Armitage Lecture 2011: The Design and Analysis of Life History Studies

JERALD F. LAWLESS

Department of Statistics and Actuarial Science,

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

E-mail: jlawless@uwaterloo.ca

Summary

Life history studies collect information on events and other outcomes during people's lifetimes. For example, these may be related to childhood development, education, fertility, health, or employment. Such longitudinal studies have constraints on the selection of study members, the duration and frequency of follow-up, and the accuracy and completeness of information obtained. These constraints, along with factors associated with the definition and measurement of certain outcomes, affect our ability to understand, model, and analyze life history processes. My objective here is to discuss and illustrate some issues associated with the design and analysis of life history studies.

Keywords: Heterogeneity, Incomplete data, Intermittent observationm Multistate models, Markov processes

1 INTRODUCTION

Information about events and other outcomes in people's lives is collected in many contexts, and well-planned studies are important in understanding life history processes and factors that influence them. Examples include studies related to employment [1], health [2, 3], and aging [4, 5]. I use the term life history studies to describe such initiatives. Objectives of life history analysis include enhancing our understanding of individual processes and of variation across individuals, groups, or populations; identifying relationships between processes and covariates; identifying risk factors associated with adverse outcomes; assessing the effects of individual-level or population-level interventions; and providing predictive models for activities such as planning, resource allocation, or patient management. The extent to which objectives can be realized depends on the representativeness of the panel members, on the events and other variables that are measured, and on the completeness and accuracy of data collection. Time, cost, and other constraints lead to a range of study designs, from ones where a randomly selected panel of individuals is followed over a designated period to observational studies based on administrative records.

Various statistical challenges arise with life history studies, and my objective is to discuss issues associated with their design and analysis. Before outlining some issues to be addressed, I introduce three studies that serve to illustrate various points.

Example 1.1 (Complications from Type 1 Diabetes)

The Diabetes Control and Complications Trial (DCCT) was a randomized study involving subjects with type 1 diabetes, which ran from 1983 to 1993. The study randomized subjects in two cohorts to a program of intensive diabetes therapy designed to achieve near-normal glucose levels or to conventional therapy designed to prevent hyperglycemic symptoms [3]. The two cohorts were a primary prevention cohort, consisting of individuals who had no retinopathy at the time of study entry, and a secondary intervention cohort, whose members had some degree of retinopathy. Upon completion, the trial showed that the intensive therapy led to a significant reduction in the onset and progression of diabetic retinopathy and nephropathy. After the DCCT was terminated, most subjects (1375 of 1441) joined the observational Epidemiology of Diabetes Interventions and Complications (EDIC) Study, which has been ongoing since 1994. Its objectives are to study complications from diabetes, and biological and genetic factors associated with progression. Those conducting the study took eye and kidney measurements approximately every 6 months during the DCCT, but only every 2-4 years during EDIC. They expressed the degree of retinopathy (ranging from none to severe) on the ordinal Early Treatment Diabetic Retinopathy Study (ETDRS) scale [6]. Measurements on renal function included urinary albumin excretion rate, and we can base disease states on this [7].

Numerous covariates were measured at entry to the DCCT and at the start of EDIC, and several time-varying covariates were measured at the intermittent visits. The most important of these is glycosylated hemoglobin, which measures average blood glucose over 3-4 months preceding the measurement. A high glycosylated hemoglobin value indicates poor control of blood glucose and is associated with diabetic complications. The DCCT Research Group [3, 8, 9], Al-Kateb *et al.* [7], and Cook and Lawless [10] described some analyses of DCCT and EDIC data.

Example 1.2 (The Canadian Longitudinal Study on Aging)

The Canadian Longitudinal Study on Aging (CLSA) is a national longitudinal study of adult development and aging, with an initial stratified random sample of 50,000 persons ages 45-85 years. Recruitment began in 2009, and individuals are to be followed for at least 20 years or to death, with formal assessments scheduled every 3 years. The broad objectives of the CLSA are to foster research into 'understanding how biological, physical, psychological, social, and environmental factors individually, and in combination, influence the health and well-being of aging individuals' [5]. Information on demographic, social, physical/clinical factors and on health service utilization are to be collected on all panel members. In addition, 30,000 members will be asked to provide physical and biological measurements at scheduled visits, but they may refuse to provide the biological specimens (blood and urine). The website for the study is at www.clsa-elcv.ca.

Example 1.3 (Canadian Observational Cohort on HIV)

The Canadian Observational Cohort on HIV (CANOC) is composed of several observational Canadian cohorts of HIV-positive individuals who initiated combination antiretroviral therapy (cART) since January 1, 2000 [11]. Biomarkers that are measured at follow-up visits (approximately every 3 months) for

each individual include viral load, CD4 and CD8 cell counts and other measures such as blood lipid levels. Clinical events that are recorded include AIDS-defining illnesses, death, and other events related to heart disease and cancer. I will restrict discussion here to the British Columbia cohort of CANOC, consisting of 2325 individuals with an average follow-up time of 3.6 years. The broad objectives of CANOC are to study disease processes of HIV-positive individuals and their relationship to risk factors and cART treatment.

These three studies represent a range of objectives and background conditions, but in all cases, information on the study individuals is collected intermittently, at visits ranging from 3 months to 3 years apart. This affects design and analysis in important ways, including the following: (i) decisions concerning which variables and events to record and whether the times of events occurring between successive visits can be (accurately) ascertained; (ii) decisions concerning the frequency and duration of followup; (iii) consideration of ways to avoid bias resulting from selection effects or nonignorable losses to follow-up (LTFs); and (iv) consideration of statistical models for planning, analysis, and prediction that can deal with partially observed life histories. At the planning stage, consideration must also be given to the method of cohort formation and to the level of baseline information to be obtained. In particular, this may include life history information prior to an individual's admission to the study. Some studies are purely retrospective, and the full data on individuals exist at the time they are selected. I will focus here on prospective studies. For convenience, I refer to the study group as a cohort or panel. Inclusion in the study sample may depend on observed covariates or prior life history for an individual, but conditional on such factors, the process data during the study's follow-up period are assumed independent of selection.

My objective is to discuss and comment on these issues. I organized the remainder of the paper as follows. Section 2 describes a statistical framework for life history analyses. Because of their importance and for reasons of brevity, I will focus on multistate models for life history events, along with concomitant variables. Section 3 contains some general discussion of the preceding issues of design and analysis, and Section 4 provides some technical development and illustrations. Section 5 contains some concluding remarks.

2 Statistical Framework

2.1 Models for life history analysis

Several types of variables arise in life history contexts. We can broadly categorize outcomes of interest as follows:

(i) Events that, at least in theory, occur at a specific instant in time, for example, giving birth, getting a job, or being diagnosed with a disease. We can use counting process notation for such events: assuming that an individual i can experience a specified event beginning at a time origin $t = 0$, we let $N_i(t)$ denote the number of events experienced up to time t. The process $\{N_i(t), t \geq 0\}$ is called a counting process [12]. This deals with events that can occur repeatedly or just once and is extended to deal with $R \geq 2$ types of events by letting $N_{ir}(t)$ denote the number of events of type r $(r = 1, 2, \ldots, R)$.

(ii) Categorical variables $Y_i(t)$ that denote the status of an individual at time t; for example, a woman may have given birth to y children $(y = 0, 1, 2, \ldots)$ by age t, or she may have attained any one of a number of educational levels. A multistate framework can be used in such cases, with $Y_i(t)$ allowed to take values in a set $\{1, 2, \ldots, a\}$ of distinct states. This framework is closely connected

to counting processes since a transition from one specified state to another can be considered a type of event.

(iii) Fixed variables such as birth year, sex, genotype. I will use (vectors) x_i or z_i to denote such features for individual i.

(iv) Time-varying variables, denoted by $X_i(t)$. These can be specific to an individual, for example, internal biological variables such as blood pressure, weight, viral loads or blood cell counts, medication, or external factors such as air quality measures.

Data on panel members are typically collected intermittently. In many studies, this occurs at scheduled visits whose frequency and spacing may vary both within and between individuals; I denote the data collection times for a generic individual as $t_0, t_1, \ldots, t_k$. Baseline conditions and covariates are obtained at t_0, and at t_j $(j = 1, \ldots, k)$, information D_j pertaining to the time interval $(t_{j-1}, t_j]$ is obtained. Time-varying covariates are typically measured only at the times t_j, but the occurrence times of events in $(t_{j-1}, t_j]$ can sometimes be retrospectively ascertained at t_j. A design issue discussed later is whether to ascertain such times when doing so is subject to errors of measurement. In some studies, the exact times of certain events such as death may also be ascertainable in some other way. Other forms of incomplete or inaccurate information can also arise, and panel members may be lost to follow-up before the designated end of the study.

