Published in final edited form as:
Statistics in Medicine (2019), **39**(4): 387–408
DOI: https://doi.org/10.1002/sim.8411

Multistate analysis from cross-sectional and auxiliary samples

LEILEI ZENG

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

RICHARD J. COOK

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

JOOYOUNG LEE

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

Summary

Epidemiological studies routinely involve cross-sectional sampling of a population comprised of individuals progressing through life history processes. We consider features of a cross-sectional sample in terms of the intensity functions of a progressive multistate disease process under stationarity assumptions. The limiting values of estimators for regression coefficients in naive logistic regression models are studied, and simulations confirm the key asymptotic results that are relevant in finite samples. We also consider the need for and the use of data from auxiliary samples, which enable one to fit the full multistate life history process. We conclude with an application to data from a national cross-sectional sample assessing marker effects on psoriatic arthritis among individuals with psoriasis.

Keywords: auxiliary data, cross-sectional sample, intensity function, Markov model, multistage disease process

1 Introduction

1.1 Background

Chronic disease processes that feature distinct stages of development can often be naturally characterized using multistate models (Hougaard, 1999; Andersen and Keiding, 2002; Cook and Lawless, 2014). Illness-death processes, for example, are fundamental to studies of disease onset as well as the consequence of disease on risk of death (Xu et al., 2010). More general multistate processes can be useful for modeling the course of progressive diseases such as hepatitis (Sweeting et al., 2006), retinopathy (Marshall and Jones, 1995), dementia (Tyas et al., 2007) or arthritis (Kobelt et al., 2002). When a representative sample of individuals is available and life histories are observed prospectively from their respective time origins, intensity-based models offer a convenient framework for modeling the life course and studying the effect of covariates (Andersen et al., 1993; Aalen et al., 2008; Cook and Lawless, 2018). For rare or slowly progressing diseases, however, prospective studies may be impractical since extended follow-up of large samples is required to obtain useful information. Prevalent cohort designs offer an appealing approach to learning about the course of chronic diseases in affected individuals. In such studies individuals with a disease of interest are selected through cross-sectional sampling of a population. Sampled individuals are then followed prospectively in order to record the occurrence of disease related complications, the development of co-morbidities, and death. Samples chosen in this way are sometimes called length-biased since diseased individuals are sampled proportionally to their lifetime with disease, so naive analyses will under-estimate the mortality rate among diseased individuals. Much research has been directed at methods which address this length-bias, either through joint modeling of the retrospective and prospective data (Asgharian and Wolfson, 2005; Luo and Tsai, 2009; Huang and Qin, 2011; Qin et al., 2011). or by conditioning on the retrospective data and using likelihoods accommodating delayed entry (Keiding and Moeschberger, 1992). When resources preclude followup of individuals, information is restricted to the state occupied by individuals and the time of their respective sampling. This is often the case in fertility research (Keiding et al., 2012) where the goal may be to estimate the distribution of the time to pregnancy among couples trying to conceive who are recruited from cross-sectional sampling of a reference population. Keiding (2006) gives numerous other examples of this situation and discusses assumptions, likelihood construction, and estimability issues. He also highlights the utility of the Lexis diagram for understanding and communicating the consequences of process-related sample selection criteria, and points out the need for supplementary data (e.g. on mortality) to fully characterize life history processes of interest; this reference is foundational to the work we discuss here, as is Kraemer et al. (2000).

Our interest lies in studying the information about general progressive multistate processes available from cross-sectional samples. While the distinction between incidence and prevalence of disease has been clearly articulated (Keiding, 1991) and there has been considerable discussion about the different interpretations of covariate effects in intensity-based and marginal prevalence-based analyses, researchers continue to report the results of naive application of logistic regression analyses based on cross-sectional samples. Therefore we first consider the consequences of naive use of logistic regression when modeling state occupancy in cross-sectional samples from a population with a stationary disease process. We use large sample theory for misspecified likelihoods (White, 1982) to obtain insight into the factors determining the estimands for covariate effects in this setting. We next consider the assumptions needed to justify construction of a valid likelihood based on a multistate model, and the use of auxiliary data to facilitate estimation. The model of primary interest involves a healthy state, a state representing the onset and early stage of a disease, successive transient states that are entered upon disease progression, and a death state that can be entered from any of the non-terminal states.

The remainder of this paper is organized as follows. In the following sub-section we introduce the disease process that motivates this work. In Section 2 we define notation and intensity functions for multistate processes, and derive expressions for features of cross-sectional samples in terms of the properties of a stationary multistate disease process. The relation between disease status and fixed covariates is routinely assessed by logistic regression using cross-sectional samples. We derive the limiting value of the regression coefficients in terms of the intensities of the multistate disease process and point out that the typical adjustment for age at risk can lead to increased bias and uninterpretable results. Illustrative calculations are then given for a two-stage disease process. Issues of non-identifiability are then discussed in Section 3 and we show how auxiliary data can facilitate estimation of the model parameters based on an augmented likelihood which synthesizes data from sources using different sampling schemes. Section 4 contains an application to the motivating problem involving the development of psoriatic arthritis among patients with psoriasis and concluding remarks are given in Section 5.

1.2 The Risk of Psoriatic Arthritis in Psoriasis

Psoriasis is a chronic immune-mediated dermatological condition affecting approximately 3.2% of adults in North America (Rachakonda et al., 2014). It is characterized by the development of red raised patches of skin with silvery-white plaques, often located on the knees, elbows, scalp and trunk. These regions are often painful and itchy and are the cause of distress which affects quality of life of affected individuals. Approximately one third of patients with psoriasis develop a more serious condition called psoriatic arthritis (Gladman et al., 2005), which involves the development of painful joint inflammation, joint stiffness and a consequent decreased range of motion which in turn decreases functional ability. With time this inflammation will ultimately lead to joint destruction (Cresswell et al., 2011) and disability and can significantly impact quality of life (Husted et al., 2001).

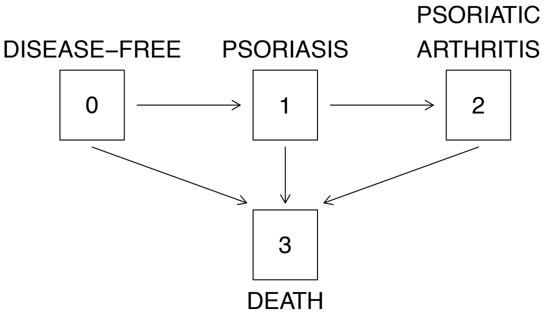


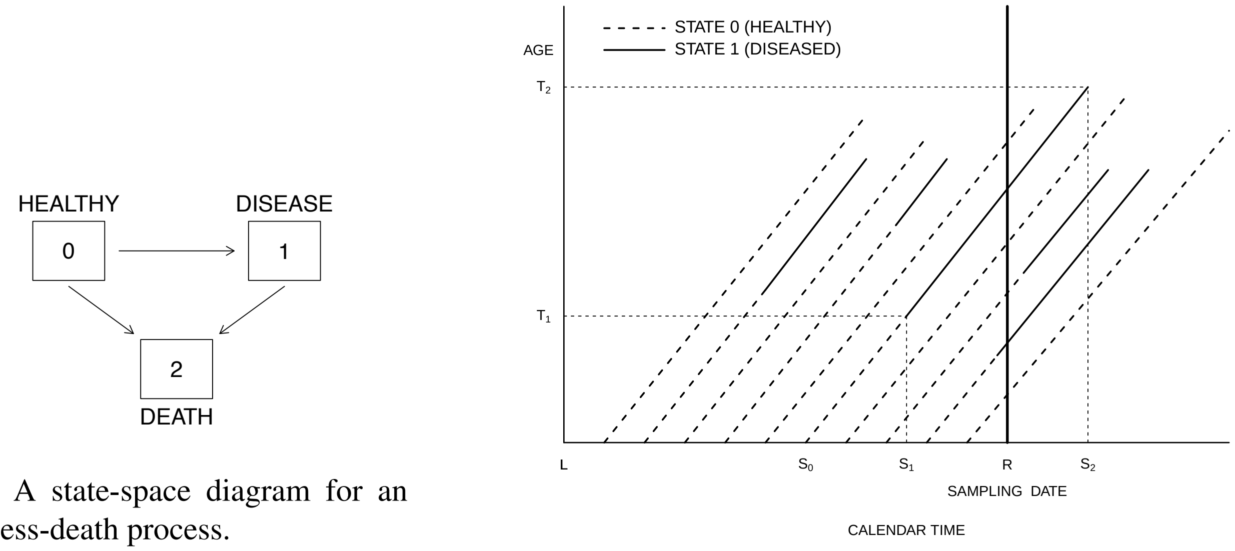
Figure 1: A four-state model for a two-stage disease process involving psoriasis and psoriatic arthritis

This work is motivated by a collaborative research project directed at understanding the risk of developing psoriatic arthritis in individuals with psoriasis. The multistate model we consider to characterize the process of interest involves fours states: a healthy state (state 0), psoriasis (state 1), psoriatic arthritis (state 2) and death (state 3); see Figure 1. Data available for analysis include the results of a cross-sectional survey conducted in 2001 by the National Psoriasis Foundation (Gelfand et al., 2005). Available information includes demographic variables such as age and gender for 14351 individuals, along with their disease status with respect to psoriasis and psoriatic arthritis. Auxiliary data are available from a) data for a sample of 657 patients in a registry of individuals with psoriasis at the Center for Prognosis Studies in Rheumatic Disease at the University of Toronto (Gladman and Chandran, 2010), b) data from a registry of 1314 psoriatic arthritis patients from the same centre (Gladman, 1991) and c) population mortality rates in Canada from 1921 to 2011 (Robert, 2017). We consider the synthesis of the data from these various sources in the analysis of Section 4 where we aim to learn about the four state process in Figure 1.

2 Cross-sectional Sampling of Mutlistate Disease Processes

2.1 Cross-Sectional Sampling and Illness-Death Processes

The illness-death model facilitates modeling the risk of disease in a healthy individual as well as the effect of disease occurrence on risk of death (Fix and Neyman, 1951) and therefore has a central role in epidemiology. The state space for an illness-death process is given in Figure 2 (a) with healthy, diseased and death states labeled 0, 1 and 2 respectively and transitions between states are governed by intensity functions (Andersen et al., 1993). The cumulative incidence function for disease is governed by the intensities for disease onset and disease-free death. The lethality of the disease is likewise reflected by the relative magnitude of the death intensities from the disease and disease-free states. For conditions with relatively little impact on mortality, it is often reasonable to model the intensity for death following disease onset on a Markov timescale and model risk in terms of age, but for more serious diseases the time since disease onset may be a more natural timescale and semi-Markov models can be used. Hybrid time scales incorporating age and disease duration are also possible, but we emphasize Markov models which appear reasonable for the setting of interest. To study features of a population of processes originating over a period of time it is necessary to consider trends in the disease process in the population over calendar time; for this the transition intensities may also depend on calendar time.



(a) A state-space diagram for an illness-death process.

(b) A Lexis diagram representing the calendar times (horizontal axis) of birth, disease onset and death for a sample of individuals; the ages of disease onset and deaths are represented on the vertical axis.

Figure 2: A state-space and Lexis diagram for a population of illness-death processes

Attributes of a population of processes at an instant of calendar time are governed by the process leading to the births of individuals, the intensities of the multistate process, and any calendar time trends. Keiding (1991) gives a general derivation of the relationship between the parameters for individual stochastic illness-death processes generated over time, and the age-specific prevalence and incidence rates at a calendar time; see also Brinks and Hoyer (2018) who study the same phenomenon but based on differential equations. It is common to assume that births arise from a stationary process (Brillinger, 1986) and that transition intensities do not depend on calendar time, although in some settings scientific interest may lie in studying related trends. Under the additional assumptions that the transition intensities do not depend on age and that the disease has no impact on mortality one obtains the well-known result that the prevalence of a disease is equal to the incidence rate times the mean lifetime with the disease (Diamond and McDonald, 1992).

The Lexis diagram (Lexis, 1875; Keiding, 1990; Keiding, 2011) in Figure 2 (b) gives a graphical representation of the relationship between births recorded in calendar time, disease onset, and death. Each of the 45-degree lines indicates the life course of an individual in the population. For the sixth individual to be born for example, the line starts at the calendar time of birth (S_0) on the horizontal axis; calendar times S_1 and S_2 represent the dates of disease onset and death. The vertical axis conveys time in terms of age, so T_1 and T_2 represent their age at disease onset and death respectively. A cross-sectional sample taken at calendar time R is restricted to individuals who are alive. Among selected individuals, the proportion in the "diseased" state at time R is an estimate of the disease prevalence; likewise the proportion of "healthy" individuals at calendar time R who develop the disease one calendar time unit later is the disease incidence (Ahrens and Pigeot, 2014). The prevalence and the incidence of the disease are features of the population which are determined only in part by the illness-death process acting at the individual level. Keiding (1991) established the mathematical relationship between the epidemiological concepts of prevalence and incidence based on a cross-sectional sample.

