Topics in the Design of Life History Studies

by

Nathalie Moon

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics - Biostatistics

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:
Dr Zeny Feng
Associate Professor, Department of Mathematics and Statistics
University of Guelph

Supervisor(s):
Dr Leilei Zeng
Associate Professor, Department of Statistics and Actuarial Science
University of Waterloo
Dr Richard Cook
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal Member:
Dr Kun Liang
Assistant Professor, Department of Statistics and Actuarial Science
University of Waterloo
Dr Mary Thompson, Distinguished Professor Emerita
Department of Statistics and Actuarial Science
University of Waterloo

Internal-External Member: Dr Suzanne Tyas

Associate Professor, School of Public Health and Health Systems

University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Substantial investments are being made in health research to support the conduct of large cohort studies with the objective of improving understanding of the relationships between diverse features (e.g. exposure to toxins, genetic biomarkers, demographic variables) and disease incidence, progression, and mortality. Longitudinal cohort studies are commonly used to study life history processes, that is patterns of disease onset, progression, and death in a population. While primary interest often lies in estimating the effect of some factor on a simple time-to-event outcome, multistate modelling offers a convenient and powerful framework for the joint consideration of disease onset, progression, and mortality, as well as the effect of one or more covariates on these transitions.

Longitudinal studies are typically very costly, and the complexity of the follow-up scheme is often not fully considered at the design stage, which may lead to inefficient allocation of study resources and/or underpowered studies. In this thesis, several aspects of study design are considered to guide the design of complex longitudinal studies, with the general aim being to obtain efficient estimates of parameters of interest subject to cost constraints. Attention is focused on a general K state model where states $1, \ldots, K-1$ represent different stages of a chronic disease and state K is an absorbing state representing death.

In Chapter 2, we propose an approach to design efficient tracing studies to mitigate the loss of information stemming from attrition, a common feature of prospective cohort studies. Our approach exploits observed information on state occupancy prior to loss-to-followup, covariates, and the time of loss-to-followup to inform the selection of individuals to be traced, leading to more judicious allocation of resources. Two settings are considered. In the first there are only constraints on the expected number of individuals to be traced,

and in the second the constraints are imposed on the expected cost of tracing. In the latter, the fact that some types of data may be more costly to obtain via tracing than other types of data is dealt with.

In Chapter 3, we focus on two key aspects of longitudinal cohort studies with intermittent assessments: sample size and the frequency of assessments. We derive the Fisher information as the basis for studying the interplay between these factors and to identify features of minimum-cost designs to achieve desired power. Extensions which accommodate the possibility of misclassification of disease status at the intermittent assessments times are developed. These are useful to assess the impact of imperfect screening or diagnostic tests in the longitudinal setting.

In Chapter 4, attention is turned to state-dependent sampling designs for prevalent cohort studies. While incident cohorts involve recruiting individuals before they experience some event of interest (e.g. onset of a particular disease) and prospectively following them to observe this event, prevalent cohorts are obtained by recruiting individuals who have already experienced this event at some point in the past. Prevalent cohort sampling yields length-biased data which has been studied extensively in the survival setting; we demonstrate the impact of this in the multistate setting. We start with observation schemes in which data are subject to left- or right-truncation in the failure-time setting. We then generalize these findings to more complex multistate models. While the distribution of state occupancy at recruitment in a prevalent cohort sample may be driven by the prevalences in the population, we propose approaches for state-dependent sampling at the design stage to improve efficiency and/or minimize expected study cost.

Finally, Chapter 5 features an overview of the key contributions of this research and outlines directions for future work.

Acknowledgements

I would like to thank my supervisors, Drs Leilei Zeng and Richard Cook, whose guidance, support, and thoughtful suggestions have been invaluable in the preparation of this thesis. Your mentorship and encouragement have been a constant throughout my graduate studies and I am very grateful for that.

Thank you also to Drs Kun Liang, Mary Thompson, Zeny Feng, and Suzanne Tyas for their helpful suggestions and insightful comments, and to Ker-Ai Lee for her assistance with statistical computing.

The friends I made over the course of my graduate studies were a great source of support, motivation, and at times much needed distraction, in particular my officemates (Narges, Meaghan, and Di), departmental colleagues (Mirabelle, Reza, Shu, and Emily), and friends outside of statistics (Gabrielle and Lindsey).

I would also like to thank my parents Jeff and Anne-Marie Moon and brothers Jean-Luc and Patrick, who have always been there for me. My sincere thanks also to Mike, who was a pillar of calm and support as I completed my thesis.

This work was supported partially by an Alexander Graham Bell Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada and Ontario Graduate Scholarships.

To my family

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Cohort Studies for Life History Processes

Investigating the association between risk factors such as exposure to toxins, genetic biomarkers, and demographic variables and disease incidence, progression, and mortality is of great interest in health research. Cohort studies where individuals are followed-up over time are particularly well suited to studying questions of this nature. Birth cohorts are typically directed at measuring the impact of maternal exposures on neonatal and early life outcomes [Kobayashi et al., 2016], whereas studies in infants and young children may be directed at the impact of early diet and care on cognitive and physical development [Lakshman et al., 2015]. The Canadian Longitudinal Study on Aging [Raina et al., 2009] focuses on disease processes in later life; 50,000 individuals aged 45-85 were recruited and are to be followed for 20 years to gain insight into the complex relationships between behaviour, biomarkers and disease incidence. The EPIC Norfolk study [Riboli, 1992] and many others have broadly similar objectives. In other settings attention may be directed at diseased individuals and interest lies in studying the incidence of complications or comorbidities in affected individuals; studies in diabetics are particularly ubiquitous [Early

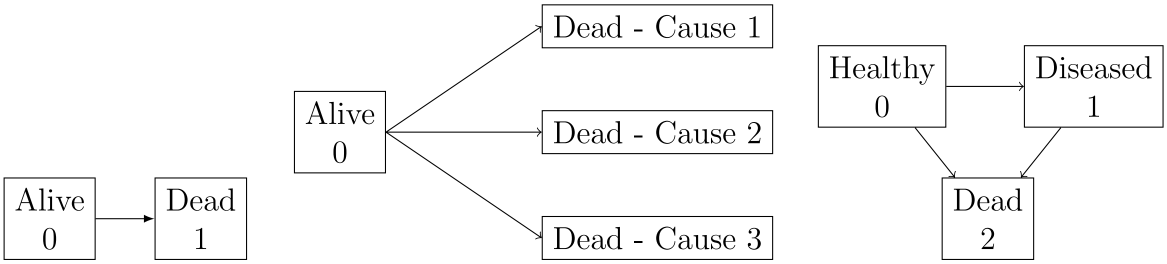
Treatment Diabetic Retinopathy Study Research Group, 1991]. While interest may lie primarily in biomarkers and their effect on the development of complications from disease and the onset of comorbidities, mortality rates may be appreciable and joint models incorporating survival times are required for valid inferences. Multistate models offer a convenient and powerful framework for the joint consideration of disease incidence, progression, and mortality.

We consider the setting in which individuals are recruited and followed prospectively to learn about the disease process and identify risk factors for the occurrence of disease complications and the development of comorbidities; retrospective information about the course of disease prior to recruitment may be available in some situations. Clinically important events are often self-evident (e.g. strokes, heart attacks, and death) but their observation times are subject to right censoring. Some complications, however, are asymptomatic and so will only be detected at the time of clinical examination or radiographic assessment. For example, asymptomatic fractures among individuals with osteoporosis are only detected upon radiographic examination [Kreiger et al., 1999], progression in retinopathy in diabetics is only detected upon examination by an ophthalmologist [Diabetes Control and Complications Trial Research Group, 1993], and progression in fibrosis of the liver among individuals with hepatitis C infection is only assessable by biopsy [Sweeting et al., 2006]. In settings where interest lies in the development of conditions or complications which are not self-evident, data become available at periodic clinic visits, giving rise to so-called panel data [Kalbfleisch and Lawless, 1985] where transition times are subject to interval censoring. Multistate models for such data are generally based on the Markov assumption and likelihoods can be easily constructed which accommodate a mixture of right-censored and interval-censored transition times [Zeng et al., 2018].

The cost of conducting a longitudinal study is often appreciable, in great part due to the expense of repeatedly assessing individuals, often via in-person examination by a physician and/or expensive clinical tests; as such, there is great interest in the design of longitudinal studies which allocate resources most efficiently [Moskowitz et al., 2017, Timmons and Preacher, 2015, Collins and Graham, 2002, Singer and Willett, 1991]. Design considerations are very much dependent on the objectives of the study, and in particular on the response of interest and which features of its distribution are key. At the planning stage, it is natural to consider the effect of various design factors on efficiency, such as sample size, duration of follow-up, assessment schedule, and the distribution of states occupied at recruitment. Albert and Hendricks Brown [1991] consider different sampling schemes and schedules for the assessments in a two-state process. Cook [2000] assessed the impact of the assessment schedule on the precision of estimates of transition intensities and occupancy probabilities; Lawless and Rad [2015] studied this and more general three-state processes. Mehtälä et al. [2015] consider sample size and the optimal scheduling of assessments for time-homogeneous two-state Markov processes; Hwang and Brookmeyer [2003] consider similar issues for strictly progressive $K-$state processes. In each chapter of this thesis, we consider different aspects of design in the context of prospective cohort studies aiming to collect data with which to model multistate life history processes.

1.2 Multistate Models

Over the course of an individual's lifetime, many attributes are subject to variation over time (e.g. disease status, physiological markers, etc). The space of all possible combinations of such attributes can be partitioned into a possibly infinite set of states $\Omega = \{0, 1, 2, \ldots\}$ where each state is defined by combinations of attribute values; in this thesis, we will restrict our attention to finite state spaces. The attributes characterizing these states are typically



(a) Survival model (b) Competing risks model (c) Illness-death model

Figure 1.1: Examples of multistate models

time-varying so individuals can transition between states over time. Directed graphs are used to depict multistate processes as they conveniently display the state space and the permissible transitions between these states. Figure 1.1 shows three examples of common multistate models: (a) the standard time-to-event survival model, (b) the competing risks model where a transition to one of two or more states precludes transitions to other states, and (c) the illness-death model modelling progression from a healthy state to death, with possible onset of some disease of interest prior to death. For more examples of multistate processes, see Hougaard [1999] and Cook and Lawless [2018]. States in a multistate process can be characterized in terms of features of the process. States from which no transitions are possible (e.g. all 'Death' states in Figure 1.1) are called absorbing states while states from which it is possible to move to one or more other state (e.g. 'Alive' and 'Diseased' states in Figure 1.1) are called transient states [Ross, 2014].

Multistate models are a natural way to characterize chronic disease processes with multiple stages. Let $\{Z(t), t > 0\}$ be a continuous time stochastic process with state space $\Omega = \{0, 1, 2, \ldots, K\}$. Let $\mathcal{H}(t) = \{Z(s), 0 \le s \le t\}$ be the history for the multistate process, with the intensity for $k \to \ell$ transitions defined as

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

where $k \neq \ell \in \Omega$. Markov models are among the most commonly used types of multistate models. Under such models, all dependence of transition intensities on the history of the process is encompassed in the current state, that is $\lambda_{k\ell}(t|\mathcal{H}(t^-)) = \lambda_{k\ell}(t)$. The transition probabilities $p_{k\ell}(s,t) = P(Z(t) = \ell | Z(s) = k)$ are obtainable from the intensities via the Kolmogorov forward differential equation

$$\frac{\partial}{\partial t}\mathbb{P}(s,t) = \mathbb{P}(s,t)\mathbb{A}(t) \qquad s < t \; , \tag{1.1}$$

where $\mathbb{P}(s,t)$ is the transition probability matrix with entries $[\mathbb{P}(s,t)]_{k\ell} = p_{k\ell}(s,t)$, and $\mathbb{A}(t)$ is the transition intensity matrix with entries $[\mathbb{A}(t)]_{k\ell} = \lambda_{k\ell}(t)$ for $k \neq \ell \in \Omega$ and $[\mathbb{A}(t)]_{kk} = -\sum_{\ell \neq k} \lambda_{k\ell}(t)$ [Cox and Miller, 1965]. A time-homogeneous model in which transition intensities are independent of t (i.e. $\lambda_{k\ell}(t) = \lambda_{k\ell}$ for all $k \neq \ell$) is the simplest model to consider. In this case, the transition probability matrix is written as a matrix exponential of the constant intensity matrix $\mathbb{A}(t) = \mathbb{A}_0$,

$$\mathbb{P}(s,t) = \exp\{(t-s)\mathbb{A}_0\} = \sum_{n=0}^{\infty} \mathbb{A}_0^n (t-s)^n / n! \; .$$

Non-homogeneous Markov models can be adopted by specifying piecewise-constant transition intensities so that $\mathbb{A}(t) = \mathbb{A}_r$ if $t \in \mathcal{B}_r = [b_{r-1}, b_r)$, $r = 1, \ldots, R$, with a sequence of pre-defined cut-points $0 = b_0 < b_1 < \ldots < b_{R-1} < b_R = \infty$. Under such models, probabilities $p_{k\ell}(s,t)$ can be obtained by multiplying a sequence of transition probabilities over the constant segments of the interval $[s,t)$ and then summing over the unobserved disease status at the cut-points. More specifically, if $r_s = \{r; s \in \mathcal{B}_r, r = 1, \ldots, R\}$ and $r_t = \{r; t \in \mathcal{B}_r, r = 1, \ldots, R\}$, then

$$\begin{aligned} \mathbb{P}(s,t) &= \prod_{r=r_s}^{r_t} \mathbb{P}\big(\max\{s, b_{r-1}\}, \min\{t, b_r\}\big) \\ &= \prod_{r=r_s}^{r_t} \exp\big\{\big(\min\{t, b_r\} - \max\{s, b_{r-1}\}\big)\mathbb{A}_r\big\} \end{aligned} \tag{1.2}$$

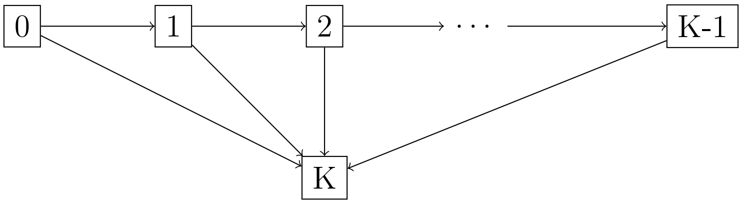


Figure 1.2: Progressive model with transient states $0, \ldots, K-1$ and one absorbing state K

where the matrix exponential is used to obtain transition probabilities within each piece intersecting the interval of interest (s, t).

Multiplicative intensity-based models can be used to characterize the effect of prognostic variables on the dynamics of the disease process. Modulated Markov models are obtained by specifying

$$\lambda_{k\ell}(t|X) = \lambda_{k\ell 0}(t) \exp(X'\beta_{k\ell}), \quad k < \ell \in \Omega,$$

where $\lambda_{k\ell 0}(t)$ is the baseline transition intensity, $X = (X_1, \ldots, X_p)'$ is a $p \times 1$ covariate vector, and $\beta_{k\ell} = (\beta_{k\ell 1}, \ldots, \beta_{k\ell p})'$ is a vector of regression coefficients specific to $k \to \ell$ transitions. We let θ be the vector of parameters indexing all baseline intensity functions and regression coefficients.

1.3 Incomplete Observations

1.3.1 Observations in Continuous Time

Incomplete observations are a common feature of studies where individuals contribute information over time, whether it be a time-to-event response or a multistate process observed intermittently. Time-to-event responses (in survival analysis or as part of a more general multistate process) are generally subject to two right censoring mechanisms: administrative censoring and loss-to-follow-up. A response T is administratively censored if the event of interest has not yet occurred by the end of the follow-up period τ, that is if $T > \tau$; this is sometimes also referred to as type I censoring [Kalbfleisch and Prentice, 2011]. Alternatively, individuals may also withdraw from the study at some random time $C < \tau$ after which no information about their responses is available. In general, we let $T^\dagger = \min(T, \tau, C)$ be the right-censored observation and $\delta = I(T = T^\dagger)$ a non-censoring indicator. Suppose $f(t;\theta)$ and $g(c;\rho)$ are the probability densities for the time-to-event response and censoring time respectively. The censoring process is said to be non-informative if the parameters in $f(t;\theta)$ and $g(c;\rho)$ are functionally independent; in this case we can factor the joint likelihood and inference for θ is solely based on the observed right-censored observations through

$$L(\theta) \propto \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} P(T_i > t_i^\dagger;\theta)^{1-\delta_i} \tag{1.3}$$

without modelling the censoring process. This assumption is reasonable if it can be argued that individuals withdraw from the study for reasons unrelated to the severity of their disease and/or their prognosis, resulting in a non-censored subpopulation which is representative of the whole population of interest. When this assumption fails to hold, censoring is deemed informative and inference based on (1.3) is incomplete as it fails to incorporate the censoring contributions; methods to handle informative right-censoring involve accounting for the censoring distribution explicitly.

1.3.2 Intermittent Observations

In some settings, it is only possible to observe the disease process at discrete times $\{a_0 < a_1 < a_2 < \cdots\}$; let $Z(a_j) = Z_j$ be the state occupied at assessment time a_j. In such panel data [Kalbfleisch and Lawless, 1985, Hwang and Brookmeyer, 2003], the exact time of transitions are unknown, and thus interval censored. Indeed, for non-progressive processes it is not even known how many transitions may have occurred between assessment times.

For longitudinal studies with intermittent assessments, 'incompleteness' in the responses may take different forms. If the assessments are regular (that is, scheduled in advance at times $a_0 < \cdots < a_J$), the notion of 'missing' a visit is natural; individuals may miss visits occasionally or withdraw from the study entirely. The latter is similar to the idea of right censoring discussed above, and loss-to-follow-up in a discrete assessment scheme may be induced by an underlying censoring time C in continuous time, where responses are observed at times $\{a_j : a_j < C\}$ and missing otherwise. Alternatively, an under-observation indicator $Y_j = I(Z_j \text{ is observed})$ may be defined to allow for (potentially) more flexible missing data patterns. A common assumption in this setting is the sequential missing at random (SMAR) assumption [Hogan et al., 2004], which states that when we condition on the past history $H_{j-1} = (\bar{Z}_{j-1}, \bar{Y}_{j-1}, X)$ where $\bar{Z}_{j-1} = (Z_1, \ldots, Z_{j-1})$, $\bar{Y}_{j-1} = (Y_1, \ldots, Y_{j-1})$, and X is a covariate vector, then Y_j is independent of the future responses $\{Z_j, \ldots, Z_J\}$, that is

$$P(Y_j = 0|\bar{Y}_{j-1}, \bar{Z}_K, X) = P(Y_j = 0|\bar{Y}_{j-1}, \bar{Z}_{j-1}, X).$$

When disease status is assessed intermittently at irregular intervals, particularly when the scheduling of these visits is patient-driven, it is much more difficult to define what is meant by a 'missing' observation. Cook and Lawless [2018] propose an analogue to the SMAR assumption, namely that of a *conditionally independent visit process* (CIVP), by which a random visit time A_j is assumed to be conditionally independent of the disease process since the last assessment at time A_{j-1}, given the observed history $H_{j-1} = \{(Z(A_1), A_1), \ldots, (Z(A_{j-1}), A_{j-1}), X\}$. Violations of the CIVP assumption are common when visits are patient-driven rather than pre-scheduled. Registry data are an extreme example of this, where patient data is only available when they interact with the health care system (e.g. hospital visits, insurance claims) at which time disease status is inferred

[Farzanfar et al., 2017].

1.4 Response-dependent Recruitment Schemes

Response-dependent sampling arises when selection probabilities depend on responses, for example when individuals may only be observed if some recruitment condition is satisfied. In his paper discussing the nonparametric estimation of incomplete data, Turnbull proposes a useful unifying framework for censoring and truncation [Turnbull, 1976]. Let T be a time-to-event response and X a covariate vector. An observation (T_i, X_i), $i = 1, \ldots, n$ is truncated by a set $\mathcal{B}_i$ if T_i follows the conditional distribution $F(t; \mathcal{B}_i, X_i) = P(T \leq t | T \in \mathcal{B}_i; X_i)$. If $\mathcal{B}_i = (0, \infty)$ the observation is not truncated, while $\mathcal{B}_i = (0, R_i)$ and $\mathcal{B}_i = (L_i, \infty)$ correspond to right- and left-truncation respectively. Further, if we only know T_i belongs to some set $\mathcal{A}_i \subset \mathcal{B}_i$ we say T_i is censored into $\mathcal{A}_i$. If $\mathcal{A}_i$ contains a single value, the value of T_i is known exactly (so T_i is uncensored), while other choices of $\mathcal{A}_i \subset \mathcal{B}_i$ can lead to right-, left-, and interval-censoring. Using this notation, the information about an observation $(T_i; X_i)$ can equivalently be described by $(\mathcal{A}_i, \mathcal{B}_i; X_i)$. Note that the sets $\mathcal{A}_i$ and $\mathcal{B}_i$ can be fixed or random, depending on the sampling scheme and study design.

Left truncation is the most well-studied type of response-dependent sampling. For example, let T be age at death and consider recruiting a sample of individuals for prospective follow-up to estimate the intensity for mortality. An individual recruited at age A necessarily must satisfy $T \in (A, \infty)$, as individuals must be alive to be recruited. Failing to account for the left truncation recruitment condition $T \geq A$ in the subsequent analysis leads to an overestimation of survival time. It is likely due to the great interest in length-biased data from prevalent cohort samples in the medical sciences [Wolfson et al., 2001, Gladman et al., 2005, Keiding, 1991, Simon, 1980, Duffy et al., 2008, Shen et al., 2017, Shen and Cook, 2013] as well as in other fields such as economics [Lancaster, 1992], and manufacturing [Blumenthal, 1967] that the development of methods for the analysis of truncated data has mostly focused on left truncation.

Time-to-event data, as shown in Figure 1.1(a), can be cast into the multistate framework. As discussed above, in this case left truncation amounts to recruiting individuals who are still in the 'alive' state at recruitment (or equivalently, who have not yet entered the 'dead' state). Consider a more general multistate model, such as the $K+1$ state model in Figure 1.2 which is the primary focus of this thesis. It is possible to restrict recruitment efforts to a subset $\mathcal{S} \subseteq \{0, \ldots, K-1\} \subset \Omega$ of the non-fatal states, representing particular stages of disease. If, as in the simple length-biased case, the selection probability for a given individual is proportional to their sojourn time in $\mathcal{S}$, then this recruitment condition must be considered in analyses. Prospective cohort studies intending to estimate disease incidence or related covariate effects would involve recruiting disease-free individuals at age A and following them prospectively over some duration τ to monitor disease onset, progression, and death. While analysis of such data is straightforward, estimability may be a concern if the disease of interest is rare or if progression is slow relative to the duration of the study (τ). Rather, prevalent studies may be considered, where individuals at different disease stages may be sampled. Considerations related to the choice of stages at accrual $\mathcal{S}$ in this case and the mechanisms by which individuals are screened from the population are discussed by Cook and Lawless [2018], but little work has been done on this topic. In practice, if interest lies in estimating all transition intensities in a multistate model, it may be necessary to employ auxiliary data to estimate some of these intensities (e.g. retrospective estimation of intensities for disease onset if prevalent sample is drawn), this is discussed in Keiding et al. [1989].

1.5 Motivating Studies

1.5.1 Research Program in Psoriatic Arthritis

Psoriatic arthritis (PsA) is a form of inflammatory arthritis associated with psoriasis, a condition characterized by inflammation of the skin [Eder et al., 2011a]. Estimates of the prevalence of psoriasis and psoriatic arthritis vary widely, from 0.5% to 11% [Michalek et al., 2017] and 0.05% to 0.25% [Ogdie and Weiss, 2015] respectively. The problems considered in this thesis are motivated by a research program in psoriasis and psoriatic arthritis at the Toronto Western Hospital and the University of Toronto.

Since 1976, the Centre for Prognosis Studies in Rheumatic Disease at the Toronto Western Hospital has maintained a registry of patients with psoriatic arthritis, the Toronto Psoriatic Arthritis Cohort [Gladman and Chandran, 2010], with the objective to treat patients and increase understanding of disease incidence and the course of progression. Patients in this cohort were monitored prospectively to identify progression in the severity of the disease; one such type of progression is arthritis mutilans, characterized by severe inflammation of joints in the hand and feet leading to deformity [Gladman et al., 1987]. More recently, the Toronto Psoriasis Cohort, established in 2006, began enrolling patients with psoriasis but *not* psoriatic arthritis, with a view to following them prospectively to assess the impact of demographic and environmental factors as well as biomarkers on the risk of developing psoriatic arthritis [Eder et al., 2011a]. particular interest lies in estimating the effect of key genetic markers (e.g. HLA-B27) on

(i) the incidence of psoriatic arthritis among individuals with psoriasis and

(ii) the incidence of arthritis mutilans among individuals with psoriatic arthritis while accommodating the full disease process, including death. Patients in the Toronto Psoriasis Cohort provide information through prospective follow-up, while individuals in the Psoriatic Arthritis cohort provide retrospective information on psoriasis onset times and prospective data on arthritis mutilans and mortality.

1.5.2 Canadian Longitudinal Study on Aging

The Canadian Longitudinal Study on Aging (CLSA) [Raina et al., 2009] is another large study currently underway, involving the recruitment of 50,000 Canadians aged 45-85 and follow-up over a period of 20 years. All CLSA participants furnish baseline information and are to be followed-up every three years; participants are directed into one of two subcohorts, 'tracking' and 'comprehensive'. Individuals in the tracking cohort of $\approx$ 20,000 individuals, follow-up is done via telephone interviews, while individuals in the comprehensive cohort undergo in-person interviews and examinations and provide biological specimens (e.g. blood and urine) every three years.

The objective of the CLSA is broad: to study the interplay between biological, physical, psychosocial, and societal factors affecting aging in Canada [Raina et al., 2009]. The study has already led to a number of research outputs, including a report on health and aging in Canada [Raina et al., 2018] and a study on cognitive measures using CLSA data [Tuokko et al., 2017]; an additional 100 projects related to the CLSA have also been approved [CLSA] and are ongoing

Because of the large scale of cohort studies such as the CLSA, improvements in study design can lead to meaningful savings and this motivates the work in this thesis. For example, due to the long duration of intended follow-up, the CLSA is likely to feature a high proportion of loss-to-follow-up as individuals move away, lose interest in participating in the study, etc., so approaches for cost-effective design of tracing studies in Chapter 2 would be well suited in this setting. If interest lies in constructing sub-cohorts of individuals for more intensive examination and data collection for the study of particular disease processes, the state-dependent sampling schemes of Chapter 4 could also be useful.