I will treat event histories and multistate paths as the processes of interest, with fixed and time-varying variables as explanatory factors. However, events or states are often defined according to a process $X_i(t)$. For example, we could define a 'viral rebound' (VR) event for HIV-positive individuals as occurring when the viral load of an individual moves from a nondetectable level to a specified level such as 10^3 copies per milliliter [13]. Models that accurately represent the determinants and dynamics of processes $\{N_i(t), t \geq 0\}$ or $\{Y_i(t), t \geq 0\}$ are of great interest, but developing such models is difficult, given the complexity of such processes, limitations on the thoroughness and frequency of data collection, and the possibility of bias in cohort selection and follow-up. Nevertheless, stochastic models that capture certain process dynamics can be developed and can help to improve understanding, prediction and decision-making. For the comparison of randomized interventions, on the other hand, marginal process features such as expected numbers of events or average time in a state are used (e.g., Section 8.4 in [14, 15]). Stochastic modelling of life history processes has a long history, perhaps beginning with Halley's models for life tables and mortality (e.g., [16]) but with rapid expansion from the 1950s (e.g., [17–19]). Statistical analysis based on such models has developed rapidly following seminal work on parametric, nonparametric, and semiparametric methods. Andersen *et al.* [12] gave a comprehensive survey up to about 1993, and [20] and [14] survey more recent work.

I consider continuous-time models that begin at some time origin $t = 0$, but discrete-time models are sometimes appealing. The times of events are denoted by $T_1 < T_2 < \cdots$ and the gap times between events by $W_j = T_j - T_{j-1}$ $(j = 1, 2, \ldots)$, where $T_0 = 0$. We can express full models in terms of process intensity functions. For a univariate counting process $\{N_i(t), t \geq 0\}$, these take the form

$$\lambda\left(t|H_i(t)\right) = \lim_{\Delta t \downarrow 0} \frac{\Pr\left\{N_i(t-, t+\Delta t-) = 1 | H_i(t)\right\}}{\Delta t}\,, \tag{1}$$

where $N(s,t) = N(t) - N(s)$ and $H_i(t)$ is individual i's event history up to time $t-$. External time-varying or fixed covariates can be dealt with by incorporating them into $H_i(t)$. Books on point processes (e.g., [21, 22]) and on event history analysis (e.g., [12, 14, 20]) discuss many types of models. The two most familiar are Poisson processes, for which $\lambda(t|H_i(t)) = \rho(t)$ for some nonnegative function $\rho(\cdot)$,

and semi-Markov processes, for which $\lambda(t|H_i(t)), N_i(t-) = j) = h_j(B_i(t))$ for nonnegative functions $h_j(w)$, with $B_i(t) = t - T_{N_i(t-)}$ the time since the most recent event.

2.2 MULTISTATE MODELS

I will focus here on multistate models with states $\{1, 2, \ldots, a\}$ which have transition intensity functions

$$\lambda_{rs}\left(t|H_i(t)\right) = \lim_{\Delta t \downarrow 0} \frac{\Pr\left\{Y_i(t+\Delta t-) = s|H_i(t), Y_i(t-) = r\right\}}{\Delta t}, \quad r \neq s. \tag{2}$$

Markov models have $\lambda_{rs}(t|H_i(t), Y_i(t-) = r) = \alpha_{rs}(t)$; semi-Markov models have $\lambda_{rs}(t|H_i(t), Y_i(t-) = r) = \alpha_{rs}(B_i(t))$, where $B_i(t)$ is the elapsed time since entry to the current state (r).

Under mild conditions and assuming that two or more events cannot occur simultaneously, the intensities provide a full specification of the process in question. A broad discussion of models is beyond my scope here, and I will mainly consider multistate models with intensities of modulated Markov form

$$\lambda_{rs}\left(t|H_i(t), Y_i(t-) = r\right) = \lambda^0_{rs}(t) \exp\left(\beta' Z_i(t)\right), \quad r \neq s\ , \tag{3}$$

where the $\lambda^0_{rs}(t)$ are baseline intensity functions and $Z_i(t)$ is a vector that may include selected aspects of previous life history along with fixed or time-varying covariates. These allow flexible modelling of life history dynamics, and we can fit models with the $\lambda^0_{rs}(t)$ unspecified using Cox model survival analysis software when complete data on life history paths and covariates are available (e.g., [12, 14]). Unfortunately, this is rarely the case for the types of studies discussed here, and challenges arise when trying to model such processes with incomplete data; Sections 3 and 4 discuss this issue.

2.3 AN ILLUSTRATION

Figure 1 shows a model that is used in settings where individuals can experience potentially recurring episodes of some kind. For example, state 1 may represent good health; state 2 a state of illness, disability, or hospitalization; and state 3 death. In some contexts, for example where state 2 represents hospitalization, the exact times of transitions from one state to another are ascertainable even when an individual is formally seen intermittently. In this case, models of the form (3) can readily include features of past episodes (e.g., the number of visits to state 2 or their durations) in the covariate terms. However, if state 2 represents a condition that can be confirmed only through a diagnostic test (e.g., detectable viral load in a person with HIV), then exact transition times cannot be observed in the case of intermittent observation. In this case, the duration of sojourns in state 2 and even the exact number of visits are unknown. This limits the models that it is feasible to fit and affects the precision of estimation and the possibilities for model checking.

3 SOME DESIGN AND ANALYSIS ISSUES

3.1 DEFINITIONS OF STATES AND EVENTS

States and events sometimes have relatively unambiguous definitions (e.g., in or out of hospital and giving birth), but often there is a degree of arbitrariness. For example, in the DCCT studies of Example 1.1, severity of retinopathy is measured on a 22-point ordinal ETDRS scale, which itself is based on photographs of the eye [6]; other categorizations based on photographs also exist. I will consider here five states of retinopathy based on the ETDRS measurements: 1 - ETDRS = 1 (no retinopathy); 2 -

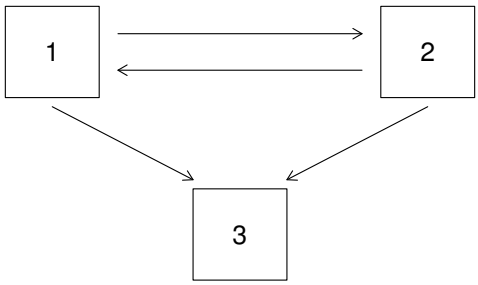

Figure 1: A model for recurrent episodes

ETDRS = 2 or 3; 3 - ETDRS = 4, 5, or 6; 4 - ETDRS = 7, 8, or 9; 5 - ETDRS $\geq$ 10 (severe retinopathy). However, other sets of states could be used. Definitions of the 'progression' of retinopathy also vary; two that have been used are the following: (i) first entry to state 3 and (ii) first occasion on which an individual is observed to be in state 3 or higher at two consecutive observation times. Definition (ii) depends on the observation schedule for an individual but has been used (e.g., [3]) because there is measurement error as well as substantial short-term variability in ETDRS scores [10]. Events defined by a prolonged sojourn in a state are used in many other areas (e.g., [23, 24]). The specification of states based on discrete or continuous measurements, along with the decision to use states rather than the raw measurements, depends on the context and objectives of analysis. States and events should be clinically meaningful and, ideally, subject to reasonably accurate measurement. It is desirable to retain the information on which individuals' states are based in order to facilitate alternative analyses and to allow comparison of analyses based on different sets of states. Section 4.1 considers some models for diabetic retinopathy.

3.2 Measurement Error

Errors of measurement can occur for states, events and covariates. Dealing properly with such errors can be difficult, and the best approach is to minimize them by careful selection and measurement of variables. To consider misclassification of states, for example, one can consider measurement error in the underlying variables on which states are based. We must also make decisions about whether to measure event times precisely. For example, if the time of an event that occurred between two successive observation points can in principle be ascertained, then we must weigh potential gains from doing this against the possibility and effects of measurement error. In longitudinal surveys, the 'seam' effect is well known: persons giving an event time based on recall tend to place it closer to data collection times than it actually is [25]. The use of the state or status at the time of measurement rather than retrospectively ascertained event times is common in many areas (e.g., [26]), and the effect on estimation of time-to-event distributions has been studied, but there has not been much investigation in the event history context. Covariate measurement error is beyond my present scope. Prentice [27] and Prentice and Huang [28] gave valuable discussion and references in the context of women's health studies.

3.3 Follow-up of panel members

The resources available for a study constrain the frequency and length of follow-up. If the visit times $t_1, \ldots, t_k$ for an individual are far enough apart that multiple events or state changes between successive visits are likely, then estimation of state duration distributions and other detailed life history features is

difficult in the case where only the $Y(t_j)$ $(j = 1, 2, \ldots, k)$ are observed. However, there may be adequate information about aspects such as the incidence or prevalence of specific conditions. For example, the CLSA in Example 1.2 collects information every three years over a long period for persons age 45 years and older. Useful information on features such as the onset and progression of cognitive impairment (CI) as a function of age and other factors can be obtained, but detailed information on the duration of hospitalization spells or the use of community health services may not be ascertainable from the basic data collection. The incorporation of auxiliary information from administrative records or from more intensive follow-up of a selective subsample of the panel is an important area that is receiving increasing attention but, so far, limited formal development.