Additional issues arise when interest lies in the association between covariates and disease status. While the intensity functions of the multistate model characterize risk of disease onset and offer the most appropriate way of formulating covariate effects on the dynamic features of the process, logistic regression is routinely carried out based on the disease status of individuals in the cross-sectional sample. Interpretation of the corresponding odds ratios is problematic since the limiting values of the resulting estimators likewise depend on the birth process, stationarity assumptions, and the intensities of the multistate process. Estimates of covariate effects obtained from logistic regression can therefore be quite misleading in terms of their effects on the intensities of the multistate process; we explore this in Section 2.4.

2.2 Multistate Models for Disease Processes

Suppose a multistage disease can be modeled by a stochastic process involving $K + 1$ states as shown in Figure 3 with finite state space $\mathcal{S} = \{0, 1, \ldots, K-1, K\}$, where 0 represents the condition of being disease-free, states $1, \ldots, K-1$ represent the stage of disease among affected individuals, and state K represents the final absorbing state of death.

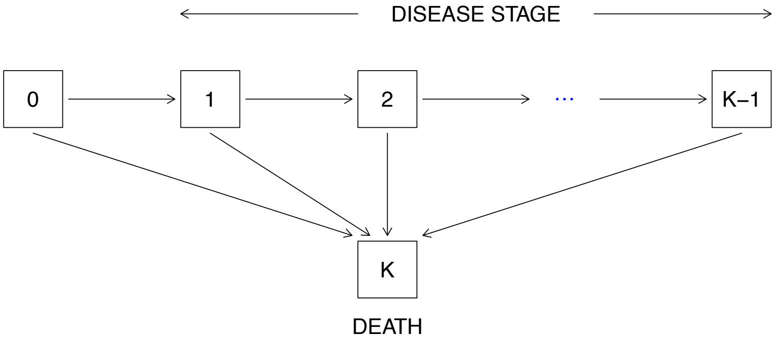


Figure 3: A multistate diagram for a multistage disease process and death.

We therefore consider processes for which individuals are at risk of progression through a sequence of states. Let t represent the age of an individual, $Z(t)$ denote the state occupied at age t (with $Z(0) = 0$) and $\bar{Z}(t) = \{Z(u); 0 \leq u < t\}$ denote the life history up to age t. Let S_0 be the birth date of an individual with $s = S_0 + t$ denoting the calendar time when they are age t. If $H(t) = \{Z(u), 0 < u < t, S_0\}$ is the complete history of the disease process at age t for an individual born at time S_0, the calendar time- and age-specific intensity function for disease progression is

$$\lim_{\Delta t \to 0} \frac{P(Z(t + \Delta t^-) = j+1 \mid Z(t^-) = j, H(t))}{\Delta t} = \lambda_j(s, t \mid H(t)) \ , \quad j = 0, 1, \ldots, K-2$$

and the intensities for death are

$$\lim_{\Delta t \to 0} \frac{P(Z(t + \Delta t^-) = K \mid Z(t^-) = j, H(t))}{\Delta t} = \eta_j(s, t \mid H(t)) \ , \quad j = 0, 1, \ldots, K-1.$$

We take it as understood in what follows that $\lambda_{K-1}(s, t \mid H(s)) = 0$ since $K-1$ is the most advanced disease state. Given the date of birth S_0, $\lambda_j(s,t \mid H(t)) = \lambda_j(s,t \mid s_0)$ and $\eta_j(s,t \mid H(t)) = \eta_j(s,t \mid s_0)$ for a Markov process, and the transition probabilities $P(Z(t_u) = k \mid Z(t_l) = j, H(t_l)) = P(Z(t_u) = k \mid Z(t_l) = j, S_0 = s_0) = P_{jk}(t_l, t_u \mid s_0)$ can be expressed simply in terms of the transition intensities. For example if the disease process is independent of calendar time (i.e. $\lambda_j(s,t \mid H(t)) = \lambda_j(t)$ and $\eta_j(s,t \mid H(t)) = \eta_j(t)$), then

$$P_{0j}(t_l, t_u) = \int_{t_l}^{t_u} P_{0,j-1}(0,t) \ \lambda_{j-1}(t) \ P_{jj}(t, t_u) \, dt \ , \ \ j = 1, \ldots, K, \tag{1}$$

where $P_{jj}(t_l, t_u) = \exp(-\int_{t_l}^{t_u} \lambda_j(t) + \eta_j(t) \, dt)$.

Intensity-based regression models can be adopted when interest lies in the covariate effects on the transition intensities. Let X denote a $p \times 1$ vector of covariates of interest and re-define $H(t) = \{Z(u), 0 \leq u < t, S_0, X\}$ by including the covariates. An intensity-based regression model can be written in the common multiplicative form as

$$\lambda_j(s, t \mid H(t)) = \lambda_j(s, t \mid s_0) \exp(X'\beta) \ ,$$

where $\lambda_j(s, t \mid s_0)$ denotes the calendar time- and age-specific baseline intensity for $j \to j+1$ transition and β is a $p \times 1$ vector of regression coefficients (Andersen et al., 1993).

2.3 CHARACTERISTICS OF CROSS-SECTIONAL SAMPLES BASED ON MULTISTATE MODELS

Consider a birth cohort composed of individuals born in a window of calendar time $[L, R]$. Let $T = R - S_0$ be the age of an individual in this cohort at calendar time R and $Y = Z(T)$ denote the state occupied by this individual at time R. The probability that an individual occupies state j at this time is given by

$$\begin{aligned} P(Y = j \mid L, R) &= \int_L^R P(Z(R - s_0) = j \mid Z(0) = 0; S_0 = s_0, L, R) \ g(s_0) \, ds_0 \\ &= \int_L^R P_{0j}(0, R - s_0 \mid s_0) \ g(s_0) \, ds_0 \ , \ \forall j \in \mathcal{S} \ , \end{aligned} \tag{2}$$

where $g(\cdot)$ is the density function of the birth time S_0. These probabilities are complex functions of the transition intensities governing the disease process as well as the distribution for birth times.

If we assume that the birth process is stationary and the disease process is Markov and independent of calendar time, the unconditional state occupancy probabilities at an arbitrary calendar time are :

$$P(Y = j) = \lim_{L \to -\infty} P(Y = j \mid L, R) \propto \int_0^\infty P_{0j}(0, t) \, dt \ , \quad \forall j \in \mathcal{S} \ , \tag{3}$$

where the transition probability $P_{0j}(0,t)$ is obtained based on (1). A cross-sectional sample selected only includes individuals who are alive (i.e. $Y \neq K$) so if let $\pi_j = P(Y = j \mid Y < K)$ denote the *overall prevalence of stage* j,

$$\pi_j = \frac{P(Y=j)}{P(Y<K)} = \frac{\int_0^\infty P_{0j}(0,t)dt}{\sum_{j=0}^{K-1} \int_0^\infty P_{0j}(0,t)dt} , \qquad j = 0, 1, \ldots, K-1 . \tag{4}$$

The conditional probability of an individual being in state $j+1$ given they are in state j or $j+1$ is then

$$\frac{\pi_{j+1}}{\pi_j + \pi_{j+1}} = \frac{\int_0^\infty \; P_{0,j+1}(0,t) \; dt}{\int_0^\infty \; P_{0j}(0,t) \; dt + \int_0^\infty \; P_{0,j+1}(0,t) \; dt} ,$$

which gives the *overall prevalence odds of stage* $j+1$ *versus stage* j as

$$\frac{\pi_{j+1}}{\pi_j} = \frac{\int_0^\infty \; P_{0,j+1}(0,t) \; dt}{\int_0^\infty \; P_{0j}(0,t) \; dt} , \quad j = 0, \ldots, K-2 . \tag{5}$$

Since $\mu_j = \int_0^\infty \; P_{0j}(0,t) \; dt$ is the mean sojourn time in disease state j over a lifetime, the *prevalence of stage* j in (4) can be equivalently viewed in terms of the ratio of mean lifetime in disease state j over the overall mean lifetime; the *prevalence odds of stage* $j+1$ *versus stage* j in (5) thus can be expressed in terms of the ratio of mean lifetimes in disease state $j+1$ and state j. The *age-specific prevalence of stage* j and *prevalence odds of stage* $j+1$ *versus stage* j can be calculated as $P_{0j}(0,t)/[\sum_{j=0}^{K-1} P_{0j}(0,t)]$ and $P_{0,j+1}(0,t)/P_{0j}(0,t)$ respectively. Expressions of this sort were derived by Hoem and Jensen (1982) who were considering the implications for applications in demography.

Further simplifications are possible when transition intensities are time homogeneous. In such cases, $\lambda_j(t) = \lambda_j$ and $\eta_j(t) = \eta_j$, $j = 0, 1, \ldots, K-1$ where again $\lambda_{K-1} = 0$. In this case the sojourn time in state j is exponentially distributed with hazard $h_j = \lambda_j + \eta_j$, mean $h_j^{-1} = (\lambda_j + \eta_j)^{-1}$, and survivor function $\mathcal{F}_j(t) = \exp(-h_j t)$, $j = 0, \ldots, K-1$. The transition probability (1) can be re-expressed in this case as

$$P_{0j}(0,t) = \prod_{k=0}^{j-1} \lambda_k \left[\sum_{k=0}^{j-1} \frac{\mathcal{F}_k(t) - \mathcal{F}_j(t)}{\prod_{l=0}^{k-1}(h_l - h_k) \; \prod_{l=k+1}^{j}(h_l - h_k)} \right] , \; j = 0, 1, \ldots, K-1, \tag{6}$$

giving

$$\mu_j = \int_0^\infty P_{0j}(0,t) \; dt = \left[\prod_{k=0}^{j-1} \psi_k \right] \frac{1}{\lambda_j + \eta_j} \tag{7}$$

where $\psi_k = \lambda_k/(\lambda_k + \eta_k)$ is the probability of a $k \to k+1$ (vs. $k \to K$) transition given state k occupancy, $k = 0, \ldots, K-2$. The product $\prod_{k=0}^{j-1} \psi_k$ in (7) is the probability of reaching disease state j and the term $1/(\lambda_j + \eta_j)$ represents the mean sojourn time in stage j; see Appendix A for further details. Thus (7) gives the result that the

mean lifetime in stage $j =$ *probability of entering stage* $j \times$ *mean sojourn time in stage* j .

The overall prevalence odds (5) can likewise be written as

$$\frac{\pi_{j+1}}{\pi_j} = \lambda_j \times \frac{1}{\lambda_{j+1} + \eta_{j+1}} , \tag{8}$$

which is a product of the risk of a $j \to j+1$ transition and mean sojourn time in stage $j+1$. For the special case of the illness-death model in Figure 2 (a) the expression in (8) gives the well-known result

$$\textit{overall prevalence odds of disease} = \textit{incidence of disease} \times \textit{mean duration of disease} \ , \tag{9}$$

which was first derived in work on the theory of *screening* for chronic disease (Zelen and Feinleib, 1969; Newman, 1988). Keiding (1991) gave a careful discussion of the relation between incidence and prevalence in terms of general models and included the time-homogenous illness-death process as a special case. Alho (1992) showed analogous results for the setting where the transition intensities for disease onset and mortality are age-dependent and showed that the prevalence odds is a weighted average of the age-dependent incidence and disease duration, averaging with respect to the age distribution of the health population at the time of sampling. The derivations here are for progressive multistage disease processes with time-nonhomogeneous Markov transition intensities with the results under the time-homogeneous setting simplifying to give (8).

