\mathcal{S}_k^p = \{i : i \in \mathcal{S}, k \in \mathcal{C}_i\}$ be the set of individuals who are known to be in class k (i.e. if $\mathcal{C}_i = k$) or may be in class k (i.e. if $k \in \mathcal{C}_i$), $k = 0, 1, 2$. To unify the notation, we label all individuals in $\mathcal{S}_k$ by i_k^p whether it is known that $Y_i = k$ or we simply know $Y_k \in \mathcal{C}_i$. Note that $I(|\mathcal{C}_i| = 1)$ indicates the outcome for individual i is observed precisely in which case $i_k^p = i$ if $Y_i = k$. If $\mathcal{C}_i = (j, k)$, then there is a pseudo-individual $i_j^p \in \mathcal{S}_j^p$ and a pseudo-individual $i_k^p \in \mathcal{S}_k^p$. More generally if $|\mathcal{C}_i| = m$, then there will be m pseudo-individuals corresponding to individual i, with each one belonging to one of m different sets $\mathcal{S}_l^p$, $l \in \mathcal{C}_i$. Following this construction, we let $\mathcal{T}^p = \{\boldsymbol{i}^p : i_k^p \in \mathcal{S}_k^p, k = 0, 1, \ldots, K\}$ denote the set of all possible $(K+1)$-tuples based on

the pseudo-individuals conceptualized corresponding to the coarsened validation sample. We assume coarsening at random in the sense of Heitjan and Rubin (1991).

Let

$$w_{ik} = P(Y_i = k|\mathcal{C}_i, X_i) = \frac{P(Y_i = k|X_i)}{\sum_{j\in\mathcal{C}_i} P(Y_i = j|X_i)} \ , \tag{8}$$

be the conditional probability individual i is in class k given their coarsened response $\mathcal{C}_i$, $k = 0, 1, 2$. If $\mathcal{C}_i = k$ then $w_{ik} = 1$ and $w_{ij} = 0$ for $j \neq k$, $k = 0, 1, 2$. We let $\boldsymbol{i}^p = (i_0^p, i_1^p, \ldots, i_K^p)$ represent a $(K + 1)$-tuple of individuals or pseudo-individuals where $i_j^p \in \mathcal{S}_j^p$ and let $D(\boldsymbol{i}^p) = I(i_j^p \neq i_k^p, j \neq k, j, k = 0, 1, \ldots, K)$ is the indicator that this $(K + 1)$-tuple of potential pseudo-individuals is comprised of distinct real individuals.

Next we let

$$A_k(\boldsymbol{i}^p) = I(\pi_{i_k^p k} > \pi_{i_j^p k} \ , \ j \neq k, \ j = 0, \ldots, K) \tag{9}$$

be the indicator that the (pseudo) individual from class k in $\boldsymbol{i}^p$ has highest predictive probability of being in class k, and define the estimating function for Δ_k as,

$$\bar{U}_k(\Delta_k) = \sum_{\boldsymbol{i}^p\in\mathcal{T}^p} D(\boldsymbol{i}^p)\{w(\boldsymbol{i}^p)(A_k(\boldsymbol{i}^p) - \Delta_k)\} \ , \tag{10}$$

where $w_i(\boldsymbol{i}^p)$ is the product $w_{i_0^p 0} w_{i_1^p 1} w_{i_2^p 2}$. Note that in the absence of coarsening,

$$\mathcal{S}_j^p \cap \mathcal{S}_k^p = \emptyset \quad \text{for} \ \ \forall\, j \neq k, \tag{11}$$

$D(\boldsymbol{i}^p) = 1$ and $w(\boldsymbol{i}^p) = 1$, and we retrieve the standard estimator of Δ_k given in equation (6). More generally, we obtain

$$\widehat{\Delta}_k = \frac{\sum_{\boldsymbol{i}^p\in\mathcal{T}^p} D(\boldsymbol{i}^p) w(\boldsymbol{i}^p) A_k(\boldsymbol{i}^p)}{\sum_{\boldsymbol{i}^p\in\mathcal{T}^p} D(\boldsymbol{i}^p) w(\boldsymbol{i}^p)} \ , \tag{12}$$

and we again estimate the overall polytomous discrimination index as $\widehat{\Delta} = \sum_{k=0}^{K} \widehat{\Delta}_k/(K + 1)$.

2.3 Simulation Studies Involving Coarsened Multinomial Data

Here we report on the results of simulation studies in which we focus on estimation of the PDI measure described in Sections 2.1 and 2.2. We consider three classes ($K = 2$) in this simulation study and express the class probabilities given the covariates as in (1). We set $p = 2$ and adopt a covariate model with $X \sim BVN(\mu, \Sigma)$ with $\mu = (\mu_1, \mu_2)'$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix} \ ; \tag{13}$$

we set $\mu = (0, 0)'$ and $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.4$. The regression coefficients satisfy $(\beta_{21}, \beta_{22})' = (2\beta_{11}, 2\beta_{12})'$ and we set $(\beta_{11}, \beta_{12})' = (\log 1.5, \log 2)'$ to represent moderate covariate effects, and $(\beta_{11}, \beta_{12})' = (\log 3, \log 4)'$ for stronger covariate effects. The intercepts β_{10} and β_{20} are chosen to give pre-specified marginal probabilities $P(Y = 0) = 0.4$, $P(Y = 1) = 0.4$ and $P(Y = 2) = 0.2$.

To avoid high dimensional integration in (4) to determine Δ_k, $k = 0, 1, 2$ and Δ, we use Monte Carlo methods to approximate them by simulating a dataset of one million individuals with complete covariate values and class membership; the resulting numerical values are reported under the column headed "Value" in Table 1. To examine the validity of the weighted estimating function for coarsened data, we consider two approaches to estimation of Δ_k based on (12). As a first pass, we use the true β value and directly evaluate the PDI in validation samples of 500 individuals. The second approach is to estimate β from an independent training sample of 1000 individuals with coarsening, using an EM algorithm (Dempster et al., 1977)

for estimation and use the resulting estimate of β to estimate the PDI in a validation sample of 500 individuals. Here if $\mathcal{D}$ is the set of indices labeling individuals in the training data and coarsening is at random, the maximum likelihood estimate $\hat{\beta}$ maximizes the observed data loglikelihood

$$\ell(\beta) = \sum_{i \in \mathcal{D}} \log P(Y_i \in \mathcal{C}_i | x_i)$$

where $P(Y_i \in \mathcal{C}_i|x_i) = \sum_{k \in \mathcal{C}_i} \pi_k(x_i; \beta)$. We then use $\hat{A}_k(\boldsymbol{i}^p)$ obtained from (9) with the estimate of β used to estimate the classification probabilities. The third approach involves estimating β based on a complete case analysis and using the resulting estimate $\widetilde{\beta}$ to compute the PDI. All three approaches are evaluated under varying degree of coarsening: 0%, 30%, and 60% where the percentages correspond to the probabilities of coarsening in the sample. In the absence of coarsening $\hat{\beta} = \tilde{\beta}$ so there is only one set of results for this setting. We specify $\mathcal{C}_i \perp Y_i | Y_i \in \mathcal{C}_i, X_i$ for coarsening at random and generate the coarsened data such that $P(|\mathcal{C}_i| \neq 1) = 0.3$ for moderate coarsening and 0.6 for more severe coarsening, $i = 1, \ldots, n$. If $|\mathcal{C}_i| = 2$ and $Y_i = 1$, for example, we consider $\mathcal{C}_i = (Y_i, j)$ with $j = 0$ or $j = 2$ to define the possible coarsening for this individual; we choose $\mathcal{C}_i = (0, 1)$ or $\mathcal{C}_i = (1, 2)$ with equal probability.

The results of the simulation study involving 1000 replicates are displayed in Table 1 where the mean estimate is reported under EST and we provide the empirical standard error (ESE), the average bootstrap standard error from 500 bootstrap samples (ASE), the empirical coverage probability of confidence intervals constructed directly on the scale of the PDI (ECP), and the corresponding ECP for confidence intervals constructed based on the logit transformation of the polytomous discrimination indices (ECP‡). The proposed weighted estimating function yields estimators with low empirical bias for all settings with this good performance maintained for the higher degrees of coarsening. We also see that when the training sample involves coarsened data and an EM algorithm is used for estimation (Dempster et al., 1977), there remains small empirical bias; the empirical standard error of the estimators increases with the increased degree of coarsening. The empirical bias of the estimator based on a complete case analysis is very small, as one would expect with data coarsened at random, but the associated empirical standard error is greater. Also as expected, there is also a larger PDI with stronger covariate effects. Additional simulation studies with 500 and 2000 individuals in the training sample lead to similar conclusions – these are reported on in Section S1.1 of the Supplemental Material.

3 Prediction with Multistate Processes Under an Intermittent Observation Scheme

3.1 Notation and model formulation

We now consider a multistate disease process with $K + 1$ states labeled $0, 1, \ldots, K$ where K is an absorbing state. Let $Z(t)$ denote the state occupied at time t and $\{Z(s), 0 < s\}$ denote the associated stochastic process. We consider a $q_2 \times 1$ covariate vector X and let $\mathcal{H}(t) = \{Z(s), 0 < s < t; X\}$ denote the history of the process at time t. The stochastic nature of the multistate process can be fully characterized via the transition intensities (Cook and Lawless, 2018) for all pairs of states where,

$$\lim_{\Delta t \downarrow 0} \frac{P(Z(t + \Delta t^-) = l | Z(t^-) = k, \mathcal{H}(t))}{\Delta t} = \lambda_{kl}(t | \mathcal{H}(t)) \,, \tag{14}$$

for $k, l = 0, 1, \ldots, K$, $k \neq l$, where t^- denotes an infinitesimal amount of time before t. For simplicity we assume that the same vector of covariates are used to model all transition intensities and restrict attention to Markov processes for which covariates act multiplicatively on

Table 1: Empirical performance of estimates of Δ_k, $k = 0, 1, 2$ and Δ with no, moderate (30%), and heavier (60%) coarsening.