Premature LTFs naturally decrease the information available, but a potentially greater concern is the possibility of bias when LTF is related to an individual's life history. If LTF is independent of future life history (after LTF), given previously observed life history, then we can avoid bias through appropriate conditioning on prior history. However, when successive visits are far apart, it is often likely that the probability of becoming LTF at time t_j depends on outcomes over the interval $(t_{j-1}, t_j]$. In studies where heavy LTF is expected, there may be a strong argument for a budget that allows some tracing of persons who are LTF (e.g., [29]). Section 4.3 provides a few details on these issues.

In observational studies, there may be a relationship between the visit times for an individual and their life history since the last visit, as when healthier individuals postpone or forego visits. Models to address this have been considered (e.g., [30, 31]), but they are often difficult to check or rely on uncheckable assumptions. In addition, labeling a person as LTF, and specification of an LTF date, is problematic when an individual's last visit was a long time before the study's administrative end date. It is important for reliable analysis to try for adherence to a visit schedule. This can allow the scheduled time for the next visit to depend on data collected up to the current visit (Section 4.3).

3.4 Panel selection and initial conditions

The collection of baseline information on individuals is crucial, regardless of whether the panel is a random sample from some population or a group defined by certain characteristics. One reason is that dynamic modelling and analysis uses previous life history (Section 2.2). Specification of models for outcomes over a follow-up period from t_0 to t_k typically relies on relevant history $H(t_0)$ before time t_0. In economics, these are referred to as initial conditions. Failure to collect relevant information may contribute to misleading conclusions or make it necessary to rely on uncheckable assumptions. Interesting examples arise when the current duration of a condition is not recorded or modeled in an analysis. Prentice *et al.* [2] and Prentice [27] discussed discrepancies in observational and randomized studies on the effect of post-menopausal hormone therapy (HT) on coronary heart disease in women. These were largely due to the fact that the effects were time dependent and that women in the randomized studies were followed from the initiation of HT, whereas many in the observational studies were not. If insufficient attention is given to the duration of prior HT usage, misleading inferences might be drawn from the observational studies. Glymour [32] discussed similar issues in longitudinal studies on aging and cognitive impairment. Section 4.4 gives some additional discussion of initial conditions.

In studies where panel members are randomly selected, problems can still arise because of refusals to participate. In the CLSA, it was estimated that 152,000 people would have to be approached to obtain a panel of 50,000 [5]. A concern is that those agreeing to join differ in significant ways from those who refuse. This can sometimes be addressed by the collection of relevant baseline information on both joiners and refusers. In an empirical study on employment histories, Pyy-Martikainen and Rendtel [33] demonstrated the biases that initial refusals and dependent LTF can produce by comparing analyses of

European longitudinal survey data with complete administrative data on the panel members and refusers. They found significant biases in survey-based estimates related to unemployment. There is a large literature on refusals and non-response in certain areas, including survey sampling (e.g., [34]) and case-control studies (e.g., [35]).

4 Some Technical Issues

4.1 Models Fitting and Analysis

Suppose that individual i in a study is observed at times $t_{i0} < t_{i1} < \cdots < t_{ik_i}$. At time t_{ij} information $D_i(t_{ij})$ on events and $X_i(t_{ij})$ on external covariates over the time interval $(t_{i,j-1}, t_{ij}]$ is obtained $(j = 1, 2, \ldots, k_i)$; baseline information is given by $D_i(t_{i0})$ and $X_i(t_{i0})$. Under conditional independence assumptions concerning the t_{ij} and LTF (Section 4.3), the probability distribution of $\{D_i(t_{i1}), \ldots, D_i(t_{ik_i})\}$ given the t_{ij}, $D_i(t_{i0})$ and external covariates is proportional to

$$\prod_{j=1}^{k_i} \Pr\left\{D_i(t_{ij}) | \bar{D}_i(t_{i,j-1}), \bar{X}_i(t_{ij})\right\}. \tag{4}$$

In (4) and henceforth, 'Pr' denotes a probability density or mass function, and $\bar{D}_i(t_{ij}) = \{D_i(t_{i0}), \ldots, D_i(t_{ij})\}$ and $\bar{X}_i(t_{ij}) = \{X_i(t_{i0}), \ldots, X_i(t_{ij})\}$ denote observed event and covariate histories. The times t_{ij} are treated as fixed in (4), although they are allowed to depend on previous observations (Section 2.2 in [10]). Detailed modelling of transition intensities as functions of previous event history and time-varying covariates is feasible only if sufficiently detailed information is collected. On the other hand, a common situation is where $D_i(t_{ij})$ consists only of event counts over $(t_{i,j-1}, t_{ij}]$ or, in the case of multistate models, only of the state $Y_i(t_{ij})$ occupied at t_{ij}. In the latter case, (4) becomes

$$\prod_{j=1}^{k_i} \Pr\left\{Y_i(t_{i,j-1}) | \bar{Y}_i(t_{i,j-1}), \bar{X}_i(t_{ij})\right\}. \tag{5}$$

The timescale in (4) could be calendar time or, more commonly, an individual-specific scale such as age or time in study. The key challenge is to specify models, such as (3), that represent the life history process adequately but allow computation of (4), so as to serve as a basis for estimation. Cook and Lawless (Section 3 in [10]) discussed models in some detail, and I summarize a few salient points needed for later developments. Markov models dominate statistical practice, in part due to their tractability. When all covariates are fixed, so that transition intensities $\lambda_{rs}(t|H(t), x)$ in (2) are of the form $q_{rs}(t; x)$, the terms in (5) are transition probabilities

$$P_{rs}\left(t_{i,j-1}, t_{ij}; x\right) = \Pr\left\{Y_i(t_{ij}) = s | Y_i(t_{i,j-1}) = r, x\right\}. \tag{6}$$

For time-homogeneous models for which $q_{rs}(t; x) = q_{rs}(x)$, the $a \times a$ transition probability matrix $P(u, u+t; x) = (P_{rs}(u, u+t; x))$ is given by the matrix exponential function [36],

$$P\left(u, u+t; x\right) = P(t; x) = \exp\left\{tQ(x)\right\}, \tag{7}$$

where $Q(x)$ is the $a \times a$ matrix with entries $q_{rs}(x)$ for $r \neq s$ and $q_{rr}(x) = -\sum_{s \neq r} q_{rs}(x)$. Nonhomogeneous models are harder to handle, but refer to Section 3.2 in [10] and [37] for computational methods. The msm package in R [38] provides convenient software for fitting time-homogeneous models and will

also handle models for which the intensities $q_{rs}(t;x)$ are piecewise constant. It should be noted that, except for the case of time-homogeneous models, the specification of the time origin $(t = 0)$ is of critical importance in Markov modeling.

Models with time-varying covariates observed only at visit times require assumptions for tractability. If inter-observation times are reasonably similar, we often assume $X_i(t) = X_i(t_{i,j-1})$ over the time interval $(t_{i,j-1}, t_{ij})$. This may not reflect the full effect of $X_i(t)$ on transition probabilities, but it is a good way to specify models for which history up to time $t_{i,j-1}$ is used to predict outcomes at t_{ij}. When times between visits vary substantially, a preferable approach is to model the covariate process jointly with the $Y_i(t)$; refer to Sections 3.6 and 5 in [10] and [39] for discussion and illustrations.

Sojourn times in states or waiting times until entry to a given state are often of interest, for example the length of a spell in which an individual is disabled. We can obtain distributions for such variables from Markov models. However, models such as $q_{rs}(t;x) = q^0_{rs}(t)e^{\beta' x}$, where covariates have a simple effect on transition intensities, do not translate into simple covariate effects for sojourns or waiting times. An alternative is to use models for which sojourn times play a central role, as with semi-Markov models. However, such models are hard to fit when observation is intermittent except in simple cases, even when there are no covariates (e.g., [37]). If a specific distribution is of interest, it is often preferable to model it directly in terms of covariates. Andersen *et al.* [40] considered linking this to a multistate model.

In some situations, the observed life histories of individuals are more heterogeneous than can be accounted for by observed covariates within a specific process. For discussion of unobserved random-effects and mixture models that can address this, refer to Section 3 in [10]. Goodness of fit for Markov models is usually checked by comparing observed and expected (model-based) transitions among states. The R package msm provides a number of options for doing this. However, the frequency of observation and incompleteness of the data often constrain model checking, and comprehensive validation of a model is often impossible. Titman and Sharples [41] gave an excellent discussion of model checking, and Section 4.2 makes some additional remarks.

Example 4.1 (Progression of diabetic retinopathy)

As an illustration of Markov multistate modelling and the choice of states, I consider some analyses of diabetic retinopathy discussed by Cook and Lawless (Section 5.2 in [10]) and based on the DCCT trial introduced in Example 1.1. I focus on the conventional therapy group within the primary prevention cohort and on the white subjects, all of whom had no retinopathy (ETDRS = 1) and had diabetes durations of 5 years or less at enrolment. Two models, each with five states as defined in Section 3.1, are considered, with t representing time since enrolment. Figure 2 shows the state diagrams, and the models are denoted as

M1: all transitions to adjacent states are allowed, giving the set of transition intensities $\{q_{12}(t), q_{21}(t), q_{23}(t), q_{32}(t), q_{34}(t), q_{43}(t), q_{45}(t), q_{54}(t)\}$,

M2: only transitions to adjacent higher states are allowed, giving the set of transition intensities $\{q_{12}(t), q_{23}(t), q_{34}(t), q_{45}(t)\}$.