1.6 Outline of Thesis

The remainder of this thesis is organized as follows.

1.6.1 Tracing Studies in Cohorts with Attrition

Attrition is a common and generally unavoidable occurrence when conducting a cohort study. When the time to withdrawal from the cohort is conditionally independent of the disease process, the primary consequence is a loss of precision for the estimation of model parameters. This loss can sometimes be mitigated by the conduct of tracing studies in which a subsample of individuals lost to follow-up are contacted and some information is obtained on their disease and survival status. In Chapter 2, we describe the use of selection models to sample individuals for tracing, which will yield more efficient estimators and/or more cost-effective subsampling than simple random sampling.

1.6.2 Cohort Study Designs for Markov Processes

In Chapter 3, we focus attention on the design of longitudinal cohort studies with a set number of intended assessments over the study period, at which times disease status is to be ascertained; here we assume the exact time of deaths occurring over the study period are available. Both sample size and the frequency of assessments are drivers of precision in estimates of transition intensities and/or covariate effects on these intensities, and we present a framework to evaluate this tradeoff by taking into account expected study cost. We present a closed form expression for the Fisher information, allowing for misclassification in the observed disease status.

1.6.3 State-dependent Sampling Designs for Prevalent Cohort Studies

In Chapter 4, we consider the impact of state-dependent sampling on efficiency in the time-to-event and multistate settings. We first demonstrate the bias induced by length-biased sampling on transition probabilities in an illness-death model. We then compare the efficiency of estimators of a regression coefficient on a time-to-event response, subject to a cost constraint. Finally, we consider the design of prevalent cohort studies in the multistate framework. We consider two approaches: in the first we consider selecting individuals for follow-up on the basis of the state they occupy at the time of recruitment, and in the second we assume a population is screened until desired state-specific sample sizes are recruited for follow-up. In the latter setting, all individuals who are screened furnish current-status data and this is exploited in the estimation procedure. In both cases, we derive the Fisher information and use this as the basis for study design. minimum-cost designs achieving a desired level of power are also defined and the relationship between disease process parameters and features of these designs are studied.

Chapter 2

Tracing Studies in Cohorts with Attrition: Selection Models for Efficient Sampling

2.1 Introduction

Attrition is a common feature of longitudinal cohort studies, wherein some of the individuals initially recruited into the study become lost to follow-up before the planned end of the study. Failing to account for attrition at the design stage may result in significantly underpowered studies; this can be counteracted by increasing the size of the initial sample size for follow-up, although this may not be feasible in cases where the rate of loss to follow-up is higher than was expected. Rather, tracing studies may be conducted, where individuals who have been lost to follow-up are tracked down, or 'traced', to recover some information about their course of disease.

This work is particularly motivated by a research program at the Centre for Prognosis Studies in Rheumatic Diseases at the Toronto Western Hospital; see Section 1.5.1. Recruitment in the Toronto Psoriatic Arthritis Cohort has been ongoing since 1976 and while assessments are intended to occur on an annual basis, a large proportion of individuals have not been seen for 2+ years and hence deemed to be lost to follow-up. Tracing studies have been done on an ad-hoc basis in the past, and the work in this chapter aims to provide a rigorous framework to design cost-effective tracing studies.

We consider the setting in which disease or vital status is determined at intermittent pre-scheduled visit times until death or loss-to follow-up. At the planned study endpoint, a subset of individuals who have been lost to follow-up are selected for tracing at which their disease status is obtained.

Likelihood inference based on available data yield consistent but less efficient estimators when data satisfy the sequential missing at random (SMAR) assumption [Hogan et al., 2004]. The loss of efficiency can be mitigated somewhat through the conduct of tracing studies whereby a subset of the individuals who have withdrawn from the cohort are contacted to obtain information on their survival and disease status [Farewell et al., 2003]. Despite the considerable appeal of enhancing information from such efforts, relatively little attention has been given to the design of tracing studies. We address this here by sampling individuals who are lost to follow-up using selection models which exploit information in the observed history prior to withdrawal from the cohort. Within a given class of selection models, sampling probabilities can be chosen to increase efficiency of estimators of parameters of primary interest (e.g. incidence rates for complications or comorbidities, marker effects, etc.). Such models are appealing when resource constraints mean that not all individuals lost to follow-up can be traced.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the multistate model of interest and the likelihood for panel data with attrition under a SMAR mechanism, define the tracing selection model, and describe the idea of optimal selection for tracing. Asymptotic calculations demonstrating the efficiency gains from optimal tracing compared to simple random sampling are also given. When the cost of securing information on disease progression status is different from the cost of simply obtaining survival status, the cost implications of optimal tracing are also provided. In Section 2.3, a more general optimization process is described with cost constraints, which leads to different optimal selection models; the efficiency gains are also illustrated in this setting based on asymptotic results. An application of the proposed methodology to data collected from a cohort study conducted at the University of Toronto Psoriatic Arthritis Clinic is presented in Section 2.4 and general remarks are given in Section 2.5.

2.2 Model Formulation and Design of Tracing Studies

2.2.1 A Multistate Markov Model for Disease Progression

Consider a progressive multistate model as in Figure 1.2, where states $\{0, \ldots, K-1\}$ represent increasingly severe stages of disease progression and state K represents death. We assume the process is Markov, so the intensity of a transition from state k to state ℓ is

$$\lim_{\Delta t \downarrow 0} \frac{P(Z((t+\Delta t)^-) = \ell | Z(t^-) = k, \mathcal{H}(t))}{\Delta t} = \lambda_{k\ell}(t), \qquad k = 0, \ldots, K-1, \quad \ell \in \{k+1, K\}.$$

As in Section 1.2, we assume covariates X have a multiplicative effect on the intensities, with $\lambda_{k\ell}(t|X) = \lambda_{k\ell 0}(t)\exp(X'\beta_{k\ell})$ where $\lambda_{k\ell 0}(t)$ is the baseline transition intensity, $X = (X_1, \ldots, X_p)'$ a $p \times 1$ covariate vector, and $\beta_{k\ell} = (\beta_{k\ell 1}, \ldots, \beta_{k\ell p})'$ a vector of regression coefficients for the $k \to \ell$ transition. Let θ be the vector of parameters indexing all baseline intensities and regression coefficients.

2.2.2 Intermittent Assessment with Dropout

It is often not possible to monitor disease status continuously in cohort studies but rather only examine individuals at periodic assessment times. Consider an inception cohort of individuals recruited and examined at the time of disease onset ($t = 0$, say), and let $a_0 = 0$ and a_j, $j = 1, \ldots, J$ represent common planned assessment times measured from the time of disease onset; in this case $a_J = \tau$ is a common administrative censoring time. To simplify the notation we consider the contributions from a generic individual and let $Z_j = Z(a_j)$ denote the state occupied at the jth assessment, where $Z_0 = Z(a_0) = Z(0) = 0$. Let $\bar{Z}_j = \{Z_0, Z_1, \ldots, Z_j\}$ denote the history of the process up to the jth assessment, $j = 0, 1, \ldots, J$; $\bar{Z}_J$ then represents the complete response vector we aim to observe.

Let $Y_j = I(Z_j \text{ is observed})$ be an under-observation indicator and $\bar{Y}_j = \{Y_0, Y_1, \ldots, Y_j\}$ be the history of the missing data process up to and including the jth assessment. Our focus here is on the loss of data due to early withdrawal which corresponds to a monotone missing data pattern whereby $Y_j = 1$ implies $Y_1 = \cdots = Y_{j-1} = 1$ and $Y_j = 0$ implies $Y_{j+1} = \cdots = Y_J = 0$. Let $C = \max\{j : Y_j = 1, j = 0, \ldots, J\}$ record the last assessment at which the individual was observed so $\bar{Z}_C$ represents the observed part of the full response vector $\bar{Z}_J$. The likelihood function of the observed data $(\bar{Z}_C, \bar{Y}_J, X)$ from a single individual is

$$\begin{aligned} P(\bar{Z}_C, \bar{Y}_J \mid X; \theta, \gamma) &= \sum_{Z_{C+1}, \ldots, Z_J} P(\bar{Z}_J \mid X; \theta) P(\bar{Y}_J \mid \bar{Z}_J, X; \gamma) \\ &= \sum_{Z_{C+1}, \ldots, Z_J} \left[\prod_{j=1}^{J} P(Z_j \mid \bar{Z}_{j-1}, X; \theta) \prod_{j=1}^{C+1} P(Y_j \mid \bar{Y}_{j-1}, \bar{Z}_J, X; \gamma) \right] \qquad (2.1) \end{aligned}$$

where $P(\bar{Y}_J \mid \bar{Z}_J; X; \gamma)$ is the conditional probability of the under-observation indicators $\bar{Y}_J$ given the full response vector $\bar{Z}_J$ and covariate vector X, parameterized in terms of γ.

Note that if $Z_C = K$ then the process has been observed to completion and the sum in (2.1) is degenerate. With a monotone SMAR mechanism [Hogan et al., 2004] the probability of becoming lost to follow-up at a given assessment only depends on an individual's disease status and covariates observed at the previous assessments, that is

$$P(Y_j|\bar{Y}_{j-1}, \bar{Z}_J, X; \gamma) = P(Y_j|Y_{j-1} = 1, \bar{Z}_{j-1}, X; \gamma) \ .$$

In other words, under a SMAR mechanism, the probability of dropout at the j^{th} visit depends only on the observed data up to and including the $(j-1)^{th}$ visit. However, under a more general MAR mechanism as described by Rubin [1976] and Little and Rubin [1987], the probability of failing to make an observation at time a_j may depend on all observed data, both before and after a_j. In the event that death occurs over (a_{j-1}, a_j), the individual will by definition *not* be present at visit j, but his/her vital status could still be ascertained, either from family members or other sources (e.g. death registries/newspapers). We assume here that the probability that vital status is ascertained is governed by a SMAR mechanism, in which case (2.1) can be factored as the product of two terms of the form

$$P(\bar{Z}_C, \bar{Y}_J|X; \theta, \gamma) = \prod_{j=1}^{C} P(Z_j|\bar{Z}_{j-1}, X; \theta) \prod_{j=1}^{C+1} P(Y_j|Y_{j-1} = 1, \bar{Z}_{j-1}, X; \gamma),$$

where the first term involves only disease process parameters θ and the second term only missing data parameters γ. If θ and γ are functionally independent then the withdrawal process is non-informative and inference about θ can be based solely on the likelihood constructed using the first term,

$$L_1(\theta) = \prod_{j=1}^{C} P(Z_j|\bar{Z}_{j-1}, X; \theta) \ , \tag{2.2}$$

where $P(Z_j|\bar{Z}_{j-1}, X; \theta) = P(Z_j|Z_{j-1}, X; \theta)$ under a Markov model. These imply that standard likelihood methods based on the observed multistate data for the cohort, $\{\bar{Z}_C, X\}$, will result in consistent estimation of θ when the loss-to-follow-up process is ignored. We use the subscript 1 on this likelihood because we think of this data as arising from phase I of a two-phase study where phase I involves routine approach to follow-up and data collection; additional data are obtained in phase II by tracing selected individuals and we describe how this is done next.

Let $\mathcal{D} = \{\bar{Z}_C, \bar{Y}_J, X, C, \Delta\}$ represent the observed phase I data obtained from the regular follow-up process where $\Delta = I(C = J)$ indicates that follow-up was complete. Individuals with incomplete follow-up (i.e. with $\Delta = 0$) are eligible to be selected for a phase II tracing study which we take to be conducted at time a_J. Let $R = 1$ indicate that an eligible individual is selected for tracing which happens according to the selection model

$$P(R = 1|\mathcal{D}, \Delta = 0) = P(R = 1|\bar{Z}_C, X, C, \Delta = 0; \rho) \ , \tag{2.3}$$

indexed by ρ. We presume that individuals who are traced furnish information on the state occupied at a_J but alternative formulations may be considered in which retrospective data are collected as well. Conditional on the phase I data, the likelihood contribution from phase II is then

$$\Big[P(Z_J|R, \mathcal{D})^R P(R|\mathcal{D})\Big]^{1-\Delta} = \Big[P(Z_J|R, \bar{Z}_C, \bar{Y}_J, X, C, \Delta)^R P(R|\bar{Z}_C, \bar{Y}_J, X, C, \Delta; \rho)\Big]^{1-\Delta} \ .$$

We assume

$$P(Z_J|R, \bar{Z}_C, \bar{Y}_J, X, C, \Delta) = P(Z_J|\bar{Z}_C, X, C; \theta) \ , \tag{2.4}$$

so the disease status at the time of tracing (Z_J) is conditionally independent of the attrition time and tracing selection outcome given the observed responses. This enables us to write the above likelihood as a product of a term involving response parameters θ only and a term involving selection model parameters ρ only. If parameters θ and ρ are functionally independent, then we can restrict attention to the partial likelihood pertaining to θ from a traced individual, which takes the form

$$L_2(\theta) = P(Z_J|\bar{Z}_C, X; \theta)^{R(1-\Delta)} \ . \tag{2.5}$$

We can then augment the likelihood $L_1(\theta)$ in (2.2) by incorporating data from the tracing study and use

$$L(\theta) = L_1(\theta)L_2(\theta) \ . \tag{2.6}$$

The incorporation of extra information obtained from the tracing study through $L_2(\theta)$ enables one to enhance the efficiency of estimation for θ. We discuss next how tracing can be done to ensure a large gain in efficiency for parameters of key interest.

2.2.3 Optimal Designs for Tracing

Now consider a sample of size n where we use the subscript i to label individuals, $i = 1, \ldots, n$. Let $\mathcal{D}_i = \{\bar{Z}_{iC}, \bar{Y}_{iJ}, X_i, C_i, \Delta_i\}$ denote the observed data from individual i from the regular follow-up process in phase I and $\mathcal{D} = \{\mathcal{D}_i, i = 1, \ldots, n\}$ denote the phase I data. Then we write $L_1(\theta) = \prod_{i=1}^{n} L_{i1}(\theta)$ where $L_{i1}(\theta) = \prod_{j=1}^{C_i} P(Z_{ij}|\bar{Z}_{i,j-1}, X_i; \theta)$ as in (2.2) and let $\tilde{\theta}$ be the MLE of θ obtained by maximizing the likelihood $L_1(\theta)$ from the phase I data. The observed information matrix from phase I is

$$I_1(\tilde{\theta}) = \sum_{i=1}^{n} I_{i1}(\tilde{\theta}) = \sum_{i=1}^{n} \left(-\frac{\partial^2 \log L_{i1}(\theta)}{\partial\theta\partial\theta'} \right) \Big|_{\theta=\tilde{\theta}} \ .$$

If $L_{i2}(\theta) = P(Z_{iJ}|\bar{Z}_{iC}, X_i; \theta)^{R_i(1-\Delta_i)}$ is the contribution from individual i from (2.5), then conditioning on their phase I data, their contribution to the expected information matrix from tracing is

$$\mathcal{I}_{i2}^{\dagger}(\theta, \rho) = E\left[-\frac{\partial^2 \log L_{i2}(\theta)}{\partial\theta\partial\theta'} \Big| \mathcal{D}_i, \Delta_i = 0 \right]$$

which over all n individuals gives expected information matrix

$$\mathcal{I}_2^{\dagger}(\theta,\rho) = \sum_{i=1}^{n}(1-\Delta_i)P(R_i=1|\bar{Z}_{iC},X_i,C_i,\Delta_i=0;\rho) \\ \times \sum_{Z_{iJ}=1}^{K}\left[P(Z_{iJ}\mid\bar{Z}_{iC},X_i;\theta)\cdot\left(-\frac{\partial^2\log P(Z_{iJ}|\bar{Z}_{iC},X_i;\theta)}{\partial\theta\partial\theta'}\right)\right]$$

under the assumption in (2.4). Consider a hybrid information matrix defined as the sum of the observed information matrix from the phase I data, and the expected information matrix arising from a phase II tracing study, given by

$$I_H(\theta,\rho) = I_1(\theta) + \mathcal{I}_2^{\dagger}(\theta,\rho)\ . \tag{2.7}$$

We propose to use (2.7) with θ replaced by the estimate $\tilde{\theta}$ from phase I to set the value of ρ for the selection model. If interest lies in making inference for a particular parameter θ_k, for example, the so-called 'optimal' tracing selection parameters ρ^{opt} may be obtained by minimizing $[I_H^{-1}(\tilde{\theta},\rho)]_{kk}$, the (k,k) element of the inverse of (2.7), subject to a constraint on $\pi = P(R=1|\Delta=0)$, the overall proportion of individuals lost to follow-up who are traced. This can be implemented by minimizing

$$\big[I_H^{-1}(\tilde{\theta},\rho)\big]_{kk} + \zeta\Big[\sum_{i:\Delta_i=0}P(R_i=1|\mathcal{D}_i,\Delta_i=0;\rho)/(n-\dot{\Delta})-\pi\Big] \tag{2.8}$$

with respect to ρ to get ρ^{opt}, where ζ is a Lagrange multiplier, the first term in square brackets is the empirical expectation of the selection probabilities averaging over the observed data with $\dot{\Delta}=\sum_{i=1}^{n}\Delta_i$, and the entire term in square brackets is a constraint which ensures the expected proportion of individuals lost to follow-up to be traced is satisfied. The delta method may be used to consider situations when estimation of a function $g(\theta)$ is the focus. The optimality criterion in (2.8) can be generalized to involve any linear function $h(\cdot)$ of the elements of $I_H(\tilde{\theta},\rho)$. In particular analogs of A-optimality and C-optimality

[Emery and Nenarokomov, 1998] can be achieved, but we do not pursue this here as we focus on the case the tracing study is conducted with a specific scientific question in mind.

Let $\widehat{\theta}$ denote the final estimates obtained based on the augmented likelihood (2.6) once the tracing study is completed. The asymptotic variance of $\widehat{\theta}$ is thus $\text{asvar}\big(\sqrt{n}(\widehat{\theta}-\theta)\big) = \mathcal{I}^{-1}(\theta,\gamma,\rho)$ where

$$\mathcal{I}(\theta,\gamma,\rho) = E\Big[-\frac{\partial^2 \log L_i(\theta)}{\partial\theta\partial\theta'}\Big] = E\big[I_{i1}(\theta)\big] + E\big[(1-\Delta_i)\mathcal{I}^{\dagger}_{i2}(\theta,\rho)\big] \ . \qquad (2.9)$$

The above expectation is taken with respect to the phase II tracing information by first conditioning on the phase I (incomplete) data and then taking the unconditional expectation. Note that to determine ρ^{opt} in applications, as in the analysis of Section 2.4, we use (2.8); but for the calculation of the asymptotic relative efficiency that follows we use (2.9) in lieu of $I_H(\tilde{\theta},\rho)$ in (2.8) for computational efficiency; the results agree extremely well with the more computationally demanding results based on (2.8).

2.2.4 Assessing the Efficiency Gains from Optimal Tracing

We now study the properties of estimators obtained following the proposed tracing procedure, highlighting the efficiency gains over selection models involving simple random sampling (SRS). We consider a time-homogeneous three-state illness-death model with $K = 2$ and $\Omega = \{0, 1, 2\}$. We assume a binary covariate X with $P(X = 1) = 0.25$ modulates the $0 \to 1$ transition intensity, which gives a parameter vector $\theta = (\lambda_{010}, \lambda_{020}, \lambda_{120}, \beta_{01})'$. For an inception cohort, without loss of generality we consider the period of observation $[0, \tau]$ with $\tau = 1$. We let $N_{01}(\tau)$ indicate that a $0 \to 1$ transition occurred over $[0, \tau]$. We set $\beta_{01} = \log 1.5$ and the values of the baseline intensities to satisfy the following constraints:

(i) $P_1 = P(N_{01}(\tau) = 1|X = 0) = \lambda_{010}/(\lambda_{010} + \lambda_{020})\{1 - e^{-(\lambda_{010}+\lambda_{020})\tau}\} = \{0.25, 0.75\}$;

(ii) $\lambda_{120}/\lambda_{020} = 1.5$; and

(iii) $P_2 = P(Z(\tau) = 2|X = 0) = \{0.1, 0.5\}$.

We assume the progression status is assessed intermittently at $J = 5$ equally spaced scheduled assessments over $[0, \tau]$. For the dropout process, we set the under-observation indicator $Y_0 = 1$ at baseline (e.g. time V_0) for all individuals and generate Y_j given (Y_{j-1}, Z_{j-1}) sequentially for $j = 1, 2, \ldots, J$. As described in Section 2.2.2, $P(Y_j = 1|Y_{j-1} = 0) = 0$ and $P(Y_j = 1|Y_{j-1} = 1, Z_{j-1} = 2) = 1$. For the SMAR mechanism, we set $\operatorname{logit} P(Y_j = 0|Y_{j-1} = 1, \bar{Z}_{j-1}, X; \gamma) = \gamma_0 + \gamma_1 I(Z_{j-1} = 1)$, that is the odds of drop-out at a given assessment depends on the disease status at the previous assessment. The value of the parameters (γ_0, γ_1) are set to achieve an overall percentage of dropout of $P(\Delta = 0) = \{0.4, 0.8\}$ and an odds ratio of dropout for individuals with previous disease status $Z_{j-1} = 1$ vs $Z_{j-1} = 0$ of $\exp(\gamma_1) = 2$.

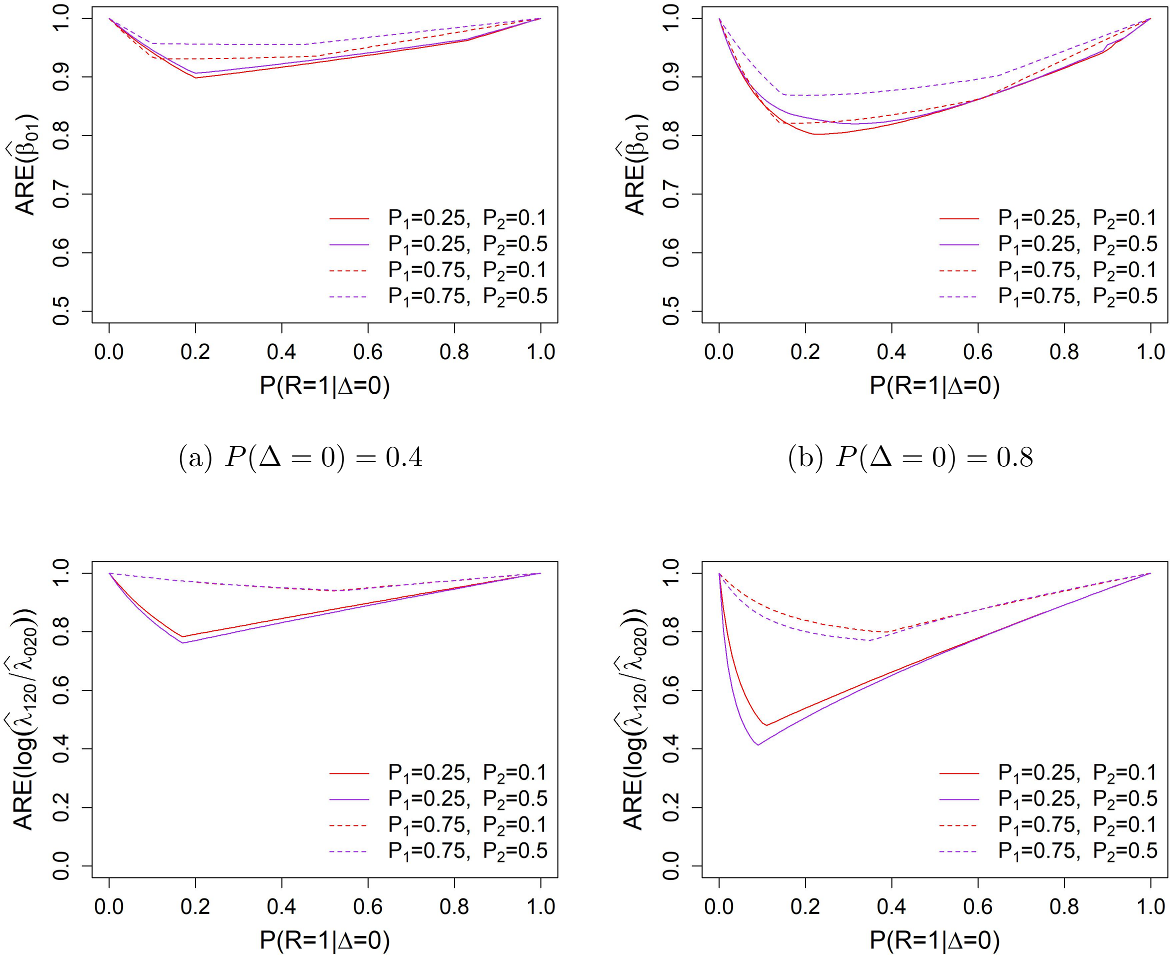
We adopt the following model for the selection of individuals for tracing

$$\operatorname{logit} P(R = 1|Z_C, X, \Delta = 0; \rho) = \rho_0 + \rho_1 I(Z_C = 1) + \rho_2 X + \rho_3 I(Z_C = 1)X \tag{M1}$$

where X is the same binary covariate related to the $0 \to 1$ transition. To illustrate the magnitude of potential efficiency gains from tracing as well as influential factors, we compare the asymptotic variance of estimates of response parameters based on an optimal design versus a simple random sampling (SRS) design (which is equivalent to setting the tracing model parameters to be $\rho^{srs} = (\rho_0, 0, 0, 0)$). The optimal tracing parameter ρ^{opt} results in the minimal asymptotic relative efficiency

$$ARE(\widehat{\theta}_k) = \frac{[\mathcal{I}^{-1}(\theta, \gamma, \rho^{opt})]_{kk}}{[\mathcal{I}^{-1}(\theta, \gamma, \rho^{srs})]_{kk}} , \tag{2.10}$$

subject to a pre-specified proportion of tracing $\pi = P(R = 1|\Delta = 0)$.