Next we examine the age distribution of individuals according to their disease status at a fixed (calendar) sampling time R. Recall S_0 is the birth date of an individual and S_j denotes the calendar time they entered state j ($j \neq K$); we also let S_{j+1}^* denote the calendar time of leaving state j (by either entering state $j+1$ or the absorbing state K). The subpopulation of individuals occupying state j at calendar time R must therefore satisfy the condition $L < S_0 < S_j < R$ and $S_{j+1}^* > R$, so the age distribution in this subpopulation is

$$f(s_0 \mid L < S_0 < S_j < R, S_{j+1}^* > R) = \frac{P(L < S_0 < S_j < R, S_{j+1}^* > R \mid S_0 = s_0, L, R)\; g(s_0)}{\int_L^R P(L < S_0 < S_j < R, S_{j+1}^* > R \mid S_0 = s_0, L, R)\; g(s_0)\; ds_0} \ .$$

If the birth process is stationary with $S_0 \sim \text{Unif}(L, R)$, it can be shown that the age distribution given state j occupancy is

$$f_T(t \mid Y = j) = \lim_{L \to -\infty} f(s_0 \mid L < S_0 < S_j < R, S_{j+1}^* > R) = \frac{P_{0j}(0,t)}{\int_0^\infty P_{0j}(0,u)\; du} \ , \tag{10}$$

where the denominator on the right is the mean lifetime in disease state j, $j = 0, \ldots, K-1$. For survival models with $K = 1$, states 0 and 1 represent the conditions of being alive and dead respectively. For a cross-sectional sample restricted to individuals who are alive at time R, the age distribution is the length-biased density of the form

$$f_T(t \mid Y = 0) = \frac{P_{00}(0,t)}{\int_0^\infty P_{00}(0,u)\; du} = \frac{\mathcal{F}_0(t)}{\mu_0} \ .$$

2.4 Modeling Covariate Effects via Logistic Regression with Cross-Sectional Samples

Studies are often conducted with the objective of identifying markers associated with occupancy of one disease state among a set of possible states. For the example in Section 1.2 one may be naively asked for the association between a marker and presence of psoriatic arthritis among individuals with psoriasis (i.e. occupancy of state 2 given occupancy of states 1 or 2). More generally consider the question about the association between a binary marker X and occupancy of disease state $j+1$ versus state j. A common naive approach is to collect a cross-sectional sample, construct a subsample of individuals in disease state j or state $j+1$ at the time of sampling, and to carry out a logistic regression analysis with a binary response $Y = I(Z(T) = j+1)$ in a sample of individuals for whom $Z(T) \in \{j, j+1\}$. Suppose the disease process is simple and can be characterized by an illness-death model as in Figure 2(a). In this setting, Aljied et al. (2018) assessed the association between fixed covariates and the presence of visual impairment in a cross-sectional sample. Jung et al. (2014) examined risk factors associated with rectal neuroendocrine tumors based on a cross-sectional sample of Koreans who underwent colonoscopy. Østbye et al. (2005) studied the association between covariates and depression at a fixed point in time in a cross-sectional sample. Finally, Toperoff et al. (2015) studied the association between premature DNA methylation aging and type 2 diabetes in a cross-sectional sample of the East Jerusalem Palestinian (EJP) Arab population. For progressive disease process like the one depicted in Figure 1, Winchester et al. (2012), Haroon et al. (2013) and Loft et al. (2018) modelled the relationship between human leukocyte antigen (HLA) markers and the presence of psoriatic arthritis among individuals with psoriasis. We characterize the logist regression analyses in these articles as naive since they fail to address the central dynamic aspect of the disease process characterized by the intensity functions of the multistate formulation.

If $V = (1, X, W)'$ is a vector of covariates including the binary marker X of interest and additional covariates W, a logistic regression model

$$\log\left(\frac{\pi(V;\gamma)}{1-\pi(V;\gamma)}\right) = V'\gamma$$

can be fitted where $\pi(V;\gamma) = P(Y = 1|V;\gamma)$ with γ being a vector of regression parameters. In the terms of White (1982), a quasi-likelihood can be constructed for this logistic regression model based on the assumption that $Y \mid V$ is Bernouilli. Given a sample of n independent individuals and data $\{(Y_i, V_i), i = 1, \ldots, n\}$, the maximum quasi-likelihood estimate of γ, denoted $\widehat{\gamma}$, is obtained by solving the quasi-score equation

$$\sum_{i=1}^{n} U_i(Y_i, V_i;\gamma) = \sum_{i=1}^{n} [Y_i - \pi(V_i;\gamma)]V_i = 0 \; . \tag{11}$$

White (1982) showed that the limiting value of $\widehat{\gamma}$, denoted $\gamma^{\dagger}$, can be obtained by computing the solution to $E(U_i(Y_i, V_i;\gamma)$ where here, and below, the expectation is taken with respect to the true distribution of the random vector (Y_i, V_i). This true distribution is a complex function of the birth process, the multistate process, the cross-sectional sampling scheme and the population covariate distribution. Moreover White (1982) shows that

$$\sqrt{n}(\widehat{\gamma} - \gamma^{\dagger}) \sim MVN(0, \mathcal{A}^{-1}(\gamma^{\dagger})\mathcal{B}(\gamma^{\dagger})\mathcal{A}^{-1}(\gamma^{\dagger}))$$

with $\mathcal{A}(\gamma) = -E\{\partial U(Y, V;\gamma)/\partial\gamma'\}$ and $\mathcal{B}(\gamma) = E\{U(Y, V;\gamma)U(Y, V;\gamma)'\}$. The key point is that the complexities of the true data generating process are not addressed with the simple logistic model and the estimand (i.e. the limiting value $\gamma^{\dagger}$) is in general an uninterpretable function of the parameters indexing the underlying multivariate distribution.

It is often recommended to include age at the time of sampling as a covariate in such logistic regression models to account for the different lengths of time individuals in the cross-sectional sample have been at risk for transitions in the multistate process. A simple adjusted logistic regression model takes the form

$$\log\left(\frac{\pi(X, T;\gamma)}{1-\pi(X, T;\gamma)}\right) = \gamma_0 + \gamma_1 X + \gamma_2 T \; , \tag{12}$$

where T is the age of an individual in the sample and $\gamma = (\gamma_0, \gamma_1, \gamma_2)'$. The limiting value $\gamma^{\dagger}$ computed by solving $E\{U(Y, V;\gamma)\} = 0$ can be used to obtain the *age-adjusted odds ratio of state* $j+1$ *versus state* j, given by $\exp(\gamma_1^{\dagger})$. To obtain $\gamma^{\dagger}$, we require the expectation of the score function in (11) with respect to the binary indicator Y, the covariate X and the age T at sampling; the required distributions were given in Section 2.3. More specifically since the cross-sectional sub-sample will only contain subjects whose disease state is either j or $j+1$, the joint distribution of (Y,X) is as follows

$$P(Y=y,X=x)=\frac{\mu_{j+y}(x)P(X=x)}{\mu_j(x)+\mu_{j+1}(x)}\ ,$$

and following (10) the age distribution stratified on (y,x) is

$$f_T(t\mid y,x)=\frac{P_{0,j+y}(t;x)}{\mu_{j+y}(x)}\ .$$

The expectation of the score functions $U(Y,X,T;\gamma)$ thus takes the form

$$\begin{aligned} E\big[U(Y,X,T;\gamma)\big] &= E\big\{E\big[U(Y,X,T;\gamma)\mid Y,X\big]\big\} \\ &= \sum_{x=0}^{1}\sum_{y=0}^{1}\frac{P(X=x)}{\mu_j(x)+\mu_{j+1}(x)}\int_0^\infty U(y,x,t;\gamma)P_{0,j+y}(t;x)\,dt\ . \end{aligned} \tag{13}$$

Solving $E[U(Y,X,T;\gamma)]=0$ enables one to explore how the *age-adjusted odds ratio* $\exp(\gamma_1^\dagger)$ varies according to the parameters of the multistate process. Closed-form solutions are not available in general so the equations must be solved numerically. Consider an illustrative calculation involving a simple logistic regression model with a binary marker as the only covariate

$$\log\left(\frac{\pi(X;\gamma)}{1-\pi(X;\gamma)}\right)=\gamma_0+\gamma_1X\ , \tag{14}$$

where $\exp(\gamma_1)$ would be interpreted as a standard odds ratio characterizing the association between X and disease state $j+1$ (vs. j). Suppose the binary covariate acts on the transition intensities via the models

$$\lambda_j(t\mid H(t))=\lambda_j\exp(X\beta_j)\ ,\quad j=0,1,\ldots,K-2, \tag{15a}$$
$$\eta_j(t\mid H(t))=\eta_j\exp(X\alpha_j)\ ,\quad j=0,1,\ldots,K-1, \tag{15b}$$

where $\exp(\beta_j)$ is the relative risk of a $j\rightarrow j+1$ transition for a subject with covariate value $X=1$ versus $X=0$, and λ_j and η_j are baseline intensities for disease progression and death. Under a time-homogeneous process the limiting value of the odds ratio estimator $\exp(\widehat{\gamma}_1)$ based on (14) and obtained from (13) has the form

$$\exp(\gamma_1^\dagger)=\frac{\exp(\beta_j)}{\psi_{j+1}\exp(\beta_{j+1})+(1-\psi_{j+1})\exp(\alpha_{j+1})}\ . \tag{16}$$

Interestingly this is a function of the “true” multiplicative effect $\exp(\beta_j)$ on the $j\rightarrow j+1$ intensity in (15a) and other parameters of the multistate model given in (15). When the covariate X is not associated with the risk of subsequent transitions out of state $j+1$ (i.e. $\beta_{j+1}=\alpha_{j+1}=0$), we obtain $\gamma_1^\dagger=\beta_j$ in which case the cross-sectional analysis yields a consistent estimator of the relative risk associated with a $j\rightarrow j+1$ transition. However, when either the probability of a $j+1\rightarrow j+2$ transition (ψ_{j+1}) is high or the covariate has a much stronger effect on $j+1\rightarrow j+2$ transition than on $j\rightarrow j+1$ transition, one may obtain $\exp(\beta_j)\leq 1<\exp(\gamma_1^\dagger)$ or $\exp(\gamma_1^\dagger)<1\leq\exp(\beta_j)$. That is, if the covariate has less impact on the $j\rightarrow j+1$ progression than on the subsequent transition out of state $j+1$, a naive binary analysis could suggest an association between the covariate and disease in the opposite direction to the one manifest in the corresponding intensity of the multistate model.

2.5 An Illustrative Example of a Two-stage Disease Process

We illustrate the findings in the previous sections using the study of psoriatic arthritis among psoriasis patients as an example, where the disease process has four states with state space $\mathcal{S} = \{0, 1, 2, 3\}$. Under the time-homogeneous model and following the results given in Section 2.2 we obtain

$$\begin{aligned}
&\textit{mean disease free lifetime:} \quad &\mu_0 = \int_{t=0}^{\infty} P_{00}(0,t)dt = \frac{1}{\lambda_0 + \eta_0} \\
&\textit{mean diseased lifetime in state 1:} \quad &\mu_1 = \int_{t=0}^{\infty} P_{01}(0,t)dt = \psi_0 \left[\frac{1}{\lambda_1 + \eta_1}\right] \\
&\textit{mean diseased lifetime in state 2:} \quad &\mu_2 = \int_{t=0}^{\infty} P_{02}(0,t)dt = \psi_0 \psi_1 \frac{1}{\eta_2}
\end{aligned}$$

and the overall prevalence of disease (in stage 1 or 2) can be obtained based on the formula given in (4). In addition, when interest lies in progression to stage 2 from stage 1, we find

$$\textit{overall prevalence odds of stage 2 vs 1} = \lambda_1 \frac{1}{\eta_2} \ .$$

Suppose the association between X and the development of psoriatic arthritis is investigated via a logistic regression model (14) and a sub-sample of the full cross-sectional sample includes individuals with psoriasis or psoriatic arthritis. Based on the result in (16) we see that when age is not controlled for,

$$\gamma_1^{\dagger} = \beta_1 - \alpha_2 \ .$$

Recall that β_1 reflects the effect of a marker on the risk of psoriatic arthritis among those with psoriasis (a $1 \to 2$ transition) and α_2 is the marker effect on mortality from the psoriatic arthritis state (a $2 \to 3$ transition) based on the multistate model. When the marker is not associated with disease mortality (i.e. $\alpha_2 = 0$), we have $\gamma_1^{\dagger} = \beta_1$ so a logistic regression analysis based on the binary disease status provides a valid estimate of log relative risk of the psoriasis $\to$ psoriatic arthritis intensity. When the marker has similar effects on the risk of psoriatic arthritis and subsequent death (i.e. $\beta_1 \approx \alpha_2$), then $\gamma_1^{\dagger} \approx 0$ and one would not tend to see evidence of an association between the marker and psoriatic arthritis. Finally, if there is a much stronger marker effect on disease mortality compared to disease onset (i.e. $\beta_1 < \alpha_2$), evidence from the logistic regression analysis results may suggest an association in the opposite direction to the one from the multistate intensity.