			Percentage of Individuals with Coarsened Observations														
			0%					30%					60%				
Parameter	Value	Method†	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡
moderate covariate effects; $(\beta_{11}, \beta_{12})' = (\log 1.5, \log 2)', (\beta_{21}, \beta_{22})' = (2\beta_{11}, 2\beta_{12})'$																	
Δ_0	0.670	True	0.665	0.028	0.029	0.967	0.956	0.665	0.024	0.024	0.949	0.943	0.663	0.022	0.023	0.961	0.954
		EM	0.665	0.028	0.030	0.961	0.956	0.665	0.030	0.031	0.950	0.953	0.666	0.034	0.032	0.943	0.942
		CC						0.666	0.032	0.033	0.954	0.941	0.669	0.045	0.044	0.947	0.949
Δ_1	0.404	True	0.405	0.032	0.032	0.953	0.954	0.402	0.029	0.030	0.960	0.951	0.400	0.024	0.023	0.944	0.952
		EM	0.406	0.033	0.031	0.926	0.930	0.407	0.031	0.030	0.941	0.938	0.411	0.031	0.029	0.934	0.946
		CC						0.408	0.037	0.037	0.952	0.950	0.418	0.052	0.051	0.947	0.947
Δ_2	0.677	True	0.663	0.034	0.033	0.944	0.946	0.663	0.037	0.038	0.958	0.949	0.662	0.025	0.026	0.959	0.951
		EM	0.664	0.034	0.032	0.934	0.941	0.665	0.037	0.036	0.958	0.954	0.665	0.045	0.044	0.942	0.945
		CC						0.664	0.043	0.044	0.957	0.952	0.671	0.056	0.056	0.950	0.955
Δ	0.583	True	0.578	0.023	0.022	0.951	0.957	0.576	0.020	0.019	0.941	0.949	0.575	0.018	0.019	0.955	0.954
		EM	0.578	0.031	0.030	0.962	0.957	0.579	0.033	0.031	0.963	0.953	0.580	0.037	0.035	0.940	0.951
		CC						0.579	0.037	0.036	0.947	0.948	0.586	0.040	0.039	0.954	0.943
strong covariate effects; $(\beta_{11}, \beta_{12})' = (\log 3, \log 4)', (\beta_{21}, \beta_{22})' = (2\beta_{11}, 2\beta_{12})'$																	
Δ_0	0.828	True	0.831	0.021	0.021	0.952	0.949	0.832	0.018	0.019	0.961	0.953	0.832	0.016	0.016	0.944	0.956
		EM	0.832	0.021	0.022	0.955	0.952	0.832	0.022	0.023	0.958	0.956	0.833	0.025	0.026	0.948	0.959
		CC						0.833	0.024	0.023	0.953	0.942	0.834	0.032	0.031	0.959	0.945
Δ_1	0.553	True	0.566	0.034	0.033	0.943	0.951	0.565	0.031	0.032	0.948	0.956	0.564	0.028	0.028	0.956	0.947
		EM	0.567	0.033	0.033	0.946	0.942	0.569	0.034	0.032	0.942	0.951	0.569	0.036	0.035	0.948	0.954
		CC						0.567	0.039	0.038	0.948	0.941	0.574	0.053	0.053	0.942	0.956
Δ_2	0.821	True	0.825	0.026	0.027	0.962	0.957	0.825	0.023	0.023	0.944	0.958	0.823	0.020	0.019	0.943	0.948
		EM	0.825	0.025	0.026	0.964	0.951	0.826	0.028	0.029	0.960	0.954	0.826	0.033	0.033	0.946	0.950
		CC						0.825	0.032	0.033	0.943	0.952	0.828	0.040	0.040	0.959	0.952
Δ	0.734	True	0.741	0.021	0.020	0.947	0.953	0.741	0.019	0.019	0.945	0.952	0.740	0.016	0.017	0.948	0.957
		EM	0.742	0.026	0.027	0.954	0.954	0.743	0.028	0.029	0.946	0.950	0.743	0.031	0.030	0.946	0.951
		CC						0.742	0.030	0.030	0.951	0.957	0.745	0.035	0.034	0.943	0.945

Note: The ASE is the average of bootstrap standard errors based on 500 bootstrap samples created for each simulated data; the ECP is the empirical coverage probability of nominal 95% confidence intervals constructed based on the normal approximation of the estimator using the bootstrap standard error while ECP‡ is the corresponding empirical coverage probability when the confidence intervals are constructed based on the logit transformation; training samples are of 1000 observations; validation samples involve 500 individuals; $nsim = 1000$.

† Prediction is based on true parameter values (True), as well as maximum likelihood estimates based on an expectation-maximization algorithm (EM) and complete-case analysis (CC); the "EM estimate" is the usual MLE obtained by fitting a standard multinomial regression model when there is no coarsening.

baseline transition intensities via

$$\lambda_{kl}(t|\mathcal{H}(t)) = \lambda_{kl}(t)\exp(X'\beta_{kl}) ,$$

with $\beta_{kl} = (\beta_{kl1}, \ldots, \beta_{klq_2})'$ a $q_2 \times 1$ vector of regression coefficients for $k \to l$ transitions (Andersen et al., 2012). If α_{kl} is a $q_1 \times 1$ parameter vector indexing $\lambda_{kl}(t)$ and $q = q_1 + q_2$, then the $k \to l$ transition intensity is indexed by the $q \times 1$ parameter vector $\theta_{kl} = (\alpha'_{kl}, \beta'_{kl})'$ and θ is the full vector containing all θ_{kl} for $k \neq l = 0, 1, \ldots, K$. We adopt a common dimension for the parameters indexing the different baseline intensities, but this is for notational convenience and can be easily relaxed.

In what follows we consider a four state process, illustrated in Figure 1 (a), with $0 \to 1$, $0 \to 2$, $1 \to 3$, and $2 \to 3$ transitions possible. The 4×4 transition probability matrix $\mathbb{P}(s, t|x)$ can be computed by product integration as discussed in Section 2.2 of Cook and Lawless (2018). We let $\mathbb{Q}(t|x)$ denote the 4×4 matrix of cumulative transition intensities

$$\mathbb{Q}(t|x) = \begin{pmatrix} -\Lambda_{01}(t|x) - \Lambda_{02}(t|x) & \Lambda_{01}(t|x) & \Lambda_{02}(t|x) & 0 \\ 0 & -\Lambda_{13}(t|x) & 0 & \Lambda_{13}(t|x) \\ 0 & 0 & -\Lambda_{23}(t|x) & \Lambda_{23}(t|x) \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

where $\Lambda_{kl}(t|x) = \int_0^t \lambda_{kl}(s|x)ds$. Then let $d\mathbb{Q}(t|x)$ be a 4×4 matrix with (k, l) entry $d\Lambda_{kl}(t|x) = \lambda_{kl}(t|x)dt$ if $k \neq l$ and $-d\Lambda_{k\cdot}(t|x)$ if $k = l$ with "$\cdot$" representing summation over the corresponding index. Then if $\mathbb{I}$ is a 4×4 identity matrix the transition probability matrix $\mathbb{P}(s, t|x)$ with (k, l) entry

$$p_{kl}(s, t|x) = P(Z(t) = l|Z(s) = k, x) ,$$

is obtained by

$$\mathbb{P}(s, t|x) = \prod_{(s,t]} \{\mathbb{I} + d\mathbb{Q}(t|x)\} .$$

We now consider a sample of n individuals under intermittent observation labeled $i = 1, \ldots, n$ and let $0 = a_{i0} < a_{i1} < \cdots < a_{im_i}$ denote the m_i visit times at which data are available for individual i. To formalize this observation process, we let τ_i denote a loss to follow-up time and $Y_i(t) = I(t \leq \tau_i)$ indicate that individual i is still on study. We also let $dA_i(s) = 1$ if individual i has a visit at time s with $dA_i(s) = 0$ otherwise. Let $A_i(t) = \int_0^t Y_i(s)dA_i(s)$ record the cumulative number of visits over $(0, t]$, and let $\{A_i(s), 0 < s\}$ denote the counting process for visits which is terminated upon censoring. We let $Z_i(a_{i0}) = 0$ with probability 1 for $a_{i0} = 0$, and let the observed process history be denoted by $\bar{\mathcal{H}}_i(t) = \{Y_i(s), dA_i(s), 0 < s < t, (Z_i(a_{im}), a_{ir}), r = 0, 1, \ldots, A_i(t^-)\}$. We assume a conditionally independent and non-informative censoring time and that the visit process is conditionally independent in the sense of Cook and Lawless (2018, 2021) – this is akin to the sequential missing at random assumption characterized by Hogan et al. (2004) for longitudinal data with drop-outs. Then under a Markov model the observed data partial likelihood is

$$L(\theta) = \prod_{i=1}^{n} \prod_{r=1}^{m_i} P(Z_i(a_{ir})|Z_i(a_{i,r-1}), X_i) , \tag{15}$$

and maximum partial likelihood estimates can be obtained by a Fisher-scoring algorithm (Kalbfleisch and Lawless, 1985) or direct maximization using the `msm` function (Jackson, 2011). This is a brief discussion of the likelihood construction for Markov processes under intermittent observation – predictive models can be constructed based on this likelihood by simple model fitting if covariates are specified, or through use of penalization. Our primary interest however, is to discuss assessment of predictive accuracy based on any particular predictive model based on a validation sample. We describe how to do this in the next section.

3.2 Estimating predictive accuracy with coarsened multistate data

We now consider the problem in which interest lies in predicting state occupancies for an individual at a time horizon denoted by $t_\circ > 0$. If we consider a 4-tuple $\boldsymbol{i} = (i_0, \ldots, i_3)$ where individual i_j is in state j at $t_\circ$, let $\pi_{i_jk}(t_\circ) = p_{0k}(0, t_\circ | X_{i_j})$ be the conditional probability that an individual with their covariate vector is in state k at $t_\circ$. We then define

$$A_k(\boldsymbol{i}; t_\circ) = I(\pi_{i_kk}(t_\circ) > \pi_{i_jk}(t_\circ),\ j \neq k, j = 0, \ldots, 3) \tag{16}$$

and define the PDI for category k at $t_\circ$ as $\Delta_k(t_\circ) = E\{A_k(\boldsymbol{i}; t_\circ)\}$ where again the expectation is take over the distributions of the covariate vectors for members of the 4-tuple.

Since the continuous-time multistate disease process is under intermittent observation, the state occupied at $t_\circ$ will be unknown for individuals who were censored in a transient state prior to $t_\circ$ and those whose recorded states at visits immediately before and after $t_\circ$ differ. The observed data for individual i is denoted by $\bar{\mathcal{H}}_i(\infty)$ with the key elements being $\mathcal{D}_i = \{(Z_i(a_{ir}), a_{ir}), r = A_i(t_\circ^-), A_i(t_\circ^-) + 1, X_i\}$ under the assumptions of Section 3.1. Note that if $A_i(t_\circ^-) = m_i$ then we let $a_{i,m_i+1} = \infty$ and $Z_i(a_{i,m_i+1}) = 3$. As in Section 2.2, we use a subscript to label individuals according to the state they are in at $t_\circ$ with the superscript p used to indicate that these may be pseudo-individuals, conceptualized to represent the represent all possible states occupied by an individual when their true state is unknown.

The weight $w_{ij}(t_\circ) = P(Z_i(t_\circ) = j | \mathcal{D}_i)$ under the multistate process is then given by,

$$w_{ij}(t_\circ) = \frac{P(Z_i(t_\circ) = j | Z_i(a_{iA_i(t_\circ^-)}), X_i) P(Z_i(a_{i,A_i(t_\circ^-)+1}) | Z_i(t_\circ^-) = j, X_i)}{P(Z_i(a_{i,A_i(t_\circ^-)+1}) | Z_i(a_{iA_i(t_\circ^-)}), X_i)} . \tag{17}$$

We use $\mathcal{S}_k^p(t_\circ)$ to denote the set of pseudo-individuals who may occupy state k at $t_\circ$, $k = 0, 1, 2, 3$. Moreover, we let $\boldsymbol{i}^p = (i_0^p, i_1^p, i_2^p, i_3^p)'$ represent a 4-tuple of individuals or pseudo-individuals where $i_j^p \in \mathcal{S}_j^p(t_\circ)$ and let $\mathcal{T}^p(t_\circ) = \{\boldsymbol{i}^p : i_k^p \in \mathcal{S}_k^p(t_\circ), k = 0, 1, 2, 3\}$ denote the set of all possible 4-tuples at $t_\circ$. We let $\pi_{i_k^p k}(t_\circ) = p_{0k}(0, t_\circ | X_{i_k^p})$ and define

$$A_k(\boldsymbol{i}^p; t_\circ) = I(\pi_{i_k^p k}(t_\circ) > \pi_{i_j^p k}(t_\circ),\ j \neq k, j = 0, \ldots, 3) \tag{18}$$

as the indicator that the (pseudo) individual from class k among the 4-tuple has highest predictive probability of being in class k. We then define the weighted estimating function for $\Delta_k(t_\circ)$ as,