(a) $P(\Delta = 0) = 0.4$

(b) $P(\Delta = 0) = 0.8$

(c) $P(\Delta = 0) = 0.4$

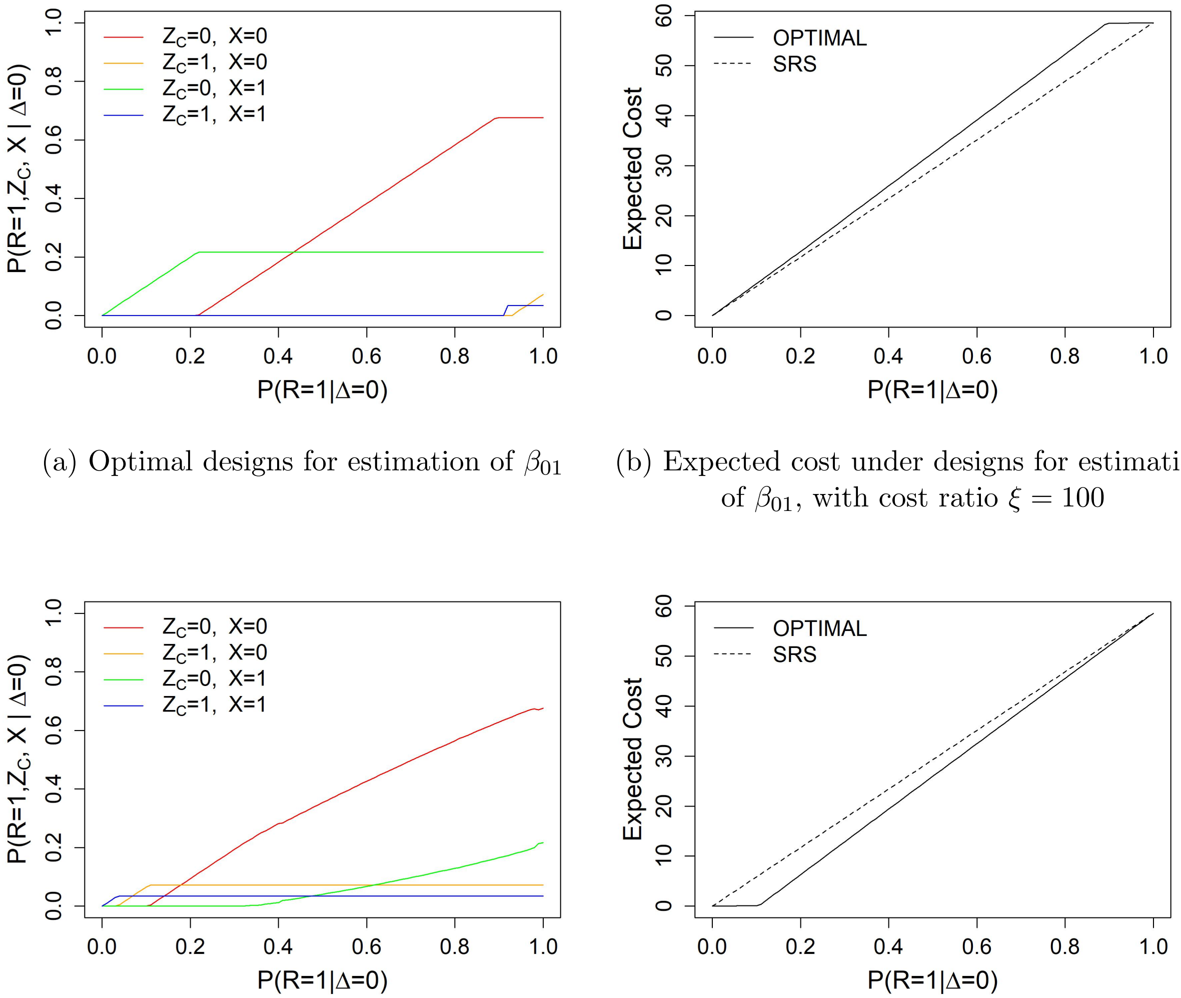
(d) $P(\Delta = 0) = 0.8$

Figure 2.1: Asymptotic relative efficiency (2.10) of the estimator $\widehat{\beta}_{01}$ (top panels) and $\log(\widehat{\lambda}_{120}/\widehat{\lambda}_{020})$ (bottom panels) with a tracing study under an optimal design versus a SRS design of the same expected size; $P_1 = P(N_{01}(\tau) = 1 \mid X = 0)$, $P_2 = P(Z(\tau) = 2 \mid X = 0)$, $\lambda_{120}/\lambda_{020} = 1.5$, $\beta_{01} = \log 1.5$, $P(X = 1) = 0.25$, and $J = 5$

As expected, the optimal tracing designs lead to more precise estimates than the SRS designs across all scenarios. This is depicted in Figure 2.1 for the estimation of covariate effect β_{01} (top panels) and $\log(\lambda_{120}/\lambda_{020})$ (bottom panels). Across all parameter configurations considered, the gain in efficiency increases with the probability of dropout $P(\Delta = 0)$. The magnitude of the gain in efficiency also varies as a function of the parameters of the disease process (as represented by the multiple curves in each panel) and the marginal tracing probability π. While these relationships are complex and dependent on properties of the disease process, we describe some general trends apparent in the present examples. When interest lies in estimating the covariate effect modulating the $0 \to 1$ transition (β_{01}), the smaller the percentage of progression by the administrative censoring time τ (i.e. P_1), the greater the gain in efficiency achieved by the optimal tracing scheme relative to the SRS approach. This is due to the fact that the optimal design for estimation of β_{01} prioritizes tracing progression-free individuals ($Z_C = 0$) over those who have already progressed ($Z_C = 1$) as the former may potentially provide new information on the $0 \to 1$ transition directly; this can be seen in panels (a) and (b) of Figure 2.1 when contrasting the solid ($P_1 = 0.25$) and dashed ($P_1 = 0.75$) lines of the same colours with the fixed P_2. This trend is much clearer for estimating the relative risk of death $\log(\lambda_{120}/\lambda_{020})$ as shown in Figure 2.1 (c) and (d). The percentage of death observed by the administrative censoring time (i.e. P_2) also has some impact on the estimation of a covariate effect on progression, β_{01}. The lower P_2 is (e.g. 0.1 versus 0.5) the bigger the gain in efficiency by adopting the optimal design for tracing, although such a difference is only appreciable when the percentage of progression is high ($P_1 = 0.75$) as shown in Figure 2.1 (a) and (b). Interestingly, P_2 seems to have a different impact on efficiency gain for the estimation of the relative risk of death $\log(\lambda_{120}/\lambda_{020})$. When the drop-out rate is high ($P(\Delta = 0) = 0.8$), Figure 2.1 (d) shows slightly greater benefit of the optimal tracing scheme over SRS as the percentage of death P_2 increases, but such a pattern is only noticeable if the percentage of tracing is low to moderate (e.g. $\pi < 0.4$).

In summary, efficiency gains for the estimation of both a covariate effect on progression β_{01} and the relative risk of death $\log(\lambda_{120}/\lambda_{020})$ are primarily driven by observing instances of disease progression, e.g P_1. The probability of death during the follow-up period, P_2, has some additional impact depending on which quantity is of interest for estimation. Slightly larger gains in efficiency for the estimation of β_{01} can be obtained when P_2 is low, because as P_2 increases the likelihood of gaining information about progression at the time of tracing decreases. However, when interest lies in estimating $\log(\lambda_{120}/\lambda_{020})$, observation of death events are more informative and so larger gains in efficiency are achieved by the proposed approach when P_2 is higher.

Figure 2.2 focuses on the setting with $P_1 = 0.25$, $P_2 = 0.1$, and $P(\Delta = 0) = 0.8$ (e.g. the solid red line in the right-hand panels of Figure 2.1), again considering estimation of β_{01} and $\log(\lambda_{120}/\lambda_{020})$ in the top and bottom panels respectively. The left-hand panels contain plots of the joint probability $P(R = 1, Z_C, X|\Delta = 0)$ against the marginal probability of tracing π under an optimal design. As will be discussed in Section 2.3, it is generally reasonable to assume that the cost of tracing individuals for vital status (C_s) is substantially lower than that of assessing disease status (C_d), so $\xi = C_d/C_s \geq 1$. In the right-hand panels of Figure 2.2 we fix $\xi = 100$ and observe that the expected cost of the proposed optimal tracing scheme (solid line) is greater than that of a SRS tracing scheme (dashed line) for the estimation of β_{01}, whereas it is lesser for the estimation of $\log(\lambda_{120}/\lambda_{020})$. This follows directly from the patterns exhibited in the left-hand panels: the optimal tracing scheme for β_{01} preferentially selects individuals with $Z_C = 0$ (more expensive) over those with $Z_C = 1$ (less expensive), while the optimal scheme for $\log(\lambda_{120}/\lambda_{020})$ prioritizes individuals with $Z_C = 1$ over those with $Z_C = 0$. It is interesting to note that for the latter case, the



(a) Optimal designs for estimation of β_{01}

(b) Expected cost under designs for estimation of β_{01}, with cost ratio $\xi = 100$

(c) Optimal designs for estimation of $\log(\lambda_{120}/\lambda_{020})$

(d) Expected cost under designs for estimation of $\log(\lambda_{120}/\lambda_{020})$, with cost ratio $\xi = 100$

Figure 2.2: Optimal tracing design (left-hand panels) and expected cost (right-hand panels) under an optimal vs a SRS design for the estimation of β_{01} (top panels) and $\log(\lambda_{120}/\lambda_{020})$ (bottom panels), with $P_1 = 0.25$ and $P_2 = 0.1$, $\lambda_{120}/\lambda_{020} = 1.5$, $\beta_{01} = \log 1.5$, $P(X = 1) = 0.25$, $J = 5$, and $P(\Delta = 0) = 0.8$

optimal design not only leads to substantial gains in efficiency, but is also more economical than the SRS design of the same size. In addition, the optimal scheme for β_{01} sequentially draws upon individuals with $(Z_C = 0, X = 1)$, $(Z_C = 0, X = 0)$, $(Z_C = 1, X = 1)$, and $(Z_C = 1, X = 0)$; preferring the former subgroups to the exclusion of the latter, as the marginal probability of selection for tracing (π) increases. However, this is not true to the same extent in the optimal scheme for $\log(\lambda_{120}/\lambda_{020})$; in this case, the proposed tracing scheme allows for the optimal equilibrium to be identified, which would not otherwise be possible. The results of extensive simulation studies (not shown) demonstrate excellent agreement between the asymptotic and empirical efficiency gains.

2.2.5 Selection Incorporating the Time of Study Withdrawal

When constructing selection models for tracing it is desirable to balance the inclusion of key factors with the need for parsimony in order to minimize the computational burden at the optimization step of the selection model. Here we illustrate the potential gains in efficiency from adopting a more general class of selection models compared to (M1), which included only the information on the state occupied at the last assessment (denoted Z_C) and a binary covariate X. Specifically here we consider a selection model of the form

$$\operatorname{logit} P(R = 1 | Z_C, X, \Delta = 0; \rho) = \rho_0 + \rho_1 I(Z_C = 1) + \rho_2 X + \rho_3 I(Z_C = 1)X + \rho_4 D \quad \text{(M2)}$$

to allow tracing selection probabilities to further depend on $D = \tau - A_C$, the time from loss-to-follow-up to tracing. Since the tracing selection model in (M1) is nested in (M2), greater efficiency gains may be realized under the latter model. The benefit of including time since loss-to-follow-up in the tracing selection model is most appreciable for the estimation of relative risk of death with, versus without, progression given by $\log(\lambda_{120}/\lambda_{020})$. A summary of results comparing asymptotic efficiency gains under these two tracing selection models

		Size Constraint (π)			Cost Constraint (ξ)		
Estimand	Tracing Model	0.05	0.25	0.50	5	20	100
β_{01}	M1	0.908	0.803	0.838	0.836	0.852	0.951
	M2	0.881	0.787	0.811	0.795	0.800	0.918
$\log(\lambda_{120}/\lambda_{020})$	M1	0.620	0.571	0.721	0.664	0.487	0.418
	M2	0.543	0.570	0.719	0.664	0.487	0.418

Table 2.1: Asymptotic relative efficiency (2.10) of estimators (optimal versus SRS tracing design) under tracing selection models in (M1) and (M2); $\pi = P(R = 1 \mid \Delta = 0)$ is the marginal probability of selection for tracing and $\xi = C_d/C_s$ is the relative cost of determining disease status compared to survival status; with $P_1 = 0.25$, $P_2 = 0.1$, $\lambda_{120}/\lambda_{020} = 1.5$, $\beta_{01} = \log 1.5$, and $P(\Delta = 0) = 0.8$

is presented in the left-hand columns of Table 2.1 under the heading Size Constraint; we defer the discussion of the right-hand side under the heading Cost Constraint to Section 3.2. Here we find the efficiency gains can be appreciable for both β_{01} and $\log(\lambda_{120}/\lambda_{020})$. We also see a non-monotonic trend in relative efficiency of "optimal" versus simple random sampling when viewed as a function of the marginal selection probability for tracing, which are similar to the trends of the red solid curves in Figure 2.1 (b) and (d).

2.3 Design with a Budgetary Constraint

2.3.1 Formulation of the Optimization Problem

In general, the cost associated with tracing individuals known to be diseased before loss-to-follow-up (i.e. those with $Z_C = 1$) is lower than that for individuals without the disease (i.e. $Z_C = 0$); in this section we exploit this fact to design optimal tracing schemes subject to more realistic budget constraints. For the former group, the only information that we can learn is the survival status at the time of tracing, but for the latter group, disease status may also be ascertained for individuals who are still alive at tracing. Let C_s and C_d denote the cost for tracing survival status and disease status respectively, and let $\xi = C_d/C_s$ be the cost ratio; we assume $\xi \geq 1$ in general.

Suppose we have a fixed budget for conducting the tracing study where we plan to trace the survival status among all the selected individuals first, and then the disease status among those who were disease-free at their last assessment and are alive at tracing. Based on a Poisson sampling process with a tracing selection model, the expected cost of tracing is

$$\begin{aligned}\dot{C}(\rho; C_s, \xi) = nP(\Delta = 0) \sum_{Z_J, \mathcal{D}} & P(\mathcal{D}|\Delta = 0) P(R = 1 \mid \mathcal{D}, \Delta = 0; \rho) \\ & \times P(Z_J | R = 1, \mathcal{D}, \Delta = 0; \theta) C_s \Big[1 + \xi I(Z_J \neq K, Z_C = 0) \Big].\end{aligned}$$

Note that the right side of this equation depends on the parameter ρ in the tracing selection model, whereas the expected number of individuals eligible for tracing, $nP(\Delta = 0)$, and the distribution of observed data among the eligible individuals, $P(\mathcal{D}|\Delta = 0)$, are known after collection of phase I data. In addition, under the assumption (2.4) the probability $P(Z_J|R, \mathcal{D}, \Delta = 0; \theta)$ can be estimated by $P(Z_J \mid \bar{Z}_C, X; \tilde{\theta})$ where $\tilde{\theta}$ is the MLE obtained from phase I. This implies that if one is interested in precise estimation of θ_k, for a given fixed total budget B, cost C_s and ratio ξ, we can optimize the selection model by minimizing

$$\big[I_H^{-1}(\tilde{\theta}, \rho) \big]_{kk} + \zeta \big[\dot{C}(\rho; C_s, \xi) - B \big] \tag{2.11}$$

which is like (2.8) but with a cost constraint in place of a constraint simply on the expected sample size.

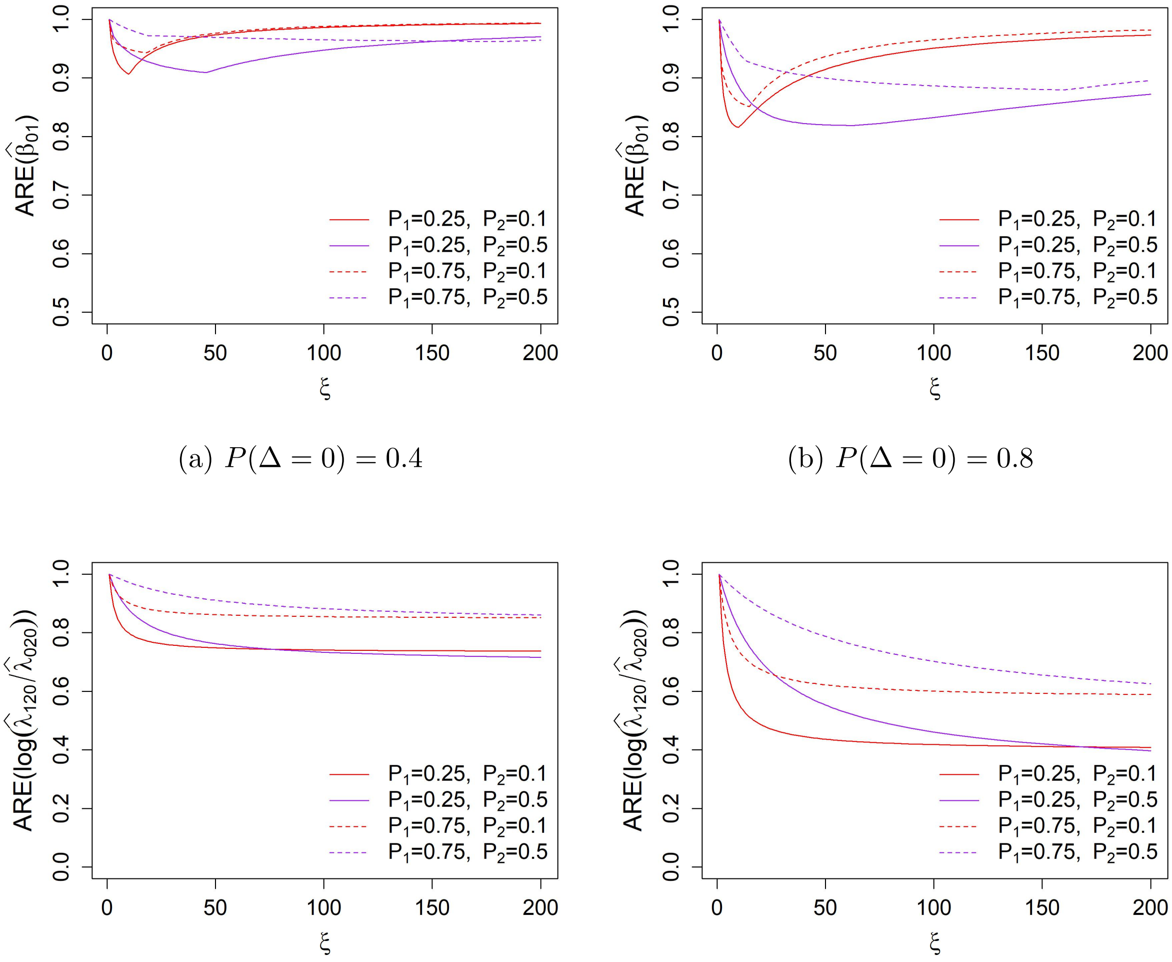
2.3.2 Efficiency Gains from Optimal Tracing with Cost Constraints

The study setting here parallels that of Section 2.2.4, with the exception that the constraint is imposed on the budget rather than the size of the sample selected for tracing. We set the maximum budget $B = \dot{C}(\rho; C_s, \xi = 1)$ to equal the expected cost of tracing all eligible individuals when $\xi = 1$. The budget constraint in (2.11) then becomes

$$\dot{C}(\rho; C_s, \xi) - B \propto \sum_{Z_J, \mathcal{D}} P(\mathcal{D}|\Delta = 0) P(R = 1 \mid \mathcal{D}, \Delta = 0; \rho) \\ \times P(Z_J|R, \mathcal{D}, \Delta = 0; \theta)(\xi - 1) I(Z_J \neq K, Z_C = 0)]$$

which only depends on the cost ratio ξ and the selection parameter ρ. We consider values of ξ from 1 to 200 and the same values of $(\theta', \gamma')'$ as in Section 2.2.4, where $\theta = (\lambda', \beta')'$.

Figure 2.3 displays the patterns of relative efficiency exhibited by the optimal tracing selection probabilities under a cost constraint with selection model (M1), which are similar to those observed under the size constraint in the previous section (see Figure 2.1). In fact, in some sense this cost constraint amounts to a transformation of the size constraint. That is, due to the choice of budget constraint B, setting $\xi = 1$ implies that all eligible individuals may be traced (e.g. $\pi = P(R = 1|\Delta = 0) = 1$); thus, the left-most points in each panel of Figure 2.3 correspond to the right-most points in the analogous panels of Figure 2.1. On the other hand, as $\xi \to \infty$, the cost of tracing individuals with $Z_C = 0$ becomes prohibitively expensive, and $\lim_{\xi \to \infty} P(R = 1|\Delta = 0, Z_C = 0) = 0$. Thus, if individuals with $Z_C = 0$ furnish more information upon tracing, as is the case for β_{01} (see Figure 2.4 (a)), then $\lim_{\xi \to \infty} P(R = 1|\Delta = 0) = 0$. On the other hand, if the optimal scheme prioritizes tracing individuals with $Z_C = 1$, as is the case for $\log(\lambda_{120}/\lambda_{020})$ (see Figure 2.4 (b)) , $\lim_{\xi \to \infty} P(R = 1|\Delta = 0) = P(Z_C = 1|\Delta = 0)$.



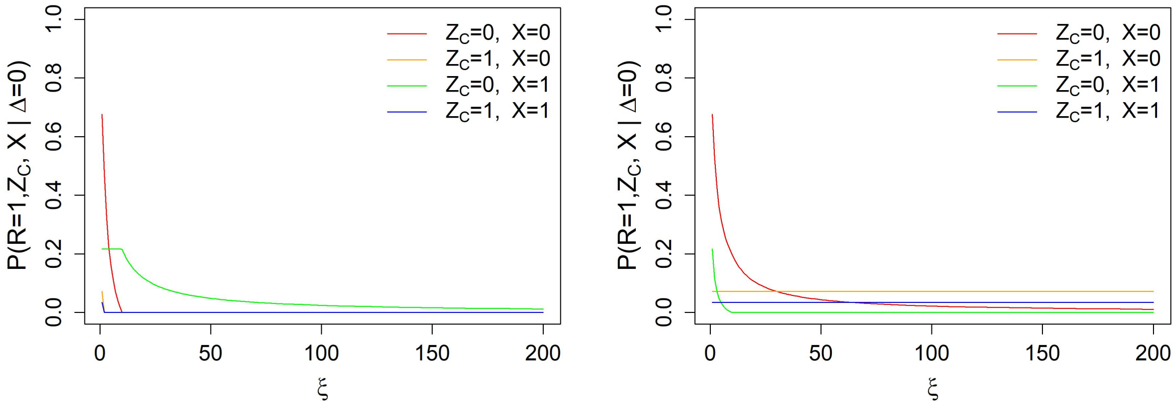
(a) $P(\Delta = 0) = 0.4$

(b) $P(\Delta = 0) = 0.8$

(c) $P(\Delta = 0) = 0.4$

(d) $P(\Delta = 0) = 0.8$

Figure 2.3: Asymptotic relative efficiency (2.10) of estimators for biomarker effect $\widehat{\beta}_{01}$ (top panels) and $\log(\widehat{\lambda}_{120}/\widehat{\lambda}_{020})$ (bottom panels) with a tracing study under an optimal design vs a SRS design of the same expected cost; $P_1 = P(N_{01}(\tau) = 1 \mid X = 0)$, $P_2 = P(Z(\tau) = 2 \mid X = 0)$, $\lambda_{120}/\lambda_{020} = 1.5$, $\beta_{01} = \log 1.5$, $P(X = 1) = 0.25$, $J = 5$, and cost ratio $\xi = C_d/C_s$



(a) Optimal Design for Estimation of β_{01}

(b) Optimal Design for Estimation of $\log(\lambda_{120}/\lambda_{020})$

Figure 2.4: Optimal tracing design under a fixed budget constraint for the estimation of β_{01} (left panel) and $\log(\lambda_{120}/\lambda_{020})$ (right panel), with $P_1 = 0.25$, $P_2 = 0.1$, $\lambda_{120}/\lambda_{020} = 1.5$, $\beta_{01} = \log 1.5$, $P(X = 1) = 0.25$, $J = 5$, and $P(\Delta = 0) = 0.8$.

To compare the two selection models (M2) and (M1), the right-hand columns of Table 2.1 contain the asymptotic relative efficiencies under both models with the cost constraints. We see that using either of these two models results in appreciable efficiency gains, where the gain decreases as the cost ratio ξ increases from 5 to 100 for β_{01} but it increases as ξ increases for $\log(\lambda_{120}/\lambda_{020})$. These are consistent with the red solid curves showed in Figure 2.3 (b) and (d). We also see the efficiency gains under (M2) are greater than those under (M1) for the estimation of β_{01} in most cases and they are very similar for $\log(\lambda_{120}/\lambda_{020})$. This is because for the latter, the optimal tracing scheme prioritizes tracing individuals with $Z_C = 1$, and the optimal selection probability of those is 1 under the settings considered here. As the cost for tracing disease status becomes more expensive (e.g. ξ increases), the optimal selection probability for individuals with $Z_C = 0$ quickly approaches 0. As a consequence the time from loss-to-follow-up to tracing has very little room to influence the selection probabilities under the optimal design, leading to comparable results under the two selection models.

2.4 Application to the University of Toronto Psoriatic Arthritis Cohort Study

Scientists at the University of Toronto Psoriatic Arthritis Clinic have created and maintained a registry of individuals with psoriatic arthritis which continues to be an invaluable resource in deepening understanding of the progression of psoriatic arthritis and related comorbidities; see Section 1.5.1. A scientific question of primary interest is in estimating the incidence of arthritis mutilans in individuals with psoriatic arthritis, and estimating the effect of the marker HLA-B27 on the disease progression taking into account the full disease process including death. This process may be viewed as an illness-death process as in Figure 1.1(c) where states 0, 1, and 2 represent psoriatic arthritis, arthritis mutilans, and death respectively.

The cohort we focused on consists of 870 individuals with psoriatic arthritis and they are scheduled to come to the clinic for assessments on an annual basis. We take December 2016 as the end of phase I follow up, and use the patients records until then as phase I data. While variability arises in practice, this protocol informs the decision to view individuals who have not been seen for 2+ years as being lost to follow-up; this leads to 72% of the cohort being eligible for tracing. In total 152 (17.5%) are observed to develop arthritis mutilans and 147 deaths are recorded (16.9%), including 36 among individuals whose disease progressed. Further, 56 individuals (6.4%) are positive for the HLA-B27 marker. Phase-I maximum likelihood estimates were obtained using the `R` package `msm`

[Jackson, 2011]. We assume the visit times are uninformative and that data are missing sequentially at random. The proposed approach is applied to demonstrate possible optimal designs for a tracing study conducted in January 2017.

