Analysis of interval-censored recurrent event processes subject to resolution

HUA SHEN

Department of Statistics and Actuarial Science,

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

RICHARD J COOK

Department of Statistics and Actuarial Science,

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

E-mail: rjcook@uwaterloo.ca

Summary

Interval-censored recurrent event data arise when the event of interest is not readily observed but the cumulative event count can be recorded at periodic assessment times. In some settings, chronic disease processes may resolve, and individuals will cease to be at risk of events at the time of disease resolution. We develop an expectation-maximization algorithm for fitting a dynamic mover-stayer model to interval-censored recurrent event data under a Markov model with a piecewise-constant baseline rate function given a latent process. The model is motivated by settings in which the event times and the resolution time of the disease process are unobserved. The likelihood and algorithm are shown to yield estimators with small empirical bias in simulation studies. Data are analysed on the cumulative number of damaged joints in patients with psoriatic arthritis where individuals experience disease remission.

Keywords: EM algorithm; Interval censoring; Mover-stayer model; Piecewise-constant rate function; Recurrent events

1 INTRODUCTION

1.1 LITERATURE REVIEW AND GENERAL MOTIVATION

There are many chronic diseases for which affected individuals experience recurrent adverse events. Often it is apparent when the events occur, as is the case in respiratory disease when individuals experience exacerbations of symptoms (Grossman et al., 1998), neurologic disease where the events may be epileptic seizures (Pledger et al., 1994), migraine where the events are headaches (Pascual et al., 2000), and cardiology where the events may be acute angina attacks (Peters et al., 2003).

Statistical methods for recurrent event analysis in such settings include those reliant on intensity-based models (Andersen et al., 1993), random effect models (Lawless, 1987), and marginal methods (Lawless and Nadeau, 1995, Lin et al., 2000). Cook and Lawless (2007) give an account of the various frameworks for analysis.

The occurrence of an event is not evident in settings where they are not immediately symptomatic, but often they can be determined to have occurred at a subsequent clinic visit, radiographic examination, or blood test. Examples of such events include the development of new superficial tumours in bladder cancer patients (Byar et al., 1986), the occurrence of asymptomatic fractures in patients with osteoporosis (Riggs et al., 1981), or the development of new skeletal metastases in patients with cancer metastatic to bone (Hortobagyi et al., 1996). A considerable amount of statistical research has taken place in recent years on the analysis of this type of data which is referred to as panel count data, grouped recurrent event data, or interval-censored recurrent event data. Two-sample tests were developed by a number of authors including Thall and Lachin (1988), Sun and Fang (2003), Zhang (2006), Park et al. (2007) and Balakrishnan and Zhao (2009). Mean function estimation was considered by Sun and Kalbfleisch (1995), and their estimator was later shown by Wellner and Zhang (2000) to be a pseudo-maximum likelihood estimator under a nonhomogeneous Poisson model; these authors proved consistency, along with their proposed maximum likelihood estimator which did not depend on the Poisson assumption. Sun and Wei (2000), Cheng and Wei (2000), Zhang (2002) and Wellner and Zhang (2007) developed methods for semiparametric regression with interval-censored recurrent event data. Lawless and Zhan (1998) consider multiplicative recurrent event models with piecewise-constant baseline rate functions fitted via estimating functions as well as fully specified random effect models fitted using maximum likelihood. Such piecewise-constant models share the advantages of parametric models and yet provide some robustness to misspecification of the parametric form of rate functions. Chen et al. (2005) extend these methods to deal with multi-type recurrent events. Sun and Zhao (2013) give an excellent account of the recent developments on methods for recurrent event analysis when data are subject to interval-censoring.

In some settings the chronic condition generating the events can resolve and from the point of resolution individuals will no longer be at risk of events. Establishment of suitable medications, removal of stressors in mental health studies (Kessing et al., 2004), or other lifestyle changes may minimize risk of future events, but it can be difficult to determine if and when such changes have taken place. In other settings, the disease process resolves naturally. Polymyalgia rheumatica (Salvarani et al., 2002), for example, is an autoimmune disease and in the most active phase patients experience recurrent acute episodes of pain in the shoulder and pelvic joints. This active phase is of variable length but upon completion the acute episodes cease to arise (Healey, 1984). Patients with systemic lupus erythematosus experience flares of lupus nephritis, but this condition can go into remission such that patients stop experiencing acute flares (Barber et al., 2006). The disease in many rheumatological conditions can go into remission (Gladman et al., 2001, Zochling and Braun, 2006). In these and other settings one is faced with the challenge of interval-censored recurrent event times with the need to accommodate the possibility that the underlying condition has resolved. In the next sub-section, we describe a registry of patients with psoriatic arthritis and the scientific objectives of the associated research program which motivated the developments in this paper.

1.2   Diease progression and the University of Toronto Psoriatic Arthritis Cohort

Psoriatic arthritis is an inflammatory arthritis and an autoimmune disease that commonly occurs among patients with psoriasis. Patients with psoriatic arthritis may experience swelling, pain and inflammation in the affected joints. The University of Toronto Psoriatic Arthritis Clinic is the largest center in the world for specialized care and research in this disease. The clinic, started in 1978, has been recruiting and following patients continuously since then. Data collected at clinic entry and reg-

ular follow-up clinic visits arise from a complete history, physical examination, blood and urine tests, and radiographic examination. Over 1100 patients have been closely followed over the years.

The development of joint damage is of primary interest to clinicians since this damage impairs functional ability and quality of life. Much of the scientific research on this condition is directed at understanding the risks of rapid onset and accumulation of damage (Gladman et al., 1995). Factors studied include information on family history of psoriatic arthritis and genetic information based on human leukocyte antigen (HLA) markers, for example. Radiological examinations of the hands, feet and spine are scheduled every two years, but the actual assessment times vary considerably. Moreover, there are some patients who experience no joint damage over the entire course of follow-up, and others who develop damaged joints for some time but then have long periods of time in which no further damage is observed. One possible explanation for the latter scenario is that these patients experience remission and hence are no longer at risk for further damage. A key point is that individuals may transition from the mover (susceptible to damage) to stayer (remission) subgroup as time passes. Figure 1 displays the timing of the assessments (hatch marks) and the number of additional damaged joints detected (red numbers) over the respective intervals for a sample of 27 individuals; here we restrict attention to patient data over the first 30 years from disease onset. The variability in the frequency of visits is apparent, as is the variation in the event counts between patients and within patients overtime.
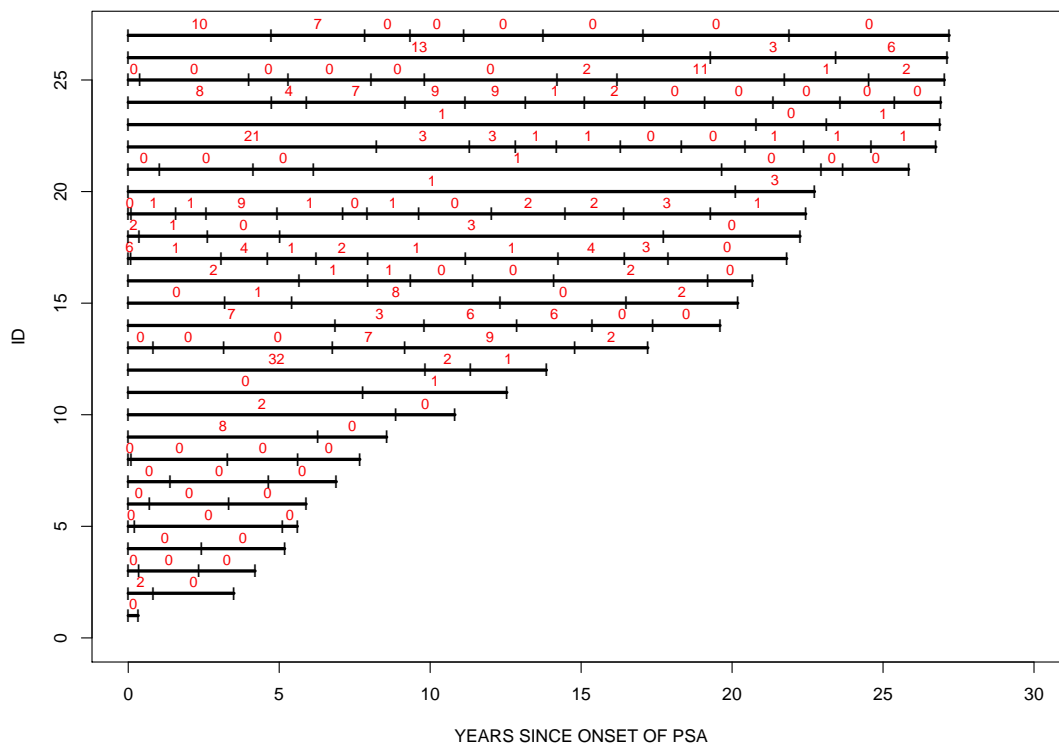
Figure 1: Plot of assessment times (hatch marks) and the number of damaged joints determined to have arisen between assessments (red numbers) by radiological examination for a selected sample of patients in the University of Toronto Psoriatic Arthritis Clinic; follow-up is restricted to the first 30 years following disease onset