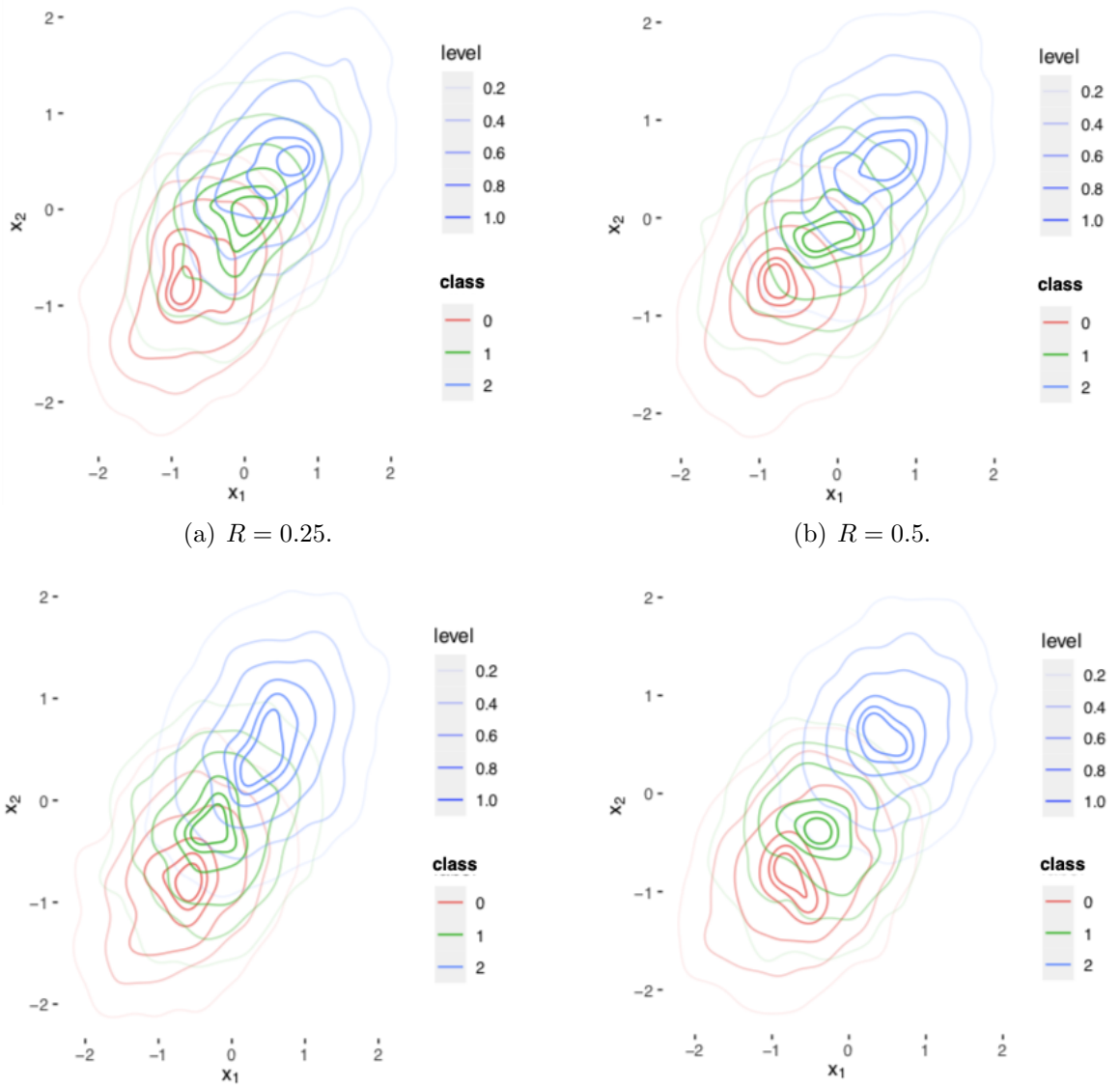
$$\bar{U}_k(\Delta_k(t_\circ); t_\circ) = \sum_{\boldsymbol{i}^p \in \mathcal{T}^p} D(\boldsymbol{i}^p; t_\circ)\{w_i(\boldsymbol{i}^p; t_\circ)(A_k(\boldsymbol{i}^p; t_\circ) - \Delta_k(t_\circ))\} , \tag{19}$$

where $D(\boldsymbol{i}^p; t_\circ) = I(\boldsymbol{i}^p \in \mathcal{T}^p(t_\circ), i_j^p \neq i_k^p, j \neq k, j, k = 0, 1, 2, 3)$ is the indicator that this 4-tuple is comprised of distinct individuals and $w_i(\boldsymbol{i}^p; t_\circ) = \prod_{k=0}^3 w_{i_k^p k}(t_\circ)$. The estimate $\widehat{\Delta}_k(t_\circ)$ is the solution to $D(\boldsymbol{i}^p; t_\circ) = 0$ and an overall PDI denoted by $\widehat{\Delta}(t_\circ)$ can be estimated by averaging the $K + 1$ category-specific PDI values obtained at $t_\circ$.

3.3 Simulation Studies Involving Multistate Processes

To mimic the data from the motivating study, we specify the parameter setting as follows. The transition intensities are set under the constraint $\lambda_{01} = \lambda_{02}$ so that the baseline intensity for the onset of unilateral damage is the same for the left and right SI joints, and $\lambda_{13} = \lambda_{23} = 2\lambda_{01}$ so that the intensity for the onset of axial disease is twice as high as it was for the onset of unilateral damage. We consider covariates $X = (X_1, X_2)'$, with $X \sim BVN(\mu, \Sigma)$, where $\mu = (\mu_1, \mu_2)'$ and Σ in a similar fashion as (13) with $\mu_k = 0$, $\sigma_k^2 = 1$, $k = 1, 2$, and $\rho = 0.5$. We

set $P(Z(t_\circ) = 3|Z(0) = 0; X = 0) = 0.5$ at $t_\circ = 1$, so the prevalence of axial disease at time $t_\circ$ is 0.5 among individuals with $X_1 = X_2 = 0$. Further, we set $\beta_{01} = \beta_{02}$ so that the covariates have the same effect for the onset of unilateral damage, and $\beta_{13} = \beta_{23}$ so that the covariates have the same effect for the onset of axial disease among those individuals with unilateral damage. Specifically, we let $\beta_{01} = \beta_{02} = (\log 1.5, \log 2.0)'$ and $\beta_{13} = \beta_{23} = R\beta_{01}$, where $R = 0.25, 0.5, 1.0$, and 2.0. Given the covariates and transitional intensities, we can then generate the multistate data $Z(s), 0 < s < 2|x$.



(a) $R = 0.25$. (b) $R = 0.5$.

(c) $R = 1$. (d) $R = 2$.

Figure 2: Contour plots of the empirical covariate distributions by class membership for the four simulation scenarios under the multistate processes with random samples of 10,000 individuals.

For the visit process we consider the follow-up period of 2 units duration, and set the time horizon for prediction to $t_\circ = 1$. We adopt a time homogeneous Poisson process for the visit times with rate $\rho = 5$ or 10 giving $A(2) \sim$ Poisson(mean 10 or 20) with $a_{i1} < \cdots < a_{im_i}$ the realized visit times. Given the intermittent observation process, the observed data is thus composed of $\{(a_{im}, Z_i(a_{im})), m = 0, 1, \ldots, m_i, X_i\}$, for $i = 1, \ldots, n$. In line with our motivating application, we are interested in estimating the prediction accuracy for state occupancy for unilateral damage (states 1 or 2), regardless of sides, as well as the state occupancy for axial

Table 2: Finite sample properties of estimates of $\Delta_k(t_\circ)$, $k = 0, 1, 2$ and $\Delta_k(t_\circ)$ with $t_\circ = 1$, for different values of R with an average of 10 or 20 visits over the period $(0, 2]$.

			$E(M) = 10$					$E(M) = 20$				
Parameter	Value	Method†	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡
			$\beta_{01} = \beta_{02} = (\log 1.5, \log 2.0)'; \beta_{13} = \beta_{23} = R\beta_{01}, R = 0.25$									
$\Delta_0(t_\circ)$	0.708	True	0.709	0.029	0.030	0.962	0.957	0.709	0.029	0.030	0.958	0.959
		MLE	0.708	0.029	0.031	0.952	0.950	0.706	0.029	0.029	0.951	0.959
$\Delta_1(t_\circ)$	0.416	True	0.428	0.029	0.028	0.959	0.957	0.429	0.031	0.030	0.940	0.943
		MLE	0.419	0.031	0.030	0.956	0.954	0.418	0.030	0.031	0.935	0.935
$\Delta_2(t_\circ)$	0.651	True	0.637	0.026	0.026	0.941	0.943	0.639	0.027	0.028	0.962	0.958
		MLE	0.638	0.027	0.028	0.942	0.940	0.637	0.027	0.027	0.946	0.951
$\Delta(t_\circ)$	0.592	True	0.591	0.021	0.020	0.942	0.949	0.592	0.021	0.020	0.942	0.943
		MLE	0.588	0.020	0.018	0.932	0.941	0.587	0.020	0.020	0.940	0.945
			$\beta_{01} = \beta_{02} = (\log 1.5, \log 2.0)'; \beta_{13} = \beta_{23} = R\beta_{01}, R = 0.5$									
$\Delta_0(t_\circ)$	0.697	True	0.690	0.030	0.031	0.952	0.946	0.689	0.030	0.031	0.953	0.959
		MLE	0.690	0.030	0.029	0.938	0.944	0.691	0.030	0.031	0.940	0.958
$\Delta_1(t_\circ)$	0.435	True	0.431	0.032	0.031	0.940	0.954	0.432	0.032	0.032	0.957	0.955
		MLE	0.427	0.031	0.030	0.942	0.939	0.428	0.031	0.032	0.962	0.945
$\Delta_2(t_\circ)$	0.702	True	0.703	0.025	0.026	0.956	0.957	0.702	0.026	0.026	0.941	0.954
		MLE	0.702	0.024	0.025	0.964	0.954	0.701	0.023	0.026	0.951	0.957
$\Delta(t_\circ)$	0.613	True	0.608	0.021	0.019	0.954	0.943	0.608	0.021	0.020	0.950	0.947
		MLE	0.606	0.020	0.019	0.958	0.950	0.607	0.020	0.021	0.960	0.958
			$\beta_{01} = \beta_{02} = (\log 1.5, \log 2.0)'; \beta_{13} = \beta_{23} = R\beta_{01}, R = 1$									
$\Delta_0(t_\circ)$	0.674	True	0.659	0.032	0.031	0.942	0.943	0.659	0.032	0.033	0.952	0.951
		MLE	0.659	0.032	0.032	0.961	0.950	0.660	0.031	0.032	0.963	0.951
$\Delta_1(t_\circ)$	0.461	True	0.456	0.033	0.032	0.956	0.958	0.457	0.031	0.032	0.954	0.957
		MLE	0.451	0.034	0.033	0.935	0.948	0.453	0.032	0.033	0.930	0.938
$\Delta_2(t_\circ)$	0.794	True	0.789	0.021	0.022	0.949	0.941	0.789	0.022	0.022	0.958	0.944
		MLE	0.784	0.022	0.023	0.952	0.947	0.787	0.021	0.022	0.937	0.941
$\Delta(t_\circ)$	0.643	True	0.635	0.021	0.019	0.939	0.945	0.635	0.019	0.020	0.941	0.945
		MLE	0.631	0.021	0.022	0.946	0.955	0.633	0.019	0.019	0.932	0.947
			$\beta_{01} = \beta_{02} = (\log 1.5, \log 2.0)'; \beta_{13} = \beta_{23} = R\beta_{01}, R = 2$									
$\Delta_0(t_\circ)$	0.640	True	0.624	0.034	0.033	0.942	0.944	0.622	0.032	0.033	0.949	0.953
		MLE	0.622	0.033	0.032	0.934	0.941	0.622	0.033	0.032	0.926	0.941
$\Delta_1(t_\circ)$	0.513	True	0.513	0.034	0.033	0.947	0.958	0.514	0.031	0.032	0.939	0.947
		MLE	0.503	0.034	0.033	0.956	0.960	0.503	0.033	0.032	0.953	0.950
$\Delta_2(t_\circ)$	0.871	True	0.869	0.017	0.017	0.953	0.953	0.870	0.017	0.017	0.949	0.942
		MLE	0.868	0.017	0.018	0.941	0.944	0.868	0.016	0.017	0.949	0.951
$\Delta(t_\circ)$	0.670	True	0.669	0.019	0.020	0.940	0.945	0.669	0.018	0.019	0.950	0.954
		MLE	0.665	0.018	0.020	0.956	0.952	0.664	0.017	0.018	0.951	0.945

Note: The ASE is approximated with 500 bootstrap samples with replacement within each simulated data. Note that ECP is the empirical coverage probability of nominal 95% confidence intervals constructed based on the normal approximation of the estimator using ASE as the standard error while ECP‡ is the corresponding empirical coverage probability when the confidence intervals are constructed on using the logit transformation; training samples are of 1000 observations; validation samples involve 500 individuals, $nsim = 1000$.