Table 2.2 reports the optimal tracing probabilities $P(R = 1|Z_C, X, \Delta = 0)$ arising from selection model (M1) under the constraint of a fixed sample size or cost respectively. It is apparent that if interest lies in estimating β_{01} one should first select all individuals who were not observed to progress before they withdrew from the study in phase I (i.e. $Z_C = 0$) and then individuals who have progressed (i.e. $Z_C = 1$), as long as the fixed sample size permits; this trend also holds when the budget is constrained. On the other hand, when interest lies in estimating $\log(\lambda_{120}/\lambda_{020})$, it always prioritizes individuals known to have progressed (e.g. with $Z_C = 1$) under both the sample size and budget constraints, since only survival status, which is less expensive, needs to be determined. We also considered using tracing selection model (M2), which leads to very similar gains in efficiency as when using model (M1). We did not report the optimal tracing probability here as it varies continuously with respect to time since loss-to-follow-up, $D = \tau - a_C$. For the psoriatic arthritis cohort, the proposed optimal tracing study design can lead to gains in efficiency of 10-30% relative to using a SRS design.

	Size Constraint						Cost Constraint					
		Strata (Z_C, X)						Strata (Z_C, X)				
	π	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	RE(%)	ξ	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	RE(%)
β_{01}	5%	0.00	0.95	0.00	0.00	71.1	3	0.54	1	0.00	0.00	85.4
	25%	0.25	1	0.00	0.00	76.2	5	0.34	1	0.00	0.00	80.1
	50%	0.57	1	0.00	0.00	84.0						
$\log(\lambda_{120}/\lambda_{020})$	5%	0.00	0.00	0.37	0	89.7	3	0.55	0.00	1	1	87.3
	25%	0.12	0.00	1	1	80.5	5	0.37	0.00	1	1	82.4
	50%	0.43	0.00	1	1	87.1						
Stratum size		495	33	84	14			495	33	84	14	

Table 2.2: Optimal selection probabilities by strata based on model (M1) for tracing psoriatic arthritis/mutilans cohort, where $\pi = P(R = 1|\Delta = 0)$ is the marginal probability of selection for tracing, RE is the relative efficiency of adopting the proposed tracing design as opposed to SRS, and $\xi = C_d/C_s$ is the relative cost of determining disease status compared to survival status

2.5 Discussion

In this chapter we consider the framework of an inception cohort study with regularly scheduled assessments. We consider the implications of loss-to-follow-up and the idea of conducting a tracing study to track down individuals who have withdrawn to obtain updated information on their health; this is planned at the end of phase I of a study. We discuss the utility of attempts to optimally select individuals lost to follow-up for the tracing study in order to maximize the value of the information gained. In our multistate setting, the optimization may be carried out with a view to maximizing the precision of transition intensities, state occupancy probabilities, or the effects of fixed (e.g. genetic) markers on disease progression. Less focused criteria can also be employed which minimize functions of information matrices. We have focused on progressive processes, but settings with reversible or alternating processes are also common and the methods can in principle be extended to deal with these types of data. Due to the complexity of the function to be optimized (e.g. the inverse of the information matrix), we suggest that care be taken to select several plausible initial values for the ρ vector to ensure the global minimum is identified. For example, when some of the strata induced by the tracing selection model are small, it is advisable to set ρ corresponding to tracing all and none of the individuals in the strata as initial values; this is due to the fact that variation in the corresponding ρ values are unlikely to have a large effect on the target of optimization, which may make optimization challenging.

We have assumed a time-homogeneous Markov model with regularly scheduled assessment times. Tracing studies can of course be designed for non-homogeneous Markov models using piecewise-constant baseline intensities. The regularity of scheduled assessments makes it reasonably straightforward to determine which individuals are lost to follow-up and therefore eligible for tracing. In settings where assessment times are less regular and left to the discretion of patients, it is more challenging to define the set of individuals who are lost to follow-up and eligible for tracing. One can discretize time in such settings and declare individuals not seen in several potential periods as lost to follow-up. We also assumed a progressive disease process in which all states can only be entered once. It is well-known that transition intensities involving recurring states are more poorly estimated under panel observation schemes [Lange and Minin, 2013, Ma et al., 2016]. Moreover when assessments are far apart in time (relative to dynamic features of the process of interest) estimates of transition intensities are less efficient compared to when the assessments are closer in time; the effect of widely spaced assessment times is smaller on other features such as state occupancy probabilities. These issues pertain to the conduct of tracing studies for non-progressive processes so one should expect a smaller gain in efficiency from "optimally" tracing individuals in reversible processes.

The likelihood we constructed presumes that individuals selected for tracing do, in fact, furnish the required information. With respect to survival status, death records can be searched and so this can be acquired independently of family engagement, but it may ultimately not be possible to determine even survival status for individuals who have moved away. In such situations the realized gain in precision may be less than anticipated. Information on progression status, which is more dependent on individual participation, may not be readily available because of initial refusals, or may require a number of attempts to secure data. In such cases it may be necessary to build and integrate more elaborate models for the tracing process which characterize the data acquisition process. Interestingly, even if a SMAR mechanism governs attrition, data may become missing not at random if the individuals responding to tracing comprise a biased subset of those selected for tracing. Thus if tracing is incompletely executed, modelling the success of the tracing process may be important to make suitable adjustments to the likelihood. Data on the outcome of each attempt to contact individuals should be recorded to facilitate fitting of models for the response process in tracing studies. Similar modelling exercises have been done in settings where the tracing selection mechanism is non-ignorable due to truncation in the cohort using likelihood and pseudo-likelihood approaches [Titman et al., 2011].

Chapter 3

Cohort Study Designs for Markov Processes

3.1 Introduction

In the previous chapter, we considered the design of tracing studies to recover information on individuals lost to follow-up. We now turn attention to the initial stage of design for longitudinal studies, prior to recruitment and prospective follow-up. We develop design criteria for a longitudinal study with a three-state illness-death process (see Figure 1.1(c)) in which individuals are under intermittent observation according to a protocol. We consider the case in which disease progression status is observed intermittently, but transitions into the death state are observed subject to right censoring. The proposed methods can be applied to a more general framework of multistate processes, as shown in Figure 1.2. The remainder of this chapter is organized as follows. In Section 3.2, we consider the design of such cohort studies, examine the interplay between the design factors such as the sample size and the frequency of assessments and disease process parameters, and study their effect on statistical power. We also define minimum-cost designs which achieve a desired level of power for the estimation of a regression coefficient, and look at the relationship between features of the disease process and of minimum-cost designs. In Section 3.3, we derive the form of the Fisher information matrix for longitudinal studies in which there is misclassification in the states recorded at inspection times and use this to evaluate the impact of misclassification on study design subject to cost constraints. In Section 3.4 we discuss extending the proposed approach to more response-adaptive protocols, where the assessment protocol may vary as a function of the observed longitudinal responses. Concluding remarks are made in Section 3.5.

3.2 Prospective Cohort Studies

3.2.1 Maximum Likelihood Estimation for Markov Models

Prospective cohort studies are commonly employed to collect data on life history processes. This involves acquiring a sample of size n from a population of individuals and tracking the occurrence of the event of interest (generally referred to as disease event) longitudinally over a certain follow-up period. It is generally infeasible to monitor individuals' disease status continuously, thus assessments are made intermittently at J specified time points $0 = a_0 < a_1 < \cdots < a_J = \tau$ over the study period $(0, \tau]$ although the exact event times are not available. On the other hand, the vital status is often tracked in continuous time and the exact death time is typically known or subject to right censoring if the participants become lost to follow-up at time C before the end of study. As such, the multistate data arising from longitudinal cohort studies may be mixed in its nature: disease status data may be available under a panel observation scheme, along with exact or right-censored death data. Let T_1 be the time to disease progression, T_2 be the time to death, $T^\dagger = \min(T_2, C, \tau)$ denote the minimum of the time to death and censoring and $\delta = I(T^\dagger = T_2)$ indicate that death is observed (see Figure 3.1). Let $\bar{Z}_j = (Z(a_1), \ldots, Z(a_j))$ denote the

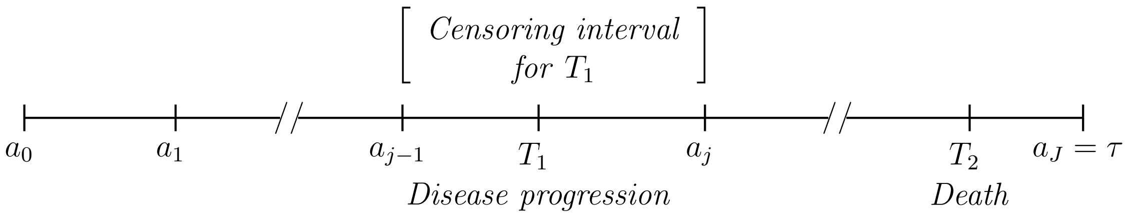


Figure 3.1: Schematic for mixed observation scheme, where the time of disease progression (T_1) is subject to interval-censoring and the time of death (T_2) is subject to right censoring.

history of the observed disease status up to and including assessment j, $j = 1, \dots, J$ and $M = \max\{j\,;\, a_j < T^\dagger, j = 0, \dots, J\}$ be the random number of assessments for an individual prior to right censoring or death. Under a Markov model indexed by the parameter vector θ in general, the likelihood contribution from a single subject is written as

$$L(\theta) = P(\bar{Z}_m, t^\dagger, \delta) = \prod_{j=0}^{m-1} P\big(Z(a_{j+1}) \mid Z(a_j)\big) \sum_{\ell=0}^{1} P\big(Z(t^\dagger) = \ell \mid Z(a_m)\big) \lambda_{\ell 2}^{\delta}(t^\dagger) \tag{3.1}$$

where the summation accounts for the fact that the disease status right before death or censoring may not be known due to the intermittent observation scheme. The estimates of θ can be obtained by maximizing the product of terms having the form of (3.1) over a sample of independent subjects. Instead of using a Newton-Raphson algorithm, a simple Fisher scoring method was proposed by Kalbfleisch and Lawless [1985] for obtaining the MLEs in which only first derivatives are required; this can be adapted to deal with observed times of death as shown by Zeng et al. [2018].

We assume censoring is independent of the disease processes. We let $Y_k(t) = I(Z(t) = k)$ indicate that an individual is in state k at time t, $Y^\dagger(t) = I(t \leq T^\dagger)$ indicate they are under observation (i.e. alive and uncensored), and $Y_k^\dagger(t) = Y^\dagger(t)Y_k(t)$ indicate that they are under observation and in state k at time t, $k < 2$. For an individual who is under observation at a_{j-1} with $Z(a_{j-1}) = k$ (i.e. $Y_k^\dagger(a_{j-1}) = 1$), the partial log-likelihood contribution pertaining to the disease process for the jth interval $\mathcal{A}_j = [a_{j-1}, a_j)$ is

$$\ell_{kj} = \sum_{\ell=0}^{1} Y_\ell^\dagger(a_j) \log \Big[p_{k\ell}(a_{j-1}, a_j)\Big] + \big(1 - Y^\dagger(a_j)\big) \log \left[\sum_{\ell=0}^{1} p_{k\ell}(a_{j-1}, t^\dagger)\lambda_{\ell 2}^{\delta}(t^\dagger)\right],$$

where $k < 2$. The Fisher information matrix thus takes the form

$$\mathcal{I} = \sum_{x=0}^{1} \sum_{j=1}^{J+1} \int_{a_{j-1}}^{a_j} \sum_{q=1}^{j} \sum_{k=0}^{1} E\left[Y_k(a_{q-1}) \frac{\partial \ell_{kj}}{\partial \theta} \frac{\partial \ell_{kj}}{\partial \theta'} \mid C = c, X = x\right] dG(c) P(X = x) \tag{3.2}$$

where $G(\cdot; \rho)$ is the distribution function for censoring time C indexed by parameter ρ and $\mathcal{G}(t) = 1 - G(t)$. The calculation details of the conditional expectation in the inner summation of (3.2) can be found in Zeng et al. [2018]. Note that the construction of the Fisher information relies on transition probabilities and their first derivatives. Under the piecewise-constant model, the transition probability matrix can be obtained using (1.2), and its first derivatives can be taken in a straightforward manner. For the illness-death process for example, suppose the derivatives are taken with respect to the vector of constant transition intensities associated with the rth piece $\mathcal{B}_r$, $\lambda^{(r)} = (\lambda_{01}^{(r)}, \lambda_{02}^{(r)}, \lambda_{12}^{(r)})'$. Then we will simply have

$$\frac{\partial}{\partial \lambda^{(r)}} p_{k\ell}(s,t) \quad \sum_{z_{r-1}, z_r} p_{k, z_{r-1}}(s, v_{r-1}) \left[\frac{\partial}{\partial \lambda^{(r)}} p_{z_{r-1}, z_r}(v_{r-1}, v_r)\right] p_{z_{r-1}, \ell}(v_r, t)$$

where $v_{r-1} = \max\{s, b_{r-1}\}$, $v_r = \min\{t, b_r\}$, and $p_{k\ell}(v_{r-1}, v_r) = 0$ if $v_{r-1} > v_r$. The time-homogeneous model can be viewed as a special case with constant transition intensities over the whole time span, and the above calculations can be much further simplified.

3.2.2 Design Choices: Sample Size and Number of Assessments

Prospective cohort studies are generally very costly, so careful consideration should be given to the design of such studies in multiple dimensions such as sample size, frequency of assessments, timing of the assessments and duration of follow-up. These design factors jointly affect both the estimation precision and the cost of the study itself. In practice, the choices for these design factors are often driven by logistical reasons. While several authors have suggested that the assessment frequency should be justified carefully [Collins and Graham, 2002, Nesselroade, 1991], this is not commonly done in the clinical literature [Timmons and Preacher, 2015]. In the present framework, we present a more formal approach to choose the sample size and frequency of assessments, by deriving the asymptotic variance of the estimates of interest and using this as the basis for study design.

Suppose the primary interest of a cohort study lies in the estimation of the effect of a covariate on the 0 → 1 transition (e.g. disease progression). As described in Section 1.2, we assume a binary covariate X has a multiplicative effect on the $0 \rightarrow 1$ transition, with intensity $\lambda_{01}(t) = \lambda_{010}(t)\exp(X'\beta)$ under a Markov model. The estimator obtained from fitting the Markov models described in Section 3.2.1 has the following asymptotic distribution

$$\sqrt{n}(\widehat{\beta} - \beta) \sim N\big(0, \mathcal{I}^{\beta\beta}(\theta, \rho, J, \tau)\big) \ , \tag{3.3}$$

where n is the sample size, $\mathcal{I}^{\beta\beta}(\theta, \rho, J, \tau) = [\mathcal{I}^{-1}(\theta, \rho, J, \tau)]_{\beta\beta}$ is the asymptotic variance of $\hat{\beta}$, which is the (β, β) element of the inverse of the Fisher information $\mathcal{I}(\theta, \rho, J, \tau)$ given in (3.2). The asymptotic variance depends on parameters θ and ρ from the disease and censoring processes respectively, as well as on the design factors including the number of assessments (J), the assessment times $(a_j,\ j = 1, \ldots, J)$, and the administrative censoring time (τ). The dependence on the actual assessment times is suppressed for convenience since we assume here that the assessment times are fixed and evenly scheduled over the interval $(0, \tau]$. It is straightforward to extend this work to irregular assessment times as long as the visit process is independent of the disease process. Response-dependent assessment processes and their impact on estimation and study design are further discussed in Section 3.5. The parameters (θ, ρ, J, τ) have an impact on the power when the asymptotic distribution (3.3) is used for inference. Following the argument of Demidenko [2007], the power for a two-sided Wald-test of $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$ at a significance level of α_1 for detecting an effect of size $\beta = \beta_A$ is

$$\text{power}(\beta) = \Phi\left(-z_{\alpha_1/2} - \frac{\beta_0 - \beta_A}{\sqrt{\mathcal{I}_A^{\beta\beta}(\theta, \rho, J, \tau)/n}}\right) + \Phi\left(-z_{\alpha_1/2} + \frac{\beta_0 - \beta_A}{\sqrt{\mathcal{I}_A^{\beta\beta}(\theta, \rho, J, \tau)/n}}\right) \quad (3.4)$$

where $\mathcal{I}_A^{\beta\beta}(\cdot)$ is the asymptotic variance evaluated at $\beta = \beta_A$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Clearly, power is a function of all the design factors namely

(i) sample size (n),

(ii) number of evenly scheduled assessments (J),

(iii) maximum duration of follow-up (τ).

When we restrict attention to the situation where the duration of follow-up τ is fixed and assessments are evenly scheduled, the study design in terms of (i) sample size n and (ii) the frequency of assessments J can be determined for different desired levels of power for testing a parameter of interest, and different pairs of design factors (n, J) may achieve the same power. Furthermore, if either n or J is fixed, the other can be solved by using (3.4). For example, if the number of assessments J is fixed the required sample size for a Wald-test at significance level α_1 and power $1 - \alpha_2$ to detect an effect $\beta = \beta_A$ is then

$$n = \left(\frac{z_{\alpha_1/2} + z_{\alpha_2}}{\beta_A}\right)^2 \mathcal{I}_A^{\beta\beta}(\theta, \rho, J, \tau). \quad (3.5)$$

We provide empirical examples of the sample size calculation and relationship between power, sample size and the number of scheduled assessments for prospective cohort studies targeting the estimation of the effect of a binary covariate X on disease incidence. We assume all subjects are in state 0 (i.e. progression-free) at the time origin (i.e. $Z(0) = 0$), disease status is determined at J equally spaced assessments, and survival status is monitored continuously over the study period $(0, \tau]$ subject to random right censoring. Without loss of generality, we let $\tau = 1$. For simplicity, we consider a time-homogeneous disease process with transition intensities $\lambda_{jk}(t|X) = \lambda_{jk} \exp\{X\beta_{jk}\}$ where λ_{01}, λ_{02}, and λ_{12} are baseline transition intensities and there is a covariate effect on disease progression denoted by β_{01} but no covariate effect on death (i.e. $\beta_{02} = \beta_{12} = 0$). Let $\beta_{01} = \log 0.75$ indicate a preventive covariate effect on disease progression, and $P(X = 1) = \{0.05, 0.25\}$. The values for parameters $(\lambda_{01}, \lambda_{02}, \lambda_{12})$ are set to satisfy the following constraints:

(i) $P_1 \;=\; P(T_1 < \tau \mid X = 0) \;=\; \int_0^\tau p_{00}(0, t; x)\lambda_{01}(t; x)dt \;=\; \{0.10, 0.25, 0.50\}$,

(ii) $P_2 \;=\; P(T_2 < \tau \mid X = 0) \;=\; \int_0^\tau p_{02}(0, t; x)dt \;=\; \{0.10, 0.25, 0.50\}$, and

(iii) $\lambda_{12}/\lambda_{02} \;=\; \{1.10\}$.

We assume individuals may become lost to follow-up at a random time C which follows an exponential distribution with a rate ρ and the value of ρ is set to satisfy $P(T_2 < \min(C, \tau)|X = 0) \;=\; \{0.05, 0.20\}$, where

$$P(T_2 < \min(C, \tau)|X = 0) \;=\; P(T_2 < \tau \mid X = 0)(1-G(\tau)) + \int_0^\tau P(T_2 < c \mid X = 0)g(c)dc \, .$$

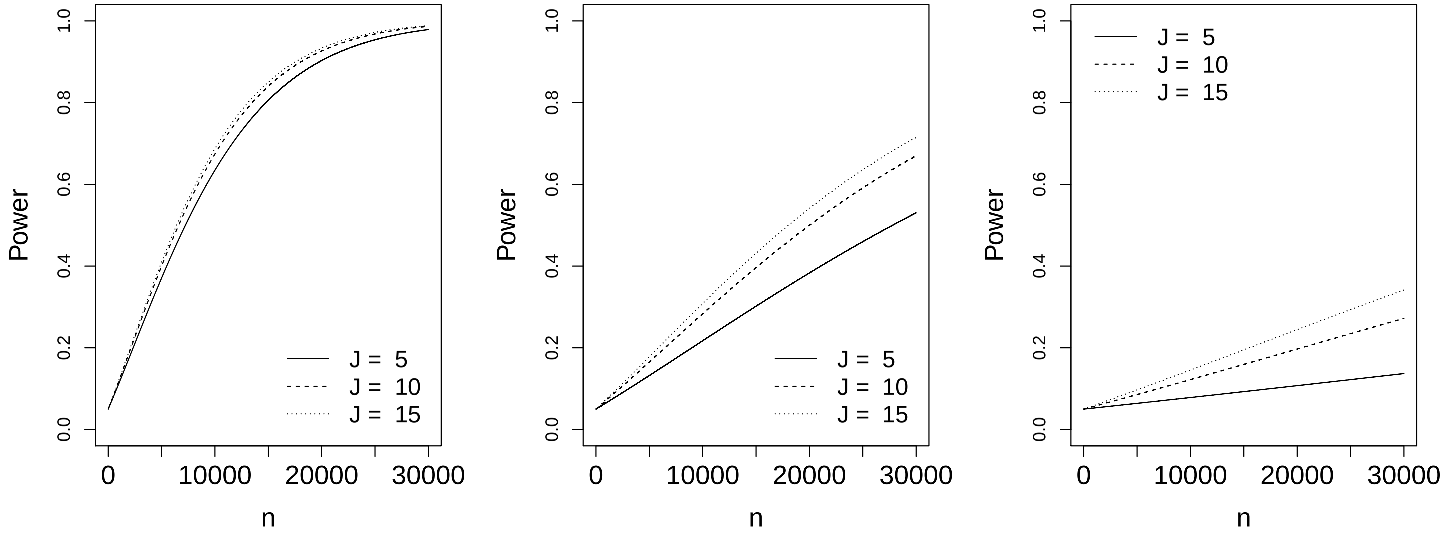
In Table 3.1, we report the sample size n for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$ calculated using formula (3.5), when the frequency of the assessments is fixed at $J = \{5, 10\}$, power at $\{80\%, 90\%\}$, significance level $\alpha_1 = 0.05$, $P_1 = \{0.1, 0.25\}$, $P_2 = 0.25$, $P(X = 1) = \{0.25, 0.05\}$, $\beta_{01} = \log 0.75$, $\lambda_{12}/\lambda_{02} = 1.1$, and $P(T_2 < \min(C, \tau)|X = 0) = 0.2$. To validate these sample size calculations, for each scenario we simulate $2,000$ datasets

			$P(X=1)=0.25$		$P(X=1)=0.05$	
Power	P_1	J	n	EP%	n	EP%
80%	0.10	5	8,442	81.3	35,025	82.0
		10	8,102	82.0	33,615	81.9
	0.25	5	3,290	82.0	13,601	81.0
		10	3,157	82.1	13,052	82.5
90%	0.10	5	11,301	90.8	46,888	91.4
		10	10,846	91.4	45,001	90.5
	0.25	5	4,405	91.7	18,208	92.1
		10	4,227	92.1	17,472	92.2

Table 3.1: Empirical power (EP%) for detecting an effect of covariate X on disease progression at the significance level $\alpha = 0.05$, when $\beta_{01} = \log 0.75$, $\lambda_{12}/\lambda_{02} = 1.1$, $P_2 = P(T_2 < \tau \mid X = 0) = 0.25$, $P(T_2 < \min(C, \tau)|X = 0) = 0.2$, and $P(X = 1) = \{0.05, 0.25\}$, based on $2,000$ simulated datasets of size n and sample sizes n are calculated as in (3.5)

following Cook and Lawless [2018, Appendix B], get point estimates $\hat{\beta}_{01}$ and their variance estimates using the `msm` package in R [Jackson, 2011], and report the empirical power. The empirical power achieves the nominal level across all scenarios, thereby validating the formula and computations.

The figures in the remainder of this section are based solely on the asymptotic variance. Figures 3.2, 3.3, and 3.4 display power curves to illustrate the impact of features of the process (P_1 and P_2) and of the study design (n and J) on power; these three figures feature $P_1 = \{0.10, 0.25, 0.50\}$ respectively, while the three panels within each figure contain the results for $P_2 = \{0.10, 0.25, 0.50\}$ respectively. Across all three figures, we have $\beta_{01} = \log 0.75$. As before, these figures focus on testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$, a two-sided test at a significance level of $\alpha = 0.05$. As expected, the power increases monotonically with the frequency of scheduled assessments over $[0, \tau]$. More generally, the power for detecting a covariate effect on progression is also higher when more precise information about disease progression is available, which, by comparing across the plots, can be seen to be driven by factors such as the proportion of disease progression events (P_1) and deaths (P_2) over $[0, \tau]$. As the proportion of deaths over $[0, \tau]$ increases, the number of realized clinical visits (M) decreases, and with it the extent of information on disease progression is reduced which leads to a large reduction in power; this can be seen by comparing across the three panels from left to right. For example, when $\beta_{01} = \log 0.75$ and $P_1 = P_2 = 0.10$, a sample of size $n = 15,000$ with $J = 5$ planned assessments over $(0, \tau]$ yields approximately 80% power for rejecting $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$, but if P_2 increases to 0.25 and 0.50, the power decreases substantially to 30% and 10% respectively. When interest lies in estimating the effect of a covariate on disease progression (β_{01}), we intuitively expect that an increase in the probability of progression should lead to an increase in power and these figures confirm this. When P_2 is fixed at 0.10, prospectively following a sample of $n = 5,000$ individuals for $J = 5$ planned assessments over $(0, \tau]$ leads to approximately 40% power when $P_1 = 0.10$, and this increases to 80% and 95% for $P_1 = 0.25$ and 0.50 respectively (comparing across analogous panels in Figures 3.2, 3.3, and 3.4).