Model M2 is, as stated, inconsistent with the observed data as many individuals experience downward transitions over successive observation times, which are approximately 6 months apart; Table 1 gives the total transitions of each type for model M1, across all subjects and pairs of successive observation times. For M2, we therefore use an operational definition of state occupancy; an individual is considered to never revisit a state that they have left. For example, if the actual observed states over times $t_0 = 0$, t_1, t_2,

t_3 for an individual were 1, 2, 1, and 3, then for model M2, the sequence of states would be amended to 1, 2, 2, and 3. This has the undesirable features of tying the definition of states to the observation times and being inconsistent with the observed data on state occupancy. However, it can be considered a reasonable way to model progression of retinopathy in the DCCT and is one of the ways this was carried out in the DCCT Research Group [3]. A second approach in that paper assumed that progression to a higher state occurred only if it was sustained over two successive observation times. Table 2 shows the total numbers of transitions for this 'sustained progression' model M3.

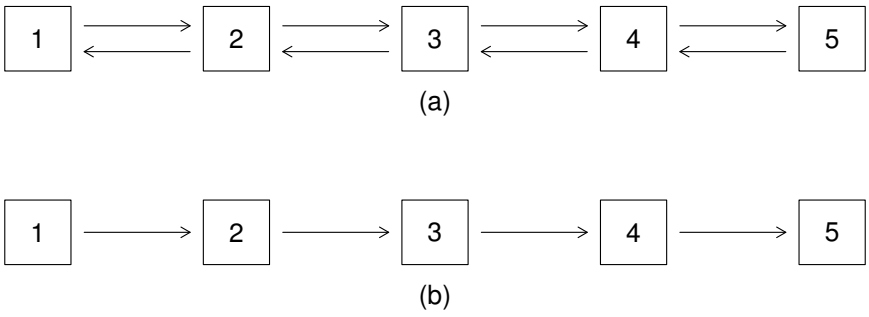

Figure 2: Two models for diabetic retinopathy: (a) M1, (b) M2

Table 1: Observed transition counts for five-state model M1 of retinopathy, conventional treatment

		To state				
		1	2	3	4	5
From state	1	1764	464	52	0	0
	2	260	743	169	1	2
	3	16	108	199	11	3
	4	0	0	6	6	0
	5	1	0	1	1	3

Section 5.2 in [10] presents detailed analysis of these models, with transition intensities taken to be piecewise constant. Several clinical papers have used entry to state 3 in models M2 and M3 as 'progression of retinopathy' outcomes and based treatment comparisons on them. M2 and M3 estimate 5-year progression (i.e. the probability of entry to state 3 by 5 years) as about 0.34 (M2) and 0.28 (M3) for the treatment group considered here. Model M1 estimates the probability of first entry to state 3 by 5 years as 0.50, however. The discrepancy with models M2 and M3 is due to the fact that transitions into state 3 (and other states) under M2 or M3 include only a portion of the actual transitions. For example, an individual who is observed in state 2 on two consecutive 6-monthly visits may have moved to state 3 and then back to state 2 between visits. The higher estimates of progression to state 3 under M1 are because it allows this possibility, whereas M2 and M3 do not. This discrepancy would become small if observation times were close together. Treatment effects based on M1-M3 are a bit less discrepant. For example, odds ratios (experimental over conventional treatment) for 5-year progression in this primary prevention cohort are 0.51, 0.41, 0.47 for models M1, M2, and M3, respectively.

Table 2: Observed transition counts for five-state 'sustained' progressive model of retinopathy, conventional treatment

		To state				
		1	2	3	4	5
	1	1790	263	26	0	0
	2	0	1153	96	0	1
From state	3	0	0	447	7	2
	4	0	0	0	5	0
	5	0	0	0	0	20

These models reflect the distinction between modeling the dynamics of a disease process and making simple treatment comparisons. Model M1 represents the actual observed longitudinal retinopathy measures well (although it can be improved through the inclusion of covariates and biomarkers) but is less satisfying for comparing the two treatment groups. In clinical studies, simpler progressive models like M2 and M3 or just models for the time to some event are usually used for comparisons, even though they may contradict certain features of the observed data. This, however, may inject an extra degree of measurement error into the treatment comparison.

4.2 Effects of Intermittent Observation in Markov Models

4.2.1 Precision of estimation

The lengths of time between visits affects the precision of parameter estimates and the possibilities for model assessment. Kalbfleisch and Lawless [36] provided some general discussion, and Hwang and Brookmeyer [42] presented a limited numerical study of a progressive three-state, time-homogeneous Markov model. Aside from this, rather little seems available, and in particular, results for bi-directional multistate models. I consider here some simple calculations for time-homogeneous Markov models; a more detailed study will appear in [43].

We can gain insight from two-state models M1 ($q_{12} > 0, q_{21} > 0$) and M2 ($q_{12} > 0, q_{21} = 0$). Model M1 is bi-directional and M2 corresponds to an exponential survival distribution for the duration of a spell in state 1. For model M1, it can be found that (Section 4 in [36])

$$P(s, s+t) = P(t) = \begin{pmatrix} 1-\pi\,(1-e^{-\alpha t}) & \pi\,(1-e^{-\alpha t}) \\ (1-\pi)\,(1-e^{-\alpha t}) & \pi + (1-\pi)e^{-\alpha t} \end{pmatrix}, \tag{8}$$

where $\alpha = q_{12} + q_{21}$ and $\pi = q_{12}\alpha^{-1}$. As t increases, both rows of (8) approach the limiting distribution $(1-\pi, \pi)$ and it follows that if inter-visit times Δt are sufficiently large, we may be able to estimate π precisely but not q_{12} or q_{21}. Table 3(a) shows ratios of the asymptotic standard deviations for panel observation with designated Δt divided by those for continuous observation ($\Delta t = 0$), for $\hat{q}_{12}$, $\hat{q}_{21}$, $\hat{P}_{11}(1)$ and $\hat{P}_{11}(4)$, for the case where $q_{12} = q_{21} = 1$. Individuals are seen at times $0, \Delta t, 2\Delta t, \ldots, 4$, with all individuals in state 1 at $t = 0$. The asymptotic standard deviations are based on Fisher information, whose calculation for panel data is described in Section 3 in [36].

Table 3 shows that estimation of q_{12} and q_{21} in M1 is very imprecise as Δt becomes larger than about the average sojourn time in states 1 or 2 ($q_{12}^{-1} = q_{21}^{-1} = 1$). However, estimation of transition probabilities

Table 3: Ratios of asymptotic standard deviations for estimators of transition intensities and probabilities in two-state Markov models with (a) $q_{12} = q_{21} = 1$ (M1) (b) $q_{12} = 1, q_{21} = 0$ (M2)

	Δt	q_{12}^a	q_{21}	$P_{11}(1)$	$P_{11}(4)$
(a) M1	0.5	2.07	1.30	1.64	1.56
	1	3.73	4.07	2.00	1.63
	2	18.6	18.7	7.64	1.94
(b) M2	0.5	1.01	-	1.01	1.01
	1	1.04	-	1.04	1.04
	2	1.86	-	1.86	1.86

Values of $P_{11}(t)$ in (a) are $P_{11}(1) = 0.568$, $P_{11}(4) = 0.500$ and in (b) $P_{11}(1) = 0.368$, $P_{11}(4) = 0.018$.
[a] Values for each parameter θ are $SD(\hat{\theta})$ based on panel data with the given Δt divided by $SD(\hat{\theta})$ based on continuous observation.

$P_{11}(t)$ (which for large t are almost equal to $1 - \pi$, as is $P_{21}(t)$) is much less affected. A caveat is that the estimates depend on the validity of the assumed Markov model. Table 3(b) shows ratios of asymptotic standard deviations for model M2; estimation of q_{12} is much less affected by larger Δt.

Bi-directional models are important for describing (repeatable) transient spells that may occur in life history processes, for example, periods of unemployment, disability or hospitalization. Our ability to estimate transition intensities and other features such as sojourn time distributions is limited in studies with widely spaced follow-up times, unless (partial or full) information on transition times can be retrospectively determined at each visit. However, as for progressive models, we can estimate the marginal probabilities of being in specific states reasonably well; this is valuable for predicting population-level outcomes and associated costs. Qualitatively similar results apply to more complex models [43].