HLA-B27 (human leukocyte antigen B27) is a protein found on the surface of white blood cell that has been found to be strongly associated with inflammatory autoimmune diseases such as ankylosing

spondylitis and psoriatic arthritis. Interest here lies in examining the effect of this HLA marker on the occurrence of damage as determined by radiographic examination of patients with psoriatic arthritis. When radiographs are taken, the extent of damage is assessed in each of 42 joints, including 30 hand joints (wrists, metacarpophalangeals, proximal interphalangeals and distal interphalangeals) and 12 joints in the foot (metatarsophalangeals and interphalangeal fist toes). The extent of damage in each joint is recorded on a six-point scale with numeric scores of 0 (normal), 1 (soft tissue swelling), 2 (surface erosions), 3 (joint space narrowing), 4 (disorganization, including subluxation, pencil-in-cup deformity and ankylosis), or 5 (damaged to the point of requiring surgery). A joint scoring 2 or higher is counted to be damaged to the point of impacting functional ability so the numbers in Figure 1 reflect the number of joints with at least grade 2 damage. Following the development of the proposed model and the algorithm for estimation, we return to this study in Section 3.2.

The remainder of this paper is organized as follows. In Section 2.1, the dynamic mover-stayer model is formulated. The complete data likelihood is constructed for the setting with interval-censored event times in Section 2.2 for a conditional Markov process for event occurence, and the requisite conditional expectations are derived in Section 2.3; details on how to implement the EM algorithm are given in the Appendix. The performance of the proposed model and algorithm are evaluated empirically in Section 3.1 and data from a psoriatic arthritis cohort are analysed in Section 3.2. General remarks and areas of further research are given in Section 4.

2 A DYNAMIC MOVER-STAYER MODEL UNDER INTERVAL-CENSORING

2.1 FORMULATION OF THE DYNAMIC MOVER-STAYER MODEL

Shen and Cook (2014) recently proposed a dynamic mover-stayer model for the analysis of right-censored recurrent event data that accommodates unusually long times from the last observed event to the right censoring time. If $T_0 = 0$ denotes the time of an initiating event such as the onset of a chronic disease, then T_j represents the time of the j-th subsequent event, $j = 1, 2, \ldots$. If $N(t) = \sum_{j=1}^{\infty} I(T_j \leq t)$ denotes the number of events over $(0, t]$, then $\{N(s), 0 \leq s\}$ is a counting process. The possible resolution of the process is accommodated by the introduction of Z_j, a time-dependent indicator, such that $Z_j = 1$ if the individual remains at risk of events following the occurrence of the jth event, and $Z_j = 0$ otherwise, $j = 0, 1, \ldots$. The resolution of the chronic condition upon the j-th event is reflected by a realization $Z_j = 0$ when $Z_{j-1} = 1$. Figure 2 is a schematic relating the random event times and random mover-stayer indicators for a hypothetical individual experiencing six events before resolution of the condition at T_6; the dashed line following T_6 relects the fact that this individual is no longer at risk of events. In general Z_j is a latent variable, but it is apparent that $Z_j = 1$ as soon as the $(j + 1)$-st event occurs, $j = 0, 1, \ldots$. As a consequence when events are observed in continuous time (or subject to right-censoring) the only unknown mover-stayer indicator is the one corresponding to the last observed event.

We let X denote a fixed covariate vector containing information on the characteristics of each individual observed upon recruitment to a cohort. We let $\bar{Z}(t) = \{Z_0, \ldots, Z_{N(t)}\}$ denote the history of the mover-stayer indicators for all events arising over $[0, t]$ and $\bar{Z}_j = \{Z_0, \ldots, Z_j\}$ denote the analogous history when viewed as a function of the event count. The complete process history is then denoted by $\mathcal{H}(t) = \{N(s), 0 \leq s \leq t, \bar{Z}(t), x\}$, which includes the values of the latent variables realized over $[0, t]$, and the history excluding the partially latent $\bar{Z}(t)$ is denoted by $H(t) = \{N(s), 0 \leq s \leq t, x\}$. We let t^- denote an infinitesimal amount of time before t, $N(s, t) = N(t^-) - N(s^-)$ denote the number of the events over $[s, t)$, and $\Delta N(t) = N(t, t + \Delta t)$ denote the number of the events over the interval $[t, t + \Delta t)$. The *complete data intensity function* is then

$$\lambda(t|\mathcal{H}(t^-)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N(t) = 1|\mathcal{H}(t^-))}{\Delta t} = Z_{N(t^-)} \lambda(t|H(t^-)) , \qquad (1)$$
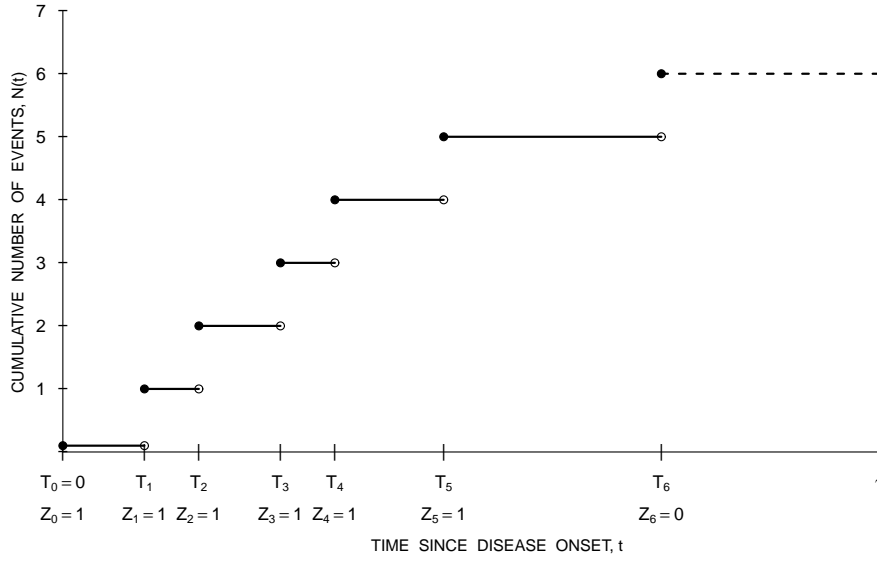
Figure 2: Schematic relating the observed and latent random variables in the dynamic mover-stayer model for a hypothetical individual experiencing six events before resolution at T_6

and the *canonical event intensity function* is

$$\lambda(t|H(t^-)) = \lim_{\triangle t \downarrow 0} \frac{P(\triangle N(t) = 1|H(t^-))}{\triangle t} \ . \tag{2}$$

The *intensity function* for the observable process (in the absence of censoring) is given by

$$E\{\lambda(t|\mathcal{H}(t^-))|H(t^-)\} = E(Z_{N(t^-)}|H(t^-)) \cdot \lambda(t|H(t^-)) \ .$$

The probability of remaining at risk following the j-th event can depend upon $\mathcal{H}(t_j^-)$, so at t_j we write this as

$$P(Z_j = 1|\mathcal{H}(t_j^-), dN(t_j) = 1) = P(Z_j = 1|H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, dN(t_j) = 1) \ , \tag{3}$$

where $dN(t) = \lim_{\triangle t \downarrow 0} \triangle N(t)$ indicates whether an event occurred at time t and 1_{j-1} is a $j \times 1$ vector of ones. This probability may therefore depend on the times of previous events and the history of the observable covariates over $[0, t_j]$ and is only relevant if $\bar{Z}_{j-1} = 1_{j-1}$. Here $j = 0, 1, \ldots, \mathcal{H}(t_0^-) = \emptyset$ and $\bar{Z}_{-1} = \emptyset$.