† Prediction is based on true parameter values (True) and maximum likelihood estimates (MLE).

disease (state 3). This results in three estimates of the PDI at time $t_\circ$. Note that although we are collapsing the unilateral damage states together when estimating the PDI, the associated parameters with multistate process are estimated under the general 4-state model to allow flexibility.

As in the multinomial setting we considered two approaches for estimating $\Delta_k(t_\circ)$. The first approach treats the full parameter vector θ as fixed at the true value and directly evaluates the PDI in a validation sample of 500 individuals. The second approach estimates θ from an independent training sample of size 1000 by maximizing the log likelihood in (15), and then evaluates the PDI in a validation sample of size 500. To visualize the setting under $R = 0.25, 0.5, 1$ and 2, we have constructed the empirical contour plots of the joint density of the covariates conditional on class membership, displayed in Figure 2. This contour plot is created by simulating a random sample of 10,000 individuals where it gives a sense of the separation of the covariate distribution between the three different classes. A complete table of simulation results repeated over 1000 simulation runs can be found in Table 2. Similar to the multinomial setting, the column named 'Value' corresponds to the true PDI value at time $t_\circ$ estimated by Monte Carlo. As in Section 2, we report the mean estimate under the column headed EST, and provide the empirical standard error (ESE) and average bootstrap standard error based on 500 bootstrap samples (ASE); the empirical coverage probability of nominal 95% confidence intervals computed on the scale of Δ_k is reported under ECP while ECP‡ reports the corresponding empirical coverage probability for confidence intervals are constructed based on the logit transformation. We can see that the proposed weighted estimating function gives estimates with small empirical bias. Notably, the empirical standard error of the estimators of $\Delta_k(t_\circ)$ and $\Delta(t_\circ)$ are only modestly affected by the estimation of θ. Thus, the increase in frequency of the visit process also had modest effect on the empirical bias and standard error. As R increased from 0.25 to 2, giving a stronger covariate effect of transitioning from the unilateral damage state to axial state, we see a big increase in $\Delta_2(t_\circ)$. This finding is in accordance with the empirical contour plots in Figure 2. Additional simulation studies with $n = 500$ and 2000 in the training sample retain similar conclusions as what we presented here. These results are in Section S1.2 of the Supplemental Material.

4 Sacroiliac Joint Damage in Psoriatic Arthritis

Here we revisit the problem of predicting sacroiliac damage in patients with psoriatic arthritis (PsA) using data from the University of Toronto Psoriatic Arthritis Clinic, where interest lies in predicting whether an individual is in a certain state at a pre-specified time horizon; see Section 1.2. This multistate disease process is depicted in Figure 1(a) and we aim to assess the accuracy of predictions for being in states defining unilateral sacroiliac damage (states 1 or 2) and axial disease (state 3). Table 3 provides the distribution of the coarsening sets at time horizons $t_\circ = 5$, 10, 15 and 20 years. From this table we can see that the state occupied for individuals is usually uncertain with increasing degrees of coarsening due to intermittent observation at later time horizons. We restrict attention to individuals recruited to the clinic that did not have any sacroiliac damage upon clinic entry (at state 0), giving a sample of 953 individuals. The baseline covariates used in this analysis included age of PsA diagnosis, gender, and several human leukocyte antigen markers including, HLA-A2, HLA-A11, HLA-B38, HLA-C12, HLA-DR8, HLA-DR14, HLA-DQ2, HLA-DQ3 and HLA-DQ5, which have been reported as important risk factors in previous work in this area. We considered four multistate models including a model with time homogeneous transition intensities and four distinct sets of regression coefficients (Model 1), a model with time homogeneous transition intensities with the regression coefficients constrained to be equal for the $0 \to 1$ and $0 \to 2$

transitions, as well as $1 \rightarrow 3$ and $2 \rightarrow 3$ transitions (Model 2). Models 3 and 4 were analogous to Models 1 and 2, respectively, but with piecewise-constant (four pieces) baseline intensities having cut-points at $t = 8$, 16, 24 years. The estimated relative risks (RR) for each of these covariates and their associated 95% confidence intervals are presented in Table 4 for all four models.

Table 3: Distribution of coarsened state occupancy data at time horizons $t_\circ = 5$, 10, 15 and 20 years from clinic entry.

$\mathcal{C}_i$	$t_\circ = 5$	$t_\circ = 10$	$t_\circ = 15$	$t_\circ = 20$
	Prediction Time Horizon			
0	611 (64.11%)	407 (42.71%)	251 (26.33%)	161 (16.89%)
1	7 (0.73%)	7 (0.73%)	6 (0.63%)	2 (0.21%)
2	12 (1.26%)	10 (1.05%)	12 (1.26%)	10 (1.05 %)
3	117 (12.28%)	188 (19.72%)	253 (26.55%)	286 (30.01%)
(0, 1)	6 (0.63%)	5 (0.53%)	1 (0.11%)	0 (0%)
(0, 2)	12 (1.26%)	6 (0.63%)	6 (0.63%)	3 (0.31%)
(1, 3)	11 (1.15%)	12 (1.26%)	18 (1.89%)	23 (2.41%)
(2, 3)	17 (1.78%)	32 (3.36%)	35 (3.67%)	42 (4.41%)
(0, 1, 2, 3)	160 (16.79%)	286 (30.01%)	371 (38.93%)	426 (44.70%)

The mean value of the PDI for each class, as well as the overall value, were computed at times $t_\circ = 5$, 10, 15, 20, with the point estimates joined over the different time horizons in Figure S1 of the Supplemental Material. All four models demonstrated much superior prediction than the null model whose value of the PDI is 1/3 (see lower dashed horizontal line). The time-homogeneous Model 1 with unconstrained regression coefficients tended to give the best predictive performance in terms of class-specific and overall discrimination. To assess the uncertainty in $\widehat{\Delta}_k(t_\circ)$ and the overall estimate $\widehat{\Delta}(t_\circ)$ for the four models, pointwise 95% confidence intervals were computed at each time horizon based on the nonparametric bootstrap. Model 1 had the best performance so we show the plots for the time-homogeneous models (Models 1 and 2) in Figure 3 and include the plots for piecewise-constant models (Models 3 and 4) in Figure S2 of the Supplemental Material.

5 Discussion

Our primary goal is to describe how to estimate the accuracy of a prediction model for state occupancy of a multistate process at a specified time horizon, in the setting where the disease process is under intermittent observation. As an initial investigation, we consider prediction with a multinomial response where outcomes may be coarsened for some individuals so that it is only known that the outcome is one of a set of possible categories. A weighted estimator of the PDI is proposed by considering a pseudo-sample of individuals accommodating the different response categories that individuals with grouped outcomes may belong to; this approach is shown to perform well with moderate to heavy completely random coarsening rates. Building on the discussion under the multinomial setup in Section 2, we then described estimation of PDI in multistate disease processes. The interest lies in predicting whether an individual is in a certain state at a pre-specified time horizon $t_\circ$. However, the state occupied by some individuals at $t_\circ$

Table 4: Estimate of regression coefficients from fitting Models 1 to 4 to data from the Univeristy of Toronto Psoriatic Arthritis Cohort; cut-points are at 8, 16 and 24 years from disease onset.

		Models 1 and 3								Models 2 and 4			
		$0 \to 1$		$0 \to 2$		$1 \to 3$		$2 \to 3$		$0 \to 1/2$		$1/2 \to 3$	
Covariate		RR	CI	RR	CI	RR	CI	RR	CI	RR	CI	RR	CI
						Time homogeneous intensity							
gender	(male vs. female)	1.96	(0.84, 4.56)	1.33	(0.75, 2.35)	1.30	(0.33, 5.14)	1.62	(0.74, 3.58)	1.52	(1.05, 2.22)	1.34	(0.85, 2.11)
age	(years)	1.00	(0.97, 1.03)	0.96	(0.94, 0.99)	1.08	(1.02, 1.14)	0.98	(0.95, 1.00)	0.97	(0.96, 0.99)	0.99	(0.97, 1.01)
a2		3.15	(1.24, 8.05)	1.58	(0.87, 2.85)	4.77	(0.89, 28.32)	0.79	(0.43, 1.44)	1.99	(1.33, 2.97)	1.34	(0.84, 2.14)
a11		1.73	(0.67, 4.41)	0.60	(0.19, 1.92)	11.51	(2.06, 64.34)	1.33	(0.47, 3.74)	0.99	(0.55, 1.80)	1.94	(1.00, 3.77)
b38		1.04	(0.30, 3.60)	1.11	(0.42, 2.95)	2.25	(0.23, 22.18)	2.52	(0.84, 7.59)	1.05	(0.53, 2.07)	2.52	(1.09, 5.83)
c12		1.05	(0.38, 2.96)	1.55	(0.72, 3.32)	2.46	(0.32, 19.07)	1.52	(0.66, 3.48)	1.35	(0.76, 2.34)	1.56	(0.85, 2.86)
dr8		0.45	(0.09, 2.12)	1.66	(0.63, 4.37)	1.30	(0.09, 18.64)	0.75	(0.24, 2.37)	1.05	(0.48, 2.30)	0.55	(0.22, 1.38)
dr14		1.05	(0.30, 3.68)	1.25	(0.43, 3.64)	0.19	(0.02, 2.15)	0.58	(0.17, 1.99)	1.17	(0.56, 2.42)	0.37	(0.15, 0.89)
dq2		0.15	(0.05, 0.46)	1.27	(0.70, 2.32)	0.17	(0.04, 0.73)	1.18	(0.57, 2.46)	0.65	(0.42, 1.00)	0.71	(0.45, 1.14)
dq3		0.58	(0.26, 1.29)	1.31	(0.72, 2.38)	0.13	(0.03, 0.57)	1.83	(0.92, 3.63)	0.94	(0.62, 1.42)	1.24	(0.76, 2.02)
dq5		0.54	(0.22, 1.30)	0.94	(0.49, 1.80)	2.31	(0.71, 7.55)	1.30	(0.58, 2.90)	0.74	(0.48, 1.15)	1.29	(0.78, 2.14)
						Piecewise constant intensity							
gender	(male vs. female)	1.56	(0.78, 3.10)	1.26	(0.75, 2.13)	0.77	(0.24, 2.46)	1.88	(0.97, 3.63)	1.35	(0.93, 1.96)	1.36	(0.86, 2.16)
age	(years)	1.00	(0.97, 1.02)	0.95	(0.93, 0.97)	1.06	(1.00, 1.11)	0.97	(0.95, 1.00)	0.97	(0.95, 0.98)	0.99	(0.97, 1.01)
a2		2.70	(1.17, 6.22)	1.50	(0.85, 2.66)	3.39	(0.66, 17.46)	0.80	(0.43, 1.49)	1.87	(1.25, 2.79)	1.35	(0.84, 2.15)
a11		1.85	(0.78, 4.34)	0.44	(0.13, 1.55)	13.62	(2.52, 73.63)	1.19	(0.42, 3.40)	0.95	(0.52, 1.73)	1.82	(0.93, 3.55)
b38		0.88	(0.24, 3.16)	1.34	(0.53, 3.40)	1.49	(0.14, 15.51)	2.77	(0.90, 8.52)	1.08	(0.55, 2.13)	2.32	(1.00, 5.41)
c12		1.04	(0.35, 3.13)	1.49	(0.69, 3.24)	2.92	(0.35, 24.27)	1.48	(0.63, 3.49)	1.33	(0.75, 2.33)	1.56	(0.84, 2.87)
dr8		0.43	(0.09, 1.95)	1.52	(0.57, 4.09)	1.12	(0.09, 14.24)	0.85	(0.27, 2.69)	0.92	(0.41, 2.05)	0.60	(0.24, 1.49)
dr14		1.03	(0.29, 3.69)	1.14	(0.31, 3.18)	0.26	(0.02, 3.61)	0.57	(0.17, 1.98)	1.18	(0.58, 2.40)	0.45	(0.18, 1.12)
dq2		0.14	(0.05, 0.41)	1.48	(0.81, 2.68)	0.14	(0.04, 0.52)	1.38	(0.70, 2.74)	0.73	(0.47, 1.13)	0.75	(0.46, 1.22)
dq3		0.65	(0.32, 1.34)	1.12	(0.63, 2.01)	0.15	(0.04, 0.57)	1.86	(0.95, 3.68)	0.92	(0.61, 1.39)	1.23	(0.75, 2.03)
dq5		0.47	(0.19, 1.13)	1.13	(0.59, 2.17)	1.73	(0.47, 6.33)	1.53	(0.71, 3.30)	0.80	(0.51, 1.25)	1.29	(0.76, 2.19)