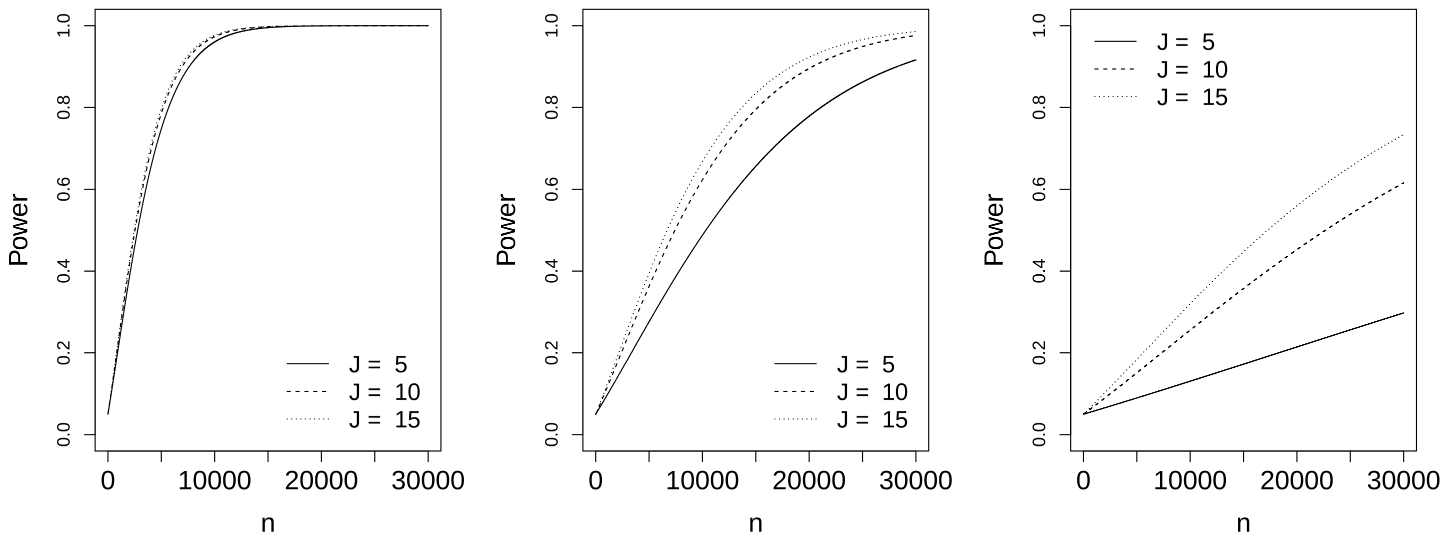


(a) $P_2 = 0.10$

(b) $P_2 = 0.25$

(c) $P_2 = 0.50$

Figure 3.2: Plots of power curves for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$, where $\beta_{01} = \log 0.75$ and type-I error rate is $\alpha = 0.05$; across all panels, we have $\lambda_{12}/\lambda_{02} = 1.1$, $P_1 = 0.10$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$

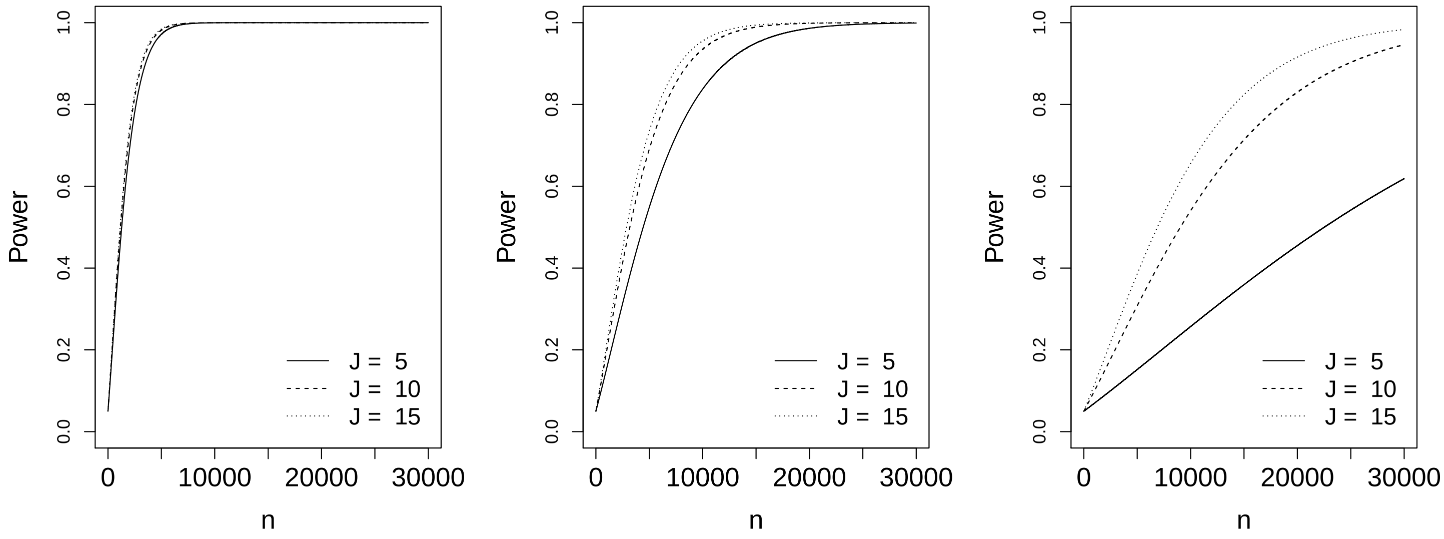


(a) $P_2 = 0.10$

(b) $P_2 = 0.25$

(c) $P_2 = 0.50$

Figure 3.3: Plots of power curves for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$, where $\beta_{01} = \log 0.75$ and type-I error rate is $\alpha = 0.05$; across all panels, we have $\lambda_{12}/\lambda_{02} = 1.1$, $P_1 = 0.25$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$



(a) $P_2 = 0.10$

(b) $P_2 = 0.25$

(c) $P_2 = 0.50$

Figure 3.4: Plots of power curves for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$, where $\beta_{01} = \log 0.75$ and type-I error rate is $\alpha = 0.05$; across all panels, we have $\lambda_{12}/\lambda_{02} = 1.1$, $P_1 = 0.50$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$

3.2.3 Cost Effective Design

Cost is a very important factor to consider when it comes to the design of prospective cohort studies. The effort of recruiting a subject into the study and assessing disease status may differ with the former being more expensive than the latter in many practical applications, therefore designs defined by different pairs (n, J) achieving the same power may lead to substantially different study costs. We consider the expected cost of cohort study designs, with a view to identify the one with the minimum cost.

Let C_0 be the cost for recruiting a subject into the study, C_1 be the cost of each follow-up assessment, and assume the assessment times are common to all individuals. The expected total cost of recruitment and follow-up of n subjects each with J intended visits over a period of τ years is then

$$E[C] = n\Big[C_0 + C_1 E(M)\Big] = n\Big[C_0 + C_1 \sum_{x=0}^{1}\sum_{j=1}^{J} jP(M = j|X = x)P(X = x)\Big] \,.$$

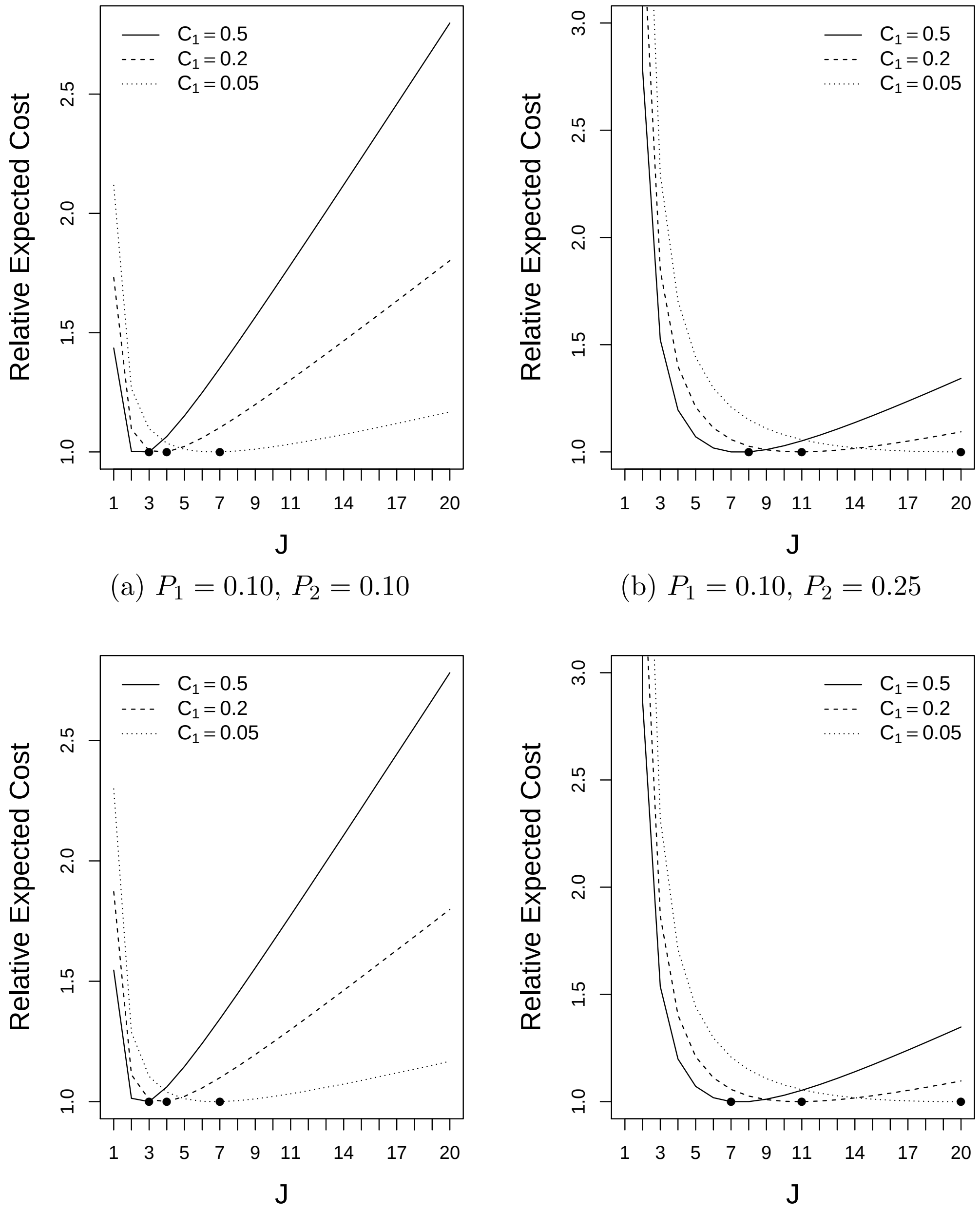
Recall M is the random number of assessments for an individual; $M = j$ implies $T^\dagger \in \mathcal{A}_{j+1} = [a_j, a_{j+1})$ and

$$P(M = j|X = x) = \begin{cases} \displaystyle\sum_{k=0}^{1} p_{0k}(0, a_j|x) \sum_{l=0}^{1} \Bigg\{ \int_{a_j}^{a_{j+1}} p_{kl}(a_j, t|x)\lambda_{l2}(t; x)dt \; (1 - G(a_{j+1})) & \text{if } j < J \\ \displaystyle\quad + \int_{a_j}^{a_{j+1}} \Big[p_{kl}(a_j, c|x) \; + \; \int_{a_j}^{c} p_{kl}(a_j, t|x)\lambda_{l2}(t; x)dt\Big] dG(c) \Bigg\} & \\ \sum_{l=0}^{1} p_{0l}(0, \tau|x)(1 - G(\tau)) & \text{if } j = J \end{cases}$$

A *minimum-cost design* is a design which minimizes expected total cost among all the designs (n, J) that achieve the same desired power to detect an effect of size $\beta_{01} = \log 0.75$. Figure 3.5 shows the relative expected cost of a design (n, J) versus the optimal one (n^{opt}, J^{opt}) represented by the dot on each line, when the power is fixed at 80%. The lines correspond to different values of cost ratio $C_1/C_0 = \{0.50, 0.20, 0.05\}$ and we set $C_0 = 1$ without loss of generality. As expected, the optimal frequency of assessments J^{opt} increases as the cost of conducting a follow-up assessment (C_1) decreases. As the probability of death over $[0, \tau]$ increases (comparing across columns in Figure 3.5), minimum-cost designs are achieved by scheduling more visits (e.g. increasing J^{opt}); this is sensible given that death terminates the observation process, and hence limits expected costs even when assessments are frequent. This observation is consistent with the power profile plots in Figures 3.2, 3.3, and 3.4. On the other hand, the probability of progression (P_1) has little effect on the determination of the frequency of assessment J^{opt} in minimum-cost designs, as can be seen by comparing across rows in Figure 3.5. While Figures 3.2, 3.3, and 3.4 demonstrated the large effect of P_1 on power for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$, J^{opt} is far less sensitive to it. However, this does imply an increase in n^{opt} as P_1 decreases, which would in turn lead to an increase in expected study cost. Finally, note that the above discussion extends to any desired (fixed) level of power, as we can easily show that

$$\frac{n_{0.8}(J_1)}{n_{0.8}(J_2)} = \frac{n_{0.9}(J_1)}{n_{0.9}(J_2)},$$

where $n_p(J)$ is the sample size obtained from (3.5) to achieve $100p\%$ power with J regular assessments over $(0, \tau]$. This implies that given the cost of follow-up assessments C_1, the value J^{opt} minimizing the expected total study cost does not change as a function of power.



(a) $P_1 = 0.10$, $P_2 = 0.10$

(b) $P_1 = 0.10$, $P_2 = 0.25$

(c) $P_1 = 0.50$, $P_2 = 0.10$

(d) $P_1 = 0.50$, $P_2 = 0.25$

Figure 3.5: Ratio of expected study cost (to achieve 80% power for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$ at a significance level of $\alpha = 0.05$ when $\beta_{01} = \log 0.75$) given J and the expected cost of the minimum-cost design; minimum-cost designs identified by dots for $C_1/C_0 = \{0.5, 0.2, 0.05\}$, $\lambda_{12}/\lambda_{02} = 1.1$, and $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$

3.3 Imperfect Assessment of Disease Status

3.3.1 Likelihood and EM Algorithm

In the previous section, we assumed that the ascertainment of disease status was always made without error, which is not often the case in practice. For example medical tests may yield false positives or false negatives, and diagnosis of many diseases may be based on subjective criterion leading to error. In some instances, while a gold standard test may exist to diagnose a condition, cost and patient burden may render the test impractical to administer in standard practice. In this section, we propose an EM algorithm [Dempster et al., 1977] for estimation in this framework and derive the Fisher information to use as the basis for investigation of study design implications, taking into account the expected cost.

Let $W(a_j)$ denote the misclassified disease status obtained from an error-prone assessment tool at assessment j, and $\bar{W}_j = (W(a_1), \ldots, W(a_j))$ be the classification history. The true disease status vector $\bar{Z}_m$ is latent and missing and the vital status is ascertained in continuous time up to $\min(C, \tau)$ without error so $T^\dagger = \min\{T_2, C, \tau\}$ and δ are observed. The likelihood of the observed data $\{\bar{W}_m, t^\dagger, \delta, x\}$ can be written as

$$L_o \propto \sum_{\bar{Z}_m} P(\bar{Z}_m, t^\dagger, \delta \mid x) P(\bar{W}_m \mid \bar{Z}_m, t^\dagger, \delta, x). \tag{3.6}$$

Note that the first term within the summation above only depends on the disease process, and the event time T_1 uniquely determines the true disease history $\bar{Z}_m$, so it is equal to

$$P(t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\dagger, \delta \mid x; \theta) = \begin{cases} p_{00}(0, a_{j-1}|x) p_{01}(a_{j-1}, a_j|x) p_{11}(a_j, t^\dagger|x) \lambda_{12}^{\delta}(t^\dagger; x) & j \leq m \\ p_{00}(0, a_m|x) \left[\sum_{k=0}^{1} p_{0k}(a_m, t^\dagger|x) \lambda_{k2}^{\delta}(t^\dagger; x) \right] & j = m+1 \end{cases},$$

where $\mathcal{A}_j = [a_{j-1}, a_j)$ is the jth intermittent observation interval as before but an extra interval is defined as $\mathcal{A}_{m+1} = [a_m, \infty)$. For the misclassification process, we assume $W(a_j)$

depends only on the current true state $Z(a_j)$ but not on the classification history or the true disease status in the past and future, thus the second term in (3.6) becomes

$$P(\bar{W}_m \mid t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\dagger, \delta, x; \pi) = \prod_{\ell=1}^{j-1} \pi_0^{1-W(a_\ell)}(1-\pi_0)^{W(a_\ell)} \prod_{\ell=j}^{m} \pi_1^{W(a_\ell)}(1-\pi_1)^{1-W(a_\ell)} \ ,$$

where $\pi = (\pi_0, \pi_1)$ with $\pi_1 = P(W(a_j) = 1|Z(a_j) = 1)$, $\pi_0 = P(W(a_j) = 0|Z(a_j) = 0)$; the misclassification rates $FP = 1 - \pi_0$ and $FN = 1 - \pi_1$ are often assumed to be known [Ma et al., 2016]. Given the above, the observed likelihood (3.6) can now be expressed as

$$L_o(\theta, \pi) = \sum_{j=1}^{m+1} P(t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\dagger, \delta | x; \theta) P(\bar{W}_m | t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\delta, x; \pi). \tag{3.7}$$

For the estimation of disease process parameters θ, direct maximization of the observed likelihood (3.7) is difficult in general. Rather, an EM algorithm can be implemented by casting matters in the missing-data framework, where the time of entry into state 1, T_1, is viewed as missing. Define the complete log-likelihood as

$$\ell_c(\theta) = \sum_{i=1}^{n} \log f(t_{i1}, t_i^\dagger, \delta_i \mid x_i; \theta)$$

where subscript i is used to index the individuals. At each iteration of the EM algorithm, an E-step computes the expected complete data log-likelihood given the observed data $\mathcal{D} = \{\bar{W}_{m_i}, t_i^\dagger, \delta_i, x_i; i = 1, \ldots, n\}$ and the current parameter estimates $\widehat{\theta}^{(r)}$, that is

$$E\left[\ell_c(\theta) \mid \mathcal{D}; \widehat{\theta}^{(r)}\right] = \sum_{i=1}^{n} \int_0^\infty \log\left[f(t_{i1}, t_i^\dagger, \delta_i \mid x_i; \theta)\right] f(t_{i1} | \bar{W}_{M_i}, t_i^\dagger, \delta_i, x_i; \widehat{\theta}^{(r)}, \pi)\, dt_{i1} \tag{3.8}$$

where the conditional distribution of T_1 given the observed data $\{\bar{W}_m, t^\dagger, \delta, x\}$ takes the form

$$f(t_1 | \bar{W}_m, t^\dagger, \delta, x; \theta, \pi) = \frac{f(t_1, t^\dagger, \delta \mid x; \theta) P(\bar{W}_m | t_1, t^\dagger, \delta, x)}{\int_0^\infty f(t_1, t^\dagger, \delta \mid x; \theta) P(\bar{W}_m | t_1, t^\dagger, \delta, x) dt_1} \ .$$

The M-step then requires maximizing the conditional expectation in (3.8) to get updated estimates of θ, and the iteration between the E- and M-steps continues until convergence. Variance estimation for the estimates $\widehat{\theta}$ from the EM algorithm is done by calculating the observed information via Louis' formula [Louis, 1982]. The details of an EM algorithm procedure for the estimation of a time-homogeneous three-state model with observed disease status subject to misclassification is provided in Appendix A.

3.3.2 Fisher Information and Design

Obtaining the Fisher information matrix with misclassified disease status requires taking derivatives of the logarithm of the observed likelihood $L_o(\bar{W}_M, T^\dagger, \delta, X)$ given in (3.6). Let $\ell_o = \log L_o(\bar{W}_M, T^\dagger, \delta, X)$ and $S_o = \partial \ell_o / \partial \theta$ be the vector of first-order derivatives. In general, the form of the observed-data score S_o may be complicated due to taking the logarithm of a sum. It is helpful to write it as an expectation of the complete-data score given the observed data,

$$S_o = \frac{\partial}{\partial \theta} \log L_o(\bar{W}_M, T^\dagger, \delta, X) = E\left[\frac{\partial}{\partial \theta} \log L_c(T_1, T^\dagger, \delta, X) \middle| \bar{W}_M, T^\dagger, \delta, X\right].$$

Under the assumption of non-informative censoring, the Fisher information $E[S_o S_o']$ is then obtained by taking the expectation of $S_o S_o'$ with respect to $\{\bar{W}_M, T^\dagger, \delta, X\}$.

$$\begin{aligned} \mathcal{I}(\theta) &= \sum_{x=0}^{1} \sum_{j=0}^{J} \int_{a_j}^{a_{j+1}} E\left[S_o S_o' \middle| C = c \in \mathcal{A}_j, X = x\right] dG(c) P(X = x) \qquad (3.9) \\ &= \sum_{x=0}^{1} \sum_{j=0}^{J} \int_{a_j}^{a_{j+1}} \left[H(j, \min(c, \tau), 0, x) + \sum_{q=1}^{j} \int_{a_{q-1}}^{\min(a_q, c)} H(q, t_2, 1, x) dt_2\right] dG(c) P(X = x) \end{aligned}$$

where

$$\begin{aligned}
H(m, t^{\dagger}, \delta, x) &= E\left[\frac{\partial}{\partial\theta}\ell_o(\bar{W}_m, t^{\dagger}, \delta, x)\frac{\partial}{\partial\theta'}\ell_o(\bar{W}_m, t^{\dagger}, \delta, x)\middle| t^{\dagger}, \delta, x\right] f(t^{\dagger}, \delta|x) \\
&= \sum_{\bar{W}_m}\left(\frac{\partial}{\partial\theta}\ell_o(\bar{W}_m, t^{\dagger}, \delta, x)\frac{\partial}{\partial\theta'}\ell_o(\bar{W}_m, t^{\dagger}, \delta, x)\right) P(\bar{W}_m|t^{\dagger}, \delta, x) f(t^{\dagger}, \delta|x) \\
&= \sum_{\bar{W}_m}\left(S_o(\bar{W}_m, t^{\dagger}, \delta, x) S'_o(\bar{W}_m, t^{\dagger}, \delta, x)\right) L_o(\bar{W}_m, t^{\dagger}, \delta, x)
\end{aligned}$$

and m satisfies $a_m \le t^{\dagger} < a_{m+1}$ and $t^{\dagger} = \min(t_2, c, \tau)$.

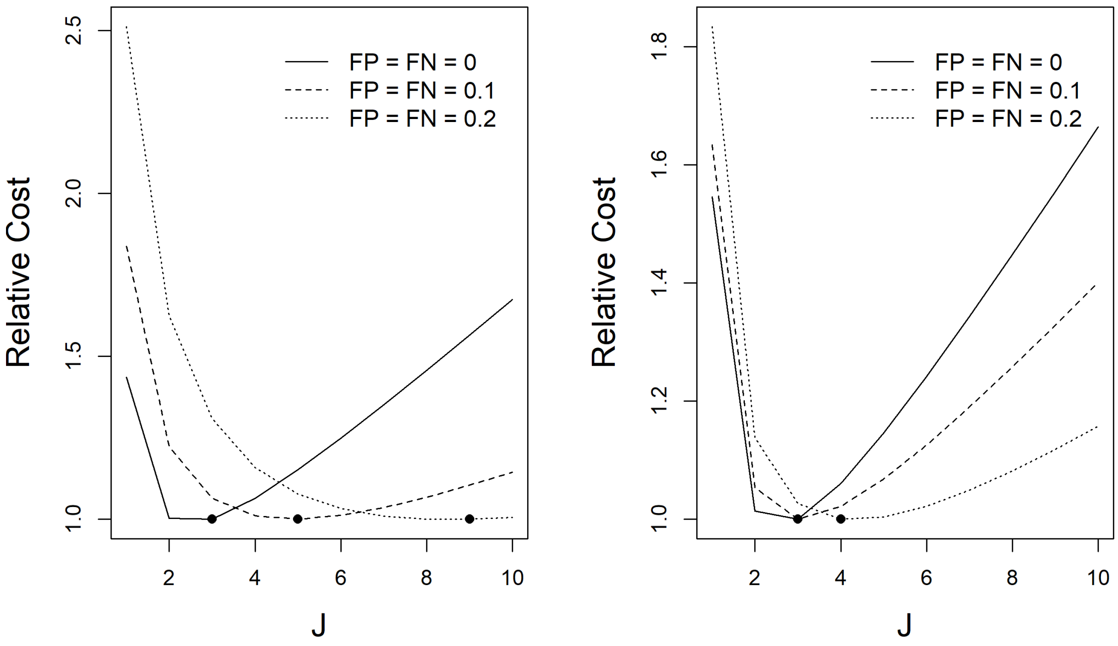
		$\log\lambda_{01}$ (-1.0928)				$\log\lambda_{02}$ (-1.2607)				$\log\lambda_{12}$ (-1.1654)				β_{01} (-0.28768)			
J	MIS	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
5	0	-0.004	0.059	0.060	94.9	3×10^{-4}	0.056	0.055	95.4	-0.005	0.156	0.154	95.4	-0.004	0.130	0.132	95.6
	0.10	8×10^{-4}	0.071	0.071	95.2	4×10^{-4}	0.057	0.058	95.4	-0.017	0.201	0.192	95.2	-0.008	0.163	0.156	94.4
	0.20	-1×10^{-4}	0.090	0.088	94.6	-1×10^{-4}	0.062	0.062	95.5	-0.029	0.264	0.239	95.4	-0.012	0.202	0.192	95.2
10	0	-0.004	0.058	0.059	95.0	4×10^{-4}	0.055	0.054	95.0	-0.004	0.148	0.145	95.2	-0.004	0.128	0.129	95.8
	0.10	-0.001	0.066	0.065	94.7	2×10^{-4}	0.056	0.056	95.6	-0.012	0.172	0.168	95.0	-0.002	0.146	0.144	94.8
	0.20	-0.001	0.076	0.075	94.5	-2×10^{-5}	0.058	0.058	95.6	-0.017	0.202	0.197	95.8	-0.006	0.166	0.164	95.2

Table 3.2: Simulation results based on $2,000$ simulated datasets, each with $n = 2,000$; EBIAS is the empirical bias, ESE is the empirical standard error, ASE is the asymptotic standard error, ECP is the empirical coverage probability expressed as a probability with nominal level of 95% and MIS$= 1-\pi$ is the misclassification probability where $\pi = \pi_0 = \pi_1$; $P_1 = 0.25$, $P_2 = 0.25$, $\lambda_{12}/\lambda_{02} = 1.1$, $\beta_{01} = \log 0.75$, $P(T_2 < \min(C,\tau)|X=0) = 0.2$, and $P(X=1) = 0.25$

We validate the asymptotic variance obtained from the Fisher information (3.9) by comparing it to the empirical variance and average estimated variance of MLEs obtained via the EM algorithm for each of $2,000$ simulated datasets. These results are reported in Table 3.2; the parameter settings mirror those in Section 3.2.2, MIS$= 1-\pi = \{0, 0.10, 0.20\}$ and $\pi = \pi_0 = \pi_1$. Note the excellent agreement between the empirical standard error and the asymptotic standard error, as well as the coverage achieving the nominal level of 95%, even in presence of slight and moderate misclassification.

Intuitively, it is clear that the scheduling of more frequent assessments mitigates, to some degree, the loss of information due to potential state misclassification. However, when considering both the cost of increasing the sample size (C_0) and the cost of follow-up assessments (C_1), it is not obvious whether it would be more cost-effective to increase n or J to achieve a desired level of power. In Figure 3.6, we see that as the degree of misclassification increases, the minimum-cost design is achieved by increasing the frequency of assessments over $[0, \tau]$. This is particularly apparent when disease progression is rare in the cohort (i.e. when P_1 is low), in which case even a modest rate of false positives/negatives has a significant impact on J^{opt}.