4.2.2 Robustness and model assessment

When transition times are fully observed or right-censored, Markov nonparametric (Aalen-Johansen) estimates also provide consistent estimation of state occupancy probabilities under non-Markov conditions [44, 45], provided LTF is independent of the multistate process. Datta and Satten [46] extended this to cover state-dependent LTF through the use of inverse probability of censoring weights (Section 4.3). These results do not hold when state occupancy is observed only intermittently, but a similar result holds for discrete-time Markov chains, which can be employed when individuals are all (potentially) seen at a common set of visit times. I will sketch a derivation and then discuss its implications for the application of Markov models.

Assume that no covariates are under consideration, and suppose that a cohort of m individuals have assigned observation times $t = 0, 1, \ldots, T$; the derivation is easily modified if observations are at arbitrary times $t_0 < t_1 < \ldots, t_T$. Let $p_r(t) = \Pr\{Y(t) = r\}$ and $p_{rs}(t, t+1) = \Pr\{Y(t+1) = s | Y(t) = r\}$ denote prevalence probabilities and one-step transition probabilities, respectively, where $t = 0, 1, \ldots, T-1$ and r, s range over states $1, 2, \ldots, a$. Let $R_i(t)$ equal 1 if individual i is seen at time t and 0 otherwise, and assume that $\{R_i(t), t = 0, 1, \ldots, T\}$ is independent of $\{Y_i(t), t = 0, 1, \ldots, T\}$. Define

$$n_{rs}(t,t+1) = \sum_{i=1}^{m} R_i(t)R_i(t+1)I\left(Y_i(t)=r, Y_i(t+1)=s\right) \tag{9}$$

and $n_{r+}(t,t+1) = \sum_{s=1}^{a} n_{rs}(t,t+1)$. A nonparametric estimate of $p_{rs}(t,t+1)$ is then given by

$$\hat{p}_{rs}(t,t+1) = \frac{n_{rs}(t,t+1)}{n_{r+}(t,t+1)} \quad r,s = 1,\ldots,a \tag{10}$$

and it is easily seen that this estimate is consistent regardless of whether the process $\{Y(t), t = 0,1,\ldots,T\}$ is Markov or not. Further, robust estimates of the $p_r(t)$ are obtained from the relationship

$$p_s(t+1) = \sum_{r=1}^{a} p_r(t)p_{rs}(t,t+1) \tag{11}$$

by starting with the robust estimates

$$\hat{p}_r(0) = \sum_{i=1}^{m} R_i(0)I\left(Y_i(0)=r\right) \Big/ \sum_{i=1}^{m} R_i(0) \tag{12}$$

and applying (11) recursively. For example, if subjects in the DCCT in Example 4.1 were seen exactly every half-year, this approach could be used to estimate the probability an individual in a given treatment group is in each specific state at times 0.5, 1.0, 1.5 years, and so on after randomization.

We can also obtain consistent estimates if individuals are followed from times that are independent of their life history and if LTF depends only on previously observed life history. This requires modeling of LTF, and is sketched in the Appendix. Markov models can thus provide robust estimates of expected state occupancy for individuals or populations, provided care is taken to allow flexible time-independent transition intensities. Although the approach outlined here applies when individuals are seen at common times, it is plausible that flexible continuous-time models based on more general intermittent observations would also give fairly robust estimates. This has implications as well for the assessment of such Markov models as model-based state occupancy probabilities will tend to agree with empirical (nonparametric) estimates, as will transition probabilities for short time intervals $(t, t+\delta)$. Thus, although we can assess the need for time-dependent Markov transition intensities, for example through likelihood ratio tests for nested models, assessment of the Markov assumption itself is more difficult. Evidence against Markov models is best assessed by comparison with alternative models. However, comparison with non-Markov models is more difficult because they are harder to fit. For example, refer to [47] and [48] concerning semi-Markov models and intermittent observation. When covariates are present, we can likewise compare different Markov models, but examination of non-Markov models is difficult.

Example 4.2 (Prediction of viral rebounds)

For HIV-positive individuals who have achieved viral suppression (reduction of the virus to non-detectable levels) through cART (see Example 1.3), the time to a VR, and factors related to it, is of interest. One approach is to use a survival model with suitably chosen covariates. In [13], a Cox model is used, and it is noted that the occurrence of viral 'blips' is associated with shorter times to VR; a blip is defined as the occurrence of a detectable viral load at a visit, preceded and followed by non-detectable viral load at adjacent visits. One difficulty with this approach is in assessing the effect of variable times between visits. An alternative approach is to use a multistate model in which VR is represented as an absorbing state.

In [49], a model is considered in which state 1 represents non-detectable viral load, states $2, ..., a-1$ represent detectable viral load ranges, and state a represents a VR state (e.g., viral load over 1000 copies per milliliter). The multistate model allows more detailed modelling of the viral load process, and Markov models readily deal with variable times between viral load measurements. Modulated Markov models in which transition intensities depend on previous observed history, such as the occurrence of temporary increases in viral load, can also be handled. Such models provide estimates of $P_{1a}(t)$, the probability of a VR by time t.

4.3 Losses to Follow-up

Letting $R_i(t_{ij})$ indicate whether individual i is seen at scheduled visit time t_{ij}, we assume in this section that an individual is LTF at the first time t_{ij} for which $R_i(t_{ij}) = 0$. Intermittent missingness, where an individual is absent at one visit but present at a later one, is harder to handle and is discussed briefly in Section 5. If $R_i(t_{ij})$ is conditionally independent of observed life and covariate history $\bar{D}_i(t)$ for $t > t_{i,j-1}$ given $\bar{D}_i(t_{i,j-1})$, then we can use the (partial) likelihoods (4) and (5) for estimation, with $k_i = \max_j(R_i(t_{ij}) = 1)$. If this 'sequential missing at random' (SMAR) condition [50] does not hold but there exists a vector $x_i^c(t_{i,j-1})$ of observed variables such that $R_i(t_{ij})$ is conditionally independent of $\{\bar{D}_i(t), t > t_{i,j-1}\}$ given $x_i^c(t_{i,j-1})$ and $\bar{D}_i(t_{i,j-1})$, then we can use inverse probability of censoring weights to adjust the log likelihoods or estimating functions on the basis of (4) or (5) [1, 51]. The main application of inverse probability of censoring weights is in fitting models with limited individual-level covariates or process history contained in $D_i(t_{ij})$; such models are typically used when interest lies in simple treatment comparisons or population level inferences (e.g., [1, 15]).

When observation times t_{ij} are widely spaced, the SMAR assumption is often violated, with the probability individual i is LTF at t_{ij} depending on life or covariate history over $(t_{i,j-1}, t_{ij}]$. Several authors have proposed not SMAR (NSMAR) models where $R_i(t_{ij})$ depends on both $H_i(t_{i,j-1})$ and $Y_i(t_{ij})$ (e.g., [30, 31, 50]). They contain assumptions that cannot be checked with the observed data but in some cases may be useful for sensitivity analysis. Multistate models that include an LTF state can be used for this purpose, but first, let us consider their use for estimation with intermittent observation.

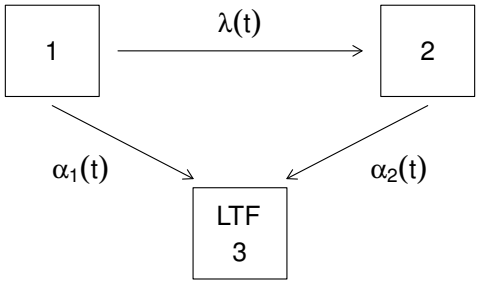

Figure 3: A model incorporating loss to follow-up (LTF)

Consider the model in Figure 3, where the transition intensity $\lambda(t)$ from state 1 to 2 is of interest. If the transition intensities $\alpha_1(t)$ and $\alpha_2(t)$ into the LTF state differ, then the SMAR condition is violated under intermittent observation. By modeling LTF, we can, however, adjust for this in the estimation of $\lambda(t)$. Note that in this model $\lambda(t)$ is the transition intensity at time t given that $R_i(t) = 1$ and that we cannot estimate the intensity for an individual who is LTF ($R_i(t) = 0$). For illustration, suppose that all individuals begin in state 1 (S_1) and that if a transition into state 2 (S_2) occurs in (t_{j-1}, t_j), then if

$R_i(t_j) = 1$, we can ascertain the transition time t; the exact times of transitions into state 3 (S_3) are not ascertainable. There are then four types of observation for an individual who is in (S_1) at t_0 and is observed at times $0 = t_0 < t_1 < \cdots < t_k$, where t_k is the LTF time: they are as follows: (i) still in S_1 at t_k; (ii) transition to S_2 at time t in $(t_{j-1}, t_j]$ and still in S_2 at t_k; (iii) transition from S_1 to S_3 in $(t_{k-1}, t_k]$; (iv) transition to S_2 at time t in $(t_{j-1}, t_j]$ and then transition to S_3 in $(t_{k-1}, t_k]$. The likelihood contributions from the four outcomes are (i) $P_{11}(0, t_k)$; (ii) $P_{11}(0, t-)\lambda(t)P_{22}(t, t_k)$; (iii) $P_{11}(0, t_{k-1})P_{13}(t_{k-1}, t_k)$; (iv) $P_{11}(0, t-)\lambda(t)P_{22}(t, t_{k-1})P_{23}(t_{k-1}, t_k)$. We can estimate all the intensities $\lambda(t)$, $\alpha_1(t)$, and $\alpha_2(t)$ for this model. Because exact entry times to S_3 are not known, it is best to use (flexible) parametric models for $\alpha_1(t)$ and $\alpha_2(t)$. We should note that what makes this approach work is that we follow individuals after they enter S_2.