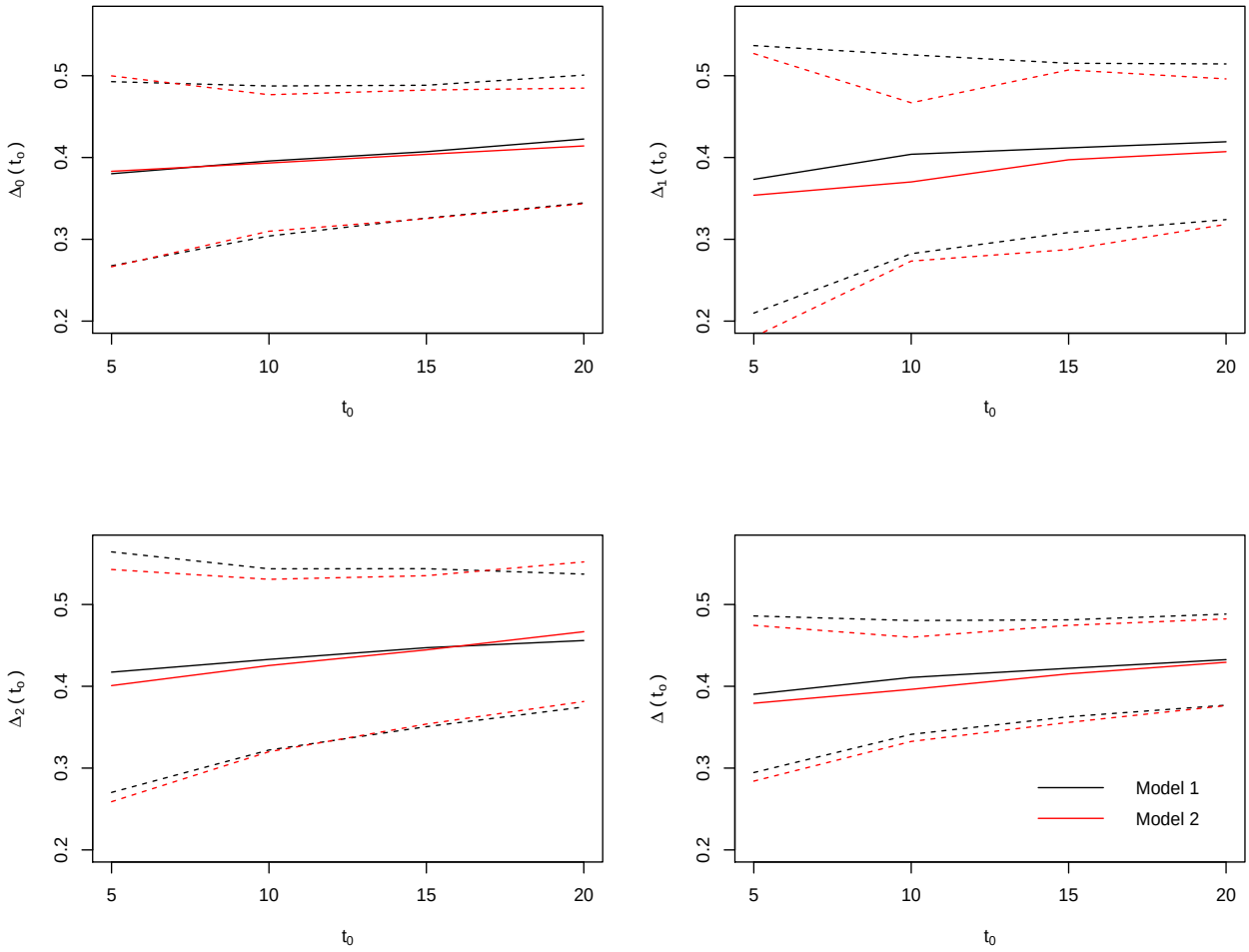


Figure 3: Plots of the predictive discrimination index estimates as a function of $t_\circ$ and the empirical 95% confidence interval (dashed lines) for a time homogeneous model with no constraints on covariate effects (Model 1), a time homogeneous model with constraints (Model 2),; constraints ensure $0 \to 1$ and $0 \to 2$ regression coefficients for the onset of unilateral damage, and $1 \to 3$ and $2 \to 3$ regression coefficients for the development of axial disease, are respectively the same. $\Delta_0(t_\circ)$ corresponds to prediction of no SI-joint involvement, $\Delta_1(t_\circ)$ corresponds to prediction of unilateral SI-joint damage, $\Delta_2(t_\circ)$ corresponds to prediction of axial disease, and $\Delta(t_\circ)$ is the overall measure.

may be unknown due to the intermittent observation scheme. A weighted estimator of the PDI considering pseudo-sample of individuals is thus proposed and has empirically shown to perform well in simulation studies. We did not consider ties in the predictive probabilities in estimating the PDI, but they can be handled easily as discussed in Van Calster et al. (2012). The proposed method is then applied to a motivating study involving data from the University of Toronto Psoriatic Arthritis Clinic. Here we fitted four models and assessed their predictive accuracy at a few different time horizons via 5-fold internal cross-validation. In general, all four models seem to have reasonable PDI values compared to the null model. The most parsimonious model with time-homogeneous transition intensities seemed often to exhibit superior performance.

For the multinomial setting coarsening completely at random (Heitjan and Rubin, 1991) implies that the presence and nature of coarsening is completely independent of the response

category. If this is not satisfied, then joint modeling, inverse probability weighting (Robins et al., 1994), or augmentation and inverse probability weighting (Bang and Robins, 2005) can be employed. In the multistate setting, response-dependent visit times can lead to bias both in terms of model fitting when building a predictive model, and in assessing predictive accuracy. Joint modeling of the multistate process and the visit process can help mitigate bias in model fitting. Use of joint disease and visit process models for prediction seem less natural and would not tend to be transportable to other clinic settings where visit schedules may differ – we are currently exploring the use of inverse-intensity weighting for this setting.

Complex disease processes often feature heterogeneity beyond that explained through covariates. Jiang and Cook (2019) describe finite mixture models of multistate processes under intermittent observation and develop score tests for effects of biomarkers on class membership. Here the use of score tests was motivated by the need to screen a large number of genetic markers for their association with the disease course combined with the difficulty in fitting such mixture models. Once a list of candidate genetic markers are identified by this approach, it is natural to incorporate them into a predictive model for the disease course. In this case, one could model covariate effects on class membership as done in Jiang and Cook (2019), as well as on the intensity functions of the multistate process in the different classes. Such a rich predictive model could then be used to predict state occupancy at $t_\circ$ – our proposed method for estimating the PDI can be readily adapted to deal with this setting.

Multistate models with hidden states may also be of interest in some disease settings. States are often based on distinct conditions such as the definition for each state is clear. There is no ambiguity in our motivating setting – whether an individual has sacroiliac joint damage on the left or right-side (or both) is typically clear from radiographic examination. In other settings it may be difficult to determine which state an individual occupies upon examination – if it can be determined that they are in a strict subset of the possible states this remains informative and the likelihood can be modified to deal with this at training stage. This would represent a hybrid coarsening process involving aspects of the settings of Sections 2 and 3 which can be dealt with using the `msm` function as described in Section 3.4 of Jackson (2011), but we do not consider this here. In other longitudinal observation schemes the states occupied may be subject to misclassification. In such cases hidden Markov models could be considered, but the general approach to estimating the PDI indices we discuss here remain applicable. Finally, as one reviewer pointed out, life history processes are often observed subject to left truncation. It is important to address this when it arises in datasets during model building and the assessment of predictive accuracy should address such complications and this, along with the development of robust standard errors for the estimators we develop, is worthy of future research.

Acknowledgements

This work has been supported by the National Cancer Institute (U01 CA195547 for SJ) and grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-04207 for RJC) and the Canadian Institutes of Health Research (FRN 159834 for RJC). Richard Cook is a Mathematics Faculty Research Chair and University Professor at the University of Waterloo. The authors thank Drs. Dafna Gladman and 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.

Data Availability Statement

The data from the University of Toronto Psoriatic Arthritis Clinic are confidential and held by the Centre for Prognosis Studies in the Rheumatic Diseases. The R code that supports the findings of this study is available upon request from the first author.

References

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media, New York.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical Models Based on Counting Processes*. Springer Science & Business Media, New York.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bergholtz, H., Lien, T. G., Swanson, D. M., Frigessi, A., Daidone, M. G., Tost, J., Wärnberg, F., and Sørlie, T. (2020). Contrasting dcis and invasive breast cancer by subtype suggests basal-like dcis as distinct lesions. *NPJ Breast Cancer*, 6(1):1–9.

Brier, G. W. (1950). The statistical theory of turbulence and the problem of diffusion in the atmosphere. *Journal of Meteorology*, 7(4):283–290.

Cook, R. J. and Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. CRC Press, New York.

Cook, R. J. and Lawless, J. F. (2021). Independence conditions and the analysis of life history studies with intermittent observation. *Biostatistics*, 22(3):455–481.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dreiseitl, S., Ohno-Machado, L., and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, 20(3):323–331.

Escolano, S., Golmard, J.-L., Korinek, A.-M., and Mallet, A. (2000). A multi-state model for evolution of intensive care unit patients: prediction of nosocomial infections and deaths. *Statistics in Medicine*, 19(24):3465–3482.

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(6):664–673.

Hanley, J. A. et al. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging*, 29(3):307–335.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics*, pages 2244–2253.

Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine*, 23(9):1455–1497.

Jackson, C. H. (2011). Multi-state models for panel data: the `msm` package for r. *Journal of Statistical Software*, 38(8):1–29.

Jiang, S. and Cook, R. J. (2019). Score tests based on a finite mixture model of Markov processes under intermittent observation. *Statistics in Medicine*, 38(16):3013–3025.

Kalbfleisch, J. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871.

Keiding, N., Klein, J. P., and Horowitz, M. M. (2001). Multi-state models and outcome prediction in bone marrow transplantation. *Statistics in Medicine*, 20(12):1871–1885.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). Applied linear statistical models, 2005. *McGraw Hill Irwin, New York. NY*, page 409.

Li, J., Feng, Q., Fine, J. P., Pencina, M. J., and Van Calster, B. (2018). Nonparametric estimation and inference for polytomous discrimination index. *Statistical Methods in Medical Research*, 27(10):3092–3103.

Li, J., Jiang, B., and Fine, J. P. (2013). Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics*, 14(2):382–394.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, 19(1):78–89.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York.

Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R., and van de Velde, C. J. (2006). Estimation and prediction in a multi-state model for breast cancer. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(3):366–380.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Schemper, M. (2003). Predictive accuracy and explained variation. *Statistics in Medicine*, 22(14):2299–2308.

Spitoni, C., Lammens, V., and Putter, H. (2018). Prediction errors for state occupation and transition probabilities in multi-state models. *Biometrical Journal*, 60(1):34–48.

Steyerberg, E. W. (2019). *Clinical Prediction Models*. Springer, New York.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L.-J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117.

Uno, H., Cai, T., Tian, L., and Wei, L.-J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537.

Van Calster, B., Van Belle, V., Vergouwe, Y., Timmerman, D., Van Huffel, S., and Steyerberg, E. W. (2012). Extending the c-statistic to nominal polytomous outcomes: the polytomous discrimination index. *Statistics in Medicine*, 31(23):2610–2626.

Supporting Information

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

Published in final edited form as:
Statistics in Medicine (2022), **41** (19): 3661–3678
DOI: 10.1002/sim.9441

Supplementary Material for *The polytomous discrimination index for prediction involving multistate processes under intermittent observation*

SHU JIANG

Division of Public Health Sciences,
Washington University School of Medicine, St. Louis, Missouri, USA

RICHARD J. COOK

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

S1 Additional Simulation Results

S1.1 Multinomial Data

Here we report on additional simulation results concerning estimation of the polytomous discrimination index for the problem considered in Section 2 of the manuscript on involving coarsened multinomial data. In Section 2.3 of the main manuscript we reported simulation results for validation samples of size $n = 1000$; here we consider $n = 500$ and 2000 in Tables S1 and S2, respectively for no coarsening, 30% and 60% coarsened observations. We report results when the true parameter values are used to estimate the category-specific polytomous discrimination indices and the overall polytomous discrimination indices, as well as methods wherein estimates were obtained from a training sample with β estimated based on an expectation-maximization algorithm or complete-case analysis; note that when there is no coarsening the results on rows labeled EM are simply maximum likelihood estimates based on complete data. As in the main body of the paper we consider moderate and stronger covariate effects; code is available from the authors upon request to facilitate exploration of other parameter configurations of interest.

As in the main body of the paper we see good agreement between the average bootstrap standard errors and the empirical standard errors, and good agreement between the empirical and nominal coverage probabilities for 95% confidence intervals based on either the original scale of the PDI or the logit transformation. We see that, paradoxically, there can be a decrease in the empirical standard errors with increasing levels of coarsening when the true parameter values are used. When estimates are used there is a tendency for the empirical standard errors to increase with increasing coarsening; the standard errors based on estimators from complete

case analyses are larger than those when estimates are obtained from more efficient analyses involving the EM algorithm. Moreover, we see that there is relatively little impact of the size of the training sample on the precision of the estimation of the PDI estimators.

S1.2 Multistate Processes

Here we report on the results on simulation studies involving multistate processes observed intermittently. In addition to the $n = 1000$ setting in Section 3.3 of the main manuscript, here we show the simulation results when the validation sample is made up of $n = 500$ or 2000 individuals in Tables S3 and S4, respectively. As in the multinomial setting, we find the size of the training sample and hence the precision of the ML estimators have little impact on the precision of the PDI estimators.

S2 Further Plots Related to PDI Estimation in the PsA Study

In Section 4 of the main manuscript, plots were provided of the estimated indices of predictive discrimination as a function of $t_\circ$ along with the empirical 95% confidence interval for the time-homogenous models (Models 1 and 2); see Figure 3. Here, we show these plots of predictive discrimination index estimates under the piecewise-constant models in Figure (S2). Here, we show the plot for predictive discrimination index estimates as a function of $t_\circ$ for all four models described under Section 4 of the main manuscript: a time homogeneous model with no constraints on covariate effects (Model 1); a time homogeneous model with constraints (Model 2); a piecewise constant intensity model with no constraints on covariate effects (Model 3); and a piecewise constant intensity model with constraints on covariate effects (Model 4). This can be found in Figure S1 where we can see that the most parsimonious Model 1, on average, tends to perform the best among the four models. Table S5 reports the estimates of the baseline intensities from the four fitted models along with 95% confidence intervals.

			Percentage of Individuals with Coarsened Observations														
			0%					30%					60%				
Parameter	Value	Method†	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡
moderate covariate effects; $(\beta_{11},\beta_{12})' = (\log 1.5, \log 2)', (\beta_{21},\beta_{22})' = (2\beta_{11}, 2\beta_{12})'$																	
Δ_0	0.670	True	0.665	0.028	0.029	0.967	0.956	0.665	0.024	0.024	0.949	0.943	0.663	0.022	0.023	0.961	0.954
		EM	0.664	0.028	0.030	0.963	0.959	0.665	0.030	0.030	0.947	0.952	0.665	0.034	0.033	0.947	0.941
		CC						0.664	0.032	0.031	0.942	0.956	0.666	0.045	0.045	0.955	0.949
Δ_1	0.404	True	0.405	0.032	0.032	0.953	0.954	0.402	0.029	0.030	0.960	0.951	0.400	0.024	0.023	0.944	0.952
		EM	0.403	0.033	0.034	0.976	0.969	0.403	0.031	0.031	0.951	0.946	0.407	0.031	0.032	0.964	0.956
		CC						0.409	0.037	0.035	0.924	0.936	0.417	0.053	0.051	0.953	0.957
Δ_2	0.677	True	0.663	0.034	0.033	0.944	0.946	0.663	0.037	0.038	0.958	0.949	0.662	0.025	0.026	0.959	0.951
		EM	0.666	0.034	0.032	0.937	0.945	0.666	0.037	0.037	0.951	0.950	0.667	0.046	0.044	0.949	0.945
		CC						0.666	0.041	0.043	0.942	0.951	0.667	0.056	0.054	0.933	0.938
Δ	0.583	True	0.578	0.023	0.022	0.951	0.957	0.576	0.020	0.019	0.941	0.949	0.575	0.018	0.019	0.955	0.954
		EM	0.582	0.032	0.030	0.952	0.963	0.581	0.035	0.034	0.961	0.957	0.582	0.037	0.036	0.953	0.952
		CC						0.579	0.037	0.036	0.944	0.952	0.586	0.040	0.039	0.956	0.961
strong covariate effects; $(\beta_{11},\beta_{12})' = (\log 3, \log 4)', (\beta_{21},\beta_{22})' = (2\beta_{11}, 2\beta_{12})'$																	
Δ_0	0.828	True	0.831	0.021	0.020	0.952	0.949	0.832	0.018	0.019	0.961	0.953	0.832	0.016	0.016	0.944	0.956
		EM	0.834	0.022	0.020	0.939	0.943	0.835	0.022	0.022	0.951	0.957	0.833	0.025	0.026	0.946	0.962
		CC						0.835	0.024	0.023	0.948	0.958	0.834	0.030	0.031	0.951	0.942
Δ_1	0.553	True	0.566	0.034	0.033	0.943	0.951	0.565	0.031	0.032	0.948	0.956	0.564	0.028	0.028	0.956	0.947
		EM	0.567	0.031	0.033	0.946	0.942	0.566	0.034	0.032	0.942	0.951	0.568	0.033	0.035	0.948	0.954
		CC						0.567	0.039	0.038	0.945	0.951	0.571	0.050	0.052	0.932	0.938
Δ_2	0.821	True	0.825	0.026	0.027	0.962	0.957	0.825	0.023	0.023	0.944	0.958	0.823	0.020	0.019	0.943	0.948
		EM	0.821	0.028	0.026	0.968	0.959	0.826	0.028	0.027	0.941	0.944	0.826	0.035	0.033	0.956	0.949
		CC						0.825	0.032	0.032	0.955	0.952	0.828	0.040	0.038	0.935	0.942
Δ	0.734	True	0.741	0.021	0.020	0.947	0.953	0.741	0.019	0.019	0.945	0.952	0.740	0.016	0.017	0.948	0.957
		EM	0.741	0.026	0.026	0.951	0.956	0.743	0.027	0.029	0.953	0.954	0.743	0.031	0.030	0.948	0.943
		CC						0.741	0.030	0.028	0.939	0.947	0.742	0.035	0.034	0.951	0.954

† Prediction is based on true parameter values (True), as well as maximum likelihood estimates based on an expectation-maximization algorithm (EM) and complete-case analysis (CC).

Table S1: Empirical performance of estimates of Δ_k, $k = 0, 1, 2$ and Δ with no, moderate (30%) and heavier (60%) coarsening; the ASE is the average of bootstrap standard errors based on 500 bootstrap samples created for each simulated data; the ECP is the empirical coverage probability of nominal 95% confidence intervals constructed based on the normal approximation of the estimator using the bootstrap standard error while ECP‡ is the corresponding empirical coverage probability when the confidence intervals are constructed based on the logit transformation; training samples are of 1000 observations; validation samples involve 500 individuals; nsim = 1000.

			Percentage of Individuals with Coarsened Observations														
			0%					30%					60%				
Parameter	Value	Method†	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡
			moderate covariate effects; $(\beta_{11},\beta_{12})' = (\log 1.5, \log 2)', (\beta_{21},\beta_{22})' = (2\beta_{11}, 2\beta_{12})'$														
Δ_0	0.670	True	0.665	0.028	0.029	0.967	0.956	0.665	0.024	0.024	0.949	0.943	0.663	0.022	0.023	0.961	0.954
		EM	0.664	0.030	0.030	0.951	0.952	0.665	0.030	0.030	0.951	0.958	0.665	0.034	0.033	0.942	0.947
		CC						0.664	0.032	0.031	0.953	0.956	0.666	0.045	0.045	0.945	0.953
Δ_1	0.404	True	0.405	0.032	0.032	0.953	0.954	0.402	0.029	0.030	0.960	0.951	0.400	0.024	0.023	0.944	0.952
		EM	0.403	0.033	0.034	0.949	0.945	0.403	0.031	0.031	0.947	0.952	0.406	0.031	0.032	0.951	0.946
		CC						0.407	0.036	0.035	0.944	0.952	0.415	0.052	0.051	0.949	0.951
Δ_2	0.677	True	0.663	0.034	0.033	0.944	0.946	0.663	0.037	0.038	0.958	0.949	0.662	0.025	0.026	0.959	0.951
		EM	0.666	0.033	0.032	0.942	0.950	0.666	0.037	0.037	0.952	0.945	0.667	0.045	0.044	0.959	0.953
		CC						0.666	0.042	0.043	0.940	0.952	0.667	0.056	0.054	0.959	0.948
Δ	0.583	True	0.578	0.023	0.022	0.951	0.957	0.576	0.020	0.019	0.941	0.949	0.575	0.018	0.019	0.955	0.954
		EM	0.582	0.031	0.030	0.952	0.957	0.581	0.035	0.034	0.954	0.949	0.582	0.037	0.036	0.961	0.954
		CC						0.580	0.037	0.036	0.947	0.953	0.582	0.040	0.039	0.946	0.941
			strong covariate effects; $(\beta_{11},\beta_{12})' = (\log 3, \log 4)', (\beta_{21},\beta_{22})' = (2\beta_{11}, 2\beta_{12})'$														
Δ_0	0.828	True	0.831	0.021	0.020	0.952	0.949	0.832	0.018	0.019	0.961	0.953	0.832	0.016	0.016	0.944	0.956
		EM	0.834	0.021	0.020	0.945	0.955	0.835	0.022	0.022	0.950	0.952	0.834	0.025	0.026	0.947	0.947
		CC						0.835	0.024	0.023	0.951	0.958	0.834	0.030	0.031	0.954	0.948
Δ_1	0.553	True	0.566	0.034	0.033	0.943	0.951	0.565	0.031	0.032	0.948	0.956	0.564	0.028	0.028	0.956	0.947
		EM	0.566	0.032	0.033	0.953	0.962	0.566	0.034	0.032	0.957	0.954	0.567	0.034	0.035	0.952	0.956
		CC						0.567	0.039	0.038	0.960	0.953	0.569	0.051	0.052	0.952	0.948
Δ_2	0.821	True	0.825	0.026	0.027	0.962	0.957	0.825	0.023	0.023	0.944	0.958	0.823	0.020	0.019	0.943	0.948
		EM	0.821	0.027	0.026	0.948	0.953	0.824	0.028	0.027	0.953	0.954	0.827	0.034	0.033	0.943	0.957
		CC						0.825	0.032	0.032	0.950	0.954	0.828	0.039	0.038	0.952	0.955
Δ	0.734	True	0.741	0.021	0.020	0.947	0.953	0.741	0.019	0.019	0.945	0.952	0.740	0.016	0.017	0.948	0.957
		EM	0.737	0.026	0.026	0.943	0.952	0.740	0.028	0.029	0.946	0.946	0.740	0.031	0.030	0.948	0.957
		CC						0.741	0.029	0.028	0.941	0.942	0.742	0.035	0.034	0.952	0.945

† Prediction is based on true parameter values (True), as well as maximum likelihood estimates based on an expectation-maximization algorithm (EM) and complete-case analysis (CC).