When fitting a logistic regression model adjusting for age as in (12), the limiting value of the estimated marker effect depends on the parameters characterizing the multistate disease process in a complicated way. To investigate this we consider the following setting and carry out asymptotic calculations. Assume births occur uniformly in $[L, R]$ where R is taken as the screening date; without loss of generality we set $L = 0$ and $R = 10$. Given a binary covariate $X \sim \text{Bernoulli}(0.5)$ and birth date S_0, transition times were simulated based on the four-state disease process as described in Figure 1 Cook and Lawless (2018). Let T_{01} and T_{03} denote the possible $0 \to 1$ and $0 \to 3$ transition times which were simulated based as exponential random variables with rates $\lambda_0 \exp(X\beta_0)$ and $\eta_0 \exp(X\alpha_0)$ respectively. If $T_{01} < T_{03}$ then T_{12} and T_{13} were simulated as exponential with rates $\lambda_1 \exp(X\beta_1)$ and $\eta_1 \exp(X\alpha_1)$. Finally if $T_{12} < T_{13}$ then T_{23} was simulated as exponential with rate $\eta_2 \exp(X\alpha_2)$. Let

$$P(Y = j) = \frac{1}{R} \sum_{x=0}^{1} \int_0^R P_{0j}(0, t \mid X = x) P(X = x) \, dt \ , j \in \mathcal{S}, \tag{17}$$

represent the probability of state j occupation at a fixed sampling time R. We set $\eta_0 = \eta_1 = \eta_2$ to consider a case where the disease state does not affect the mortality rate, and solved for λ_0, λ_1 and

λ_0 such that $P(Y = 0) = 0.1$, $P(Y = 1) = 0.01$ and $P(Y = 2) = 0.01$. The low probability of being in states 1 and 2 were chosen to correspond to diseases with relatively low prevalence and the high fraction of individuals assumed to have died ensures that the data are compatible with the assumption that the disease process has been occurring in the population for a long time. Finally, we let $\exp(\beta_0) = \exp(\alpha_0) = \exp(\alpha_1) = 1$, and allow $\exp(\beta_1)$ and $\exp(\beta_2)$ to vary. Figure 4 displays plots of $\gamma_1^\dagger - \beta_1$, the discrepancy between $\gamma_1^\dagger$ and the intensity based effect of the marker on a Ps to PsA transition. It is apparent that adjusting for age in logistic regression model does not resolve the issue, as there is still an asymptotic bias in the estimation of the marker effect. Figure 4 also shows that there can be evidence of an association in the opposite direction of the intensity-based effect.

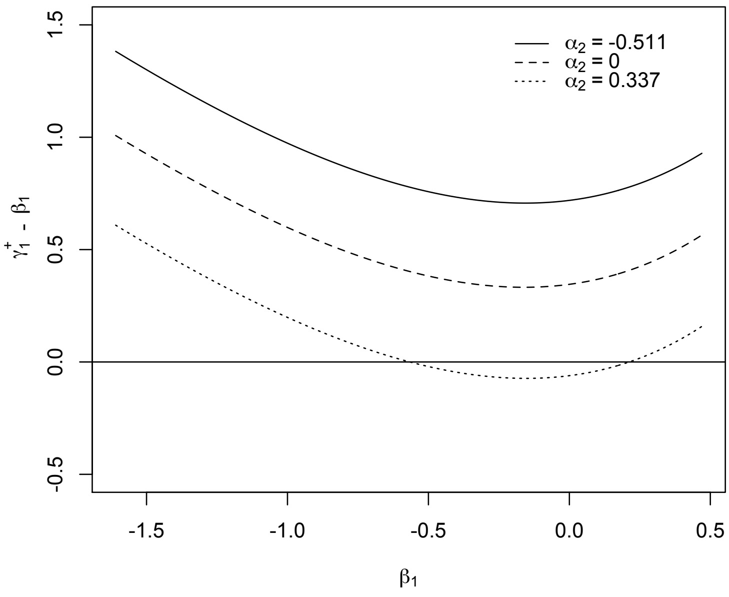


Figure 4: Plot of $\gamma_1^\dagger - \beta_1$ against β_1. $\gamma_1^\dagger$ is the limiting value of the marker effect from the logistic regression model adjusted by age as given in (12), β_1 and α_2 are the gender effects on psoriasis $\rightarrow$ psoriatic arthritis and psoriatic arthritis $\rightarrow$ death transition intensities respectively in the multistate model.

Next we report on simulation studies conducted to assess the relevance of the asymptotic calculations for finite samples. We consider the same setting as specified for the calculation of the limiting value plotted in Figure 4. The disease process is simulated for a population of size $n = 10,000$ and random cross-sectional sample is drawn from individuals in states 1 (e.g. Ps) and 2 (e.g. PsA) at the screening time R and a simple logistic regression analysis is conducted to estimate the odds ratio $\exp(\gamma_1)$. Table 2 of Appendix B reports the average estimated odds ratio from 200 cross-sectional samples of the same size and compares those with the theoretical value $\exp(\beta_1 - \alpha_2)$ under different parameter configurations; the agreement between the asymptotic and empirical results is evident. Finally, in Figure 8 of Appendix B we display histograms of the age at sampling for healthy individuals, individuals with psoriasis and individuals with psoriatic arthritis respectively. The two diseased groups have slightly different age distribution, and the healthy, disease-free group has quite a different distribution again as one would expect. Moreover, there is a strong agreement between the histogram and theoretical distribution curves.

3 Multistate Analysis with Cross-Sectional and Auxiliary Data

In this section we discuss the need for, and use of, auxiliary data for the analysis of data from a cross-sectional sample based on a multistate model. We consider the study of the psoriasis and psoriatic arthritis disease process based on the four-state model in Figure 1 with states $\mathcal{S} = \{0, 1, 2, 3\}$ for healthy, psoriasis, psoriatic arthritis and death. We assume that given the birth date S_0 and covariate X the process is Markov and consider general calendar time- and age-specific intensities $\lambda_j(s, t \mid S_0, X)$ for $j \to j + 1$ transitions, $j = 0, 1$ and $\eta_j(s, t \mid S_0, X)$ for transition to death state, $j = 0, 1, 2$. Transition intensity matrix is defined by $\mathbb{A}(s, t \mid S_0, X)$.

We further assume a common mortality rate for patients with psoriasis or psoriatic arthritis which are taken to be proportional to that for disease-free individuals giving

$$\eta_1(s, t \mid S_0, X) = \eta_2(s, t \mid S_0, X) = \eta_0(s, t \mid S_0, X) \exp(\alpha) \ .$$

Let $C_0, \ldots, C_{R_c}$ partition calendar time so $L = C_0 < C_1 < \cdots < C_{R_C} = R$ and let $\mathcal{C}_r = (C_{r-1}, C_r]$, $r = 1, \ldots, R_C$ define R_C birth cohorts. We likewise partition age according to $0 = A_0 < A_1 < \cdots < A_{R_A} = \infty$ to create R_A age strata defined by $\mathcal{A}_h = (A_{h-1}, A_h]$, $h = 1, \ldots, R_A$. Then for a disease-free individual of age t at calendar time s, the cohort- and age-specific intensity for mortality given covariate x is

$$\eta_0(s, t|s_0, x) = \sum_r \sum_h I(s \in \mathcal{C}_r) I(t \in \mathcal{A}_h) \eta_{rhx}$$

where the vector $\eta = \{\eta_{rhx}; r = 1, \ldots, R_C, h = 1, \ldots, R_A, x = 0, 1\}$ is the set of parameters indexing the disease-free mortality rate. We assume that the intensity for psoriasis and psoriatic arthritis (among psoriasis patients) does not depend on calendar time and consider multiplicative models of the form $\lambda_0(s, t \mid S_0, X) = \lambda_0(t) \exp(X\beta_0)$ and $\lambda_1(s, t \mid S_0, X) = \lambda_1(t) \exp(X\beta_1)$, where $\beta = (\beta_0, \beta_1)'$. Piecewise-constant functions are also adopted for the baseline intensities $\lambda_0(t)$ and $\lambda_1(t)$ where we let $\lambda_j(t) = \lambda_{jh}$ for $t \in \mathcal{A}_h$, $j = 0, 1$ with the full set of baseline intensities for psoriasis and psoriatic arthritis denoted by $\lambda = \{\lambda_{jh}; h = 1, \ldots R_A, j = 0, 1\}$. The full vector of parameters associated with the disease process is thus $\theta = (\eta', \alpha, \lambda', \beta')'$.

Let $r_0 = \sum_{r=1}^{R_C} r I(S_0 \in \mathcal{C}_r)$ denote the birth cohort of an individual with $\mathcal{A}^* = \{C_r - S_0;\ r = r_0, r_1, \ldots\}$ their ages at which mortality risk changes due to trends in calendar time. Finally let $\mathcal{B} = \mathcal{A}^* \cup \{A_h; h = 0, 1, \ldots\} = \{b_0, b_1, b_2, \ldots\}$ be the ages defining intervals within which the transition intensity matrix is constant; that is, within intervals $\mathcal{B}_r = (b_{r-1}, b_r]$ as $\mathbb{A}(S_0 + b_r, b_r \mid S_0, X)$, $r = 1, 2, \ldots$. If $\mathbb{P}(u, t \mid S_0, X) = [p_{jk}(u, t \mid S_0, X)]$ denotes the transition probability matrix over any arbitrary age interval $[u, t]$ given the birth date S_0 and covariate X then

$$\mathbb{P}(u, t \mid S_0, X) = \prod_r I([u, t] \cap \mathcal{B}_r \neq \emptyset) \mathbb{P}\big(\max\{u, d_{r-1}\}, \min\{t, d_r\} \mid S_0, X\big) \tag{18}$$

and $\mathbb{P}\big(b_{r-1}, b_r \mid S_0, X\big) = \exp\{\big(b_r - b_{r-1}\big)\mathbb{A}(S_0 + b_r, b_r \mid S_0, X)\}$ is calculated using the matrix exponential given a constant transition intensity matrix for the age interval $(b_{r-1}, b_r]$, $r = 1, 2, \ldots$.

The National Psoriasis Foundation (NPF) survey (Gelfand et al., 2005) was a large scale cross-sectional survey conducted to estimate prevalence of psoriasis and psoriatic arthritis in North America, we consider such a study and let $\mathcal{S}_c$ denote the subsample of individuals having psoriasis (PsC) or psoriatic arthritis (PsA) from the cross-sectional sample. The information available from an individual in this subsample includes their birth date S_0 and hence their current age T, a covariate X and their disease status $Z(T)$. The likelihood contribution based on an individual in this sample is thus

$$\mathcal{L}_c(\theta) = P(Z(T) = j \mid Z(T) \neq 3, S_0, X) = \frac{P_{0j}(0, t \mid S_0, X)}{\sum_{j=0}^{2} P_{0j}(0, t \mid S_0, X)} \tag{19}$$

if this individual is observed to be in state j at the sampling time. This likelihood is expressed in terms of the transition probabilities given the birth date and covariate.

The parameters indexing likelihood $\mathcal{L}_c(\theta)$ are of course not identifiable based on the cross-sectional sample alone. In particular there is no information on the mortality rates from the cross-sectional sample since it is restricted to individuals who are alive at the time of sampling. Moreover the current status data on the disease status are limited for estimation of the intensities for disease-related transitions. To overcome this difficulty, auxiliary data from other sources can be utilized to facilitate the estimation (Lee and Cook, 2018). We consider the case in which the covariate of interest is gender and the published national year-, age- and gender-specific mortality statistics can be used to estimate the disease-free mortality parameter η. With longitudinal follow-up of patients in a psoriasis and a psoriatic arthritis registry, some information on disease-related transition intensities and disease-specific mortality rates can be obtained by gender.

Caution is warranted when synthesizing data from the different registries because the selection conditions differ; to enter the psoriasis registry an individual is required to be in state 1 (i.e. we require them to be diseased with psoriasis, so $Z(T) = 1$). To enter the psoriatic arthritis registry upon sampling they are required to be diseased with psoriatic arthritis (i.e. we require $Z(T) = 2$). Moreover, information available from the various sources differ. Some retrospective and prospective information on the disease processes may be available but information on mortality rates are obtained prospectively. Current status data are available from the cross-sectional sample obtained through the National Psoriasis Foundation survey. Exploiting this information however requires information on mortality rates which are available from the prospective follow-up of individuals recruited to the two registries, as well as population mortality rates.