2.2   THE COMPLETE DATA LIKELIHOOD FOR INTERVAL-CENSORED DATA

We now consider the setting in which individuals are only seen intermittently but continue to focus on the contributions from a single individual. While the general formulation of the model in the previous section can be exploited for right-censored data, as in most settings involving interval-censored life history processes (Kalbfleisch and Lawless, 1985), we focus here on processes which are Markov given the mover-stayer indicators. We let $a_0 = 0$ denote the time of a baseline assessment at which a fixed covariate vector x is observed, and let $a_1 < \cdots < a_R$ denote the times of the follow-up assessments. At the r-th follow-up assessment at time a_r, the number of events over interval $\mathcal{A}_r = [a_{r-1}, a_r)$ is recorded and we denote this more compactly at times by $n_r = N(a_{r-1}, a_r)$, $r = 1, \ldots, R$, where $n = \sum_{r=1}^R n_r$. At any time $t \in \mathcal{A}_{r+1}$ the observed history is

denoted by $H_r = \{(a_\ell, n_\ell), \ell = 1, 2, \ldots, r, x\}$. The full data for an individual is then $D = H_R = \{(a_r, n_r), r = 1, \ldots, R, x\}$.

We assume individuals are observed from the time of disease onset in which case $N(0) = 0$ and H_0 simply contains the covariate data. In the present setting, this is reasonable because individuals invariably seek medical care when joints become inflamed, which occurs considerably earlier to the development of joint damage. Since the times of the assessments are random, the observed data likelihood is

$$L \propto \prod_{r=1}^{R} P(N_r = n_r, A_r = a_r | H_{r-1}) \ .$$

The assessment process is sequentially ignorable (Hogan et al., 2004, Lawless, 2013) in a likelihood based analysis if given H_{r-1}, the probability an assessment is made at a_r is independent of event occurence over $[a_{r-1}, a_r)$. Under this assumption (i.e. $A_r \perp N(a_{r-1}, a_r) | H_{r-1}$), we can write

$$P(N_r, A_r = a_r | H_{r-1}) = P(N_r | A_r = a_r, H_{r-1}) P(A_r = a_r | H_{r-1}) \ ,$$

(Cook and Lawless, 2014). Furthermore, if the inspection process is *non-informative*, we can disregard the contributions of the form $P(A_r = a_r | H_{r-1})$ and work with the resulting partial log-likelihood

$$\log L = \sum_{r=1}^{R} \log P(N_r = n_r | A_r = a_r, H_{r-1}) \tag{4}$$

without any loss of efficiency. Maximization of (4) is challenging however, since the expression for the necessary conditional probabilities are complicated and existing software cannot be used for maximization of this objective function. We develop an expectation-maximization algorithm in what follows.

Let θ_1 index the canonical event intensity (2), θ_2 index the mover-stayer model (3), and $\theta = (\theta_1', \theta_2')'$. A complete data log-likelihood corresponding to (4) can be constructed by considering the event times and the latent mover-stayer indicators as observed. The contribution to such a log-likelihood from an individual is then

$$\ell_C(\theta) = \ell_{C1}(\theta_1) + \ell_{C2}(\theta_2) \tag{5}$$

where

$$\ell_{C1}(\theta_1) = \int_0^\infty Y(u) \left[dN(u) \log \lambda(u|H(u^-)) - Z_{N(u)} \lambda(u|H(u^-)) du \right] \tag{6}$$

$$\ell_{C2}(\theta_2) = \sum_{j=0}^{n} \log P(Z_j | H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, dN(t_j) = 1) \tag{7}$$

pertain to the event process and mover-stayer model, respectively, and $Y(u) = I(u \le a_R)$.

Suppose interest lies in modeling data from individuals over the interval $[0, \tau]$ where τ is fixed. We focus here on settings with latent multiplicative Poisson processes, where $\lambda(t|H(t^-)) = \rho(t|x) = \rho_0(t) \exp(x'\beta)$ and $\rho_0(t)$ is the canonical baseline rate function. Given a set of cut-points $0 = b_0 < b_1 < \cdots < b_K = \tau$, a piecewise-constant baseline rate function is obtained by letting $\rho_0(t) = \rho_k$ for $t \in \mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. Let $u_{jk} = I(j = k)$, $j = 1, 2, \ldots, K$, $k = 1, 2, \ldots, K$, $x_k = (u_{1k}, \ldots, u_{Kk}, x')'$, $u_k(t) = I(t \in \mathcal{B}_k)$, $k = 1, 2, \ldots, K$, and $x(t) = (u_1(t), \ldots, u_K(t), x')'$. One can then write

$$\rho(t|x; \theta_1) = \rho_0(t; \alpha) \exp(x'\beta) = \exp(x'(t)\theta_1) = \prod_{k=1}^{K} [\exp(\alpha_k + x'\beta)]^{u_k(t)} \ .$$
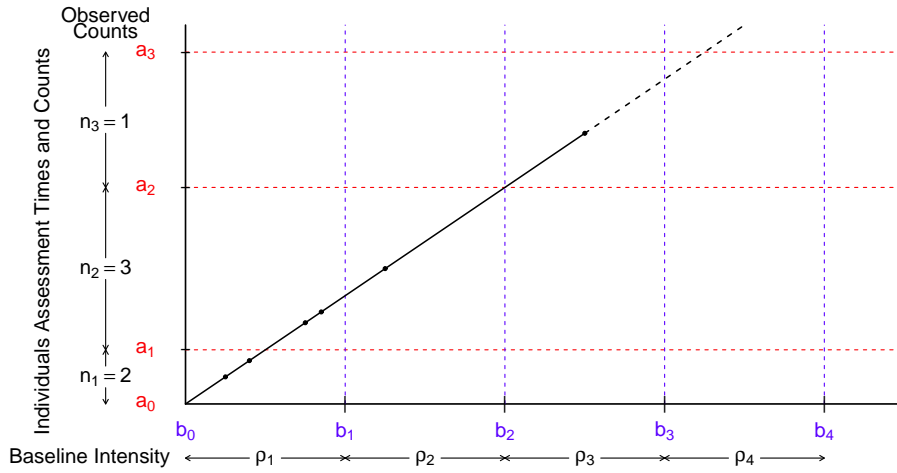
Figure 3: A two-way plot of the cut-points (horizontal axis), assessment times (vertical axis) and events (dots on the diagonal timeline); this hypothetical individual experienced six events before disease resolution at T_6

where $\alpha_k = \log \rho_k$, $k = 1, \ldots, K$, $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and $\theta_1 = (\alpha', \beta')'$. Figure 3 illustrates the observed counts and time intervals using this notation for a hypothetical individual experiencing six events before disease resolution at T_6.

We let $\mathcal{C}_{rk} = \mathcal{A}_r \cap \mathcal{B}_k = [C_{r,k-1}, C_{rk})$ and let $\mathcal{K}^r = \{k : \mathcal{A}_r \cap \mathcal{B}_k \neq \varnothing\}$ represent the labels for the q_r $(0 < q_r \leq K)$ pieces intersecting $\mathcal{A}_r$, denoted $\{k_\ell^r, \ell = 1, \ldots, q_r\}$. If we let $n_{rk} = \int I(u \in \mathcal{C}_{rk})dN(u)$ denote the number of events over $\mathcal{C}_{rk}$ and $w_k(t) = \int_0^t I(u \in \mathcal{B}_k)du$ denote the time at risk in $\mathcal{B}_k$ over $(0, t]$, then (6) can be rewritten as

$$\sum_{k=1}^K \left\{ \sum_{r=1}^R n_{rk}(\alpha_k + x'\beta) - [Z_n w_k(a_R) + (1 - Z_n)w_k(t_n)] \exp(\alpha_k + x'\beta) \right\}. \tag{8}$$

There are many discrete waiting time models suitable for modeling the resolution of the process. The model in (3) suggests that the conditional probability that the process resolves at t_j can depend on the history up to t_j^-, but this is of course not observed under an intermittent inspection process. While such dependence could be handled by treating this as a missing covariate problem we assume instead that

$$P(Z_j = 1 | H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, dN(t_j) = 1) = P(Z_j = 1 | x, \bar{Z}_{j-1} = 1_{j-1})$$

in which case we may adopt a logistic model of the form

$$\text{logit} P(Z_j = 1 | \mathcal{H}(t_j^-), dN(t_j) = 1) = \dot{x}_j' \gamma_j \tag{9}$$

where $\dot{x}_j = (1, j, x')'$.