Table S2: Empirical performance of estimates of Δ_k, $k = 0, 1, 2$ and Δ with no, moderate (30%) and heavier (60%) coarsening; the ASE is the average of bootstrap standard errors based on 500 bootstrap samples created for each simulated data; the ECP is the empirical coverage probability of nominal 95% confidence intervals constructed based on the normal approximation of the estimator using the bootstrap standard error while ECP‡ is the corresponding empirical coverage probability when the confidence intervals are constructed based on the logit transformation; training samples are of 2000 observations; validation samples involve 500 individuals; nsim = 1000.

Parameter	Value	Method†	$E(M)=10$ EST	ESE	ASE	ECP	ECP‡	$E(M)=20$ EST	ESE	ASE	ECP	ECP‡
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=0.25$									
$\Delta_0(t_\circ)$	0.708	True	0.709	0.029	0.030	0.962	0.957	0.709	0.029	0.030	0.958	0.959
		MLE	0.704	0.027	0.029	0.952	0.956	0.706	0.029	0.031	0.954	0.961
$\Delta_1(t_\circ)$	0.416	True	0.428	0.029	0.028	0.959	0.957	0.429	0.031	0.030	0.940	0.943
		MLE	0.418	0.029	0.030	0.932	0.936	0.418	0.030	0.030	0.940	0.946
$\Delta_2(t_\circ)$	0.651	True	0.637	0.026	0.026	0.941	0.943	0.639	0.027	0.028	0.962	0.958
		MLE	0.640	0.028	0.026	0.944	0.946	0.637	0.027	0.028	0.943	0.946
$\Delta(t_\circ)$	0.592	True	0.591	0.021	0.020	0.942	0.949	0.592	0.021	0.020	0.942	0.943
		MLE	0.587	0.019	0.020	0.954	0.951	0.587	0.020	0.021	0.958	0.959
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=0.5$									
$\Delta_0(t_\circ)$	0.697	True	0.690	0.030	0.031	0.952	0.946	0.689	0.030	0.031	0.953	0.959
		MLE	0.692	0.029	0.030	0.970	0.969	0.691	0.030	0.031	0.946	0.942
$\Delta_1(t_\circ)$	0.435	True	0.431	0.032	0.031	0.940	0.954	0.432	0.032	0.032	0.957	0.955
		MLE	0.426	0.031	0.030	0.944	0.944	0.428	0.031	0.031	0.940	0.944
$\Delta_2(t_\circ)$	0.702	True	0.703	0.025	0.026	0.956	0.957	0.702	0.026	0.026	0.941	0.954
		MLE	0.703	0.023	0.025	0.974	0.972	0.701	0.023	0.024	0.956	0.944
$\Delta(t_\circ)$	0.613	True	0.608	0.021	0.019	0.954	0.943	0.608	0.021	0.020	0.950	0.947
		MLE	0.607	0.020	0.020	0.948	0.940	0.607	0.020	0.019	0.958	0.951
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=1$									
$\Delta_0(t_\circ)$	0.674	True	0.659	0.032	0.031	0.942	0.943	0.659	0.032	0.033	0.952	0.951
		MLE	0.659	0.030	0.031	0.942	0.948	0.660	0.031	0.032	0.948	0.944
$\Delta_1(t_\circ)$	0.461	True	0.456	0.033	0.032	0.956	0.958	0.457	0.031	0.032	0.954	0.957
		MLE	0.452	0.034	0.032	0.920	0.931	0.453	0.032	0.032	0.938	0.942
$\Delta_2(t_\circ)$	0.794	True	0.789	0.021	0.022	0.949	0.941	0.789	0.022	0.022	0.958	0.944
		MLE	0.754	0.020	0.021	0.956	0.951	0.787	0.021	0.022	0.944	0.942
$\Delta(t_\circ)$	0.643	True	0.635	0.021	0.019	0.939	0.945	0.635	0.019	0.020	0.941	0.945
		MLE	0.632	0.019	0.019	0.942	0.939	0.633	0.019	0.020	0.946	0.956
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=2$									
$\Delta_0(t_\circ)$	0.640	True	0.624	0.034	0.033	0.942	0.944	0.622	0.032	0.033	0.949	0.953
		MLE	0.622	0.033	0.032	0.938	0.941	0.622	0.033	0.032	0.938	0.936
$\Delta_1(t_\circ)$	0.513	True	0.513	0.034	0.033	0.947	0.958	0.514	0.031	0.032	0.939	0.947
		MLE	0.503	0.030	0.032	0.966	0.960	0.503	0.033	0.033	0.938	0.942
$\Delta_2(t_\circ)$	0.871	True	0.869	0.017	0.017	0.953	0.953	0.870	0.017	0.017	0.949	0.942
		MLE	0.868	0.017	0.016	0.919	0.931	0.868	0.016	0.017	0.958	0.958
$\Delta(t_\circ)$	0.670	True	0.669	0.019	0.020	0.940	0.945	0.669	0.018	0.019	0.950	0.954
		MLE	0.665	0.018	0.017	0.934	0.940	0.664	0.017	0.017	0.946	0.955

† Prediction is based on true parameter values (True) and maximum likelihood estimates (MLE).

Table S3: Finite sample properties of estimates of $\Delta_k(t_\circ)$, $k=0,1,2$ and $\Delta(t_\circ)$ with $t_\circ=1$, for different values of R ($\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01}$) based on Markov process under intermittent observation with an average of 10 or 20 visits over the period $(0,2]$. The ASE is approximated with 500 bootstrap samples with replacement within each simulated data. Note that ECP is the empirical coverage probability of nominal 95% confidence intervals constructed based on the normal approximation of the estimator using ASE as the standard error while ECP‡ is the corresponding empirical coverage probability when the confidence intervals are constructed on using the logit transformation; training samples are of 500 observations; validation samples involve 500 individuals; nsim = 1000.

			$E(M)=10$					$E(M)=20$				
Parameter	Value	Method†	EST	ESE	ASE	ECP	ECP‡	EST	ESE	ASE	ECP	ECP‡
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=0.25$									
$\Delta_0(t_\circ)$	0.708	True	0.709	0.029	0.030	0.962	0.957	0.709	0.029	0.030	0.958	0.959
		MLE	0.709	0.029	0.030	0.940	0.952	0.709	0.030	0.030	0.952	0.952
$\Delta_1(t_\circ)$	0.416	True	0.428	0.029	0.028	0.959	0.957	0.429	0.031	0.030	0.940	0.943
		MLE	0.421	0.031	0.031	0.942	0.957	0.423	0.031	0.030	0.944	0.942
$\Delta_2(t_\circ)$	0.651	True	0.637	0.026	0.026	0.941	0.943	0.639	0.027	0.028	0.962	0.958
		MLE	0.642	0.027	0.028	0.948	0.942	0.643	0.027	0.028	0.958	0.950
$\Delta(t_\circ)$	0.592	True	0.591	0.021	0.020	0.942	0.949	0.592	0.021	0.020	0.942	0.943
		MLE	0.590	0.020	0.021	0.959	0.957	0.590	0.020	0.019	0.959	0.955
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=0.5$									
$\Delta_0(t_\circ)$	0.697	True	0.690	0.030	0.031	0.952	0.946	0.689	0.030	0.031	0.953	0.959
		MLE	0.690	0.030	0.031	0.956	0.950	0.691	0.031	0.031	0.948	0.946
$\Delta_1(t_\circ)$	0.435	True	0.431	0.032	0.031	0.940	0.954	0.432	0.032	0.032	0.957	0.955
		MLE	0.427	0.030	0.030	0.936	0.942	0.429	0.032	0.031	0.952	0.954
$\Delta_2(t_\circ)$	0.702	True	0.703	0.025	0.026	0.956	0.957	0.702	0.026	0.026	0.941	0.954
		MLE	0.703	0.024	0.024	0.952	0.958	0.702	0.025	0.024	0.962	0.956
$\Delta(t_\circ)$	0.613	True	0.608	0.021	0.019	0.954	0.943	0.608	0.021	0.020	0.950	0.947
		MLE	0.606	0.020	0.021	0.958	0.954	0.606	0.021	0.020	0.944	0.944
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=1$									
$\Delta_0(t_\circ)$	0.674	True	0.659	0.032	0.031	0.942	0.943	0.659	0.032	0.033	0.952	0.951
		MLE	0.663	0.031	0.031	0.954	0.958	0.666	0.031	0.032	0.948	0.950
$\Delta_1(t_\circ)$	0.461	True	0.456	0.033	0.032	0.956	0.958	0.457	0.031	0.032	0.954	0.957
		MLE	0.454	0.033	0.033	0.949	0.948	0.453	0.031	0.032	0.946	0.944
$\Delta_2(t_\circ)$	0.794	True	0.789	0.021	0.022	0.949	0.941	0.789	0.022	0.022	0.958	0.944
		MLE	0.787	0.022	0.022	0.956	0.959	0.787	0.021	0.022	0.956	0.958
$\Delta(t_\circ)$	0.643	True	0.635	0.021	0.019	0.939	0.945	0.635	0.019	0.020	0.941	0.945
		MLE	0.633	0.019	0.019	0.956	0.958	0.633	0.019	0.019	0.954	0.956
			$\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01},R=2$									
$\Delta_0(t_\circ)$	0.640	True	0.624	0.034	0.033	0.942	0.944	0.622	0.032	0.033	0.949	0.953
		MLE	0.625	0.033	0.032	0.942	0.948	0.626	0.032	0.032	0.940	0.941
$\Delta_1(t_\circ)$	0.513	True	0.513	0.034	0.033	0.947	0.958	0.514	0.031	0.032	0.939	0.947
		MLE	0.506	0.033	0.033	0.946	0.940	0.505	0.033	0.033	0.940	0.942
$\Delta_2(t_\circ)$	0.871	True	0.869	0.017	0.017	0.953	0.953	0.870	0.017	0.017	0.949	0.942
		MLE	0.870	0.017	0.018	0.954	0.956	0.868	0.018	0.017	0.958	0.947
$\Delta(t_\circ)$	0.670	True	0.669	0.019	0.020	0.940	0.945	0.669	0.018	0.019	0.950	0.954
		MLE	0.667	0.018	0.017	0.956	0.951	0.665	0.017	0.018	0.946	0.953

† Prediction is based on true parameter values (True) and maximum likelihood estimates (MLE).