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the sets of patients from the psoriasis and psoriatic arthritis registries accordingly. Suppose the patients from both registries are followed over time up to a random censoring age C, or age at death T_3, and let $T^\dagger = \min(T_3, C)$ and $\delta = I(T^\dagger = T_3)$. The retrospective information such as the age at onset of psoriasis (i.e. T_1) and/or psoriatic arthritis (i.e. T_2) are assumed available for both registries. For a patient in the psoriasis registry the onset of the psoriatic arthritis is assessed intermittently at age $T = v_0 < v_1 < \cdots < v_m < T^\dagger$ and let $Z_k = Z(v_k)$ for convenience, $k = 0, 1, \ldots, m$, and $Z_0 = 1$. The likelihood contribution from a patient from the psoriasis registry is then

$$\mathcal{L}_1(\theta) = P(\bar{Z}(T^\dagger) \mid Z(T) = 1, S_0, X) = \frac{P(\bar{Z}(T^\dagger) \mid S_0, X)}{P_{01}(0, t \mid S_0, X)} \tag{20}$$

where $P(\bar{Z}(T^\dagger) \mid S_0, X)$ is given by

$$\lambda_0(t_1|x)P_{00}(0,t_1|S_0,X)P_{11}(t_1,t|S_0,X)\prod_{k=0}^{m-1} P_{z_k,z_{k+1}}(v_k,v_{k+1}|S_0,X)\sum_{j=1}^{2} P_{z_m,j}(v_m,t^\dagger|S_0,X)\eta_j^\delta(t^\dagger|S_0,X) \ .$$

The patients from the psoriatic arthritis registry provide retrospective data on the age of onset for psoriasis and psoriatic arthritis as well as prospective survival information. Their likelihood contribution takes the form

$$\mathcal{L}_2(\theta) = P(\bar{Z}(T^\dagger) \mid Z(T) = 2, S_0, X) = \frac{P(\bar{Z}(T^\dagger) \mid S_0, X)}{P_{02}(0, t \mid S_0, X)} \tag{21}$$

where $P(\bar{Z}(T^\dagger) \mid S_0, X)$ is

$$\lambda_0(t_1|X)\lambda_1(t_2|X)\eta_2^\delta(t^\dagger|S_0,X)P_{00}(0,t_1|S_0,X)P_{11}(t_1,t_2|S_0,X)P_{22}(t_2,t^\dagger|S_0,X) \ .$$

The augmented likelihood based on the data from the various sources is then

$$\mathcal{L}(\theta) = \prod_{i\in\mathcal{S}_c} \mathcal{L}_{ic}(\theta) \prod_{i\in\mathcal{S}_1} \mathcal{L}_{i1}(\theta) \prod_{i\in\mathcal{S}_2} \mathcal{L}_{i2}(\theta) \tag{22}$$

where i indexes the individuals.

4 Estimating Risk of Psoriatic Arthritis in Individuals with Psoriasis

Here we use an augmented likelihood as described in the previous section to model the development of psoriatic arthritis in individuals with psoriasis. Cross-sectional data are used from a survey by the National Psoriasis Foundation (NPF) in the United States from November and December, 2001 (Gelfand et al., 2005). It includes 14351 individuals (51.8% female) who provided their disease status with respect to psoriasis and psoriatic arthritis. Among the 347 individuals with psoriasis 49 had developed psoriatic arthritis. We also consider data from the University of Toronto Psoriatic Arthritis Registry (UT-PsA) which was established in1976 (Gladman and Chandran, 2010). It is a large cohort of patients diagnosed with psoriatic arthritis who were recruited and followed at 6- to 12-month intervals according to a standardized protocol for the collection of clinical and laboratory datas. As of February 2018 there were 1341 patients with complete information on dates of birth and recruitment, retrospective reports of the onset time of both psoriasis and psoriatic arthritis; prospective followup gave some information on mortality. Another source of auxiliary data is the University of Toronto Psoriasis Registry (UT-Ps), a registry of patients with psoriasis without arthritis which was established in 2006. These individuals were followed and assessed annually for the development of psoriatic arthritis. This cohort is currently composed of 657 subjects with complete information on date of birth and recruitment date, retrospective reporting of their onset time of psoriasis, and prospective followup for the development of psoriatic arthritis (57), and death (13). Finally, population mortality rates are used to evaluate the age-specific mortality rates by 5 year birth cohorts over 1921 to 2011 (Lee and Cook, 2018); the raw data are displayed in Figure 5.

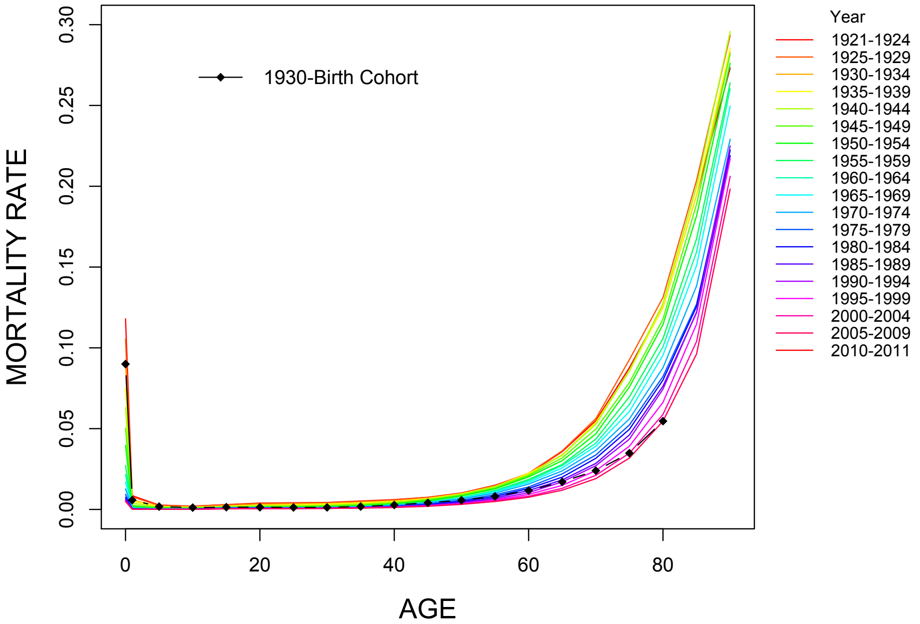


Figure 5: Age-specific population mortality rates by calendar period in Canada from 1921 to 2011 (Lee and Cook, 2018).

A multistate analysis is conducted using the augmented likelihood (22) based on the NPF survey data, data from the UT-Ps and UT-PsA registries, and population calendar time-, age- and gender-specific mortality rates. The latter are taken as fixed and used to characterize the disease-free mortality with $\eta_0(s,t|S_0,X)$ considered known. The relative risk of mortality in diseased and disease-free states is set as $\exp(\nu) = \eta_j(s,t|S_0,X)/\eta_0(s,t \mid S_0,X)$. The parameters in the multiplicative intensity models for healthy → psoriasis and psoriasis → psoriatic arthritis transitions, $\lambda_0(t)\exp(X\beta_0)$

and $\lambda_1(t)\exp(X\beta_1)$, are estimated. The baseline intensities, $\lambda_0(t)$ and $\lambda_1(t)$ are set to be piecewise constant with age cut points at $(30, 60)$. Table 1 summarizes parameter estimates from the multistate analysis. It shows increasing trends in the baseline intensities for the onset of psoriasis and psoriasis to psoriatic arthritis transition as people age. There is a significant gender difference in the risks of psoriasis and progression from psoriasis to psoriatic arthritis. Women have significantly higher risk of psoriasis ($\exp(\widehat{\beta}_0) = 1.454, 95\%$ CI:$(1.419, 1.489)$, $p < 0.0001$), but lower risk of developing psoriatic arthritis once with psoriasis ($\exp(\widehat{\beta}_1) = 0.766, 95\%$ CI:$(0.717, 0.818)$, $p < 0.0001$). From a naive logistic regression based on a cross-sectional sample of 347 patients with psoriasis or psoriasis and psoriatic arthritis from the NPF survey, the estimated odds ratio for the association between gender and psoriatic arthritis vs. psoriasis is 1.234 (95% CI: $(0.656, 2.323), p = 0.515$), which is consistent with literature (Eder et al., 2012). On the other hand, an earlier age of onset of psoriasis in females and a higher probability of severe disease in men has also been reported (Colombo et al., 2014). Our results are from a first attempt of using multistate models to capture the disease dynamics, information from different data sources are pulled together to facilitate and enhance the estimation from such elaborate models.

Table 1: Parameter estimates from multistate analysis of cross-sectional and auxiliary data on psoriasis and psoriatic arthritis

			TRANSITION			
			$j=0$ Healthy → Ps		$j=1$ Ps → PsA	
			EST.	S.E.	EST.	S.E.
Intensity	$\log\lambda_j(t)$	(0, 30]	-8.434	0.009	-5.741	0.029
		(30, 60]	-7.328	0.008	-4.192	0.019
		>60	-6.661	0.020	-3.143	0.022
Gender	β_j		0.374	0.012	-0.267	0.034

We distinguish the state of death according to the disease state occupied at the time of death as shown in Figure 6 and under this new six-state multistate diagram the state D_j indicates the death post state j, for $j = 0, 1, 2$. The transition intensity functions involved are the same as those in the previous four-state space diagram. Based on this, we define the cumulative incidence function for psoriasis and psoriatic arthritis as

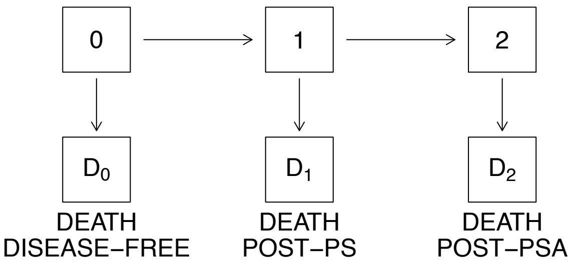


Figure 6: A state space diagram for healthy, psoriasis, psoriatic arthritis, and cause-specific death.

$$\sum_{k=j}^{2}[P(Z(t)=k \mid Z(0)=0, S_0, X) + P(Z(t)=D_k \mid Z(0), S_0, X)]\ ,\ j=1,2\ ,$$

shown in Figure 7 for different birth cohorts defined by $S_0 = 1940$ and $S_0 = 1970$. The cumulative incidence of psoriasis increases sharply as age increases, and there is a clear gender difference in the cumulative incidence function particularly for psoriasis. The much lower cumulative incidence of psoriatic arthritis is as expected because this is only among patients who have developed psoriasis first.

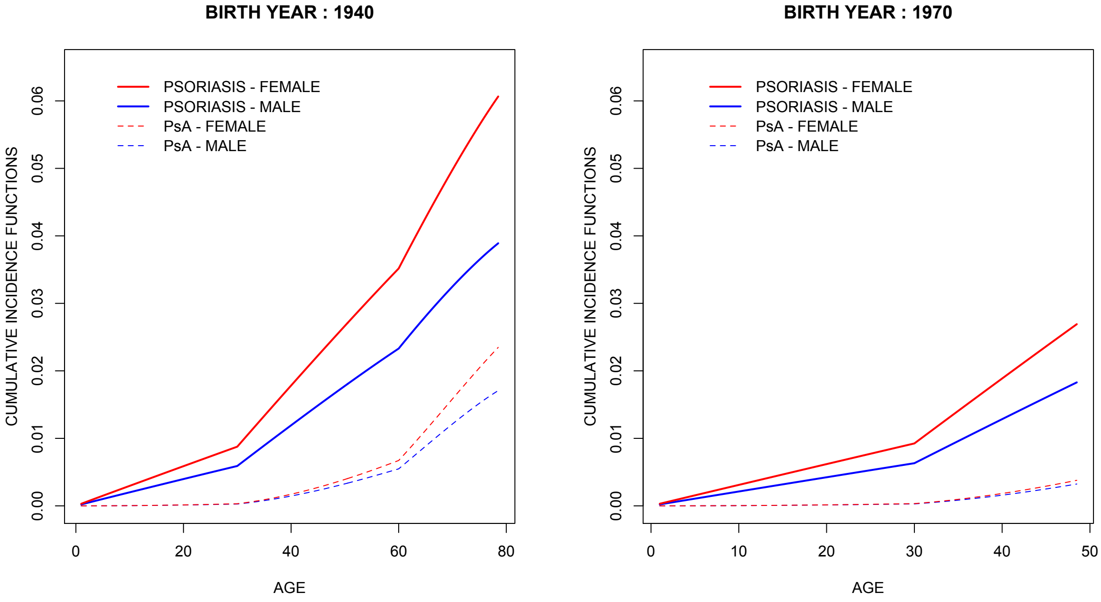


Figure 7: Cumulative incidence function estimates for psoriasis and psoriatic arthritis (PsA) for males and females based on the six-state model in Figure 6.