We can also use this model if exact entry times to S_2 are not ascertainable. Moreover, we can use it to assess the effect of naively considering LTF to be ignorable (SMAR). If a transition into S_2 is observed (at time t_j) to have occurred at time t in $(t_{j-1}, t_j]$, then we do not estimate $\lambda(t)$, but

$$\lambda^*(t) = \lambda(t) \left\{ \frac{P_{22}(t, t_j)}{P_{11}(t, t_j) + P_{12}(t, t_j)} \right\} \qquad t_{j-1} < t \leq t_j. \tag{13}$$

When $\alpha_1(t) = \alpha_2(t)$, we can see that $\lambda^*(t) = \lambda(t)$, but otherwise, estimation of $\lambda(t)$ is biased. For example, if $\lambda(t) = \lambda$, $\alpha_1(t) = \alpha_1$ and $\alpha_2(t) = \alpha_2$, then

$$\lambda^*(t) = \frac{\lambda\left(\lambda + \alpha_1 - \alpha_2\right)}{\lambda + (\alpha_1 - \alpha_2)\exp\left\{-\left(\lambda + \alpha_1 - \alpha_2\right)(t_j - t)\right\}}. \tag{14}$$

The bias is positive $(\lambda^*(t) > \lambda)$ if $\alpha_1 > \alpha_2$, negative if $\alpha_1 < \alpha_2$, and is maximal when $t = t_{j-1}$.

The preceding model does not consider the transition intensity from S_1 to S_2 in the absence of LTF, except under the SMAR assumption. Authors such as Barrett *et al.* [4] considered NSMAR models in which sensitivity analysis can be undertaken. In the context here, for example, we could consider the model in Figure 4. We cannot estimate all the four intensities from intermittent observations, and we do not know if someone entering S_2 did so from S_1 or S_3, and in some cases, whether a person is in S_3 or S_4. Note that state 2 and $\lambda(t)$ in Figure 4 are the same as in Figure 3 but that $\lambda'(t) \neq \lambda(t)$ in general. If we assume that $\lambda'(t) = r\lambda(t)$ with r known, then we can estimate $\lambda(t)$ and examine sensitivity to the value of r. This, however, makes estimation of $\lambda(t)$ dependent on r, unlike the model in Figure 3. The main advantage, and importance, of the model in Figure 4 is in situations where some individuals who are LTF can be traced, so we can determine whether they entered S_3 or S_4 and the time of entry to S_2 if that occurred. This allows us to assess whether $\lambda'(t)$ and $\lambda(t)$ differ.

4.4 Initial Conditions and Heterogeneity

In many studies, some of the panel members have begun a life history process of interest prior to enrolment and the start of follow-up. For example, in the CLSA (Example 1.2), one process of interest is cognitive decline and impairment, and a person may already have some degree of impairment at enrolment. Suppose for discussion that a process begins at age a_0 for a specific individual and that they are enrolled for follow-up at age $t_0 > a_0$. The study provides information $\{\bar{D}(t), t > t_0\}$ on the process history $\{H(t), t > t_0\}$, along with baseline process and covariate information $D(t_0)$. A prospective likelihood function for inference purposes in the case of intermittent followup is given by (4). It has been noted previously that calculation of (5) can be challenging for many models if crucial information in $H(t_0)$ or $\{H(t), t > t_0\}$ is missing from $D(t_0)$ and $\{\bar{D}(t), t > t_0\}$. In this case, model assumptions that are not fully verifiable may be needed to make progress.

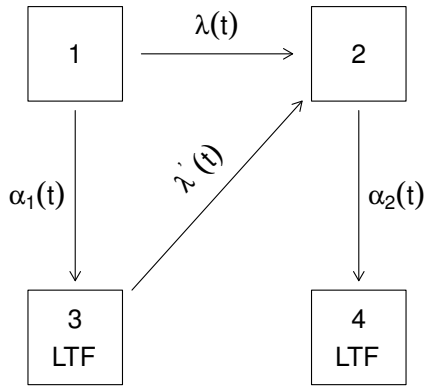

Figure 4: A model with two LTF states

A much-studied example concerns the distribution of time W from the onset of some condition (e.g., HIV infection) to a subsequent event (e.g., an AIDS-defining illness); see for example [52]. This may be described (ignoring mortality for convenience) as a three-state progressive model with states 1, 2, 3 representing absence of the condition, presence of the condition only, and presence of the condition plus the subsequent event; W is the sojourn duration in state 2. Suppose the conditional distribution of W given the time a_0 of entry to state 2 is of interest and that enrolment in a study is at time $t_0 > a_0$ for a certain individual. If a_0 is known, we can fit models with density $f(w|a_0;\theta)$ via the left-truncated likelihood based on $f(w|a_0;\theta)/S(t_0 - a_0|a_0;\theta)$, where $S(w|a_0) = \Pr(W \geq w|a_0)$. However, if a_0 is not part of the initial information $D(t_0)$, we would instead have to base a likelihood on

$$\Pr \text{ (entry to state 3 at time } a_0 + w|Y(t_0) = 2\ ) = \frac{\int_0^{t_0} f_1(a_0) f(w|a_0;\theta) da_0}{\int_0^{t_0} f_1(a_0) S(t_0 - a_0|a_0;\theta) da_0}\ ,$$

where $f_1(a)$ is the density for the time of entry to state 2. If the study includes individuals who are in state 1 at enrolment, this can support modelling of $f_1(a_0)$; we need otherwise external information.

Additional complications arise if random effects are considered (e.g., [10, 53]). In particular, if u_i is a random effect associated with the ith individual, then u_i would in general be related to $D_i(t_{i0})$, and so the distribution of u_i at t_{i0} should be conditional on $D_i(t_{i0})$ [54]. Failure to allow for this can lead to estimation biases; the following example provides an illustration.

Example 4.3 (Cognitive impairment)

The development and progression of CI as people age has received much attention (e.g., [4, 32, 55, 56]). It is apparent that there is considerable heterogeneity in both the age of onset and the progression of CI, only some of which is currently explainable by known risk factors. Consequently, we need to be careful care in the modeling and analysis of data on CI. For illustration, consider the rather simple model in Figure 5, in which state 1 represents no CI, state 2 represents mild CI, and state 3 represents severe CI. An individual is often assigned a state at any given time on the basis of their response to a test, for example, the Mini Mental State Examination. It is possible for a person's score to improve over two consecutive observations, but for simplicity I consider here a progressive model.

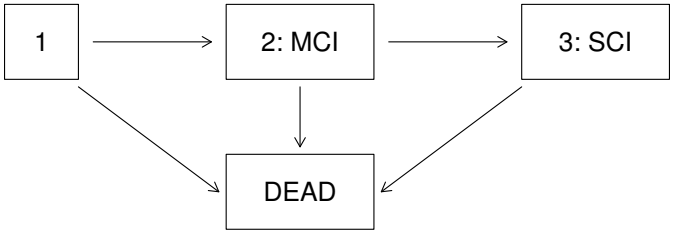

Figure 5: A simple model of cognitive impairment and death. MCI, mild cognitive impairment; SCI, severe cognitive impairment

For the sake of discussion, consider a conditional Markov model in which t represents age, x represents known risk factors, and u represents unobservable random effects; the transition intensities I shall focus on are from states 1 to 2 and 2 to 3. Multiplicative models might specify these as

$$\lambda_{12}(t|u,x) = u_{12}\lambda_{12}^0(t)e^{\beta_{12}x} \quad \text{and} \quad \lambda_{23}(t|u,x) = u_{23}\lambda_{23}^0(t)e^{\beta_{23}x}.$$

Suppose that an individual in a cohort study such as the CLSA (Example 1.2) is observed from age a_0, at which time they are in state 2. In addition, suppose that the (approximate) age t_1 at which they entered state 2 can be ascertained. With t_2 defined as the minimum of the age of entry to state 3 and the age of the individual at death or the last follow-up time, the prospective likelihood function is based on the observed data $\{Y(t), t > a_0\}$, and initial conditions (baseline information) are that $Y(a_0) = 2$ and that entry to state 2 was at age $t_1 < a_0$. This gives the likelihood

$$L = \int_0^\infty \Pr\left(t_2, \delta_2 | u, x, t_1, Y(a_0) = 2\right) g\left(u | t_1, t_2 > a_0\right) du, \tag{15}$$

where δ_2 indicates whether t_2 is the observed age of entry to state 3 ($\delta_2 = 1$) or the age at death or end of follow-up ($\delta_2 = 0$) and $g(u|t_1, t_2 > a_0)$ is the density of $u = (u_{12}, u_{23})$ given the initial conditions.