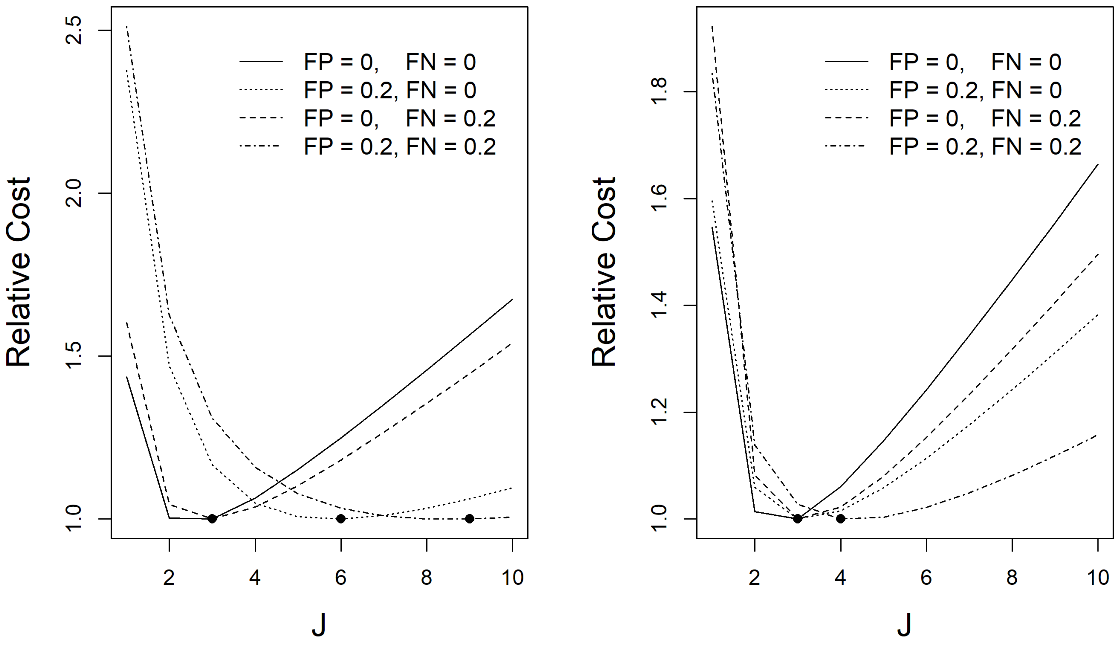
Finally, we consider the differential impact of false positive and false negative errors on features of the minimum-cost design (see Figure 3.7). For example, when P_1 is low (that is when progression events are rare in the cohort) and interest lies in detecting a covariate effect on disease progression, an increase in the rate of false positives (FP $= 1 - \pi_0$) has a much larger impact on J^{opt} than does an increase in the rate of false negatives (FN $= 1 - \pi_1$).



(a) $P_1 = 0.1, P_2 = 0.1$

(b) $P_1 = 0.5, P_2 = 0.1$

Figure 3.6: Ratio of expected study cost (to achieve 80% power for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$ at a significance level of $\alpha = 0.05$ when $\beta_{01} = \log 0.75$) given J and the expected cost of the minimum-cost design; minimum-cost designs identified by dots for different degrees of misclassification where we assume equal false positive (FP) and false negative (FN) rates; $C_1/C_0 = 0.5$, $\lambda_{12}/\lambda_{02} = 1.1$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$



(a) $P_1 = 0.1, P_2 = 0.1$

(b) $P_1 = 0.5, P_2 = 0.1$

Figure 3.7: Ratio of expected study cost (to achieve 80% power for testing $H_0 : \beta_{01} = 0$ vs $H_A : \beta_{01} \neq 0$ at a significance level of $\alpha = 0.05$ when $\beta_{01} = \log 0.75$ given J and the expected cost of the minimum-cost design; minimum-cost designs identified by dots for various combinations of false positive rate (FP$= 1-\pi_0$) and false negative rate (FN$= 1-\pi_1$); $C_1/C_0 = 0.5$, $\lambda_{12}/\lambda_{02} = 1.1$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$

3.4 Response-adaptive Designs for Prospective Cohort Studies

In the previous section, we proposed a framework for the design of prospective cohort studies where the assessment schedule is pre-determined; in practice this assumption is generally violated. Departures from the scheduled assessment times are common: if these are random and relatively modest, they may not have a large impact on design. When the timing of assessments is respondent-driven (and possibly dependent on the responses themselves), there exist methods for accounting for this dependence in estimation [Rad, 2014] although accounting for this is difficult at the design stage due to the strong assumptions required on the visit process. However, systematic protocol-driven dependence of the visit process on *observed* responses, via repeated testing or referrals to specialists for more advanced testing/examinations following a positive test result, can be considered at the design stage.

Suppose there exist both a gold standard assessment tool to ascertain disease status and an error-prone tool, where the former is more expensive and/or invasive than the latter. Due to the increased burden of the gold standard assessment, the error prone tool is employed at assessment times a_j, $j = 1, \dots, J$. If $W(a_j)$ is positive (i.e. $W(a_j) = 1$)) at an assessment j, the gold standard assessment is administered to determine the true status $Z(a_j) \in \{0, 1\}$; if $Z(a_j) = 1$, assessments of disease status terminate, and if $Z(a_j) = 0$, subsequent assessments again employ the error-prone assessment tool. In short, this study eliminates false positives, although the possibility of false negatives remains.

Let Q be the index of the first visit at which disease progression is confirmed and assessments are discontinued, $Q \in \{1, 2, \dots, J, \infty\}$; if $Q = \infty$, then progression is not confirmed by the end of the study. As before, for $Q \leq J$ we define $\bar{Z}_Q = \{Z(a_1), \dots, Z(a_Q)\}$

and $\bar{W}_Q = \{W(a_1), \ldots, W(a_Q)\}$, while $\bar{Z}_\infty = \bar{Z}_J$ and $\bar{W}_\infty = \bar{W}_J$. The observed data is $\{\bar{W}_Q, \bar{Z}^*_Q, t^\dagger, \delta, x\}$ where $\bar{W}_Q = \{W_1, \ldots, W_Q\}$ is the observed (potentially misclassified) disease status history up to visit Q and $\bar{Z}^*_Q$ is the vector of true disease statuses confirmed by gold standard assessment. Note that $\bar{Z}^*_Q$ is a subset of the full true disease status history $\bar{Z}_Q$. The observed-data likelihood is a modification of (3.6) and can be rewritten as

$$L_o(\bar{W}_Q, \bar{Z}^*_Q, t^\dagger, \delta, x) = \sum_{\bar{Z}_Q \backslash \bar{Z}^*_Q} P(\bar{Z}_Q, \bar{Z}^*_Q, t^\dagger, \delta | x) P(\bar{W}_Q | \bar{Z}_Q, \bar{Z}^*_Q, t^\dagger, \delta, x)$$

where $\bar{Z}_Q \backslash \bar{Z}^*_Q$ is the set of unobserved true disease statuses, which needs to be summed over to get the observed-data likelihood.

The Fisher information can then be obtained in a manner similar to that in Section 3.3.2. The outer expectations in (3.9), with respect to X and C, are taken in the same way here. However, while previously the observation process was terminated after M visits by the minimum of the death and censoring times, in the present adaptive design protocol it is terminated earlier by $Q \leq M$ if a positive diagnosis is obtained (i.e.if $Z(a_Q) = 1$).

3.5 Discussion

We have developed a framework for the design of cohort studies in which interest lies primarily in the effect of a covariate on the development of an intermediate event; this event could represent the onset of a disease (e.g. diabetes) in a large cohort study or the development of a complication if the cohort is comprised of disease individuals (onset of kidney damage in a diabetes cohort). This work offers a theoretical underpinning of the simulation-based study of Ma et al. [2016]. While we have examined the features that most influence the sample size requirements for a three-state illness-death model, the framework naturally accommodates progressive multistate processes with more than three states. Diseases such as dementia, hepatitis, retinopathy or nephropathy all represent progressive conditions for which several intermediate states can be introduced for a more detailed modelling of the progression. Sample size can be likewise determined if estimation of a covariate effect on any particular transition in the disease process is of primary interest.

Ma et al. [2016] considered the impact of misclassification of a genetic marker when the goal is to assess the power of a cohort study for detecting its effect. There is a large literature on the impact of covariate measurement error or misclassification [Yi, 2016, Carroll et al., 2006, Fuller, 1987] and likelihood methods can be employed to accommodate this with either external or internal validation samples. A natural extension of our work would be to base study design on a model accommodating covariate misclassification based on a prior external validation sample.

We considered the setting in which the status of individuals may be misclassified at examination times. We presumed that individuals would continue to be examined after any positive assessments suggesting the intermediate event had occurred; this corresponds to the setting in which the analysis might be done retrospectively upon completion of the cohort study. In practice, if the assessments are made by treating physicians, individuals testing positive would be referred immediately for definitive diagnostic checks. If they were found to have experienced the event based on a gold standard test then the schedule for the subsequent follow-up assessments may be modified. If they were subsequently found not to have experienced the event, but the false positive assessment was suggestive of higher risk, the subsequent assessment times might be more frequent. While this framework is outlined briefly in Section 3.4, study designs accommodating such adaptive observation schemes warrant further development.

More generally, longitudinal cohort studies are designed under idealized assumptions which may not always be realized in practice; participants may not adhere to the visit schedule specified in the protocol, some may not take treatments as directed, and responses may be misreported or missing. While some departures from design assumptions are unavoidable, it is important to capture as many realistic features of the study as possible at the design stage. In longitudinal studies where responses are collected on an individual repeatedly over time, it is reasonable to expect that the schedule of assessments and/or the treatment protocol may change over time, and this is an area that warrants future work. For example, while we assumed the assessments were scheduled at regular intervals over $(0, \tau]$, this need not be the case. The asymptotic variance can be calculated with (3.9) with unequally spaced assessment times, as long as they are scheduled in the protocol. This allows for the assessment schedule to be optimized at the design stage; this may be of particular interest if transition intensities are vary with time, for example via piecewise constant intensities, and scheduling more frequent assessments for participants in high-risk periods may lead to more cost-effective designs. Accommodating deviations from the protocol at the design stage would involve modelling the visit time process which is, in general, challenging; this is further discussed in Section 5.2.2.

Appendix A: Misclassified Disease Status - EM algorithm

For illustration purposes, we consider a time-homogeneous model, with $\lambda_{k\ell}(t;x) = \lambda_{k\ell}(x) = \lambda_{k\ell}e^{\beta_{k\ell}x}$ for $k < \ell$ and assume the misclassification parameters π are known. The transition probability matrix can be obtained via matrix exponential $\mathbb{P}(s,t) = \exp\big\{\big(t-s)\mathbb{A}\big\}$ and $\mathbb{A} = [\lambda_{k\ell}(x)]$ is the constant transition intensity matrix; we obtain explicit expressions for the individual transition probabilities

$$p_{00}(t) = \exp\{-(\lambda_{01}(x) + \lambda_{02}(x))(t)\}\ ,$$
$$p_{11}(t) = \exp\{-\lambda_{12}(x)(t)\}\ ,$$
$$p_{01}(t) = \lambda_{01}(x)/[\lambda_{01}(x) + \lambda_{02}(x) - \lambda_{12}(x)]\Big[\exp\{-\lambda_{12}(x)(t)\} - \exp\{-(\lambda_{01}(x) + \lambda_{02}(x))(t)\}\Big]\ .$$

Under the time-homogeneous model with misclassified disease status, the likelihood function of the observed data $\mathcal{D} = \{\bar{W}_m, T^{\dagger}, \delta, X\}$ can be written in closed form

$$\begin{aligned} L_o(\theta,\pi) = {} & \frac{\lambda_{01}(x)\lambda_{12}(x)^{\delta}e^{-\lambda_{12}(x)t^{\dagger}}}{\lambda^*(x)}\sum_{j=1}^{m+1}\Big(e^{-\lambda^*(x)a_{j-1}} - e^{-\lambda^*(x)\min\{a_j,t^{\dagger}\}}\Big)P(\bar{W}_m \mid t_1 \in \mathcal{A}_j, \bar{Z}_m, t^{\dagger}, \delta, x; \pi) \\ & + \lambda_{02}(x)^{\delta}e^{-(\lambda_{01}(x)+\lambda_{02}(x))t^{\dagger}}P(\bar{W}_m \mid t_1 \in \mathcal{A}_{m+1}, \bar{Z}_m, t^{\dagger}, \delta, x; \pi), \end{aligned} \tag{3.10}$$

where $\lambda^*(x) = \lambda_{01}(x) + \lambda_{02}(x) - \lambda_{12}(x)$ and

$$P(\bar{W}_m \mid t_1 \in \mathcal{A}_j, \bar{Z}_m, t^{\dagger}, \delta, x; \pi) = \prod_{\ell=1}^{j-1}\pi_0^{1-W(a_\ell)}(1-\pi_0)^{W(a_\ell)}\prod_{\ell=j}^{m}\pi_1^{W(a_\ell)}(1-\pi_1)^{1-W(a_\ell)}\ , \tag{3.11}$$

where $\pi = (\pi_0, \pi_1)$ with $\pi_1 = P(W(a_j) = 1|Z(a_j) = 1)$, $\pi_0 = P(W(a_j) = 0|Z(a_j) = 0)$ and we have $\mathcal{A}_j = [a_{j-1}, a_j)$ for $j = 1, \ldots, m$ and $\mathcal{A}_{m+1} = [a_m, \infty)$.

The objective here is to estimate the disease process parameters θ, assuming the misclassification parameters π are known. We discuss the steps of the EM algorithm considering a single subject only for convenience, but note that the generalization over all subjects is straightforward.

E-step: For a time-homogeneous model, the joint distribution of the complete data $\{T_1, T^\dagger, \delta\}$ is

$$f(t_1, t^\dagger, \delta \mid x) = \Big[p_{00}(t_1)\lambda_{01}(t_1)p_{11}(t^\dagger - t_1)\lambda_{12}^{\delta}(t^\dagger)\Big]^{I(t_1<t^\dagger)}\Big[p_{00}(t^\dagger)\lambda_{02}^{\delta}(t^\dagger)\Big]^{I(t_1>t^\dagger)}.$$

The complete log-likelihood then can be written as a linear function of event time T_1 such that

$$\begin{aligned}
\ell_c(\theta) &= \log f(t_1, t^\dagger, \delta \mid x) \\
&= I(t_1 < t^\dagger)\Big\{\log\lambda_{01}(x) - \lambda_{12}(x)t^\dagger + \delta\log\lambda_{12}(x) - \big[\lambda_{01}(x) + \lambda_{02}(x) - \lambda_{12}(x)\big]t_1\Big\} \\
&\quad + I(t_1 > t^\dagger)\Big\{\delta\log\lambda_{02}(x) - \big[\lambda_{01}(x) + \lambda_{02}(x)\big]t^\dagger\Big\}\ .
\end{aligned}$$

The conditional expectation $Q(\theta; \widehat{\theta}^{(r)}) = E\Big[\ell_c(\theta) \mid \mathcal{D}; \widehat{\theta}^{(r)}\Big]$ becomes

$$\begin{aligned}
Q(\theta;\widehat{\theta}^{(r)}) \;&=\; w_1^{(r)}\Big\{\log\lambda_{01}(x) + \delta\log\lambda_{12}(x) - \lambda_{12}(x)t^\dagger\Big\} - w_2^{(r)}\Big\{\lambda_{01}(x) + \lambda_{02}(x) - \lambda_{12}(x)\Big\} \\
&\quad + \big[1 - w_1^{(r)}\big]\Big\{\delta\log\lambda_{02}(x) - \big[\lambda_{01}(x) + \lambda_{02}(x)\big]t^\dagger\Big\} \qquad (3.12)
\end{aligned}$$

where $w_1^{(r)} = P(t_1 < t^\dagger \mid \mathcal{D}; \widehat{\theta}^{(r)}, \pi)$ and $w_2^{(r)} = \int_0^{t^\dagger} t_1 f(t_1 \mid \mathcal{D}; \widehat{\theta}^{(r)}, \pi)dt_1$. The weight functions are calculated as follows:

$$w_1(\theta,\pi) = \int_0^{t^\dagger} f(t_1 \mid \mathcal{D};\theta,\pi)dt_1 = 1 - \Big[\lambda_{02}^{\delta}e^{-(\lambda_{01}+\lambda_{02})t^\dagger}P(\bar{W}_m|t_1 \in \mathcal{A}_{m+1}, t^\dagger, \delta, x; \pi)\Big]/L_o(\theta,\pi)$$

$$w_2(\theta,\pi) = \int_0^{t^\dagger} t_1 f(t_1 \mid \mathcal{D};\theta,\pi)dt_1 = \left[\sum_{k=1}^{M+1}\int_{a_{k-1}}^{\min\{a_k, t^\dagger\}} t_1 f(t_1, t^\dagger, \delta|x;\theta)dt_1 P(\bar{W}_m|t_1 \in \mathcal{A}_k, t^\dagger, \delta, x; \pi)\right]/L_o(\theta,\pi)$$

where $P(\bar{W}_m|t_1 \in \mathcal{A}_k, t^\dagger, \delta, x; \pi)$ is given in (3.11), $L_o(\theta,\pi)$ is given in (3.10), and the integration takes the form

$$\int_a^b t_1 f(t_1, t^\dagger, \delta|x;\theta)dt_1 = \lambda_{01}(x)\lambda_{12}(x)^{\delta}e^{-\lambda_{12}t^\dagger}\Big\{\big[a\lambda^*(x) + 1\big]e^{-a\lambda^*(x)} - \big[b\lambda^*(x) + 1\big]e^{-b\lambda^*(x)}\Big\}/\lambda^*(x)^2\ ,$$

here recall $\lambda^*(x) = \lambda_{01}(x) + \lambda_{02}(x) - \lambda_{12}(x)$.

M-step: The updated estimates of θ are obtained by maximizing the conditional expectation $Q(\theta; \widehat{\theta}^{(r)})$. Note that under time-homogeneous model, the $Q(\theta; \widehat{\theta}^{(r)})$ function given in (3.12) can be re-organized as

$$\begin{aligned} Q(\theta; \widehat{\theta}^{(r)}) \;=\; & w_1^{(r)} \left[\log \lambda_{01}(x) - v^{(r)} \lambda_{01}\right] \;+\; \left(1 - w_1^{(r)}\right) \left[\delta \log \lambda_{02} - t^{\dagger} \lambda_{02}\right] \\ & + \; w_1^{(r)} \left[\delta \log \lambda_{12} - \left(t^{\dagger} - v^{(r)}\right) \lambda_{12}\right] \;+\; \left(1 - w_1^{(r)}\right) \left[-t^{\dagger} \lambda_{01}\right] \\ & + \; w_1^{(r)} \left[-v_1^{(r)} \lambda_{02}\right] \end{aligned} \tag{3.13}$$

Note that $Q(\theta; \widehat{\theta}^{(r)})$ function resembles the sum of weighted log-likelihood function of Poisson observations with offsets. This implies that the maximization can be done by generating a pseudo dataset and fitting log-linear Poisson models. More specifically for each subject i, we create five pseudo responses y_{ij} $(j = 1, \ldots, 5)$, and we assume $y_{ij} \sim \text{Poisson}(\lambda_{ij} u_{ij})$ with a pseudo offset u_{ij} and a log-linear model for the rate $\log \lambda_{ij} = \mathbb{z}'_{ij} \theta$ where $\mathbb{z}_{ij}$ is a vector of pseudo covariates associated with y_{ij}. The conditional expectation (3.13) is then equivalent to the weighted log-likelihood of a pseudo dataset

$$Q(\theta; \widehat{\theta}^{(r)}) = \sum_{i,j} \widehat{w}_{ij} \log f(y_{ij}; \theta) = \sum_{i,j} \widehat{w}_{ij} \Big\{ y_{ij} \big[\log \lambda_{ij} + \log u_{ij} \big] - \lambda_{ij} u_{ij} \Big\}$$

where we let $\theta = (\log \lambda_{01}, \beta_{01}, \log \lambda_{12}, \beta_{01}, \log \lambda_{02}, \beta_{02})'$, and the values of weights (i.e. $\widehat{w}_{ij}$), responses (i.e. y_{ij}), covariates (i.e. z_{ij}) and offsets (i.e. u_{ij}) of this pseudo dataset are given in Table 3.3

In other words, we can use the `glm` function in `R` to do the maximization by fitting a Poisson log-linear model on a pseudo dataset where each individual in the original sample will give rise to a number of 'pseudo-individuals' with their weights, responses, and associated covariates and offsets generated as described in Table 3.3.

$\widehat{w}_{ij}$	y_{ij}	$\mathbb{z}'_{ij}$	u_{ij}	$\log\lambda_{ij} = z'_{ij}\theta$
$w_{1,i}^{(r)}$	1	$(1,0,0,x,0,0)$	$w_{2,i}^{(r)}/w_{1,i}^{(r)}$	$\log\lambda_{01}(x)$
$\left(1-w_{1,i}^{(r)}\right)$	δ	$(0,1,0,0,x,0)$	$t^\dagger$	$\log\lambda_{02}(x)$
$w_{1,i}^{(r)}$	δ	$(0,0,1,0,0,x)$	$t^\dagger - w_{2,i}^{(r)}/w_{1,i}^{(r)}$	$\log\lambda_{12}(x)$
$\left(1-w_{1,i}^{(r)}\right)$	0	$(1,0,0,x,0,0)$	$t^\dagger$	$\log\lambda_{01}(x)$
$w_{1,i}^{(r)}$	0	$(0,1,0,0,x,0)$	$w_{2,i}^{(r)}/w_{1,i}^{(r)}$	$\log\lambda_{02}(x)$

Table 3.3: Pseudo-data for loglinear model

Observed Information: We calculate the observed information $I(\hat{\theta})$ by taking the expectation of derivatives of the complete data loglikelihood $l_c(\theta)$ given the observed data $\mathcal{D} = \{\bar{W}_M, T^\dagger, \delta, X\}$ as in [Louis, 1982]

$$I(\hat{\theta}) = E\left[\frac{\partial^2 l_c(\theta)}{\partial\theta\partial\theta'}\Big|\mathcal{D};\hat{\theta}\right] - E\left[\frac{\partial l_c(\theta)}{\partial\theta}\,\frac{\partial l_c(\theta)}{\partial\theta'}\Big|\mathcal{D};\hat{\theta}\right] + E\left[\frac{\partial l_c(\theta)}{\partial\theta}\Big|\mathcal{D};\hat{\theta}\right] E\left[\frac{\partial l_c(\theta)}{\partial\theta'}\Big|\mathcal{D};\hat{\theta}\right].$$

Evaluating the above amounts to calculating $w_1(\theta) = P(t_1 < t^\dagger|\mathcal{D};\hat{\theta},\pi)$, $w_2(\theta) = E\big[t_1|t_1 < t^\dagger,\mathcal{D};\hat{\theta},\pi\big]$, and $w_3(\theta) = E\big[t_1^2|t_1 < t^\dagger,\mathcal{D};\hat{\theta},\pi\big]$ for each individual, where $w_1(\theta)$ and $w_2(\theta)$ are as in the E-step of the EM algorithm and

$$w_3(\theta,\pi) = \left[\sum_{j=1}^{M+1}\int_{a_{j-1}}^{\min\{a_j,t^\dagger\}} t_1^2 f(t_1,t^\dagger,\delta|x;\theta)dt_1 P(\bar{W}_m|t_1\in\mathcal{A}_j,t^\dagger,\delta,\pi)\right]\Big/ L_o(\theta,\pi)$$

where

$$\int_a^b t_1^2 f(t_1,t^\dagger,\delta|x;\theta)dt_1 = \frac{\lambda_{01}(x)\lambda_{12}(x)^\delta e^{-\lambda_{12}t^\dagger}}{(\lambda^*)^3}\Big\{\big[(a\lambda^*)^2 + 2a\lambda^*(x) + 2\big]e^{-a\lambda^*(x)} - \big[(b\lambda^*)^2 + 2b\lambda^*(x) + 2\big]e^{-b\lambda^*(x)}\Big\},$$

and $\lambda^*(x) = \lambda_{01}(x) + \lambda_{02}(x) - \lambda_{12}(x)$.

Chapter 4

State-dependent Sampling Designs for Prevalent Cohort Studies

4.1 Introduction

Incident cohort studies of disease involve the recruitment of individuals who are healthy and their follow-up over time with the goal of observing the time of disease onset within a specific follow-up period [Wang, 1999]. In contrast, prevalent cohort studies involve the recruitment and follow-up of a sample of individuals who have already developed the disease with the goal of learning about the course of a chronic disease [Armitage and Colton, 2005]; for this goal prevalent cohort studies can result in significant cost savings compared to incident cohort studies, where many individuals may not develop the disease during the follow-up period. The effect of length-biased sampling in prevalent cohort studies has been extensively studied for failure-time data [Asgharian et al., 2002], but little attention has been given to the impact of such a sampling scheme on more general multistate processes. State-dependent sampling designs and associated analyses which account for the recruitment scheme can be exploited to optimize efficiency and/or minimize expected study cost. We explore these and related issues in this chapter.

In Section 4.2, we demonstrate the impact of ignoring length-biased sampling of multistate processes in data analyses. We find that the preferential sampling of individuals with long survival times with disease induces a conservative bias in estimators of cumulative transition intensities. This bias can be mitigated by constructing likelihood functions which account for the selection conditions appropriately. With correctly formulated likelihoods, one can then consider the impact of state-dependent selection criteria in constructing samples of individuals from a population screened for recruitment. We consider first a simple setting of a failure time process in Section 4.3. In Section 4.4 we derive a closed form expression for the Fisher information for a general multistate process when data are acquired under two reasonable sampling schemes: (i) state-dependent sampling in which data are only available on selected individuals, and (ii) state-dependent sampling in which additional current-status data are available from individuals screened but not recruited for follow-up. In Section 4.5 we demonstrate the impact of the choice of state-specific sample sizes on the precision of the estimates of interest (e.g. a biomarker effect), when subject to a realistic cost constraint. In Section 4.6 we present empirical results confirming the validity of the proposed designs. Concluding remarks are given in 4.7.

4.2 Length-biased Sampling of Multistate Processes

Consider a progressive multistate model as in Figure 1.2, with transient states $0, \dots, K-1$ representing progressive stages of disease and an absorbing state K representing death. We assume all individuals begin the process in state 0 at age 0 and let T_k denote the entry time to state k, $k = 1, \dots, K$, so T_K is the age at death.