Finally, based on the multistate analysis we estimate the prevalence of the disease for a birth cohort between 1921 and 2009 according to formula (4). Based on the birth data from Statistics Canada, we consider linear spline models for the birth time distribution, $f(s_0)$, with a constant rate of $250,000$ births per year prior 1936, an increasing trend with 10,000 more births per year during the baby boomer time from 1936 to 1960, a decreasing trend from 1960 to 1969, and a flat rate of 375,000 births per year after that. Our estimates of the prevalence are 1.22% (male) and 1.9% (female) for psoriasis; 0.27% (male) and 0.36% (female) for psoriatic arthritis. WHO Global Report on Psoriasis (World Health Organization, 2016) estimated the prevalence of psoriasis to be 1.5-5% in developed countries, and Catanoso et al. (2012) suggested that the prevalence of psoriasis in the general population is 2-3% and that for psoriatic arthritis is 0.3-1.0%. The estimated prevalence from the multistate analysis and the reported prevalences from these large scale cross-sectional studies are therefore in good alignment.

5 Discussion

This work was motivated by a research project in which collaborators had reviewed published work aiming to identify markers associated with the presence of psoriatic arthritis in individuals with psoriasis. Most articles reported on the results of fitting logistic regression models based on their respective cross-sectional samples using a binary disease status as the response and genetic markers as predictors (Winchester et al., 2012; Haroon et al., 2013; Loft et al., 2018). To explore the relation between the intensity-based effect of a marker on a transition intensity in a multistate framework and the apparent effect on state occupancy in a cross-sectional sample, we considered problems in the context of cross-sectional sampling from a population in the spirit of Keiding (1991). We adopted a progressive intensity-based model for the life history (disease) process applicable to each individual and assume births arise in calendar times according to a stationary process. We then derived the limiting value of the effect of a marker on the disease state occupancy when it is estimated based on a logistic regression model fitted to data from a cross-sectional sample. In doing this we derived i. the state prevalence at a calendar time, ii. the state-specific age distributions, and iii. age-specific state prevalences. We also study the limiting behaviour of marker effect estimates when logistic regression models adjust for the age of individuals, an approach that is routinely adopted to adjust for time at risk. Interestingly we find that following adjustment for the individuals' age in a logistic regression analysis, the bias in the odds ratio (or log odds ratio) estimator may be larger than that of the unadjusted analysis. Details on the derivations can be found in the supplementary appendices.

Most studies of life history processes involve considering the effect of covariates on transition intensities; this was raised by Cuzick (1991) and Brookmeyer (1991) in the discussion of Keiding and Moeschberger (1992). In occupational epidemiology, Thompson et al. (1998) discuss regression based on the prevalence odds ratio, prevalence ratio and the incidence rate ratio and point out that adjustment for potential confounders in these frameworks will be done differently. Miettinen (1976) discusses estimability issues in case-referrant studies and emphasizes issues of interpretability. Barros and Hirakata (2003) among others point out that simple logistic regression will typically yield estimates that are uninterpretable. The fact that there exists a setting where a binary analysis can give a consistent estimator of the intensity-based effect is surprising, but we emphasize that this is only the case when there is a stationary birth process, there are no trends in the disease process, transition intensities are time homogenous, and the conditions necessary in (X) are satisfied. More generally the two approaches to analysis will usually lead to quite different inferences about the effect of markers as demonstrated in Figure 4. Specifically, if the covariate has a stronger effect on death after disease than on the incidence of disease in either the illness-death or psoriatic arthritis processes, binary analysis would indicate an association on the opposite direction of the true marker effect.

The multistate process we describe is relatively simple (e.g. it does not allow for recurrent disease states) in order to provide a basis for discussion of the issues and calculations related to cross-sectional analyses. If interest lies in fitting this model to data certain transition intensities will be inestimable and auxiliary data are required. We construct a likelihood function which makes use of auxiliary follow-up data from the cohorts of individuals, population mortality data by year, and current status data from a national survey which facilitates fitting of the multistate model; see also Palloni and Thomas (2011) for related work. When possible it is important to assess the compatibility of the data from different sources. In cases where even limited data are available some tests could be carried out to investigate whether the data are plausibly coming from populations with similar compositions of risk factors. We have allowed the mortality rates to differ for disease and disease-free individuals but more could be done in terms of sensitivity analyses to check into the influence of compatibility assumptions on findings.

A significant challenge which we have not addressed is the complex referral process to the two registries providing retrospective and prospective data. While some individuals are recruited by way of screening the population, the real process is more complex since patients can be referred by primary care or other specialist physicians. Modeling the referral process may help mitigate biases arising from dependent delayed entry but the recruitment process is likely also tied to the severity of the symptoms which may be associated with risk of disease transitions. We acknowledge this limitation and consider it an important topic for future research given the increased emphasis on use of disease registries in recent times.

There are limitations of simply using cross-sectional data to analyse life history processes (Kraemer et al., 2000) and use of odds ratios to characterize effects (Langholz, 2010). There has been considerable discussion about the merits of using retrospectively reported transition times (e.g. dates of disease onset in prevalent cohort samples) in analyses. Diamond and McDonald (1992) discuss the poor reliability of retrospective data, and the fact that the errors are often of an unknown size and direction. In the absence of any knowledge about this measurement error process it is difficult to consider the nature of any biases that may result from it. Jewell (2016) gives a broad review of the alternative types of designs that can be employed for the study of life history processes and an extensive bibliography.

ACKNOWLEDGEMENTS

The authors thank Dr. Michael Siegel and the National Psoriasis Foundation for permission to use the data from the national survey; Dr Dafna Gladman, Dr Vinod Chandran, and Dr Lihi Eder of the Centre for Prognosis Studies inRheumatic Disease at the University of Toronto for collaboration; and Dr Jerry Lawless and Dr Niels Keiding for comments on an earlier version of this manuscript.

FUNDING

This work was supported by the Natural Science and Engineering Research Council of Canada through grants RGPIN 115928 (LZ) and RGPIN 155849 (RJC) and the Canadian Institutes for Health Research through grant FRN 13887 (RJC). Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research.

REFERENCES

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

Ahrens, W. and Pigeot, I. (2014). *Handbook of Epidemiology*. Springer, New York, NY.

Alho, J. M. (1992). On prevalence, incidence, and duration in general stable populations. *Biometrics*, pages 587–592.

Aljied, R., Aubin, M.-J., Buhrmann, R., Sabeti, S., and Freeman, E. E. (2018). Prevalence and determinants of visual impairment in canada: cross-sectional data from the canadian longitudinal study on aging. *Canadian Journal of Ophthalmology*, 53(3):291–297.

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

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Asgharian, M. and Wolfson, D. B. (2005). Asymptotic behavior of the unconditional npmle of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics*, 33(5):2109–2131.

Barros, A. J. and Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology*, 3(1):21.

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, 42(4):693–734.

Brinks, R. and Hoyer, A. (2018). Illness-death model: statistical perspective and differential equations. *Lifetime Data Analysis*, 24(4):743–754.

Brookmeyer, R. (1991). Discussion of the paper by Keiding, “Age-specific incidence and prevalence: a statistical perspective”. *Journal of the Royal Statistical Society, Series A*, 154(3):402.

Catanoso, M., Pipitone, N., and Salvarani, C. (2012). Epidemiology of psoriatic arthritis. *Reumatismo*, 64(2):66–70.

Colombo, D., Cassano, N., Bellia, G., and Vena, G. A. (2014). Gender medicine and psoriasis. *World Journal of Dermatology*, 3(3):36–44.

Cook, R. and Lawless, J. (2018). *Multistate Models for the Analysis of Life History Data*. CRC Press, Boca Raton, FL.

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

Cresswell, L., Chandran, V., Farewell, V. T., and Gladman, D. D. (2011). Inflammation in an individual joint predicts damage to that joint in psoriatic arthritis. *Annals of the Rheumatic Diseases*, 70(2):305–308.

Cuzick, J. (1991). Discussion of the paper by Keiding, “Age-specific incidence and prevalence: a statistical perspective”. *Journal of the Royal Statistical Society, Series A*, 154(3):398.

Diamond, I. D. and McDonald, J. W. (1992). Analysis of current-status data. In Trussell, J., Hankinson, R., and Tilton, J., editors, *Demographic Applications of Event History Analysis.* , pages 231–252. Oxford University Press.

Eder, L., Chandran, V., and Gladman, D. D. (2012). Gender-related differences in patients with psoriatic arthritis. *International Journal of Clinical Rheumatology*, 7(6):641.

Fix, E. and Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, 23(3):205–241.

Gelfand, J. M., Gladman, D. D., Mease, P. J., Smith, N., Margolis, D. J., Nijsten, T., Stern, R. S., Feldman, S. R., and Rolstad, T. (2005). Epidemiology of psoriatic arthritis in the population of the United States. *Journal of the American Academy of Dermatology*, 53(4):573–e1.

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

Gladman, D. D. (1991). Psoriatic arthritis. In Bellamy, N., editor, *Prognosis in the Rheumatic Diseases.* , pages 153–166. Springer.

Gladman, D. D. and Chandran, V. (2010). Observational cohort studies: lessons learnt from the university of toronto psoriatic arthritis program. *Rheumatology*, 50(1):25–31.

Haroon, M., FitzGerald, O., and Winchester, R. (2013). Epidemiology, genetics and management of psoriatic arthritis 2013: focus on developments of who develops the disease, its clinical features, and emerging treatment options. *Psoriasis Targets Therapy*, 3:11–23.

Hoem, J. and Jensen, U. (1982). Multistate life table methodology: a probabilist critique. In Land, K. and Rogers, A., editors, *Multidimensional mathematical demography.*, chapter 4, pages 155–264. New York NY Academic Press.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Analysis*, 5(3):239–264.

Huang, C. Y. and Qin, J. (2011). Nonparametric estimation for length-biased and right-censored data. *Biometrika*, 98(1):177–186.

Husted, J. A., Gladman, D. D., Farewell, V. T., and Cook, R. J. (2001). Health-related quality of life of patients with psoriatic arthritis: a comparison with patients with rheumatoid arthritis. *Arthritis Care and Research*, 45(2):151–158.

Jewell, N. P. (2016). Natural history of diseases: Statistical designs and issues. *Clinical Pharmacology & Therapeutics*, 100(4):353–361.

Jung, Y. S., Yun, K. E., Chang, Y., Ryu, S., Park, J. H., Kim, H. J., Cho, Y. K., Sohn, C. I., Jeon, W. K., Kim, B. I., et al. (2014). Risk factors associated with rectal neuroendocrine tumors: a cross-sectional study. *Cancer Epidemiology and Prevention Biomarkers*, 23(7):1406–1413.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 332(1627):487–509.

Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society, Series A*, 154(3):371–412.

Keiding, N. (2006). Event history analysis and the cross-section. *Statistics in Medicine*, 25(14):2343–2364.

Keiding, N. (2011). Age-period-cohort analysis in the 1870s: Diagrams, stereograms, and the basic differential equation. *Canadian Journal of Statistics*, 39(3):405–420.

Keiding, N., Hansen, O. K., Sørensen, D. N., and Slama, R. (2012). The current duration approach to estimating time to pregnancy. *Scandinavian Journal of Statistics*, 39(2):185–204.

Keiding, N. and Moeschberger, M. (1992). Independent delayed entry. In Klein, J. P. and Goel, P. K., editors, *Survival analysis: State of the art.* , pages 309–326. Springer Science & Business Media.

Kobelt, G., Jönsson, L., Lindgren, P., Young, A., and Eberhardt, K. (2002). Modeling the progression of rheumatoid arthritis: a two-country model to estimate costs and consequences of rheumatoid arthritis. *Arthritis & Rheumatism*, 46(9):2310–2319.

Kraemer, H. C., Yesavage, J. A., Taylor, J. L., and Kupfer, D. (2000). How can we learn about developmental processes from cross-sectional studies, or can we? *American Journal of Psychiatry*, 157(2):163–171.

Langholz, B. (2010). Case-control studies= odds ratios: blame the retrospective model. *Epidemiology*, 21(1):10–12.

Lee, J. Y. and Cook, R. J. (2018). The illness-death model for family studies. *Biostatistics*.

Lexis, W. H. R. A. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. KJ Trübner.

Loft, N. D., Skov, L., Rasmussen, M. K., Gniadecki, R., Dam, T. N., Brandslund, I., Hoffmann, H. J., Andersen, M. R., Dessau, R. B., Bergmann, A. C., et al. (2018). Genetic polymorphisms associated with psoriasis and development of psoriatic arthritis in patients with psoriasis. *PloS One*, 13(2):e0192010.

Luo, X. and Tsai, W. Y. (2009). Nonparametric estimation for right-censored length-biased data: a pseudo-partial likelihood approach. *Biometrika*, 96(4):873–886.