Table S4: Finite sample properties of estimates of $\Delta_k(t_\circ)$, $k=0,1,2$ and $\Delta(t_\circ)$ with $t_\circ=1$, for different values of R ($\beta_{01}=\beta_{02}=(\log 1.5,\log 2.0)';\beta_{13}=\beta_{23}=R\beta_{01}$) based on Markov process under intermittent observation with an average of 10 or 20 visits over the period $(0,2]$. The ASE is approximated with 500 bootstrap samples with replacement within each simulated data. Note that ECP is the empirical coverage probability of nominal 95% confidence intervals constructed based on the normal approximation of the estimator using ASE as the standard error while ECP‡ is the corresponding empirical coverage probability when the confidence intervals are constructed on using the logit transformation; training samples are of 2000 observations; validation samples involve 500 individuals; nsim = 1000.

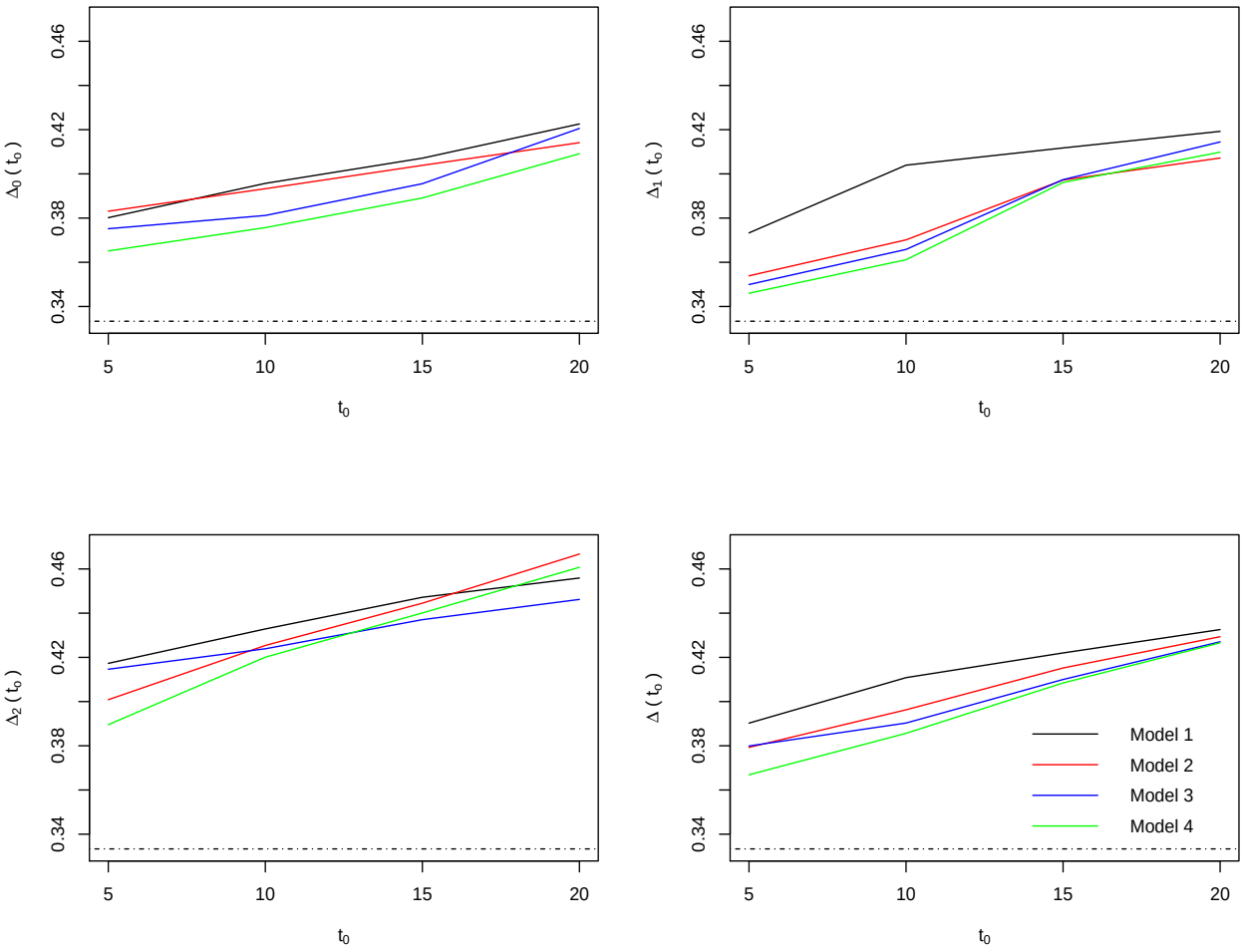


Figure S1: Plots of the predictive discrimination index estimates as a function of $t_\circ$ for a time homogeneous model with no constraints on covariate effects (Model 1), a time homogeneous model with constraints (Model 2), a piecewise constant intensity model with no constraints on covariate effects (Model 3), and a piecewise constant intensity model with constraints on covariate effects (Model 4); constraints ensure $0 \to 1$ and $0 \to 2$ regression coefficients for the onset of unilateral damage, and $1 \to 3$ and $2 \to 3$ regression coefficients for the development of axial disease, are respectively the same. $\Delta_0(t_\circ)$ corresponds to prediction of no SI-joint involvement, $\Delta_1(t_\circ)$ corresponds to prediction of unilateral SI-joint damage, $\Delta_2(t_\circ)$ corresponds to prediction of axial disease, and $\Delta(t_\circ)$ is the overall measure.

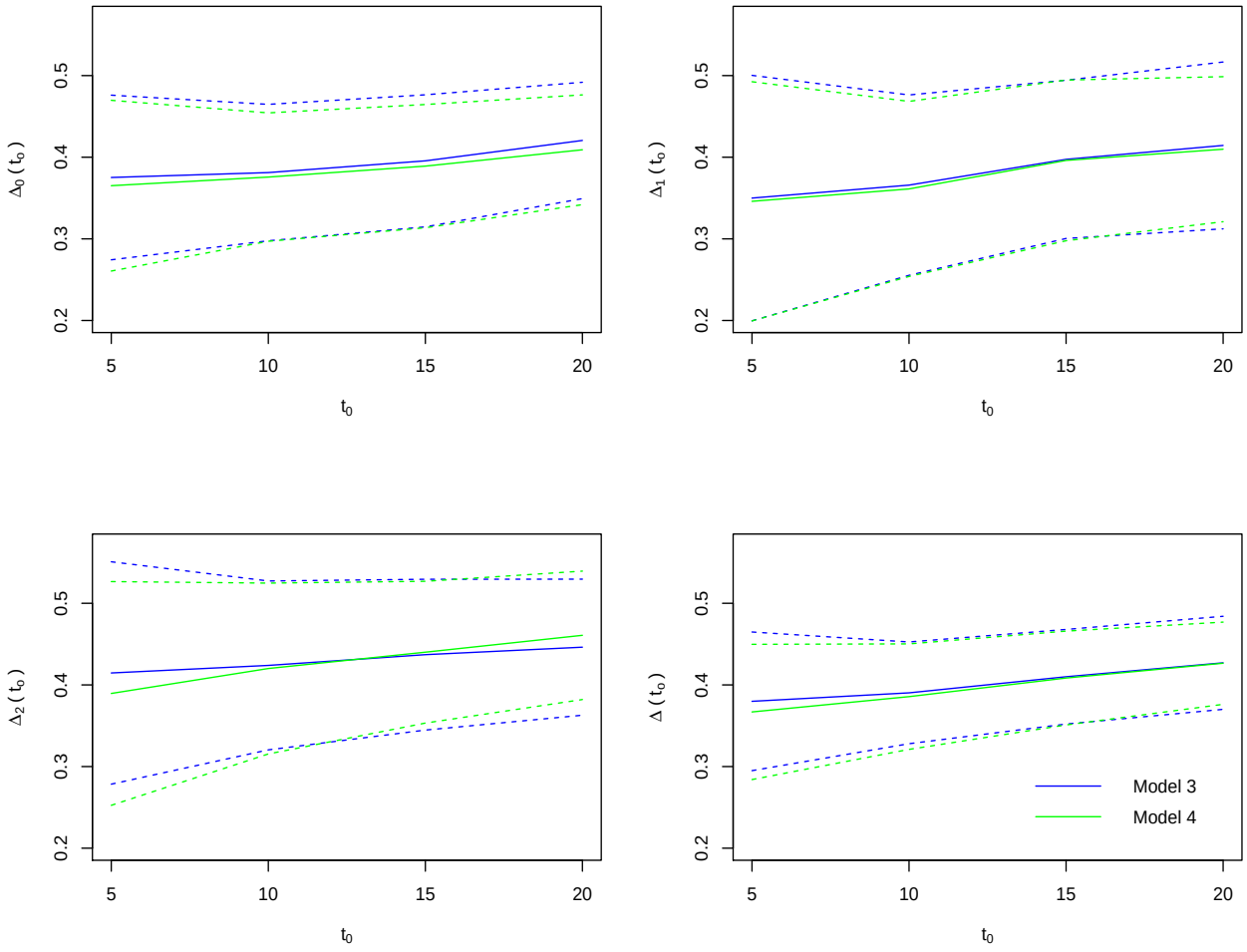


Figure S2: Plots of the predictive discrimination index estimates as a function of $t_\circ$ and the empirical 95% confidence interval (dashed lines) for a piecewise constant intensity model with no constraints on covariate effects (Model 3), and a piecewise constant intensity model with constraints on covariate effects (Model 4); constraints ensure $0 \to 1$ and $0 \to 2$ regression coefficients for the onset of unilateral damage, and $1 \to 3$ and $2 \to 3$ regression coefficients for the development of axial disease, are respectively the same. $\Delta_0(t_\circ)$ corresponds to prediction of no SI-joint involvement, $\Delta_1(t_\circ)$ corresponds to prediction of unilateral SI-joint damage, $\Delta_2(t_\circ)$ corresponds to prediction of axial disease, and $\Delta(t_\circ)$ is the overall measure.

		$0 \to 1$	$0 \to 2$	$1 \to 3$	$2 \to 3$
		$\lambda_{01}(t)$ 95% CI	$\lambda_{02}(t)$ 95% CI	$\lambda_{13}(t)$ 95% CI	$\lambda_{23}(t)$ 95% CI
Model 1		0.009 (0.003, 0.023)	0.028 (0.015, 0.050)	0.009 (0.001, 0.094)	0.187 (0.069, 0.508)
Model 2		0.009 (0.006, 0.014)	0.030 (0.022, 0.042)	0.073 (0.039, 0.136)	0.158 (0.096, 0.260)
Model 3	[0, 8)	0.015 (0.005, 0.042)	0.086 (0.039, 0.188)	0.038 (0.001, 1.119)	0.146 (0.047, 0.452)
	[8, 16)	0.007 (0.002, 0.033)	0.051 (0.014, 0.181)	0.049 (0.005, 0.537)	0.182 (0.041, 0.809)
	[16, 24)	0.005 (0.001, 0.032)	0.030 (0.008, 0.117)	0.007 (0.001, 0.102)	0.079 (0.016, 0.384)
	[24, ∞)	0.008 (0.002, 0.037)	0.011 (0.002, 0.048)	0.008 (0.001, 0.083)	0.213 (0.048, 0.947)
Model 4	[0, 8)	0.019 (0.010, 0.038)	0.072 (0.048, 0.107)	0.088 (0.031, 0.248)	0.156 (0.084, 0.291)
	[8, 16)	0.014 (0.004, 0.049)	0.088 (0.036, 0.216)	0.043 (0.009, 0.203)	0.050 (0.017, 0.151)
	[16, 24)	0.009 (0.002, 0.039)	0.023 (0.008, 0.071)	0.067 (0.011, 0.388)	0.070 (0.020, 0.245)
	[24, ∞)	0.009 (0.002, 0.034)	0.012 (0.004, 0.041)	0.036 (0.006, 0.220)	0.228 (0.073, 0.715)

Table S5: Estimate of baseline intensities from fitting Models 1 to 4 with the corresponding 95% confidence interval bands to data from the Univeristy of Toronto Psoriatic Arthritis Cohort for time homogeneous and piecewise constant models