The likelihood (15) avoids left-truncation bias, assuming the adequacy of the models for u and for the process. Within this framework, it is necessary to model the effect of u on both $\lambda_{12}(t|u,x)$ and $\lambda_{23}(t|u,x)$. If a naive analysis were carried out in which the fact that $Y(a_0) = 2$ was included but the corresponding selection effect on u was not, then bias would be incurred in the estimation of baseline transition rates and the effects of risk factors. For example, consider the simple model where there are no covariates x, $\lambda_{12}^0(t) = \lambda_{12}$, $\lambda_{23}^0(t) = \lambda_{23}$, $u_{12} = u_{23} = u$, and u has a gamma distribution in the population with mean = 1 and variance $v = \phi^{-1}$. Then, assuming that u does not affect the death intensities in Figure 5, the expected value of u given that $t_2 > a_0$ and $t_1 (< a_0)$ is

$$E\left(u | t_2 > a_0, t_1\right) = \frac{1+\phi}{\lambda_{12}t_1 + \lambda_{23}(a_0 - t_1) + \phi}. \tag{16}$$

When ϕ is very large (that is, there is little variability in u), (16) is close to the population average u-value 1. However, if u varies substantially in the population, then persons with larger value of t_1 or larger values of $a_0 - t_1$ are associated with smaller values of u. Conversely, for a given value of $a_0 - t_1$ or t_1, persons with larger values of u are underrepresented.

Random-effects modeling gives 'observable' intensity functions $\lambda_{23}(t|H(t), x)$ that depend on the age t_1 of entry to state 2. We can also consider such models without resorting to random effects. For

example, the preceding model has transition intensity

$$\lambda_{23}(t|H(t), Y(t) = 2) = \frac{(1+\phi)\lambda_{23}}{\lambda_{12}t_1 + \lambda_{23}(t - t_1) + \phi}.$$

The crucial point is that delayed observation of an individual's sojourn in state 2 (i.e., entry to state 2 occurred prior to a_0) necessitates modeling the sojourn in state 1, as well as mortality. In studies such as the CLSA, many individuals are in state 1 upon entry to the study, and there are data to support such modeling. In studies where this is not the case, we must rely on external information in order to model $\lambda_{12}(t|H(t))$ and mortality from state 1. Finally, studies in which individuals are randomized at some point to alternative treatment interventions give balance across treatment groups with respect to initial conditions, thus facilitating the use of simple outcomes for treatment comparisons. However, a detailed understanding of treatment effects may require joint consideration of initial conditions and treatment.

5 Concluding Remarks

I have emphasized multistage models here, but the issues considered arise for other models and for most studies of life history processes. In particular, data collected at intermittent times typically provide partial information on processes of interest, and this constrains the models and questions that can realistically be examined. Fairly accurate occurrence times for events or transitions are in principle feasible for outcomes that are readily observed and recorded (for example, disability spells, hospitalization episodes or diagnosis of disease), but outcomes related to biological variables, tests or questionnaires are usually observable only when an individual is seen and so process history between visit times is missing. Well-designed studies that use non-invasive measurement technology for biological variables, diaries and other ways to obtain more detailed data may become more common. In addition, linkage of life history study data to administrative data bases is increasingly common. This facilitates more detailed modeling and analysis but also raises methodological issues concerning the comparison and combination of data from different sources.

Multistate models have become very widely used for modeling and analysis in many areas of medicine and public health. They are useful when dealing with both time-varying markers and clinical events and facilitate investigation of LTF. They are also the basis for many microsimulation models that are used for planning and policy making (e.g., [55]). As discussed here, Markov models are most easily fitted to incomplete data, and they possess robustness properties (Section 4.2.2) where the prediction of state occupancy at different time points is concerned. In addition, costs or utilities can be associated with different states, thus allowing assessment of cumulative health costs, quality-of-life measures, and so on [57].

There are many methodological issues that I have not discussed in any detail. One is measurement error (Section 3.2). A second is nonignorable process-dependent selection of individuals for a study. A selection plan is ignorable if the selection of an individual is conditionally independent of their life history, given baseline information. Nonignorable plans require careful analysis (e.g., [10], Section 6.2) and this area has received rather limited attention. A third issue concerns process-dependent subsampling or measurement of specific variables ([10], Section 6.3). This has received considerable attention in the case of single event times (e.g., [58]) but has received rather little formal study for more general processes, although selective subsampling of cohort members for the measurement of expensive variables is often based on prior process information (e.g., [59]).

A final issue concerns situations where 'intermittent' missing data can occur. The simplest case is when a study involves scheduled visits (e.g., annual), but an individual may miss one or more visits

and then be seen later. If visits are SMAR, then the methods considered here apply to the observed life history data. However, if whether or not an individual is seen at time t_j is not conditionally independent of process history following the preceding visit, then the only recourse is to use NSMAR models with uncheckable assumptions (e.g., [30, 31]) or to seek supplementary data, for example through tracing individuals or administrative records [29]. Another difficult situation is where individuals appear for visits in a random fashion. Models that use a point process for the observation times $\{t_{ij}, j = 1, 2, \ldots\}$ for an individual, which is related in some way to the life history process of interest, have been proposed in special settings (e.g., [60]). Once again, these models involve uncheckable assumptions in cases where the observation time process is non-ignorable.

Appendix A. Robust Estimation of One-Step Transition and State Occupancy Probabilities

To estimate the $p_{r,s}(t, t+1)$ and $p_r(t)$ of Section 4.2.2 under state-dependent LTF, we assume that a model for

$$\pi_i(t+1) = \Pr\left\{R_i(t+1) = 1 | \bar{Y}_i(t)\right\}$$

can be specified and fitted and that $R_i(t+1)$ is independent of $\{Y_i(s), s \geq t+1\}$, given $\bar{Y}_i(t)$. The estimating equation

$$\sum_{i=1}^{m} \frac{R_i(t+1)}{\pi_i(t+1)} I\left(Y_i(t) = r\right) \left\{I\left(Y_i(t+1) = s\right) - p_{rs}(t, t+1)\right\} \tag{17}$$

is seen to have expected value zero for all r, s and t by first taking the expectation with respect to $R_i(t+1)$ given $\bar{Y}_i(t)$ and then the expectation with respect to $Y_i(t+1)$ given $Y_i(t)$. This is the Inverse Probability of Censoring approach of Robins *et al.* [51]. For consistent estimation of $p_{rs}(t, t+1)$, we require a consistent estimate $\hat{\pi}_i(t+1)$ of $\pi_i(t+1)$; refer to Hajducek and Lawless [1] for discussion of this. Then, we estimate $p_{rs}(t, t+1)$ by replacing $n_{rs}(t, t+1)$ in (10) with

$$\tilde{n}_{rs}(t, t+1) = \sum_{i=1}^{m} \frac{R_i(t+1)}{\hat{\pi}_i(t+1)} I\left(Y_i(t) = r, Y_i(t+1) = s\right) \tag{18}$$

and defining $\tilde{n}_{r+}(t, t+1)$ accordingly.

ACKNOWLEDGEMENTS

This paper is based on the 9th Armitage Lecture, presented on November 9, 2011 at the Medical Research Council Biostatistics Unit in Cambridge, UK. The support and hospitality of the Biostatistics Unit are gratefully acknowledged. The author thanks the DCCT sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases of the US National Institutes of Health for access to data on the DCCT and Janet Raboud and the CANOC for access to CANOC data. Thanks to Richard Cook for valuable comments, to Ker-Ai Lee for technical assistance, and to Narges Nazeri Rad for providing information for Table 3. The authors research was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Hajducek DM, Lawless JF. Duration analysis in longitudinal studies with intermittent observation times and losses to followup. *Canadian Journal of Statistics* 2012; **40** (1): 1–21.

[2] Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, Barad D, Curb JD, Kotchen J, Kuller L, Limacher M, Wactawski-Wende J and for the Women's Health Initiative Investigators. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative Clinical Trial. *American Journal of Epidemiology* 2005; **162** (5): 404–414.

[3] The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *The New England Journal of Medicine* 1993; **329** (14): 977–986.

[4] Barrett JK, Siannis F, Farewell VT. A semi-competing risks model for data with interval-censoring and informative observation: An application to the MRC cognitive function and ageing study. *Statistics in Medicine* 2011; **30**: 1–10.

[5] Raina PS, Wolfson C, Kirkland SA, Griffith LE, Oremus M, Patterson C, Tuokko H, Penning M, Ballion CM, Hogan D, Wister A, Payette H, Shannon H, Brazil K. The Canadian Longitudinal Study on Aging (CLSA). *Canadian Journal on Aging* 2009; **28**: 221–229.

[6] Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy: ETDRS report number 12. *Ophthalmology* 1991; **98** (Suppl): 823–833.