Marshall, G. and Jones, R. H. (1995). Multi-state models and diabetic retinopathy. *Statistics in Medicine*, 14(18):1975–1983.

Miettinen, O. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, 103(2):226–235.

Newman, S. C. (1988). Odds ratio estimation in a steady-state population. *Journal of Clinical Epidemiology*, 41(1):59–65.

Østbye, T., Kristjansson, B., Hill, G., Newman, S. C., Brouwer, R. N., and McDowell, I. (2005). Prevalence and predictors of depression in elderly canadians: The canadian study of health and aging. *Chronic Diseases and Injuries in Canada*, 26(4):93.

Palloni, A. and Thomas, J. (2011). Estimation of health status inequalities from prevalence data: A risky business (cde working paper no. 2011–09). Technical report, Center for Demography and Ecology, University of Wisconsin, Madison.

Qin, J., Ning, J., Liu, H., and Shen, Y. (2011). Maximum likelihood estimations and em algorithms with length-biased data. *Journal of the American Statistical Association*, 106(496):1434–1449.

Rachakonda, T. D., Schupp, C. W., and Armstrong, A. W. (2014). Psoriasis prevalence among adults in the united states. *Journal of the American Academy of Dermatology*, 70(3):512–516.

Robert, B. (2017). Mortality data for Canada. https://www.mortality.org/cgi-bin/hmd/country.php?cntr=CAN&level=1.

Sweeting, M., De Angelis, D., Neal, K., Ramsay, M., Irving, W., Wright, M., Brant, L., Harris, H., and Trent HCV Study Group and HCV National Register Steering Group and others (2006). Estimated progression rates in three United Kingdom hepatitis C cohorts differed according to method of recruitment. *Journal of Clinical Epidemiology*, 59(2):144–152.

Thompson, M. L., Myers, J., and Kriebel, D. (1998). Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? *Occupational and Environmental Medicine*, 55(4):272–277.

Toperoff, G., Kark, J. D., Aran, D., Nassar, H., Ahmad, W. A., Sinnreich, R., Azaiza, D., Glaser, B., and Hellman, A. (2015). Premature aging of leukocyte dna methylation is associated with type 2 diabetes prevalence. *Clinical epigenetics*, 7(1):35.

Tyas, S. L., Salazar, J. C., Snowdon, D. A., Desrosiers, M. F., Riley, K. P., Mendiondo, M. S., and Kryscio, R. J. (2007). Transitions to mild cognitive impairments, dementia, and death: findings from the nun study. *American Journal of Epidemiology*, 165(11):1231–1238.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Winchester, R., Minevich, G., Steshenko, V., Kirby, B., Kane, D., Greenberg, D. A., and FitzGerald, O. (2012). Hla associations reveal genetic heterogeneity in psoriatic arthritis and in the psoriasis phenotype. *Arthritis & Rheumatism*, 64(4):1134–1144.

World Health Organization (2016). Global report on psoriasis 2016. http://apps.who.int/iris/bitstream/10665/204417/1/9789241565189_eng.pdf.

Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3):716–725.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, 56(3):601–614.

Appendix A: Derivation of Transition Probabilities

Consider a time homogeneous $K+1$-state progressive disease process as shown Figure 3. We write the transition intensity functions as $\lambda_j(t) = \lambda_j$ and $\eta_j(t) = \eta_j$ to differentiate the risks of progression and mortality from stage j, $j = 0, 1, \ldots, K-1$ and $\lambda_{K-1} = 0$. The sojourn time in state j (the time from entry to exit out of state j) follows an exponential distribution with a hazard $h_j = \lambda_j + \eta_j$, survival function $\mathcal{F}_j(t) = \exp\{-h_j t\}$ and the mean sojourn time $1/h_j$, $j = 0, \ldots, K-1$. Under this time-homogeneous model, we have transition probabilities

$$P(Z(s+t) = k \mid Z(s) = j) = P_{jk}(s, s+t) = P_{jk}(t) \ , \forall j < k \ .$$

Let $h_j = \lambda_j + \eta_j$ and $\mathcal{F}_j(t) = \int_0^t \exp\{-h_j u\} du$ be hazard and survival functions of the sojourn time in state j, the transition probabilities can be actually expressed in terms of distributions of sojourn times. More specifically, we have $P_{00}(t) = \mathcal{F}_0(t)$, and given that

$$\int_0^t \mathcal{F}_{j_1}(u)\mathcal{F}_{j_2}(t-u)du = \frac{\mathcal{F}_{j_1}(t) - \mathcal{F}_{j_2}(t)}{h_{j_2} - h_{j_1}} \quad , \forall j_1, j_2 \in \mathcal{S} \ ,$$

we can obtain the following results

$$\begin{aligned}
P_{01}(t) &= \int_0^t P_{00}(u)\lambda_0\mathcal{F}_1(t-u)\, du = \lambda_0 \frac{\mathcal{F}_0(t) - \mathcal{F}_1(t)}{h_1 - h_0} \\
P_{02}(t) &= \int_0^t P_{01}(u)\lambda_1\mathcal{F}_2(t-u)\, du = \lambda_0\lambda_1 \left[\frac{\mathcal{F}_0(t) - \mathcal{F}_2(T)}{(h_1 - h_0)(h_2 - h_0)} - \frac{\mathcal{F}_1(t) - \mathcal{F}_2(t)}{(h_1 - h_0)(h_2 - h_1)} \right] \\
P_{03}(t) &= \int_0^t P_{02}(u)\lambda_2\mathcal{F}_3(t-u)\, du
\end{aligned}$$

where $P_{03}(t)$ can be written as

$$\lambda_0\lambda_1\lambda_2 \Big[\frac{\mathcal{F}_0(t) - \mathcal{F}_3(t)}{(h_1 - h_0)(h_2 - h_0)(h_3 - h_0)} - \frac{\mathcal{F}_1(t) - \mathcal{F}_3(t)}{(h_1 - h_0)(h_2 - h_1)(h_3 - h_1)} + \frac{\mathcal{F}_2(t) - \mathcal{F}_3(t)}{(h_2 - h_0)(h_2 - h_1)(h_3 - h_2)} \Big] \ .$$

Through mathematical induction, the transition probability from state 0 to state j is

$$P_{0j}(t) = \int_0^t P_{0,j-1}(u)\ \lambda_{j-1}\ \mathcal{F}_j(t-u)\ du = \left[\prod_{k=0}^{j-1} \lambda_k \right] \left[\sum_{k=0}^{j-1} \frac{\mathcal{F}_k(t) - \mathcal{F}_j(t)}{\left[\prod_{\ell=0}^{k-1}(h_\ell - h_k)\right]^{I(k>0)} \prod_{\ell=k+1}^{j}(h_\ell - h_k)} \right] \ ,$$

and its integration is subsequently

$$\int_0^\infty P_{0j}(t)\ dt = \left[\prod_{k=0}^{j-1} \frac{\lambda_k}{h_k} \right] \frac{1}{h_j} = \left[\prod_{k=0}^{j-1} \frac{\lambda_k}{\lambda_k + \eta_k} \right] \frac{1}{\lambda_j + \eta_j} \ , \quad \forall j = 1, \ldots, K-1 \ .$$

Appendix B: Simulation Results for a Two-stage Disease Process

Here we provide the simulation results from cross-sectional studies of a time-homogeneous four-state disease process as shown in Figure 1 and discussed in Section 2.5. We consider a single binary covariate X with $P(X = 1) = 0.5$ which is associated with $1 \to 2$ and $2 \to 3$ transitions. The intensity functions for $0 \to 1$ and $1 \to 2$ transitions are of the form $\lambda_0 \exp(X\beta_0)$ and $\lambda_1 \exp(X\beta_1)$, and the intensity function for $j \to 3$ transition is $\eta_j \exp(X\alpha_j)$, $j = 0, 1, 2$. We set $\beta_0 = \alpha_0 = \alpha_1 = 0$

and $\eta_0 = \eta_1 = \eta_2$, and allow the values of β_1 and α_2 to vary. The values of parameters η_0, λ_0 and λ_1 are solved by setting $P(Y = 0) = 0.1$, $P(Y = 1) = 0.01$ and $P(Y = 2) = 0.01$ following the formula given in (17).

We simulate the life history of such a disease process for a population of 300,000 people that were born between the time interval $(0, R]$ with a constant rate, and set $R = 10$ without loss of generality. Figure 8 shows the histogram of the age of individuals in this population by their state occupied at time R, the derived age distribution functions based on multistate models as given in (10) in Section 2.2 are superimposed, and there is a good agreement.

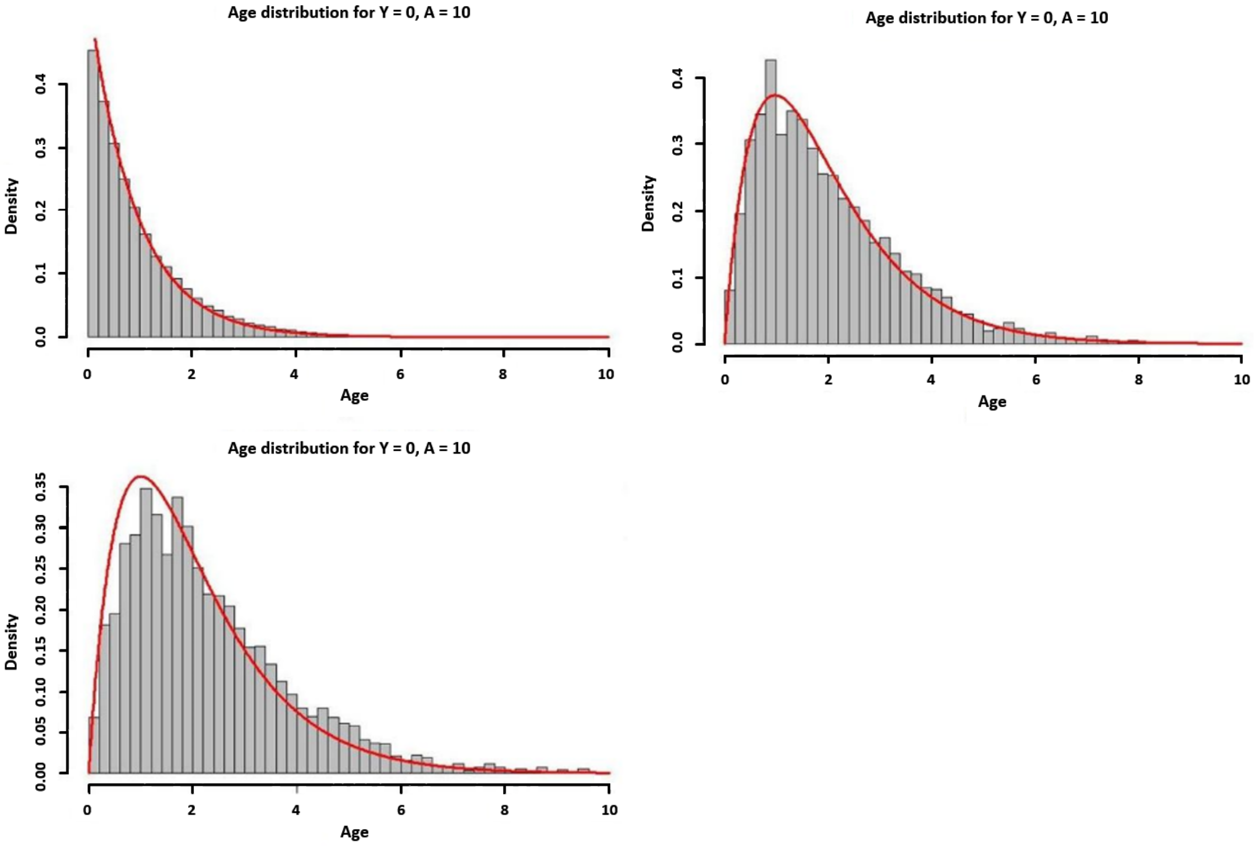


Figure 8: Age distribution of the cross-sectional sample by state occupation.

A cross-sectional sample is formed from a simulated population of size $n = 10,000$ according to their disease status at the fixed sampling time R, using the method described previously. The sample contains individuals in states 1 (e.g. Ps) and 2 (e.g. PsA), a binary response indicating state 2 is created and a logistic regression analysis is conducted with a single covariate X and its coefficient γ_1. Table 2 reports the limiting value and the average estimates of $\exp(\gamma_1)$ from the binary logistic regression analysis based on 100 iterations, they agree quite well. One can also see that the relative risk, $\exp(\beta_1)$, from the multistate model and the limiting value of the odds ratio from the cross-sectional samples, $\exp(\gamma_1)$, may imply the association between the covariate and the stage 2 disease in opposite directions under certain circumstances.