A prevalent cohort sample is obtained by drawing a random sample of individuals from a population satisfying a selection criterion $Z(A) \in \mathcal{S}$ where A is the age at contact,

$Z(A)$ indicates the state at the time of contact, and $\mathcal{S}$ is a subset of the state space $\Omega = \{0, 1, \ldots, K\}$. For example, $\mathcal{S} = \{0, \ldots, K-1\}$ corresponds to the condition that individuals must be alive to be sampled, whereas if $\mathcal{S} = \{1, \ldots, K-1\}$, healthy individuals will be excluded from the sample.

4.2.1 Multistate Model Induced by Length-biased Sampling

Let $\Delta = 1$ indicate an individual is sampled into a prevalent study and $\Delta = 0$ otherwise. Suppose $\mathcal{S} = \{0, 1, \ldots, K-1\}$ so that individuals may be sampled if they are alive at the time the population is screened. Under the typical length-biased sampling scheme, individuals are selected with a probability proportional to their lifetime. In the multistate context, transition probabilities in such a length-biased sample are obtained by conditioning on the sampling indicator, that is

$$\begin{aligned}
P_{jk}(s,t|\Delta = 1, \mathcal{H}(s)) &= P(Z(t) = k | Z(s) = j, \Delta = 1, \mathcal{H}(s)) \qquad (4.1)\\
&= \left(\frac{P(\Delta = 1 | Z(t) = k, Z(s) = j, \mathcal{H}(s))}{P(\Delta = 1 | Z(s) = j, \mathcal{H}(s))} \right) P(Z(t) = k | Z(s) = j, \mathcal{H}(s))\\
&= \left(\frac{E[T_K | Z(t) = k, Z(s) = j, \mathcal{H}(s)]}{E[T_K | Z(s) = j, \mathcal{H}(s)]} \right) P(Z(t) = k | Z(s) = j, \mathcal{H}(s)),
\end{aligned}$$

for $j \leq k$ and $s \leq t$, since

$$\begin{aligned}
&P(\Delta = 1 | Z(t) = k, Z(s) = j, \mathcal{H}(s))\\
&\qquad = E_{T_K}\Big[P\big(\Delta = 1 \,|\, T_K, Z(t) = k, Z(s) = j, \mathcal{H}(s)\big) \big| Z(t) = k, Z(s) = j, \mathcal{H}(s) \Big]\\
&\qquad = cE[T_K | Z(t) = k, Z(s) = j, \mathcal{H}(s)]
\end{aligned}$$

under the assumption that sampling probabilities are proportional to the duration in $\mathcal{S}$, that is T_K in the present setting. If the multistate process is Markov, (4.1) becomes

$$P_{jk}(s,t|\Delta = 1) = \left(\frac{E[T_K | Z(t) = k]}{E[T_K | Z(s) = j]} \right) P_{jk}(s,t) \qquad (4.2)$$

and it can be seen that the Markov property is preserved for the length-biased sample. The intensity of a $j \to k$ transition at time t given recruitment in the sample is based on (4.2) as

$$\rho_{jk}(t|\Delta = 1) = \lim_{\Delta t \downarrow 0} \frac{P(Z(t+\Delta t^-) = k | Z(t^-) = j, Z(0) = 0, \Delta = 1)}{\Delta t}, \tag{4.3}$$

for $j \in \mathcal{S}$ and $k \in \{1, \ldots, K\}$, which can differ appreciably from

$$\lambda_{jk}(t) = \lim_{\Delta t \downarrow 0} \frac{P(Z(t+\Delta t^-) = k | Z(t^-) = j, Z(0) = 0)}{\Delta t},$$

the population Markov intensity.

4.2.2 Length-bias in the Three-state Illness-death Model

Here, we consider a three-state time-homogeneous illness-death process with $K = 2$ (see Figure 1.2) to illustrate the difference between the population model and the model induced by prevalent cohort sampling with $\mathcal{S} = \{0, 1\}$. The transition probability matrix is

$$\mathbb{P}(s,t) = \begin{pmatrix} e^{-(\lambda_{01}+\lambda_{02})(t-s)} & P_{01}(s,t) & 1 - e^{-(\lambda_{01}+\lambda_{02})(t-s)} - P_{01}(t-s) \\ 0 & e^{-\lambda_{12}(t-s)} & 1 - e^{-\lambda_{12}(t-s)} \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.4}$$

where

$$P_{01}(s,t) = \lambda_{01} e^{-\lambda_{12}(t-s)} (t-s)^{I(\lambda_{01}+\lambda_{02}=\lambda_{12})} \left(\frac{1 - e^{-(t-s)(\lambda_{01}+\lambda_{02}-\lambda_{12})}}{\lambda_{01}+\lambda_{02}-\lambda_{12}} \right)^{I(\lambda_{01}+\lambda_{02}\neq\lambda_{12})} . \tag{4.5}$$

The expected lifetime with disease is $E[T_K|Z(0) = 0] = \int_0^\infty P(T_K \geq t|Z(0) = 0)dt$. More generally, we can consider the expected lifetime given that $Z(t) = k$ for some $t > 0$, and

$$E[T_K|Z(t) = k, Z(0) = 0] = t + \int_t^\infty P(T_K \geq u|Z(t) = k, Z(0) = 0)du.$$

For the illness-death model,

$$E[T_2|Z(t)=k, Z(0)=0] = t + \int_t^\infty 1 - P_{k2}(t,u)du$$

where $P_{k2}(t,u)$ is the $(k,2)$ entry of (4.4). Specifically, we have

$$E[T_2|Z(t)=0, Z(0)=0] = t + \frac{1}{\lambda_{01}+\lambda_{02}} + \frac{\lambda_{01}}{\lambda_{12}(\lambda_{01}+\lambda_{02})}$$

and

$$E[T_2|Z(t)=1, Z(0)=0] = t + \frac{1}{\lambda_{12}}.$$

We may derive the transition probability matrix for the sample obtained by prevalent cohort sampling, $\mathbb{P}(s,t|\Delta=1)$, via (4.2) as

$$\begin{pmatrix} P_{00}(s,t|\Delta=1) & P_{01}(s,t|\Delta=1) & 1-P_{00}(s,t|\Delta=1)-P_{01}(s,t|\Delta=1) \\ 0 & \frac{1+\lambda_{12}t}{1+\lambda_{12}s}\,e^{-\lambda_{12}(t-s)} & 1-\frac{1+\lambda_{12}t}{1+\lambda_{12}s}\,e^{-\lambda_{12}(t-s)} \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.6}$$

where

$$\begin{aligned} P_{00}(s,t|\Delta=1) &= \left(\frac{t\lambda_{12}(\lambda_{01}+\lambda_{02})+\lambda_{01}+\lambda_{12}}{s\lambda_{12}(\lambda_{01}+\lambda_{02})+\lambda_{01}+\lambda_{12}}\right)e^{-(\lambda_{01}+\lambda_{02})(t-s)}, \\ P_{01}(s,t|\Delta=1) &= \left(\frac{t(\lambda_{12}+1)(\lambda_{01}+\lambda_{02})}{s\lambda_{12}(\lambda_{01}+\lambda_{02})+\lambda_{01}+\lambda_{02}}\right)P_{01}(s,t), \end{aligned}$$

and $P_{01}(s,t)$ is given in (4.5).

Comparing (4.4) and (4.6), it is clear that the transition probabilities differ in the population and in the induced sample, but the magnitude of the difference, which is a function of (s,t) and the population transition intensities λ_{01}, λ_{02}, and λ_{12}, is not immediately obvious. In Figure 4.1, we see the probability of entering state 2 at time t given that state 0 is occupied at time s; $s = 0, 2, 4$ are considered in the left, middle, and right panels respectively. In each panel, the population-level probability (solid line) is overlaid with the

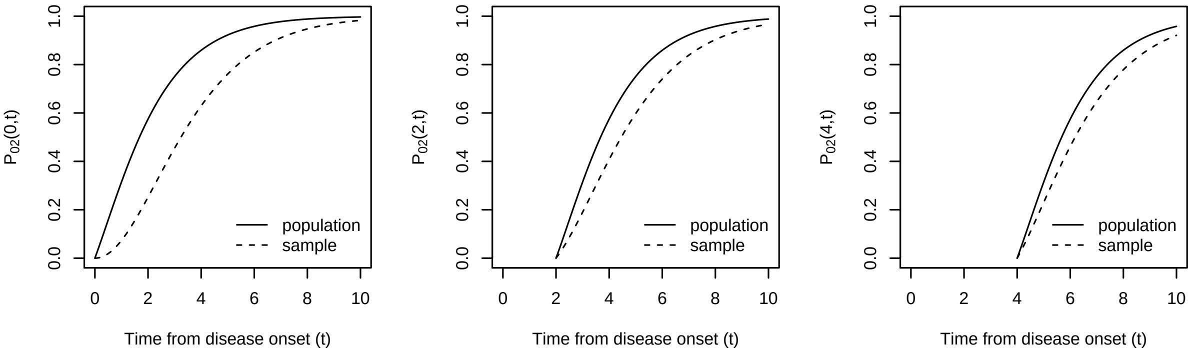


Figure 4.1: $P_{02}(s,t)$ for the population and length-biased sample with $\mathcal{S} = \{0, 1\}$, $\lambda_{01} = 0.5, \lambda_{02} = 0.3, \lambda_{12} = 0.7$

sample-level equivalent (dashed line). We see here that the cumulative distribution function is consistently lower in the sample than in the population as we vary s and t, which agrees with the results from length-biased survival analysis [Wolfson et al., 2001]. The same trends are exhibited in plots of the transition probability $P_{12}(s,t)$ vs $P_{12}(s,t|\Delta = 1)$ which are not shown in this thesis. This illustration demonstrates the necessity of conditioning on recruitment criteria for the analysis of multistate data arising from prevalent cohort studies to reflect the fact that the sample is not representative of the study population in general.

We have shown here that the well-known impact of length-biased sampling arising from prevalent cohort sampling has a corresponding effect on the transition intensities of a more general multistate process. This means that for progressive conditions such as dementia studied by Wolfson et al. [2001] if one were to model the progressive process via the decline in cognitive ability naively, one would underestimate the progression rate of cognitive decline if there were an association between cognitive ability and the risk of death. This has important public health implications. Fortunately, the correct analyses are relatively straightforward and we explore the use of biased sampling schemes in what follows with a view to efficient design of observational studies.

In the remainder of this chapter we explore designs based on state-dependent sampling, where samples are intentionally not representative, to improve efficiency and/or minimize expected study cost.

4.3 State-dependent Sampling for Failure Times

Failure-time models are the simplest form of multistate processes (see Figure 1.1(a)); we begin our discussion of state-dependent sampling by comparing the value of left- and right-truncated failure-time data.

Truncation arises when subjects are sampled subject to pre-specified criteria based on the response of interest. Left-truncation occurs when subjects must be at risk for a particular event to enter the study, and so cannot have experienced this event prior to a contact time. Right-truncation, a far less studied phenomenon, refers to the setting in which subjects must have already experienced the event of interest prior to entering the study.

As a motivating example, we consider the problem of characterizing the incidence of psoriatic arthritis among individuals with psoriasis, introduced in Section 1.5.1. Recall that two cohorts of individuals, one with psoriasis and one with psoriatic arthritis, have been established. Both of these cohorts are within the purview of the Centre for Prognosis Studies in Rheumatic Disease at the Toronto Western Hospital, and recruitment is ongoing. As such, it is natural to consider the relative value of information per individual in each cohort. Setting aside mortality, which is relatively low in both cohorts, the time from the onset of psoriasis to psoriatic arthritis is left-truncated in the psoriasis cohort and right-truncated in the psoriatic arthritis cohort. This idea will then be extended to more general state-dependent sampling schemes for multistate processes in Section 4.4.

4.3.1 Score and Information for Truncated Failure-time Processes

We derive the score vector and information matrix for a general truncated time-to-event process, allowing for right- and/or left-truncation. Suppose time T can only be observed if $T \in W$, where $W = [L, R]$ is the truncation interval; note $L = 0$ and $R > 0$ for right-truncated data and $L > 0$ and $R = \infty$ for left-truncated data; otherwise data are interval-truncated. In this general setting with a sample of size n we can write the likelihood function as

$$L(\theta) = \prod_{i=1}^{n} f(t_i | T_i \in W_i; \theta) = \prod_{i=1}^{n} \frac{f(t_i; \theta)}{P(T_i \in W_i; \theta)}$$

and the loglikelihood function

$$l(\theta) = \sum_{i=1}^{n} \{\log f(t_i; \theta) - \log P(T_i \in W_i; \theta)\}$$

Taking derivatives of the loglikelihood function with respect to a $p \times 1$ parameter vector θ indexing the survival function of the event time of interest, $\mathcal{F}(t; \theta)$, yields the score vector

$$\begin{aligned} S(\theta) = \frac{\partial l(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \left\{ \frac{\partial \log(f(t_i; \theta))}{\partial \theta} - \frac{1}{P(T_i \in W_i; \theta)} \frac{\partial P(T_i \in W_i; \theta)}{\partial \theta} \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{\partial \log(f(t_i; \theta))}{\partial \theta} + \frac{1}{P(T_i \in W_i; \theta)} \frac{\partial P(T_i \in W_i^c; \theta)}{\partial \theta} \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{\partial \log(f(t_i; \theta))}{\partial \theta} + \left[\frac{P(T_i \in W_i^c; \theta)}{P(T_i \in W_i; \theta)} \right] \frac{\partial \log P(T_i \in W_i^c; \theta)}{\partial \theta} \right\}. \end{aligned} \qquad (4.7)$$

where W_i^c denotes the complement of the truncation interval W_i.

Turnbull [1976] introduced the idea of ghosts, pseudo-individuals who are envisioned as existing in the population but who are not sampled because they did not satisfy the respective truncation conditions. We can think of J_i as the number of such candidates, similar to individual i but that were not selected for the sample because their event time fell outside the truncation region W_i. It follows that $J_i \sim \text{GEO}(p_i)$, where $p_i = P(T \in W_i^c;\theta)/P(T \in W_i;\theta)$. In light of this, we re-write (4.7) as

$$\frac{dl(\theta)}{d\theta} = \sum_{i=1}^{n} \left\{ \frac{\partial \log f(t_i;\theta)}{\partial\theta} + E[J_i|W_i;\theta] \frac{\partial \log P(T_i \in W_i^c;\theta)}{\partial\theta} \right\}. \tag{4.8}$$

We may use (4.8) as the basis for implementing an EM algorithm to estimate θ. Note that the first term in (4.8) is the score contribution for individuals who are observed to fail at time t_i, while the second term is the score contribution for ghosts whose failure times are censored in the "complete" sample because all that is known is that their failure time is in W_i^c. If T follows a Weibull distribution, (4.8) can be easily maximized in R using the `survreg` function, updating the estimates of θ at each iteration of the EM algorithm by specifying a weight for the contribution of the second term based on $E[J_i|W_i;\theta^{k-1}]$ where θ^{k-1} is the estimate at the $(k-1)st$ iteration.

We can derive the information matrix by differentiating the expression in (4.8) as follows

$$\begin{aligned} I(\theta) = -\sum_{i=1}^{n} \Bigg\{ & \frac{\partial^2 \log f(t_i;\theta)}{\partial\theta^2} + \left[\frac{P(T_i \in W_i^c;\theta)}{P(T_i \in W_i;\theta)} \right] \frac{\partial^2 \log P(T_i \in W_i^c;\theta)}{\partial\theta^2} \qquad (4.9) \\ & + \left[\frac{P(T_i \in W_i;\theta)\ \partial P(T_i \in W_i^c;\theta)/\partial\theta \ - \ P(T_i \in W_i^c;\theta)\ \partial P(T_i \in W_i;\theta)/\partial\theta}{P(T_i \in W_i;\theta)^2} \right] \frac{\partial \log P(T_i \in W_i^c;\theta)}{\partial\theta} \Bigg\}. \end{aligned}$$

Alternatively, by casting truncation as a missing data problem, we can make use of the approach from Louis [1982] for evaluating the observed information. Let $S(\theta;T)$ and $U(x;\theta)$ be the score functions based on the complete and observed data, respectively. Louis [1982] notes that if $U(\theta;T \in W) = \partial \log P(T \in W;\theta)/\partial\theta$ and $S(\theta;T) = \partial \log f(T;\theta)/\partial\theta$, one can write

$$U(\theta; T \in W) = E_T[S(\theta; T) | T \in W].$$

Moreover, if $I(\theta; T \in W) = -\partial U(\theta; T \in W)/\partial\theta'$ and $B(\theta; T) = -\partial S(\theta; T)/\partial\theta'$, then

$$I(\theta; T \in W) = E_T[\mathcal{B}(\theta; T)|T \in W] - \Big(E_T[S(\theta; T)S'(\theta; T)|T \in W] - U(\theta; T \in W)U'(\theta; T \in W)\Big). \tag{4.10}$$

4.3.2 Asymptotic Results

In this section, we illustrate the relative efficiency of estimates from left- and right-truncated samples for a time-to-event response T with probability density function $f(t; \theta)$ indexed by θ. Let V be the latent truncation time with density $g(u; \rho)$ indexed by ρ; T and V are assumed to be independent. Under a right-truncation sampling scheme, individuals may only be recruited if $T < V$, while under a left-truncation scheme, T may be sampled only if $T > V$. Let α be the probability of exclusion from the sample due to truncation, such that for right truncation we have

$$\alpha = P(T > V) = \int_0^\infty g(u; \rho)P(T > u; \theta)$$

and for left truncation we have

$$\alpha = P(T < V) = \int_0^\infty g(u; \rho)P(T < u; \theta).$$

The special case of a common truncation time V for all individuals is achieved by setting a degenerate distribution on V.

Suppose interest lies in estimating β, the effect of a covariate X on the failure-time response T. A natural question arises: what is the value of recruiting a sample of left-truncated individuals and following them prospectively for up to τ years relative to that of recruiting a sample of right-truncated individuals and retrospectively recording the response. To explore this, we assume common truncation probabilities α for the left- and right- truncated samples (in general these correspond to different truncation times V). Let C_0 denote the cost of initial sampling and C the cost per year of follow-up. Let n_L and n_R be the number of individuals in each of the left- and right-truncated samples. To facilitate fair comparisons, we equate the costs of obtaining left- and right-truncated samples,

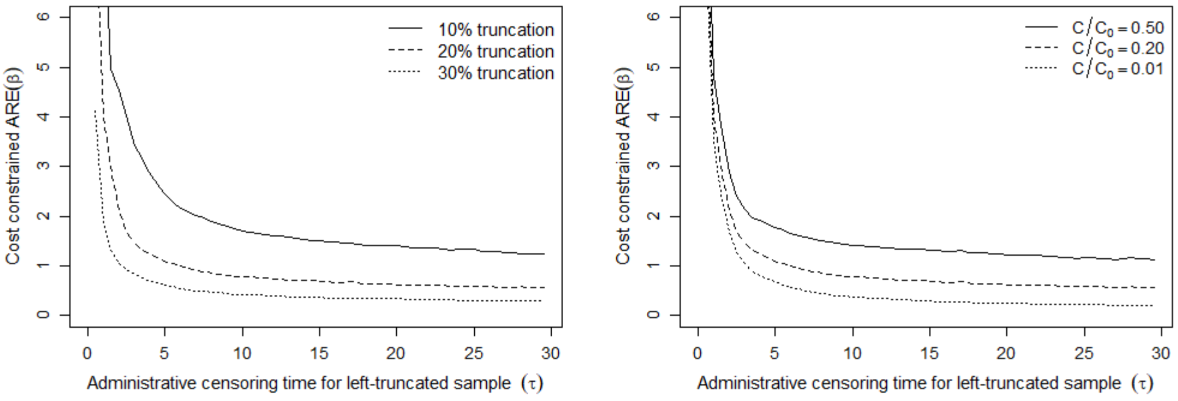
$$n_L(C_0 + DC) = n_R C_0.$$

The cost-effective asymptotic relative efficiency of left-truncated vs right-truncated samples with the same expected cost is then

$$ARE = \frac{[\mathcal{I}_L^{-1}(\theta)]_{\beta\beta} \,/\, n_L}{[\mathcal{I}_R^{-1}(\theta)]_{\beta\beta} \,/\, (1 + E[D]C/C_0)n_L} = \frac{[\mathcal{I}_L^{-1}(\theta)]_{\beta\beta}}{[\mathcal{I}_R^{-1}(\theta)]_{\beta\beta}} \left(1 + E[D]C/C_0\right)$$

where $\mathcal{I}_L(\theta)$ and $\mathcal{I}_R(\theta)$ are the Fisher information matrices for left- and right-truncated samples, obtained by taking the expectation of $I(\theta; T \in W)$ from (4.9) with $W = (V, \infty)$ and $W = (0, V)$ respectively, $[\mathcal{I}^{-1}(\theta)]_{\beta\beta}$ is the diagonal entry of $\mathcal{I}^{-1}(\theta)$ giving the variance of β, and $D = \min(T - V, \tau)$ is the duration of follow-up for an individual from a left-truncated sample subject to administrative right censoring at τ. Let $T \sim WEI(\lambda, p)$, where λ and p are the rate and shape parameters respectively, and X a binary covariate with a multiplicative effect on the hazard , so $h(t) = \lambda p t^{p-1} e^{\beta X}$. We set $p = 1.25$, choose λ such that the median of T is 15 years when $X = 0$ (e.g. $P(T > 15|X = 0; \lambda, p) = 0.5$), let $P(X = 1) = 0.5$ and let $\beta = \log 1.5$. We assume a fixed truncation time in each of the right- and left-truncated samples, and consider truncation probabilities $\alpha = \{0.10, 0.20, 0.30\}$. For costs, we consider $C/C_0 = \{0.50, 0.20, 0.01\}$.

In both panels of Figure 4.2, we plot the cost-effective asymptotic relative efficiency (ARE) of the regression coefficient β as a function of the administrative censoring time



(a) $C/C_0 = 0.2$

(b) 20% truncation

Figure 4.2: Cost-effective asymptotic relative efficiency (ARE) for the estimation of regression coefficient β on a failure time T for a sample subject to left-truncation (with up to τ years of follow-up) vs a sample subject to right-truncation (with retrospective ascertainment of T); $T \sim WEIB(\lambda, p)$ where $\lambda = 1.5$ and $P(T > 15|X = 0; \lambda, p) = 0.5$, $P(X = 1) = 0.5$, $\beta = \log 1.5$.

τ for the left-truncated sample. In Figure 4.2(a), costs are fixed at $C/C_0 = 0.2$ and the solid, dashed, and dotted lines represent increasing truncation probabilites, $0.10, 0.20, 0.30$ respectively. Values of $ARE > 1$ indicate that estimates of β from a right-truncated sample are more efficient than those from a left-truncated sample with the same expected cost, while $ARE < 1$ indicates that estimates from the left-truncated sample are more efficient. When τ is small, many of the responses T from left-truncated observations are administratively censored so it is more cost-effective to obtain a right-truncated sample. As expected, as τ increases, the relative value of left-truncated observations increases. Also, when the extent of truncation is more severe (that is when a higher proportion of the population is excluded from the sample due to the truncation mechanism), left-truncated observations with prospective follow-up yield more efficient estimators of β than do right-truncated samples. In Figure 4.2(b), the truncation probability is fixed (at 0.20) and the

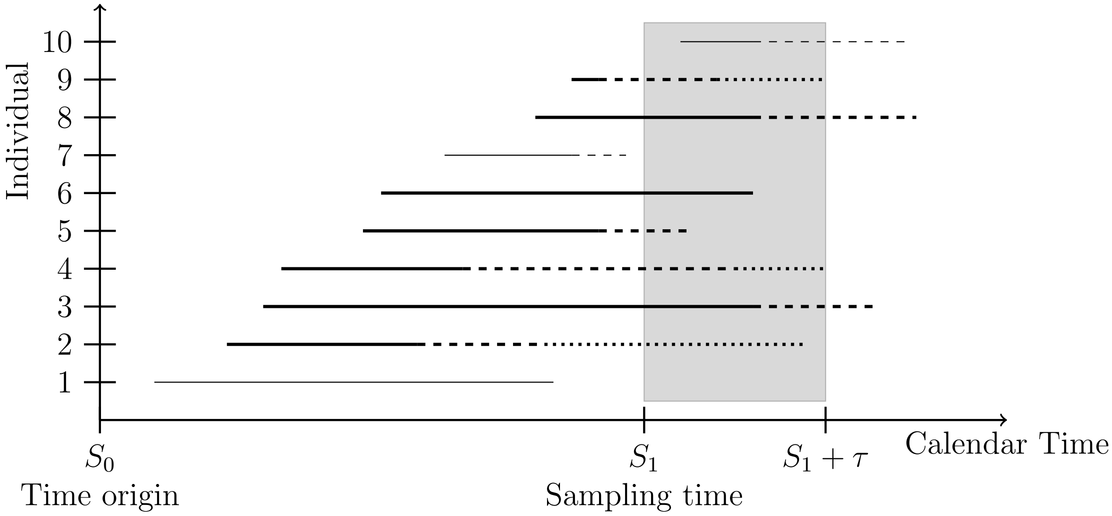


Figure 4.3: Course of disease for 10 individuals from a population; solid, dashed, and dotted lines represent duration in states 0, 1, and 2 respectively; bolded lines correspond to individuals eligible to be sampled at the time of accrual S_1 due to having $Z(S_1) \in \mathcal{S} = \{0, 1, 2\}$.

lines represent different costs of follow-up relative to that of initial recruitment, $C/C_0 =$ 0.01, 0.20, and 0.50 respectively. As expected, as the cost of follow-up decreases relative to that of recruitment, left-truncated (and right-censored) samples yield increasingly efficient estimates of β compared to those from right-truncated samples with the same expected cost.

4.4 State-dependent Recruitment for Follow-up

We now return attention to sampling issues for the more general $K + 1$ state progressive multistate model depicted in Figure 1.2.

4.4.1 Likelihood and Fisher Information

Let A denote the (random) age at contact. Suppose m_j^* individuals are recruited with $Z(A) = j$, and for each recruited individual, we obtain the value of the covariate of interest X, the exact time of transitions having occurred over $(0, A)$ (retrospectively), and the time of transitions over the prospective study period $(A, A + \tau)$, where τ is the planned duration of follow-up. We do not consider the issue of random censoring but the following results generalize easily to handle this. For example, Figure 4.3 depicts the recruitment of individuals at calendar time S_1 in states $\mathcal{S} = \{0, 1, 2\}$ from a population, and prospective follow-up over a period $(S_1, S_1 + \tau)$; individuals with bolded paths are eligible to be recruited into the study and the shaded gray box represents the prospective follow-up period. Individuals who are still alive at the end of the follow-up period are considered censored and no information about their trajectory beyond that point is known.