[7] Al-Kateb H, Boright AP, Mirea L, Xie X, Sutradhar R, Mowjoodi A, Bharaj B, Liu M, Bucksa JM, Arends VL, Steffes MW, Cleary PA, Sun W, Lachin JM, Thorner PS, Ho M, McKnight AJ, Maxwell AP, Savage DA, Kidd KK, Kidd JR, Speed WC, Orchard TJ, Miller RG, Sun L, Bull SB, Paterson AD and the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group. Multiple superoxide dismutase 1/splicing factor serine alanine 15 variants are associated with the development and progression of diabetic nephropathy: The diabetes control and complications trial/epidemiology of diabetes interventions and complications genetics study. *Diabetes* 2008; **57**: 218–228.

[8] The Diabetes Control and Complications Trial Research Group. Progression of retinopathy with intensive versus conventional treatment in the Diabetes Control and Complications Trial. *Opthalmology* 1995; **102**: 647–661.

[9] The Diabetes Control and Complications Trial Research Group. Early worsening of diabetic retinopathy in the Diabetes Control and Complications Trial. *Archives of Opthalmology* 1998; **116**: 874–887.

[10] Cook RJ, Lawless JF. Statistical issues in modeling chronic disease in cohort studies. *To appear in Statistics in Biosciences* 2013.

[11] Raboud JM, Loutfy MR, Su D, Bayoumi AM, Klein MB, Cooper C, Machouf N, Rourke S, Walsmley S, Rachilis A, Harrigan PR, Smieja M, Tsoukas C, Montaner JS, Hogg RS, CANOC Collaboration. Regional differences in rates of HIV-1 viral load monitoring in Canada: Insights and implications for antiretroviral care in high income countries. *BMC Infectious Diseases* 2010; **10**: 1–9. DOI: 10.1186/1471-2334-10-40.

[12] Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer-Verlag: New York, 1993.

[13] Grennan JT, Loutfy MR, Su D, Harrigan PR, Cooper C, Klein M, Machouf N, Montaner JS, Rourke S, Tsoukas C, Hogg RS, Raboud J, CANOC Collaboration. Magnitude of virologic blips is associated with a higher risk for virologic rebound in HIV-infected individuals: a recurrent events analysis. *Journal of Infectious Diseases* 2012; **205**: 1230–1238.

[14] Cook RJ, Lawless JF. *The Statistical Analysis of Recurrent Events*. Springer Science + Business Media, LLC: New York, 2007.

[15] Cook RJ, Lawless JF, Lakhal-Chaieb L, Lee K-A. Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone. *Journal of the American Statistical Association* 2009; **104**: 60–75.

[16] Bellhouse DR. A new look at Halley's life table. *Journal of the Royal Statistical Society A* 2011; **174**: 823–832.

[17] Fix E, Neyman J. A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology* 1951; **23**: 205–241.

[18] Blumen L, Kagan M, McCarthy PJ . *The Industrial Mobility of Labor as a Probability Process*. Cornell University Press: Ithaca NY, 1955.

[19] Coleman JS. *Introduction to Mathematical Sociology*. Free Press of Glencoe: New York, 1964.

[20] Aalen OO, Borgan O, Gjessing HK. *Survival and Event History Analysis: A Process Point of View*. Springer Science + Business Media, LLC: New York, 2008.

[21] Cox DR, Isham V. *Point Processes*. Chapman and Hall: London, 1980.

[22] Daley DJ, Vere-Jones D. *An Introduction to the Theory of Point Processes*. Springer: New York, 1988.

[23] Mandel M. Estimating disease progression using panel data. *Biometrics* 2010; **66**: 304–316.

[24] Farewell VT, Su L. A multi-state model for events defined by prolonged observation. *Biostatistics* 2011; **12**: 102–111.

[25] Callegaro M. Seam effects in longitudinal surveys. *Journal of Official Statistics* 2008; **24**: 387–409.

[26] McKeown K, Jewell NP. Current status observation of a three-state counting process with applications to simultaneous accurate and diluted HIV test data. *Canadian Journal of Statistics* 2011; **39**: 475–487.

[27] Prentice RL. Chronic disease prevention research methods and their reliability, with illustrations from the Women's Health Initiative. *Journal of the American Statistical Association* 2010; **105**: 1431–1443.

[28] Prentice RL, Huang Y. Measurement error modeling and nutritional epidemiology association analyses. *Canadian Journal of Statistics* 2011; **39**: 498–509.

[29] Farewell VT, Lawless JF, Gladman DD, Urowitz MB. Analysis of the effect of lost-to-followup on the estimation of mortality from patient registry data. *Applied Statistics* 2003; **52**: 445–456.

[30] Chen B, Yi GY, Cook RJ. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine* 2010; **29** (11): 1175–1189.

[31] Sweeting MJ, Farewell VT, De Angelis D. Muliti-state Markov models for disease progression in the presence of informative examination times. *Statistics in Medicine* 2010; **29**: 1161–1174.

[32] Glymour MM. When bad genes look good - APOE*E4, cognitive decline and diagnosis thresholds. *American Journal of Epidemiology* 2007; **165**: 1239–1246.

[33] Pyy-Martikainen M, Rendtel U. Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys. *Advances in Statistical Analysis* 2008; **92**: 293–318.

[34] Steele F, Durrant GB. Alternative approaches to multilevel modelling of survey non-contact and refusal. *International Statistical Review* 2011; **79**: 79–91.

[35] Jiang Y, Scott AJ, Wild CJ. Adjusting for non-response in population-based case-control studies. *International Statistical Review* 2011; **79**: 145–159.

[36] Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 1985; **80** (392): 863–871.

[37] Titman AC. Flexible nonhomogeneous Markov models for panel observed data. *Biometrics* 2011; **67**: 780–787.

[38] Jackson CH. Multi-state models for panel data: the msm package for R. *Journal of Statistical Software* 2011; **38** (8): 1–28.

[39] Tom BDM, Farewell VT. Intermittent observation of time-dependent explanatory variables: a multistate modelling approach. *Statistics in Medicine* 2011; **30** (30): 3520–3531.

[40] Andersen PK, Klein JP, Rosthoj S. Generalized linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 2003; **90**: 15–27.

[41] Titman AC, Sharples LD. Model diagnostics for multi-state models. *Statistical Methods in Medical Research* 2010; **19**: 621–651.

[42] Hwang W and Brookmeyer R. Design of panel studies for disease progression with multiple stages. *Lifetime Data Analysis* 2003; **9**: 261–274.

[43] Lawless JF, Nazeri Rad N. Estimation and assessment of Markov multistate models with intermittent observations on individuals, 2013. Manuscript.

[44] Aalen OO, Borgan O, Fekjaer H. Covariate adjustment of event histories estimated with Markov chains: The additive approach. *Biometrics* 2001; **57**: 993–1001.

[45] Datta S, Satten GA. Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statistics and Probability Letters*, 2001; **55**: 403–411.

[46] Datta S, Satten GA. Estimation of integrated transition probabilities for non-Markov systems under dependent censoring. *Biometrics* 2002; **58**: 792–802.

[47] Satten GS, Sternberg MR. Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics* 1999; **53**: 507–513.

[48] Titman AC, Sharples LD. Semi-Markov models with phase-type sojourn distributions. *Biometrics* 2010; **66** (3): 742–752.

[49] Lawless JF, Nazeri Rad N. Multistate modelling and predictive model assessment, with application to viral rebounds in an HIV-positive cohort, 2013. Manuscript.

[50] Hogan JW, Roy J, Korkontzelou C. Tutorial in Biostatistics: Handling drop-out in longitudinal studies. *Statistics in Medicine* 2004; **23**: 1455–1497.

[51] Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90** (429): 106–121.

[52] Brookmeyer R, Gail MH. *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press: Oxford, 1994.

[53] O'Keeffe AG, Tom BDM, Farewell VT. Mixture distributions in multi-state modelling- What to choose? How to choose? Some considerations in a study of psoriatic arthritis. *Statistics in Medicine*, 2012; **32**: 600–619.

[54] Lawless JF, Fong DYT. State duration models in clinical and observational studies. *Statistics in Medicine* 1999; **18**: 2365–2376.

[55] Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alheimer's Disease. *Alzheimers and Dementia* 2007; **3**: 186–191.

[56] Tyas SL, Salazar JC, Snowdon DA, Desrosiers MF, Riley KP, Mendiondo MS, Kryscio RJ. Transitions to mild cognitive impairments, dementia, and death: findings from the nun study. *American Journal of Epidemiology* 2007; **165** (11): 1231–1238.

[57] Cook RJ, Lawless JF, Lee K-A. Cumulative processes related to event histories. *Statistics and Operations Research Transactions* 2003; **27**: 13–29.

[58] Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* 2009; **169**: 1398–1405.

[59] Mueller PW, Rogus JJ, Clearly PA, Zhao Y, Smiles AM, Steffes MW, Bucksa J, Gibson TB, Cordovado SK, Krolewski AS, Nierras CR, Warram JH. Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors in diabetic nephropathy in Type 1 diabetes. *Journal of the American Society of Nephrology* 2006; **17**: 1782–1790.

[60] Sun J. *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer: New York, 2006.