Table 2: Asymptotic and simulation results from a four-state progressive disease process

		$\exp(\gamma_1)^a$	
$\exp(\beta_1)^b$	$\exp(\alpha_2)^c$	Limiting Value	Empirical Mean
0.35	0.35	1.000	0.981
0.35	0.50	0.700	0.695
0.35	0.75	0.467	0.477
0.50	0.35	1.429	1.393
0.50	0.50	1.000	1.003
0.50	0.75	0.667	0.685
0.75	0.35	2.143	2.180
0.75	0.50	1.500	1.487
0.75	0.75	1.000	0.994
1.25	1.25	1.000	1.027
1.25	1.50	0.833	0.845
1.25	1.75	0.714	0.736
1.50	1.25	1.200	1.221
1.50	1.50	1.000	1.043
1.50	1.75	0.857	0.864
1.75	1.25	1.400	1.440
1.75	1.50	1.167	1.218
1.75	1.75	1.000	1.033

[a] Limiting value and empirical average of odds ratio estimators under misspecified logistic regression model

[b] Multiplicative effect of marker on $1 \rightarrow 2$ transition intensity of four-state process

[c] Multiplicative effect of marker on $2 \rightarrow 3$ transition intensitiy of four-state process

Appendix C: An Extension Accommodating Direct Transitions

Instead of assuming the process is strictly one-step progressive, this section considers extensions that allow progression between two un-adjacent states. For simplicity and illustration purpose, the discussion is in the context of the Psoriatic Arthritis example where $0 \rightarrow 2$ transition is allowed as shown in Figure 9. The intensity for $j \rightarrow k$ transition takes a form $\lambda_{jk}\exp(X\beta_{jk})$, $j < k$ and $j, k \in \{0, 1, 2, 3\}$.

Note that the general result given in (6) is no longer valid for the calculation of the transition probability $P_{02}(t)$, which affects the derivation of the overall prevalence, odds and odds ratio subsequently. Under the time-homogeneous model,

$$\begin{aligned} P_{02}(t) &= \int_0^t P_{01}(u)\lambda_{12}P_{22}(t-u)\,du + \int_0^t P_{00}(u)\lambda_{02}P_{22}(t-u)\,du \\ &= \lambda_{01}\lambda_{12}\left[\frac{\mathcal{F}_0(t)-\mathcal{F}_2(t)}{(h_1-h_0)(h_2-h_0)} - \frac{\mathcal{F}_2(t)-\mathcal{F}_1(t)}{(h_0-h_1)(h_2-h_1)}\right] + \lambda_{02}\frac{\mathcal{F}_0(t)-\mathcal{F}_2(t)}{h_2-h_0} \end{aligned}$$

where h_j and $F_j(t)$ are the hazard and the survival functions of the sojourn times in state j as before, but h_0 = $\lambda_{01} + \lambda_{02} + \eta_{03}$, $h_1 = \lambda_{12} + \eta_{13}$, and $h_2 = \eta_{23}$. This leads to different expressions of mean

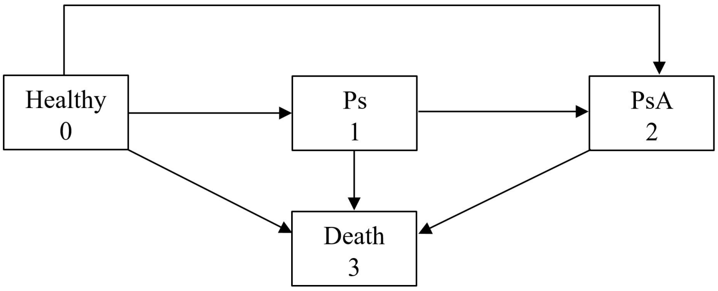


Figure 9: Multi-state diagram for the psoriasis and psoriatic arthritis disease process with $0 \rightarrow 2$ transition.

diseased lifetime with Ps or PsA such that

$$\begin{aligned}\int_0^\infty P_{01}(t)\,dt &= \frac{\lambda_{01}}{h_0}\frac{1}{h_1}\\ \int_0^\infty P_{02}(t)\,dt &= \frac{\lambda_{01}}{h_0}\frac{\lambda_{12}}{h_1}\frac{1}{h_2}+\frac{\lambda_{02}}{h_0}\frac{1}{h_2}\,.\end{aligned}$$

The overall prevalence odds of PsA against Ps now becomes

$$\begin{aligned}\mathit{ODDS} &= \frac{\lambda_{12}}{h_2}+\frac{\lambda_{02}/h_2}{\lambda_{01}/h_1}\\ &= \text{incidence of Ps to PsA} \times \text{PsA duration} + \frac{\text{incidence of healthy to PsA} \times \text{PsA duration}}{\text{incidence of Ps} \times \text{Ps duration}}\,.\end{aligned}$$

If we assume non-differential mortality due to the covariate, that is $\alpha_{03} = \alpha_{13} = \alpha_{23} = 0$, following the same discussion as in Section 2.4, the ratio of the overall odds of PsA against Ps for individuals with versus without the marker of interest can be shown to be

$$\mathit{OR} = w_1\mathit{RR} + w_2\mathit{RR}\cdot\exp\left(\beta_{02}-\beta_{01}\right) + (1-w_1-w_2)\exp\left(\beta_{02}-\beta_{01}\right)\,, \tag{23}$$

a weighted average of some sort with $w_1 = r_1r_2/(r_1r_2+r_2+1)$, $w_2 = r_2/(r_1r_2+r_2+1)$, $r_1 = \lambda_{01}/\lambda_{02}$ and $r_2 = \lambda_{12}/\eta_{13}$. Again, the OR given above is a complicated function of the actual relative risk $\exp\{\beta_{12}\}$. Suppose the marker effect is either preventive or conducive on both Ps and PsA. When the marker has a stronger association with the incidence of PsA than Ps (i.e. $|\beta_{02}| \geq |\beta_{01}|$), the *OR* estimated from binary analysis is always in the same direction as the actual relative risk $\exp\{\beta_{12}\}$ as shown in Figure 10. As the ratios r_1 and r_2 get bigger than 1, which implies a relatively higher risk of $0 \rightarrow 1 \rightarrow 2$ path than the $0 \rightarrow 2$ and $0 \rightarrow 1 \rightarrow 3$ path, $w_3 \rightarrow 0$ and hence *OR* $\rightarrow \exp\{\beta_{12}\}$. On the contrary, if r_1 and r_2 decrease from 1, $w_3 \rightarrow 1$ and OR $\rightarrow 1$ implying a null effect. That is, as the lifetime risk of a $0 \rightarrow 1 \rightarrow 2$ path (against $0 \rightarrow 1 \rightarrow 3$ and $0 \rightarrow 2$ paths) is getting higher, the odds ratio estimated from binary analysis approaches to the actual relative risk. On the other hand, the odds ratio provides a dampened underestimation of the relative risk, as the risk of the $0 \rightarrow 1 \rightarrow 2$ path decreases. Finally, when the marker has a stronger association with incidence of Ps than PsA, the *OR* estimated from binary analysis may imply an association of an opposite nature of the one based on the relative risk $\exp\{\beta_{12}\}$ from the multistate model. These findings are further illustrated by simulation results reported in Table 3.

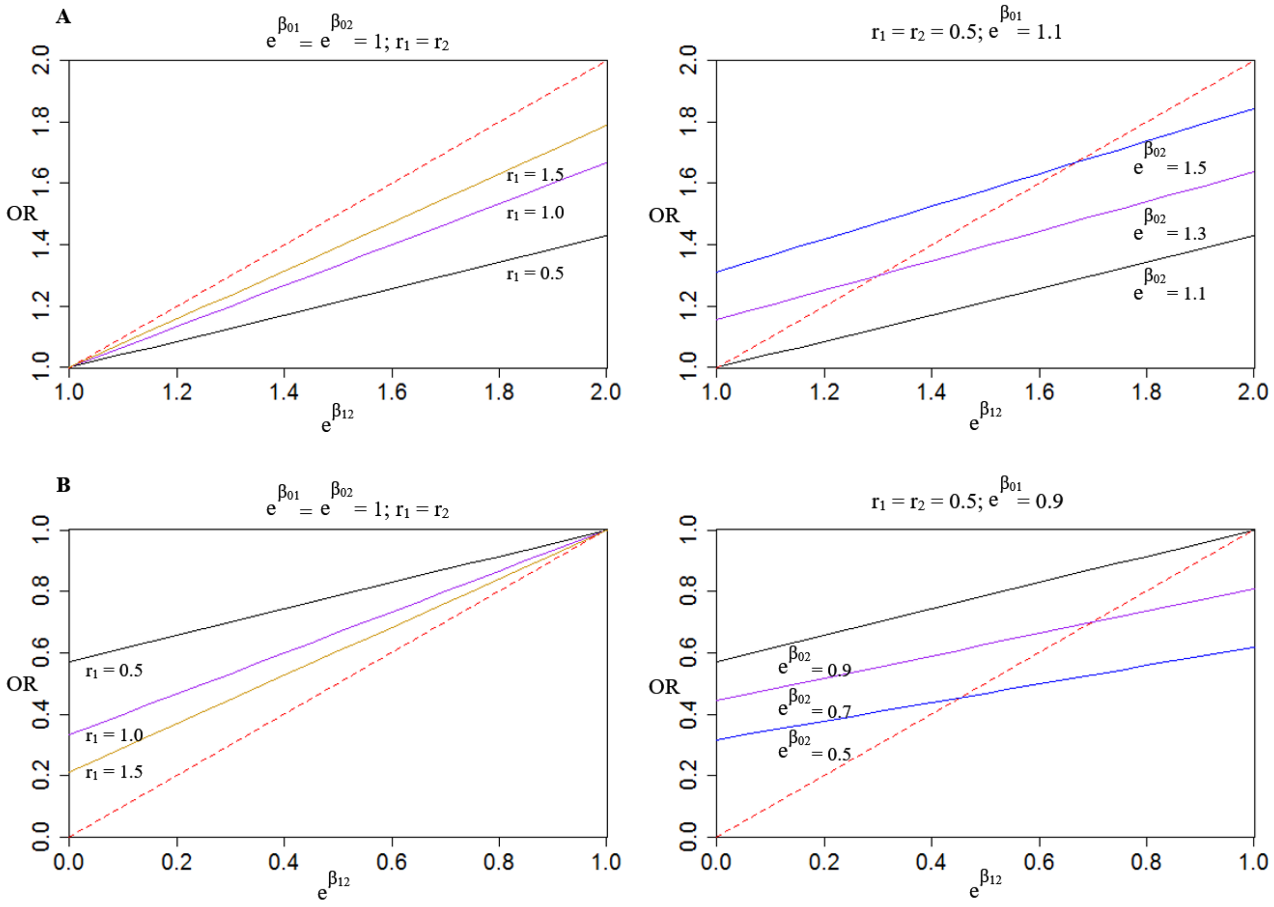


Figure 10: A: Covariate effect for varying values of r_1, r_2, β_{01} and β_{02}, where $\beta_{12} > 0$. B: Covariate effect for varying values of r_1, r_2, β_{01} and β_{02}, where $\beta_{12} < 0$.

Table 3: Simulation results for an extended four-state progressive process allow $0 \to 2$ transition: $(P(Y=0)=0.1, P(Y=1)=0.01, P(Y=2)=0.01$, R = 10, n = 10,000, 200 iterations).

			$\exp(\gamma_1)^a$	
$\exp(\beta_{01})$	$\exp(\beta_{02})$	$\exp(\beta_{12})$	Value	Mean
0.25	0.25	0.25	0.785	0.794
0.25	0.25	0.50	0.856	0.898
0.25	0.25	0.75	0.928	0.967
0.25	0.50	0.25	1.569	1.642
0.25	0.50	0.50	1.660	1.734
0.25	0.50	0.75	1.751	1.833
0.25	0.75	0.25	2.371	2.535
0.25	0.75	0.50	2.476	2.649
0.25	0.75	0.75	2.582	2.739
0.50	0.25	0.25	0.417	0.397
0.50	0.25	0.50	0.466	0.461
0.50	0.25	0.75	0.516	0.508
0.50	0.50	0.25	0.821	0.845
0.50	0.50	0.50	0.880	0.915
0.50	0.50	0.75	0.940	0.946
0.50	0.75	0.25	1.234	1.241
0.50	0.75	0.50	1.300	1.341
0.50	0.75	0.75	1.367	1.381
0.75	0.25	0.25	0.289	0.275
0.75	0.25	0.50	0.329	0.318

[a] Limiting value and empirical average of odds ratio estimators under misspecified logistic regression model