2.3 Derivation of the conditional expectations

Since the actual events times and the final mover-stayer indicator are not observed, the quantities n_{rk}, $w_k(a_R)$, $w_k(t_n)$, and Z_n in (8) are unknown and we require expressions for their conditional expectations (Dempster et al., 1977). We focus initially on the expectations given D and Z_n, and consider first the case in which $Z_n = 1$. We let

$$\eta_{rk}^{(1)} = E(n_{rk} | D, Z_n = 1) = \frac{n_r \mu_{rk}}{\sum_{k \in \mathcal{K}^r} \mu_{rk}}, \tag{10}$$

where

$$\mu_{rk} = \int I(u \in \mathcal{C}_{rk}) \rho(u|x) du = \exp(\alpha_k + x'\beta) |\mathcal{C}_{rk}| \ ,$$

and $|\mathcal{C}_{rk}|$ denotes the length of $\mathcal{C}_{rk}$, $k \in \mathcal{K}^r$, $r = 1, \ldots, R$ (Lawless and Zhan, 1998). A superscript of 1 is used in (10) to reflect the fact that the expectation is conditional on $Z_n = 1$. Since $w_k(a_R) = \sum_{r=1}^{R} w_{rk}(a_R)$ where $w_{rk}(a_R) = \int_0^{a_R} I(u \in \mathcal{C}_{rk}) du$, given $Z_n = 1$ we can write

$$\omega_{rk}^{(1)} = E(w_{rk}(a_R)|D, Z_n = 1) = |\mathcal{C}_{rk}| \ .$$

If s denotes the index for the inspection interval containing t_n, the time of the final event observed over $(0, \tau]$ (i.e. $t_n \in \mathcal{A}_s$), then for $r < s$, $E(n_{rk}|D, Z_n = 0) = E(n_{rk}|D, Z_n = 1)$ as in (10), and for $r > s$, $E(n_{rk}|D, Z_n = 0) = 0$, $k \in \mathcal{K}^r$, by the definition of $\mathcal{A}_s$. Note that $E(n_{sk}|D, Z_n = 0)$, $k \in \mathcal{K}^s$, can be obtained by conceptualizing a progressive time nonhomogeneous multistate Markov process with a finite number of states labelled $N(a_{s-1}), \ldots, N(a_s)$ where only $\ell \to \ell + 1$ transitions are allowed with a common "transition" intensity $\rho(u|x)$, $\ell = N(a_{s-1}), \ldots, N(a_s) - 1$ and $N(a_s) = n$ is an absorbing state. We let $\boldsymbol{n}_s = (n_{sk}, k = k_1^s, \ldots, k_{q_s}^s)$ denote the counts over the subintervals of $\mathcal{A}_s$, let $\bar{n}_{sk} = n(a_{s-1}) + \sum_{j \in \mathcal{K}^s} I(j \le k) n_{sj}$ denote the cumulative count at C_{sk}, and let $\bar{\boldsymbol{n}}_s = (\bar{n}_{sk}, k = k_1^s, \ldots, k_{q_s}^s)$ denote the vector of cumulative counts. We can then write $P(\boldsymbol{n}_s|D, Z_n = 0)$ as

$$P(N(C_{sk}) = \bar{n}_{sk} \text{ for all } k \in \mathcal{K}^s | N(a_{s-1}) = n(a_{s-1}), N(a_s) = n(a_s), x, Z_n = 0) \ ,$$

where $n(a_{s-1}) = n - n_s$ and $n(a_s) = n$ by the definition of $\mathcal{A}_s$. This can in turn be written as

$$P(\boldsymbol{n}_s|D, Z_n = 0) = \frac{P(N(C_{sk}) = \bar{n}_{sk} \text{ for all } k \in \mathcal{K}^s | N(a_{s-1}) = n(a_{s-1}), x, Z_n = 0)}{P(N(a_s) = n | N(a_{s-1}) = n(a_{s-1}), x, Z_n = 0)} \ , \qquad (11)$$

where the numerator is equal to

$$\prod_{k \in \mathcal{K}^s} P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, x, Z_n = 0) \qquad (12)$$

by the Markov property, and the denominator is

$$\sum_{\boldsymbol{n}_s \in \mathcal{N}_s} \prod_{k \in \mathcal{K}^s} P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, x, Z_n = 0) \ , \qquad (13)$$

where $\mathcal{N}_s = \{\boldsymbol{n}_s : n_s = \sum_{k \in \mathcal{K}^s} n_{sk}\}$ is the set of all vectors $\boldsymbol{n}_s$ compatible with observed total. To determine the terms in (12), for a given vector $\boldsymbol{n}_s$ consider the specific subinterval $\mathcal{C}_{s\ell}$ containing t_n (i.e. $t_n \in \mathcal{C}_{s\ell}$). For $k \in \mathcal{K}^s$ with $k < \ell$, we have

$$P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, x, Z_n = 0) = \mu_{sk}^{n_{sk}} \exp(-\mu_{sk})/n_{sk}! \ ,$$

where $\mu_{sk} = \exp(\alpha_k + x'\beta) |\mathcal{C}_{sk}|$, and when $k > \ell$, $P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, x, Z_n = 0) = 1$. When $k = \ell$, a time-homogeneous Markov process governors events over $\mathcal{C}_{s\ell}$ with allowable transitions $0 \to 1 \to \cdots \to N_\ell = n_{s\ell}$ occurring with rate $\exp(\alpha_\ell + x'\beta)$.

Note that the probability of making transition from state i to state j, $i, j = 0, \ldots, N_\ell$, within time t is

$$P_{ij}(t) = (\exp(\alpha_\ell + x'\beta) t)^{j-i} \exp(-\exp(\alpha_\ell + x'\beta) t)/(j-i)! \qquad (14)$$

if $0 \le i \le j < N_\ell$, with $P_{iN_\ell}(t) = 1 - \sum_{j=i}^{N_\ell - 1} P_{ij}(t)$ if $0 \le i \le N_\ell - 1$. Given this we can calculate $P(N(C_{s\ell}) = \bar{n}_{s\ell} | N(C_{s,\ell-1}) = \bar{n}_{s,\ell-1}, x, Z_n = 0)$ as $P_{0N_\ell}(|\mathcal{C}_{s\ell}|)$. Also note that (11) can be used to compute the conditional expectation for the counts in each subinterval of $\mathcal{A}_s$ since

$$\eta_{sk}^{(0)} = E(n_{sk}|D, Z_n = 0) = \sum_{n_{sk}=0}^{n_s} n_{sk} P(n_{sk}|D, Z_n = 0) \qquad (15)$$

where $P(n_{sk}|D, Z_n = 0) = \sum_{\boldsymbol{n}_s \in \mathcal{N}_s} I(N_{sk} = n_{sk})P(\boldsymbol{n}_s|D, Z_n = 0)$ for any $k \in \mathcal{K}^s$.

For $r < s$, $E(w_{rk}(t_n)|D, Z_n = 0) = |\mathcal{C}_{rk}|$ for all $k \in \mathcal{K}^r$ and for $r > s$ $E(w_{rk}(t_n)|D, Z_n = 0) = 0$ for all $k \in \mathcal{K}^r$ by the definition of $\mathcal{A}_s$. For $r = s$, $\mathcal{K}^s = \{k_\ell^s, \ell = 1, \ldots, q_s\}$, and for a given vector of counts $\boldsymbol{n}_s = (n_{sk}, k \in \mathcal{K}^s)$, we can find an ℓ such that $t_n \in \mathcal{C}_{s\ell}$. Once again, for $k \in \mathcal{K}^s$, when $k < \ell$, we have $E(w_{sk}(t_n)|D, Z_n = 0) = |\mathcal{C}_{sk}|$ and when $k > \ell$, $E(w_{sk}(t_n)|D, Z_n = 0) = 0$ by the definition of $\mathcal{C}_{s\ell}$. When $r = s$, and $k = \ell$,

$$w_{s\ell}(t_n) = \int_0^{t_n} I(u \in \mathcal{C}_{s\ell})du = \int_{\mathcal{C}_{s\ell}} I(t_n > u)du = \int_{\mathcal{C}_{s\ell}} I(N(u) < n)du \;,$$

and for $u \in \mathcal{C}_{s\ell}$, the latter expression can be helpful since $P(N(u) < n|D, Z_n = 0, \boldsymbol{n}_s)$ is

$$\sum_{j=0}^{n_{s\ell}-1} \frac{P(N(u) = j|N(C_{s,\ell-1}) = 0, x, Z_n = 0)P(N(C_{s\ell}) = n_{s\ell}|N(u) = j, x, Z_n = 0)}{P(N(C_{s\ell}) = n_{s\ell}|N(C_{s,\ell-1}) = 0, x, Z_n = 0)} \;. \tag{16}$$

Note that over $\mathcal{C}_{s\ell}$, we again have a continuous time Markov process with a time homogenous transition intensity $\exp(\alpha_\ell + x'\beta)$ and transitions from ℓ to $\ell + 1$ for $\ell = 0, \ldots, N_\ell - 1$ where $N_\ell = n_{s\ell}$, we can therefore use (14) to obtain the values of the three items in (16) for given $u \in \mathcal{C}_{s\ell}$, and obtain

$$E(w_{s\ell}(t_n)|D, \boldsymbol{n}_s, Z_n = 0) = \int_{\mathcal{C}_{s\ell}} P(N(u) < n|D, \boldsymbol{n}_s, Z_n = 0)du \tag{17}$$

via numerical integration. Finally, for any $k \in \mathcal{K}^s$,

$$E(w_{sk}(t_n)|D, Z_n = 0) = \sum_{\boldsymbol{n}_s \in \mathcal{N}_s} E(w_{sk}(t_n)|D, \boldsymbol{n}_s, Z_n = 0)P(\boldsymbol{n}_s|D, Z_n = 0) \;,$$

where $P(\boldsymbol{n}_s|D, Z_n = 0)$ is given in (11).

Let $H_s = \{(a_r, n_r), r = 1, \ldots, s, x\}$ denote the inspection times and counts up to and including the inspection interval containing the last event known to have occurred, along with the covariate vector. The remainder of the data are denoted by $H_{s+1,R} = \{(a_r, n_r), r = s + 1, \ldots, R, x\}$. Then

$$\begin{aligned} \zeta &= P(Z_n = 1|D) = P(Z_n = 1|D, \bar{Z}_{n-1} = 1_{n-1}) \\ &= \frac{P(H_{s+1,R}|H_s, Z_n = 1, \bar{Z}_{n-1} = 1_{n-1})P(Z_n = 1|H_s, \bar{Z}_{n-1} = 1_{n-1})}{\sum_{z=0}^1 P(H_{s+1,R}|H_s, Z_n = z, \bar{Z}_{n-1} = 1_{n-1})P(Z_n = z|H_s, \bar{Z}_{n-1} = 1_{n-1})} \end{aligned} \tag{18}$$

where

$$P(H_{s+1,R}|H_s, Z_n = z, \bar{Z}_{n-1} = 1_{n-1}) = \prod_{r=s+1}^R P((a_r, n_r)|H_{r-1}, Z_n = z, \bar{Z}_{n-1} = 1_{n-1}) \;.$$

Note that $P((a_r, n_r)|H_{r-1}, Z_n = z, \bar{Z}_{n-1} = 1_{n-1})$ can be written as

$$P(n_r|a_r, H_{r-1}, Z_n = z, \bar{Z}_{n-1} = 1_{n-1})P(a_r|H_{r-1}, Z_n = z, \bar{Z}_{n-1} = 1_{n-1}) \;.$$

If $a_r \perp Z_n|H_{r-1}, \bar{Z}_{n-1} = 1_{n-1}$, under the conditionally (given $Z_n = z$) sequentially missing at random assumption (Hogan et al., 2004), the second term on the right-hand side does not depend on z. Moreover,

$$P(n_r = 0|a_r, H_{r-1}, Z_n = z, \bar{Z}_{n-1} = 1_{n-1}) = \begin{cases} \exp(-\mu_r) & \text{if } z = 1 \\ 1 & \text{if } z = 0 \;, \end{cases}$$

where $\mu_r = \int_{\mathcal{A}_r} \rho(u|x)du = \sum_{k \in \mathcal{K}^r} \mu_{rk}$ for $r = s+1, \cdots, R$. Under the conditionally sequentially MAR assumption, $P(Z_n = z | H_s, \bar{Z}_{n-1} = 1_{n-1})$ is given by

$$\frac{P(n_s | a_s, H_{s-1}, Z_n = z, \bar{Z}_{n-1} = 1_{n-1}) P(Z_n = z | H_{s-1}, \bar{Z}_{n-1} = 1_{n-1})}{\sum_{z=0}^{1} P(n_s | a_s, H_{s-1}, Z_n = z, \bar{Z}_{n-1} = 1_{n-1}) P(Z_n = z | H_{s-1}, \bar{Z}_{n-1} = 1_{n-1})} ,$$

where $P(Z_n = 1 | H_{s-1}, \bar{Z}_{n-1} = 1_{n-1}) = P(Z_n = 1 | x, \bar{Z}_{n-1} = 1_{n-1})$ is given by (9) and $P(n_s | a_s, H_{s-1}, Z_n = 0, \bar{Z}_{n-1} = 1_{n-1})$ is given by (13).

3 EMPIRICAL STUDIES AND AN APPLICATION

3.1 DESIGN AND INTERPRETATION OF SIMULATION STUDIES

In this section, a simulation study is conducted to evaluate the performance of the method proposed to deal with interval-censored recurrent event data with disease resolution. We let X_i denote a binary covariate for individual i which takes on the value of 1 or 0 with equal probability to reflect, for example, whether an individual is on an experimental or control treatment. We then generate the initial mover-stayer indicators Z_{ij}, $j = 0, \cdots, n_i$, following the model

$$\text{logit} P(Z_{ij} = 1 | X_i, \bar{Z}_{i,j-1} = 1_{j-1}) = \gamma_0 + \gamma_1 j + \gamma_2 X_i ,$$

where n_i is the total number of events observed over the entire study period $(0, \tau]$ if there is administrative censoring only. For simplicity, we set $\tau = 1$ for all subjects. γ_1 and γ_2 are set as $\log 0.95$ and $\log 0.75$ respectively, so that for given treatment X_i, the odds of being a mover decreases by 5% with the occurrence of each additional event for the same treatment group, and the odds of being a mover is 25% lower if $X_i = 1$ compared to when $X_i = 0$ for a given $N_i(t)$. For the purpose of illustration, we assume the events are generated according to a homogenous Poisson process among movers, in which case the gap times follow an exponential distribution with rate $\lambda \exp(X_i \beta)$; here λ is the baseline intensity and β is reflects the treatment effect on the event rate. We set $\beta = \log 0.75$ so that there is a 25% reduction in the event rate among movers on treatment and solve for λ such that the expected number of events over $(0, \tau]$ is 6 or 12 among individuals for whom the disease process does not resolve. We then solve for γ_0 to ensure the expected number of events over $(0, \tau]$ is 1.5, 3 or 6. A piecewise-constant baseline intensity model was adopted wherein cut-points were set by dividing the observation period $(0, \tau]$ evenly into $K = 3$ intervals.

For simplicity, we assume here that each subject has the same number of prescheduled evenly spaced assessments with $R_i = R = 4$ or 8; as described in Section 2 we observe only the number of events occuring between clinic visits, along with the treatment indicator. If $\vartheta = (\alpha_1, \alpha_2 - \alpha_1, \cdots, \alpha_K - \alpha_1, \beta, \gamma_0, \gamma_1, \gamma_2)$, we terminate the EM algorithm of the previous section when $\max(|\vartheta_j^{new} - \vartheta_j^{old}|) < \epsilon$, where $\epsilon = 10^{-6}$. A total of 2000 simulations were conducted for each parameter configuration with a sample size of $m = 500$ or 2000. The empirical biases (EBIAS) and empirical standard errors (ESE) are reported in Table 1.

The simulation results in Table 1 show that the empirical biases are generally small, and decrease as the sample size increases. Moreover for a given sample size, the estimators have smaller empirical biases and smaller empirical standard errors the more frequent the assessments. When the expected numbers of events among the movers are the same, the covariate effects on the recurrent event process have smaller empirical biases and smaller empirical standard errors when the marginal expected number of events is larger; in such cases, the empirical standard errors in the estimates of the event rate are also smaller. For a given marginal expected numbers of events, both the association between the explanatory variables and the mover-stayer indicator and the covariate effect on the event process have smaller standard errors and negligible biases. When the assessment times are governed by a Poisson process, similar results are seen.

Table 1: Simulation results where the gap times follow an exponential distribution by adopting a piecewise-constant model when the data are subject to administrative censoring only, $m = 500$ or $m = 2000$, $nsim = 2000$, $\epsilon = 10^{-6}$

| | | R=4 | | | | R=8 | | | |
| | | m=500 | | m=2000 | | m=500 | | m=2000 | |
Parameter	Value	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE
		$E(N(1)) = 1.5$, $E(N(1)|Z_n = 1) = 6$							
γ_0	0.7091	0.0041	0.1009	0.0001	0.0499	0.0039	0.1005	0.0002	0.0497
γ_1	-0.0513	-0.0014	0.0510	0.0005	0.0256	-0.0022	0.0481	0.0000	0.0237
γ_2	-0.2877	-0.0038	0.1237	0.0004	0.0628	-0.0040	0.1224	0.0001	0.0624
α_1	1.9253	0.0009	0.0654	0.0004	0.0329	0.0003	0.0621	0.0000	0.0306
α_2	1.9253	-0.0113	0.1035	-0.0023	0.0504	-0.0036	0.0897	-0.0006	0.0441
α_3	1.9253	-0.0095	0.2668	-0.0069	0.1310	-0.0086	0.2052	-0.0043	0.1016
β	-0.2877	0.0005	0.0896	-0.0019	0.0431	0.0005	0.0871	-0.0017	0.0410
		$E(N(1)) = 3$, $E(N(1)|Z_n = 1) = 6$							
γ_0	1.7331	0.0100	0.1174	0.0038	0.0584	0.0105	0.1165	0.0037	0.0580
γ_1	-0.0513	0.0008	0.0424	-0.0007	0.0206	0.0002	0.0398	-0.0006	0.0192
γ_2	-0.2877	-0.0068	0.1395	-0.0030	0.0684	-0.0076	0.1376	-0.0029	0.0677
α_1	1.9253	-0.0017	0.0500	-0.0014	0.0252	-0.0013	0.0474	-0.0012	0.0240
α_2	1.9253	-0.0004	0.0661	0.0009	0.0326	0.0001	0.0589	0.0010	0.0291
α_3	1.9253	-0.0021	0.1163	0.0003	0.0594	-0.0014	0.1003	-0.0006	0.0499
β	-0.2877	-0.0004	0.0641	0.0002	0.0313	-0.0001	0.0631	0.0002	0.0307
		$E(N(1)) = 3$, $E(N(1)|Z_n = 1) = 12$							
γ_0	1.4123	0.0032	0.0944	-0.0002	0.0484	0.0036	0.0938	-0.0002	0.0481
γ_1	-0.0513	-0.0018	0.0221	-0.0003	0.0113	-0.0022	0.0211	-0.0003	0.0108
γ_2	-0.2877	0.0010	0.1089	0.0004	0.0542	0.0005	0.1085	0.0003	0.0541
α_1	2.6184	0.0032	0.0447	0.0001	0.0221	0.0024	0.0415	0.0000	0.0205
α_2	2.6184	-0.0026	0.0779	-0.0000	0.0389	0.0007	0.0654	0.0004	0.0329
α_3	2.6184	-0.0039	0.2066	0.0004	0.0999	-0.0004	0.1426	0.0004	0.0701
β	-0.2877	-0.0042	0.0592	-0.0012	0.0289	-0.0040	0.0560	-0.0011	0.0274
		$E(N(1)) = 6$, $E(N(1)|Z_n = 1) = 12$							
γ_0	2.4275	0.0037	0.1145	0.0003	0.0575	0.0046	0.1125	0.0005	0.0568
γ_1	-0.0513	-0.0005	0.0187	-0.0001	0.0091	-0.0009	0.0175	-0.0002	0.0086
γ_2	-0.2877	0.0041	0.1194	0.0007	0.0599	0.0039	0.1186	0.0007	0.0591
α_1	2.6184	0.0020	0.0342	0.0006	0.0168	0.0013	0.0321	0.0005	0.0157
α_2	2.6184	-0.0008	0.0482	0.0003	0.0234	0.0005	0.0409	0.0004	0.0204
α_3	2.6184	-0.0011	0.0783	-0.0011	0.0386	0.0004	0.0623	-0.0006	0.0312
β	-0.2877	-0.0026	0.0410	-0.0008	0.0197	-0.0024	0.0402	-0.0008	0.0192

3.2 Application to joint damage in psoriatic arthritis

We consider a subcohort of patients with psoriatic arthritics from University of Toronto Psoriatic Arthritis Clinic. These 207 selected patients have disease onset time and HLA-B27 information available. They entered the clinic and were followed-up between 1978 and 2013. The observation period is limited to the first 30 years following disease onset. The reported age of disease onset is taken as the time origin and dates of radiological assessments and numbers of new damaged joints were recorded at the following assessment visits. The average time since disease onset to first radiological assessment is 5.54 years (S.D. 6.02, range 0.03 to 27.23). The average number of radiological assessments within 30 years of disease onset is 3.63 (S.D. 2.83, range 1 to 13). A total of 32 (15.5%) patients are HLA-B27 positive.

The data suggested that some patients experience remission during the follow-up. We fit the proposed algorithm on the interval-censored recurrent event data to study the occurrence of joint damage. Piecewise-constant baseline rate functions are adopted to model the recurrent event process not subject to resolution with one fixed covariate HLA-B27 ($X = 1$ if HLA-B27 positive, $X = 0$ if HLA-B27 negative); The canonical baseline rate are assumed to be constant for every 10 years. A dynamic mover-stayer model is fitted with the cumulative number of damaged joints (j) and HLA-B27 (X) being the explanatory variables in the logistic regression. A recurrent event model treating all patients as susceptible for joint damage is fitted for comparison, in which a piecewise-constant baseline rate model is assumed as well and the effect of HLA-B27 on event rate is also of interest.

The estimation results, presented in Table 2, demonstrate that the event rate is noticeably underestimated when all the patients are assumed to experience joint damage over the entire observation period. The effect of HLA-B27 on event rate is also underestimated though it is not significant in either model. The increased number of damaged joints is associated with higher odds of continuing to have new damaged joints.

Table 2: Results of fitting piecewise-constant baseline rate model and dynamic mover-stayer model to study the occurrence of joint damage among patients with psoriatic arthritis whose follow-ups are within 30 years of disease onset; $m = 207$, standard errors based on 100 bootstrap samples

	Recurrent Event Model		Dynamic Mover-Stayer Model	
	EST	SE	EST	SE
Mover-Stayer Model				
γ_0	—	—	1.636	0.257
γ_1	—	—	0.153	0.107
γ_2	—	—	-0.201	0.317
Recurrent Event Model				
α_1	-0.467	0.097	-0.012	0.109
α_2	-0.879	0.149	-0.509	0.188
α_3	-0.952	0.365	-0.133	0.537
β	-0.163	0.269	-0.069	0.357

4   DISCUSSION

We developed an expectation-maximization algorithm to analyze interval-censored recurrent event data for disease processes subject to resolution. While there is a large class of intensity functions that can be adopted when the event times are only right-censored, under interval-censoring the Markov assumption for the conditional event process leads to a highly tractable model; for progressive disease processes this is a reasonable framework for analysis. The algorithm makes use of existing softwarwe at the maximization step and is shown to produce estimators with small empirical bias for frequent and less frequent assessments.

Some degree of robustness to misspecification is achieved through use of a piecewise-constant baseline rate function, but extensions to deal with semiparametric models would be worthy of development. Possible avenues include adapting the pseudo-likelihood estimator proposed by Sun and Kalbfleisch (1995) for the mean function, or the semiparametric maximum likelihood approach of Wellner and Zhang (2000). Here, however, one might expect more challenges in maximization of the observed data likelihood whether by direct maximization or an extension of the algorithm we present here.

In this paper, we assumed the observation times and the event process are conditionally independent. It is important to highlight that the assumption of sequential ignorability accommodates the setting in which the next assessment is scheduled based on the observed history up to and including the current assessment. This reflects the scenario in which physicians schedule appointments based on the current state of the patient they are examining. Huang et al. (2006), Sun et al. (2007) and Zhao and Tong (2011) developed methods to deal with more complex assessment processes, which could be utilized in conjuction with the current model. These methods generally require joint modeling of the process of interest and the assessment process and hence require stronger modeling assumptions.

Cook et al. (2002) describe a generalized mover-stayer model for multistate data under interval censoring, which is somewhat similar in spirit to what we have described. In their model, conditional on the mover-stayer indicators, subjects move according to time-homogeneous Markov transition intensities. Here however, the first time an individual enters a state, a latent mover-stayer indicator is realized which can render it an absorbing state. Thus individuals can make transitions between a number of states before finally entering their absorbing state.

Often recurrent events arise in settings where the event process is terminated by some event. For example in transplant studies recurrent graft rejection episodes arise when recipients are experiencing graft versus host disease (Cole et al., 1994). This condition resolves at a latent time when the graft is accepted, but the process can also end upon in severe cases by total graft rejection or death of the patient. Adapting these methods to handle this situation is feasible but may again be more naturally addressed by casting the process into a multistate framework as in Conlon et al. (2014). Extensions of these methods would be useful for this setting as well.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

SUPPLEMENTARY MATERIAL

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site.

APPENDIX: THE EXPECTATION-MAXIMIZATION ALGORITHM

For the EM algorithm, we have

$$Q(\theta; \widetilde{\theta}) = Q_1(\theta_1; \widetilde{\theta}) + Q_2(\theta_2; \widetilde{\theta}) \,,$$

where $Q(\theta; \widetilde{\theta}) = E(\ell_C(\theta)|D, \widetilde{\theta})$ and $Q_j(\theta_j; \widetilde{\theta}) = E(\ell_{Cj}(\theta_j)|D, \widetilde{\theta})$, $j = 1, 2$, $\theta_1 = (\alpha', \beta')'$, $\theta_2 = \gamma$ and $\theta = (\theta_1', \theta_2')'$. Given the expressions in Section 2.3, adopting the notation $\widetilde{\eta}_{irk}^{(1)} = E(n_{irk}|D_i, Z_{in_i} = 1; \widetilde{\theta})$, $\widetilde{\eta}_{irk}^{(0)} = E(n_{irk}|D_i, Z_{in_i} = 0; \widetilde{\theta})$, $\widetilde{\omega}_{irk}^{(1)} = E(w_{irk}(a_{iR_i})|D_i, Z_{in_i} = 1; \widetilde{\theta})$, $\widetilde{\omega}_{irk}^{(0)} = E(w_{irk}(t_{in_i})|D_i, Z_{in_i} = 0; \widetilde{\theta})$ and $\widetilde{\zeta}_i = E(Z_{in_i}|D_i; \widetilde{\theta})$, we write $Q_1(\theta_1; \widetilde{\theta})$ as

$$
\begin{aligned}
Q_1(\theta_1; \widetilde{\theta}) = \sum_{i=1}^{m} & \left\{ \widetilde{\zeta}_i \sum_{k=1}^{K} \sum_{r=1}^{R_i} \left[\widetilde{\eta}_{irk}^{(1)}(\alpha_k + x_i'\beta) - \exp(\alpha_k + x_i'\beta + \log \widetilde{\omega}_{irk}^{(1)}) \right] \right. \\
& + \left. (1 - \widetilde{\zeta}_i) \sum_{k=1}^{K} \sum_{r=1}^{R_i} \left[\widetilde{\eta}_{irk}^{(0)}(\alpha_k + x_i'\beta) - \exp(\alpha_k + x_i'\beta + \log \widetilde{\omega}_{irk}^{(0)}) \right] \right\} \,.
\end{aligned}
$$

One can construct a pseudo-data frame for each individual as in Table A.1, which can be written more concisely as in Table A.2, with the call

$$\mathrm{glm}(\mathrm{n} \sim \mathrm{X} + \mathrm{factor}(\mathrm{k}) + \mathrm{offset}(\log \mathrm{w}), \mathrm{weight} = \mathrm{wt}, \mathrm{family} = \mathrm{poisson}, \mathrm{link} = \log)$$

used to obtain updated estimates of θ_1.

In addition, the function $Q_2(\theta; \widetilde{\theta})$ is

$$
\begin{aligned}
Q_2(\theta; \widetilde{\theta}) = \sum_{i=1}^{m} & \Bigg[\sum_{j=0}^{n_i-1} \log P(Z_{ij} = 1|H(t_{in_i}^-), \bar{Z}_{i,j-1} = 1_{j-1}, dN(t_{in_i}) = 1) \\
& + \widetilde{\zeta}_i \log P(Z_{in_i} = 1|H(t_{in_i}^-), \bar{Z}_{i,n_i-1} = 1_{n_i-1}, dN(t_{in_i}) = 1) \\
& + (1 - \widetilde{\zeta}_i) \log P(Z_{in_i} = 0|H(t_{in_i}^-), \bar{Z}_{i,n_i-1} = 1_{n_i-1}, dN(t_{in_i}) = 1) \Bigg] \,.
\end{aligned}
$$

Again we construct a pseudo-data frame as in Table A.3 and call the glm function

$$\mathrm{glm}(\mathrm{Z} \sim \mathrm{j} + \mathrm{x}, \mathrm{weight} = \mathrm{wt}, \mathrm{family} = \mathrm{quasi} - \mathrm{binomial}, \mathrm{link} = \mathrm{logit})$$

to obtain updated estimates of θ_2. Variance estimation (Louis, 1982) is described in the Supplementary Material.

Table A.1: Pseudo-data frame for recurrent event process for one subject using the proposed method, where r is the index of the inspection interval, k is the index of the piece of the baseline rate, X is the covariate, w is the expected time at risk, n is the expected number of events, and wt is the weight

r	k	X	w	n	wt
1	k_1^1	x	$\widetilde{\omega}_{1k_1^1}^{(1)}$	$\widetilde{\eta}_{1k_1^1}^{(1)}$	$\widetilde{\zeta}$
⋮	⋮	⋮	⋮	⋮	⋮
1	$k_{q_1}^1$	x	$\widetilde{\omega}_{1k_{q_1}^1}^{(1)}$	$\widetilde{\eta}_{1k_{q_1}^1}^{(1)}$	$\widetilde{\zeta}$
1	k_1^1	x	$\widetilde{\omega}_{1k_1^1}^{(0)}$	$\widetilde{\eta}_{1k_1^1}^{(0)}$	$1-\widetilde{\zeta}$
⋮	⋮	⋮	⋮	⋮	⋮
1	$k_{q_1}^1$	x	$\widetilde{\omega}_{1k_{q_1}^1}^{(0)}$	$\widetilde{\eta}_{1k_{q_1}^1}^{(0)}$	$1-\widetilde{\zeta}$
⋮	⋮	⋮	⋮	⋮	⋮
R	k_1^R	x	$\widetilde{\omega}_{Rk_1^R}^{(1)}$	$\widetilde{\eta}_{Rk_1^R}^{(1)}$	$\widetilde{\zeta}$
⋮	⋮	⋮	⋮	⋮	⋮
R	$k_{q_R}^R$	x	$\widetilde{\omega}_{Rk_{q_R}^R}^{(1)}$	$\widetilde{\eta}_{Rk_{q_R}^R}^{(1)}$	$\widetilde{\zeta}$
R	k_1^R	x	$\widetilde{\omega}_{Rk_1^R}^{(0)}$	$\widetilde{\eta}_{Rk_1^R}^{(0)}$	$1-\widetilde{\zeta}$
⋮	⋮	⋮	⋮	⋮	⋮
R	$k_{q_R}^R$	x	$\widetilde{\omega}_{Rk_{q_R}^R}^{(0)}$	$\widetilde{\eta}_{Rk_{q_R}^R}^{(0)}$	$1-\widetilde{\zeta}$

Table A.2: Pseudo-data frame for recurrent event process for one subject using the proposed method (simplified version), where r is the index of the inspection interval, k is the index of the piece of the baseline rate, X is the covariate, w is the expected time at risk, n is the expected number of events, and wt is the weight

r	k	X	w	n	wt
1	k_1^1	x	$\widetilde{\omega}_{1k_1^1}^{(1)}$	$\widetilde{\eta}_{1k_1^1}^{(1)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$k_{q_1}^1$	x	$\widetilde{\omega}_{1k_{q_1}^1}^{(1)}$	$\widetilde{\eta}_{1k_{q_1}^1}^{(1)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s-1$	k_1^{s-1}	x	$\widetilde{\omega}_{s-1,k_1^{s-1}}^{(1)}$	$\widetilde{\eta}_{s-1,k_1^{s-1}}^{(1)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s-1$	$k_{q_{s-1}}^{s-1}$	x	$\widetilde{\omega}_{s-1,k_{q_{s-1}}^{s-1}}^{(1)}$	$\widetilde{\eta}_{s-1,k_{q_{s-1}}^{s-1}}^{(1)}$	1
s	k_1^s	x	$\widetilde{\omega}_{sk_1^s}^{(1)}$	$\widetilde{\eta}_{sk_1^s}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	$k_{q_s}^s$	x	$\widetilde{\omega}_{sk_{q_s}^s}^{(1)}$	$\widetilde{\eta}_{sk_{q_s}^s}^{(1)}$	$\widetilde{\zeta}$
s	k_1^s	x	$\widetilde{\omega}_{sk_1^s}^{(0)}$	$\widetilde{\eta}_{sk_1^s}^{(0)}$	$1-\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	$k_{q_s}^s$	x	$\widetilde{\omega}_{sk_{q_s}^s}^{(0)}$	$\widetilde{\eta}_{sk_{q_s}^s}^{(0)}$	$1-\widetilde{\zeta}$
$s+1$	k_1^{s+1}	x	$\widetilde{\omega}_{s+1,k_1^{s+1}}^{(1)}$	$\widetilde{\eta}_{s+1,k_1^{s+1}}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s+1$	$k_{q_{s+1}}^{s+1}$	x	$\widetilde{\omega}_{s+1,k_{q_{s+1}}^{s+1}}^{(1)}$	$\widetilde{\eta}_{s+1,k_{q_{s+1}}^{s+1}}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	k_1^R	x	$\widetilde{\omega}_{Rk_1^R}^{(1)}$	$\widetilde{\eta}_{Rk_1^R}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$k_{q_R}^R$	x	$\widetilde{\omega}_{Rk_{q_R}^R}^{(1)}$	$\widetilde{\eta}_{Rk_{q_R}^R}^{(1)}$	$\widetilde{\zeta}$

Table A.3: Pseudo-data frame for mover-stayer model for one subject using the proposed method, where Z is the mover-stayer indicator, j is the number of events, X is the covariate, wt is the weight

Z	j	X	wt
1	0	x	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$n-1$	x	1
1	n	x	$\widetilde{\zeta}$
0	n	x	$1-\widetilde{\zeta}$

REFERENCES

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, NY.

Balakrishnan, N. and Zhao, X. (2009). New multi-sample nonparametric tests for panel count data. *The Annals of Statistics*, 37:1112–1149.

Barber, C., Geldenhuys, L., and Hanly, J. (2006). Sustained remission of lupus nephritis. *Lupus*, 15:94–101.

Byar, D., Kaihara, S., Sylvester, R., Freedman, L., Hannigan, J., Koiso, K., Oohashi, Y., and Tsugawa, R. (1986). Statistical analysis techniques and sample size determination for clinical trials of treatments for bladder cancer. *Progress in Clinical and Biological Research*, 221:49–64.

Chen, B., Cook, R., Lawless, J., and Zhan, M. (2005). Statistical methods for multivariate interval-censored recurrent events. *Statistics in Medicine*, 24:671–691.

Cheng, S. and Wei, L. (2000). Inferences for a semiparametric model with panel data. *Biometrika*, 87:89–97.

Cole, E., Cattran, D., Farewell, V., Aprile, M., Bear, R., Pei, Y., Fenton, S., Tober, J., and Cardella, C. (1994). A comparison of rabbit antithymocyte serum and OKT3 as prophylaxis against renal allograft rejection. *Transplantation*, 57:60–67.

Conlon, A., Taylor, J., and Sargent, D. (2014). Multi-state models for colon cancer recurrence and death with a cured fraction. *Statistics in Medicine*, 33:1750–1766.

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

Cook, R. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York, NY.

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

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

Gladman, D., Farewell, V., and Nadeau, C. (1995). Clinical indicators of progression in psoriatic arthritis: multivariate relative risk model. *The Journal of Rheumatology*, 22:675–679.

Gladman, D., Hing, E., Schentag, C., and Cook, R. (2001). Remission in psoriatic arthritis. *The Journal of Rheumatology*, 28:1045–1048.

Grossman, R., Mukherjee, J., Vaughan, D., Cook, R., LaForge, J., Lampron, N., and Eastwood, C. (1998). A 1-year community-based health economic study of ciprofloxacin vs usual antibiotic treatment in acute exacerbations of chronic bronchitis: the Canadian Ciprofloxacin Health Economic Study Group. *CHEST Journal*, 113:131–141.

Healey, L. (1984). Long-term follow-up of polymyalgia rheumatica: evidence for synovitis. *Seminars in Arthritis and Rheumatism*, 13:322–328.

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

Hortobagyi, G., Theriault, R., Porter, L., Blayney, D., Lipton, A., Sinoff, C., Wheeler, H., Simeone, J., Seaman, J., Knight, R., Heffernan, M., Reitsma, D., Kennedy, I., Allan, S., and Mellars, K. (1996). Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases. *New England Journal of Medicine*, 335:1785–1792.

Huang, C., Wang, M., and Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika*, 93:763–775.

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

Kessing, L., Hansen, M., and Andersen, P. (2004). Course of illness in depressive and bipolar disorders naturalistic study, 1994-1999. *The British Journal of Psychiatry*, 185:372–377.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15:209–225.

Lawless, J. (2013). The design and analysis of life history studies. *Statistics in Medicine*, 32:2155–2172.

Lawless, J. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37:158–168.

Lawless, J. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, 26:549–565.

Lin, D., Wei, L., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62:711–730.

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

Park, D.-H., Sun, J., and Zhao, X. (2007). A class of two-sample nonparametric tests for panel count data. *Communications in Statistics - Theory and Methods*, 36:1611–1625.

Pascual, J., Falk, R., Piessens, F., Prusinski, A., Docekal, P., Robert, M., Ferrer, P., Luria, X., Segarra, R., and Zayas, J. (2000). Consistent efficacy and tolerability of almotriptan in the acute treatment of multiple migraine attacks: results of a large, randomized, double-blind, placebo-controlled study. *Cephalalgia*, 20:588–596.

Peters, R., Mehta, S., Fox, K., Zhao, F., Lewis, B., Kopecky, S., Diaz, R., Commerford, P., Valentin, V., Yusuf, S., and Clopidogrel in Unstable angina to prevent Recurrent Events (CURE) Trial Investigators. (2003). Effects of aspirin dose when used alone or in combination with clopidogrel in patients with acute coronary syndromes observations from the clopidogrel in unstable angina to prevent recurrent events (cure) study. *Circulation*, 108:1682–1687.

Pledger, G., Sackellares, J., Treiman, D., Pellock, J., Wright, F., Mikati, M., Sahlroot, J., Tsay, J., Drake, M., Olson, L., Handforth, C., Garnett, W., Schachter, S., Kupferberg, H., Ashworth, M., McCormick, C., Leiderman, D., Kapetanovic, I., Driscoll, S., O'Hara, K., Torchin, C., Gentile, J., Kay, A., and Cereghino, J. (1994). Flunarizine for treatment of partial seizures results of a concentration-controlled trial. *Neurology*, 44:1830–1836.

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

Salvarani, C., Cantini, F., Boiardi, L., and Hunder, G. (2002). Polymyalgia rheumatica and giant-cell arteritis. *New England Journal of Medicine*, 347:261–271.

Shen, H. and Cook, R. (2014). A dynamic mover-stayer model for recurrent event process subject to resolution. *Lifetime Data Analysis*, 20:404–423.

Sun, J. and Fang, H.-B. (2003). A nonparametric test for panel count data. *Biometrika*, 90:199–208.

Sun, J. and Kalbfleisch, J. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, 5:279–290.

Sun, J., Tong, X., and He, X. (2007). Regression analysis of panel count data with dependent observation times. *Biometrics*, 63:1053–1059.

Sun, J. and Wei, L. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62:293–302.

Sun, J. and Zhao, X. (2013). *Statistical Analysis of Panel Count Data*. Springer, New York, NY.

Thall, P. and Lachin, J. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, 83:339–347.

Wellner, J. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Annals of Statistics*, 28:779–814.

Wellner, J. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics*, 35:2106–2142.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika*, 89:39–48.

Zhang, Y. (2006). Nonparametric k-sample tests with panel count data. *Biometrika*, 93:777–790.

Zhao, X. and Tong, X. (2011). Semiparametric regression analysis of panel count data with informative observation times. *Computational Statistics & Data Analysis*, 55:291–300.

Zochling, J. and Braun, J. (2006). Remission in ankylosing spondylitis. *Clinical and Experimental Rheumatology*, 24:88–92.