We adopt a Markov model with piecewise-constant intensities, a flexible model allowing for transition intensities to vary with age. As presented in Section 1.2, if $\mathcal{H}(t) = \{Z(s), 0 < s < \tau\}$ then under a Markov assumption the baseline transition intensities are

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

for $\ell = \{k + 1, K\}$, $k = 0, \ldots, K - 1$. Under a piecewise-constant model for baseline intensities $\lambda_{kl}(t)$ and assuming a multiplicative effect of a covariate X, we write

$$\lambda_{k\ell}(t) e^{x'\beta_{k\ell}} = \lambda_{k\ell r} e^{x'\beta_{k\ell}} = e^{\alpha_{k\ell r} + x'\beta_{k\ell}}, \qquad \text{for } t \in \mathcal{B}_r = [b_{r-1}, b_r),\ r = 1, \ldots, R.$$

Let $\theta = (\alpha', \beta')$ index this piecewise-constant multistate process (Figure 1.2), where $\alpha = (\alpha_1', \ldots, \alpha_{K-1}')'$ with $\alpha_k = (\alpha_{k,k+1}', \alpha_{kK}')'$, $k = 0, \ldots, K - 2$, $\alpha_{K-1} = \alpha_{K-1,K}'$, and $\alpha_{k\ell} = (\alpha_{k\ell 1}, \ldots, \alpha_{k\ell R})'$ is the vector of baseline log transition intensities and $\beta = (\beta_0', \ldots, \beta_{K-1}')'$

is the vector of regression coefficients, with $\beta_k = (\beta'_{k,k+1}, \beta'_{kK})'$, $k = 0, \dots, K-2$ and $\beta_{K-1} = \beta_{K-1,K}$.

Conditional on the age at screening and recruitment, the likelihood contribution from a single subject takes the form

$$\begin{aligned} L(\theta) &= \prod_{j \in \mathcal{S}} P(Z(s), 0 < s < a+\tau | A = a, X = x, Z(A) = j; \theta)^{I(Z(A)=j)} \\ &= \prod_{j \in \mathcal{S}} \left(\frac{P(Z(s), 0 < s < a+\tau | A = a, X = x; \theta)}{P(Z(a) = j | A = a, X = x; \theta)} \right)^{I(Z(A)=j)} . \end{aligned} \tag{4.11}$$

If $n_{k\ell r} = I\big[\text{subject has a } k \to \ell \text{ transition over } [b_{r-1}, b_r)\big]$ and

$$w_{kr} = \int_{b_{r-1}}^{\min(b_r, a+\tau)} I(Z(u) = k)\; du$$

is the duration that the subject is under observation in state k over $[b_{r-1}, b_r)$, the loglikelihood is

$$\log L(\theta) = \sum_{j \in \mathcal{S}} I(Z(A) = j) \left\{ \sum_{r=1}^{R} \sum_{k,\ell \le K} \left[n_{k\ell r} \log \lambda_{k\ell r} e^{x'\beta_{k\ell}} - w_{kr} \lambda_{k\ell r} e^{x'\beta_{k\ell}} \right] - \log P_{0j}(0, a | x; \theta) \right\} \tag{4.12}$$

where m_j^* is the sample size recruited from state j and $m^* = \sum_{j \in \mathcal{S}} m_j^*$ is the total sample size. Maximum likelihood estimates of θ will be obtained by directly maximizing the summation of the loglikelihood in (4.12) over all subjects in the sample.

To facilitate investigation of the impact of state dependent sample sizes on the precision of estimates from a prevalent cohort study, we consider the Fisher information. When m_j^* individuals are directly recruited from each state $j \in \mathcal{S}$ this takes the form

$$\mathcal{I}(\theta) = E\left[-\sum_{i=1}^{m^*} \frac{\partial^2 \log L(\theta)}{\partial \theta_u \partial \theta_v} \right] = \sum_{j \in \mathcal{S}} m_j^* \, E\left[- \frac{\partial^2 \log L(\theta)}{\partial \theta_u \partial \theta_v} \middle| Z(A) = j \right] . \tag{4.13}$$

This expectation is taken by conditioning on the state at enrollment (i.e. $Z(A) = j$) and averaging over the distribution of $(X, A) \mid Z(A) = j$,

$$
\begin{aligned}
E\left[-\frac{\partial^2 \log L}{\partial\theta_u\partial\theta_v}\middle| Z(A) = j\right] = \sum_x \int_0^\infty \Bigg\{ \Bigg[\sum_{r=1}^{R} \sum_{k=0}^{K-1} E\big(W_{kr} | A = a, Y = j, X = x\big) \left(\sum_{\ell=0}^{K} \frac{\partial^2}{\partial\theta_u\partial\theta_v} \lambda_{k\ell r} e^{x'\beta_{k\ell}} \right) \\
+ \frac{\partial^2}{\partial\theta_u\partial\theta_v} \log\left(\sum_{\ell\in\mathcal{S}} P_{0\ell}(0, a|x;\theta) \right) \Bigg] f(A = a|X = x, Z(A) = j) \Bigg\} da \\
\times P(X = x|Z(A) = j) \qquad (4.14)
\end{aligned}
$$

Assuming the birth process in the population is stationary and there are no trends in the disease process over time, the probability of screening an individual and finding them in state j (i.e. with $Z(A) = j$) is equal to the prevalence of being in state j among $\mathcal{S}$ in the population. As shown in Zeng and Cook [2018], this probability is

$$
P(Z(A) = j|Z(0) = 0) = \frac{\sum_{x=0}^{1} \left[\int_0^\infty P_{0j}(0, u|x)du\right] P(X = x)}{\sum_x \left(P_{0j}(0, u|x)du\right) P(X = x)}, \qquad (4.15)
$$

the density for the age at sampling in state j given $X = x$ is

$$
f_A(a|Z(A) = j, X = x) = \frac{P_{0j}(0, a|x)}{\int_0^\infty P_{0j}(0, u|x)du}, \qquad (4.16)
$$

and

$$
P(X = x, Z(A) = j|Z(0) = 0) = \frac{\left[\int_0^\infty P_{0j}(0, u|x)du\right] P(X = x)}{\sum_x \left(P_{0j}(0, u|x)du\right) P(X = x)}. \qquad (4.17)
$$

Based on these results, under a piecewise-constant model we have

$$E[W_{kr}|A=a,Z(a)=j,X=x]$$
$$= \begin{cases} \int_{b_{r-1}}^{a} \frac{\sum_{h=0}^{K-1} p_{0h}(0,b_{r-1}|x)p_{hk}(b_{r-1},u|x)p_{kj}(u,a|x)}{p_{0j}(0,a|x)}du \; + \; \int_{a}^{\min(b_r,a+\tau)} p_{jk}(a,u|x)du & \text{if } a \in [b_{r-1},b_r), \\ \sum_{h=0}^{K-1} p_{jh}(a,b_{r-1}|x) \int_{b_{r-1}}^{\min(b_r,a+\tau)} p_{hk}(b_{r-1},u|x)du & \text{if } a < b_{r-1} \\ \int_{b_{r-1}}^{b_r} \left(\frac{\left(\sum_{h,\ell=0}^{K-1} p_{0h}(0,b_{r-1}|x)p_{hk}(b_{r-1},u|x)\right)\left(p_{k\ell}(u,b_r|x)p_{\ell j}(b_r,a|x)\right)}{p_{0j(0,a|x)}} \right) du & \text{if } a \geq b_r \end{cases}, \tag{4.18}$$

Derivatives of $P_{0j}(0,a|x;\theta)$ for $j \in \mathcal{S}$ can be taken analytically or using matrix formulations as in Kalbfleisch and Lawless [1985] and Kosorok and Chao [1995, 1996].

4.4.2 Efficient Sampling with Expected Cost Constraints

Here we consider the cost of recruiting and following individuals in different states. In reality, the duration of follow-up will be random and depend on the survival distribution. The expected time at risk in state k over the study period for an individual sampled with $Z(A)=j$ is obtained by averaging over the age at enrollment A and covariate X. Thus $E\left[W_k \middle| Z(A)=j;\theta\right]$ is

$$\sum_{x=0}^{1} \left\{ \int_0^\infty \left[\int_a^{a+\tau} p_{jk}(a,u|x)du \right] f_A(a|Z(A)=j,X=x)da \right\} P(X=x|Z(A)=j) \tag{4.19}$$

where $P(X=x|Z(A)=j)$ is given by the ratio of (4.17) to (4.15) from the previous section.

Let C_{0j} be the cost of initial accrual for an individual with $Z(A)=j$. In general, we assume the cost of follow-up varies as a function of the state occupied (e.g. the cost may be higher when physical examinations or testing is required to assess disease status than when only vital status is to be monitored); let C_{1k} be the cost of one year of follow-up in state k. The expectation of the total cost of recruiting and following up an individual with $Z(A) = j$ for up to τ years (denoted by C_j) is

$$E[C_j | Z(A) = j; \theta] = C_{0j} + \sum_{k=0}^{K-1} C_{1k} E[W_k | Z(A) = j; \theta], \tag{4.20}$$

which depends on the planned follow up time τ, costs C_{0j} and C_{1j} for $j = 0, \ldots, K-1$, and the disease process parameters.

4.4.3 Illustrative Design Setting

To examine the impact of the choice of state-specific sampling targets on the design of prevalent cohort studies, we focus on a piecewise-constant model with $R = 4$ pieces to accommodate different transition intensities for different age groups: 0-20 years, 20-50 years, 50-80 years, and greater than 80 years of age. This illustration is motivated by the research program in psoriatic arthritis presented in Section 1.5.1. We focus on a four-state model where states 0, 1, and 2 represent 'healthy', 'psoriasis', and 'psoriatic arthritis' states respectively, and state $K = 3$ represents death; see Figure 4.4. Only the individuals who are either in state 1 (psoriasis) or state 2 (psoriatic arthritis) will be considered for follow-up, so the set of sampling states is $\mathcal{S} = \{1, 2\}$.

> *Constraints on mortality:* There is little agreement in the literature regarding the extent to which psoriasis and psoriatic arthritis affect mortality rates [Ogdie et al., 2014, Gladman, 2008, Gelfand et al., 2007]; we assume here that the death intensity in age interval r is independent of the state occupied so $\alpha_{j3r} = \gamma_r$ for $Z(A) = j = 0, 1, 2$. Values of γ_r are chosen to reflect mortality in the Canadian population based on rates found in reports from Statistics Canada [2018] through the following

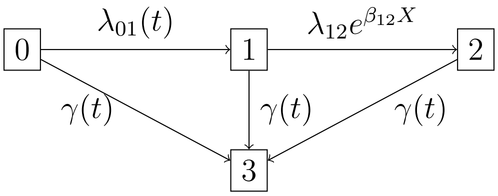


Figure 4.4: Parameterization of four-state reduced model with $R = 4$, where $\lambda_{01}(t) = \exp(\alpha_{011})$ for $t < b_2$ and $\lambda_{01}(t) = \exp(\alpha_{013})$ otherwise, $\lambda_{j3}(t) = \exp(\gamma(t)) = \exp(\gamma_r)$ for $t \in [b_{r-1}, b_r)$, $r = 1, 2, 3, 4$, and $\lambda_{12}(t|x) = \exp(\alpha_{12} + \beta_{12}x)$.

constraints: $P(Z(20) = 3) = 0.008$, $P(Z(50) = 3) = 0.036$, $P(Z(80) = 3) = 0.342$, and $P(Z(100) = 3) = 0.997$

Constraints on incidence of Ps: We assume $\alpha_{011} = \alpha_{012}$, $\alpha_{013} = \alpha_{014}$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P(0 \to 1 \text{ in lifetime}) = 0.03$ [Eder et al., 2011b].

Constraints on incidence of PsA: We assume the rate of psoriatic arthritis onset is constant (i.e. $\alpha_{12r} = \alpha_{12}$ for $r = 1, 2, 3, 4$) and $P(1 \to 2 \text{ in lifetime}) = 0.30$ is the probability that individuals with psoriasis develop psoriatic arthritis in their lifetime [Eder et al., 2011b].

Covariate effects: Finally, we assume a binary covariate X has an effect on the $1 \to 2$ transition only, with $\beta_{12} = \log 2$. We let $P(X = 1) = \{0.50, 0.05\}$, and there are no covariate effects on the other transitions.

In this illustration, interest lies in estimating $\theta_1 = (\alpha_{12}, \beta_{12})$, while assuming values for $\theta_2 = (\alpha_{011}, \alpha_{013}, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ are known and $\theta = (\theta_1, \theta_2)$; see Figure 4.4 for a graphical representation of the parameters in this reduced mode. More generally, auxiliary data could be used to aid in the estimation of these additional parameters; this will be discussed further in Section 4.6.

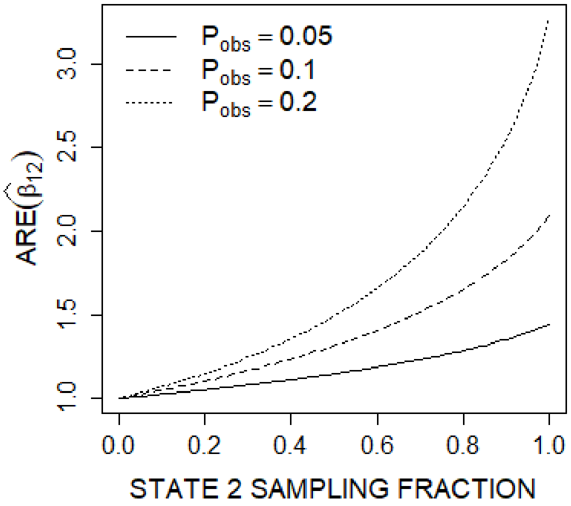
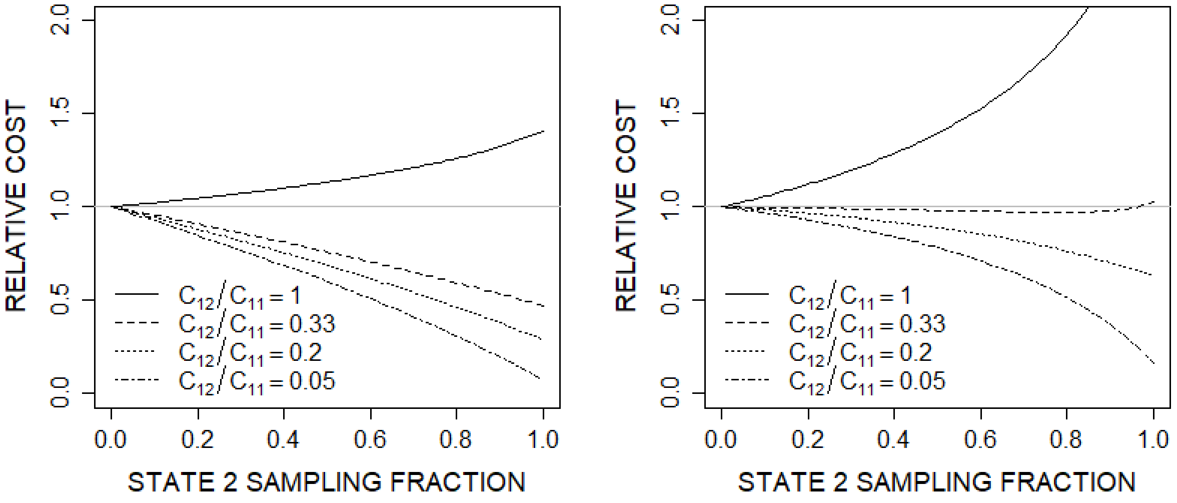


Figure 4.5: Asymptotic relative efficiency (ARE) of the estimator $\hat{\beta}_{12}$ given state 2 sampling fractions $m_2^*/(m^*)$ vs $m_2^*/m^* = 0$, for a fixed sample of size m^* when jointly estimating $\theta_1 = (\alpha_{12}, \beta_{12})$; $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$, $\beta_{12} = \log 2$, $P(0 \to 1) = 0.03$, $P(1 \to 2) = 0.30$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P(X = 1) = 0.05$

When the total sample size $m^* = m_1^* + m_2^*$ is fixed, the asymptotic variance of $\hat{\beta}_{12}$ is minimized when $m^* = m_1^*$, that is when a sample is drawn from state 1 and followed prospectively for entry into states 2 and/or 3; this is shown in Figure 4.5. As β_{12} modulates the 1 → 2 transition, this demonstrates that prospective follow-up carries more information than retrospective information (as would be provided by individuals with $Z(A) = 2$). Further, as the probability of observing 1 → 2 during prospective follow-up (e.g. $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$) increases, the efficiency gap between prospective and retrospective observation of 1 → 2 widens, as expected. Note that given θ, large values of P_{obs} imply long durations of follow-up τ; in Figure 4.5, values $P_{obs} = \{0.05, 0.10, 0.20\}$ map to $\tau = \{4.5, 10.1, 29.1\}$ years respectively. However, the cost of prospective monitoring for disease progression is higher than that of retrospective inspection so comparisons on the basis of a fixed sample size do not convey the whole picture. Figure 4.6 shows the



(a) $P_{obs} = 0.05$

(b) $P_{obs} = 0.20$

Figure 4.6: Expected cost of studies designed given the state 2 sampling fraction π_2 and achieving 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$ with $\alpha = 0.05$, relative to that of the design with $m_2^* = 0$, when jointly estimating $\theta_1 = (\alpha_{12}, \beta_{12})$; C_{1k} is the cost of one year of follow-up in state k, $C_{0k} = 0$ for all k, $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$, $\beta_{12} = \log 2$, $P(0 \to 1) = 0.03$, $P(1 \to 2) = 0.30$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P(X = 1) = 0.05$

impact of the cost of prospective follow-up on expected study cost, when the sample sizes (m_1^*, m_2^*) are chosen to achieve 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$ with significance level $\alpha = 0.05$, while constrained to various values for the state 2 samping fraction $\pi_2 = m_2^*/(m_1^* + m_2^*)$. We assume the cost of follow-up is lower for individuals in state 2 than in state 1 (i.e. $C_{12} < C_{11}$) as only vital status is to be monitored. In this figure, the relative expected cost is plotted against the state 2 sampling fractions defining these designs, where

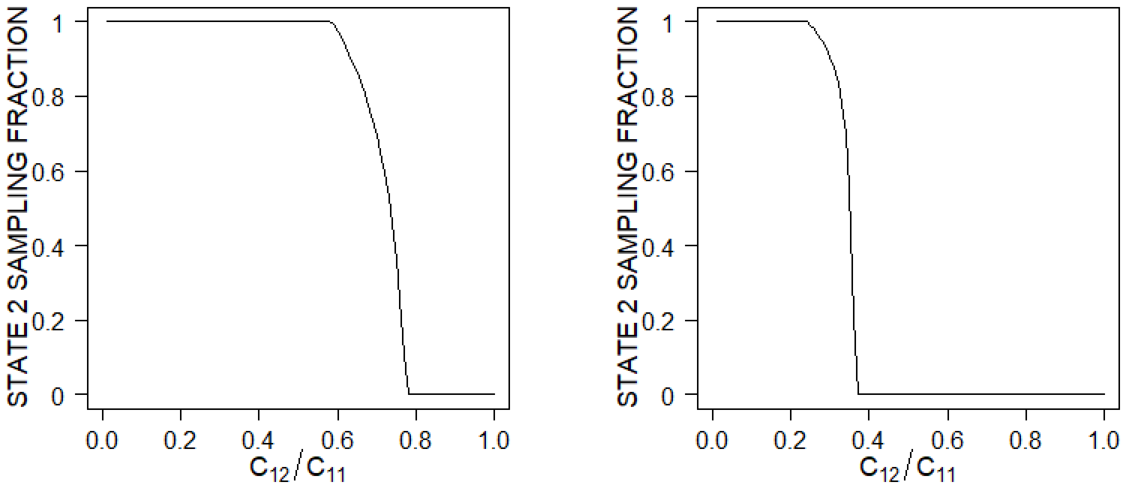
$$\text{RELATIVE COST } = \frac{m^*(\pi_2)\big[(1-\pi_2)E[C_1] + \pi_2 E[C_2]\big]}{m^*(\pi_2 = 0)E[C_1]},$$

where $m^*(\pi_2)$ is the total sample size required to achive a desired power (for example, 80%) for β_{12} with a sampling fraction π_2 and the expectations in the numerator and denominator are given in (4.20). Each line represents a family of designs (m^*, π_2) achieving 80% power for estimation of β_{12} with a given yearly follow-up cost ratio C_{12}/C_{11}; as π_2 varies, the expected study cost varies and we define the *minimum-cost design (for β_{12})* to be the design which minimizes expected study cost while achieving 80% power for β_{12}. For many values of C_{12}/C_{11}, these minimum-cost designs have $\pi_2 = 0$ or $\pi_2 = 1$, the latter preferable when the cost of prospective monitoring of vital status alone is much lower than that of monitoring for entry into state 1 (e.g. when C_{12}/C_{11} is small). However, in some instances the minimum-cost design is achieved by sampling from both states 1 and 2. This can be seen more clearly in Figure 4.7, where the state 2 sampling fraction π_2 corresponding to minimum-cost designs for β_{12} is plotted against the cost ratio C_{12}/C_{11}. Here we see that within each panel (corresponding to increasing durations of follow-up τ), there is a range of cost ratios C_{12}/C_{11} for which the minimum-cost designs for β_{12} involve sampling from both states. While it can be seen that minimum-cost designs feature $m_2^* \gg m_1^*$ when $C_{12} \ll C_{11}$, it is important to note that due to the lower information 'per-individual' for those sampled from state 2, the absolute sample sizes for these designs (to achieve 80% power for β_{12}) are much larger than those of minimum-cost designs when $m_1^* \gg m_2^*$. For example, designs $(m_1^* = 965, m_2^* = 0)$ and $(m_1^* = 0, m_2^* = 3152)$ both achieve 80% power for β_{12} when jointly estimating $\theta_1 = (\alpha_{12}, \beta_{12})$ with $P_{obs} = 0.2$, $\beta_{12} = \log 2$, $P(0 \to 1) = 0.03$, $P(1 \to 2) = 0.30$, and $P(X = 1) = 0.05$.

4.5 Screening and State-dependent Sampling

4.5.1 Likelihood and Fisher Information

The feasibility of the state-dependent sampling scheme presented in the previous section relies on the assumption that there exist populations from which it is possible to directly



(a) $P_{obs} = 0.05$

(b) $P_{obs} = 0.20$

Figure 4.7: Optimal sampling fraction π_2 for minimum-cost designs achieving 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$, with $C_{12}/C_{11} \in (0, 1]$; when jointly estimating $\theta = (\alpha_{12}, \beta_{12})$; $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$, $\beta_{12} = \log 2$, $P(0 \rightarrow 1) = 0.03$, $P(1 \rightarrow 2) = 0.30$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P(X = 1) = 0.05$

recruit individuals from each state. To do so, individuals must be screened from the population at large until the desired number of individuals from each state has been identified. Augmenting the likelihood L in (4.11) with current status information $(Z(A), A, X)$ for all screened individuals furnishes additional information about the multistate process. This additional contribution from an individual simply screened and found to be in state j is

$$\begin{aligned} L_A(\theta) &= \prod_{j \in \mathcal{S}} P(Z(a) = j | A = a, X = x, Z(a) \in \mathcal{S})^{I(Z(A)=j)} \\ &= \prod_{j \in \mathcal{S}} \left(\frac{P(Z(a) = j | A = a, X, x)}{\sum_{\ell \in \mathcal{S}} P(Z(a) = \ell | A = a, X = x)} \right)^{I(Z(A)=j)} \end{aligned}$$

Let $\eta = 1$ for screened individuals who are selected for follow-up, and $\eta = 0$ otherwise, let m_j^* the desired number of individuals to be recruited in state $j \in \mathcal{S}$ in the follow-up sample, and M the total number of screened individuals (note $m_j^* = \sum_{i=1}^{M} I(Z(A_i) = j)^{\eta_i}$). The full likelihood is

$$\begin{aligned}
L_2(\theta) &= \prod_{i=1}^{M} L(\theta)^{\eta_i} \times L_A(\theta) \\
&= \prod_{i=1}^{M} \prod_{j \in \mathcal{S}} \Bigg\{ \left(\frac{P(Z(s), 0 < s < a_i + \tau | A_i = a_i, X_i = x_i)}{P(Z(a_i) = j | A_i = a_i, X_i = x_i)} \right)^{\eta_i} \\
&\qquad \times \left(\frac{P(Z(a_i) = j | A_i = a_i, X_i, x_i)}{\sum_{\ell \in \mathcal{S}} P(Z(a_i) = \ell | A_i = a_i, X_i = x_i)} \right) \Bigg\}^{I(Z(A_i)=j)}
\end{aligned}$$

For the Fisher information, we combine the contributions for individuals selected for follow-up and those who are only screened, to write

$$E\left[-\frac{\partial^2 \log L_2(\theta)}{\partial\theta_u \partial\theta_v}\right] = \sum_{j \in \mathcal{S}} \left\{ m_j^* E\left[-\frac{\partial^2 \log L(\theta)}{\partial\theta_u \partial\theta_v} \Big| Z(A) = j\right] + E[M_j | \bar{m}^*] E\left[-\frac{\partial^2 \log L_A(\theta)}{\partial\theta_u \partial\theta_v} \Big| Z(A) = j\right] \right\} \tag{4.21}$$

where

$$E\left[-\frac{\partial^2 \log L(\theta)}{\partial\theta_u \partial\theta_v} \Big| Z(A) = j\right] = E\Bigg[- \frac{\partial^2}{\partial\theta_u \partial\theta_v} \log \left(\frac{P(\bar{Z}(a+\tau) | A = a, X = x)}{P(Z(a) = j | A = a, X = x)} \right) \Big| Z(A) = j \Bigg], \tag{4.22}$$

$$E\left[-\frac{\partial^2 \log L_A(\theta)}{\partial\theta_u \partial\theta_v} \Big| Z(A) = j\right] = E\left[-\frac{\partial^2}{\partial\theta_u \partial\theta_v} \log \left(\frac{P(Z(a) = j | A = a, X = x)}{\sum_{\ell \in \mathcal{S}} P(Z(a) = \ell | A = a, X = x)} \right) \Big| Z(A) = j \right], \tag{4.23}$$

and we let $\bar{Z}(t) = \{Z(s), 0 < s < t\}$ be the history of the disease process up to time t, M_j the number of individuals screened and found to be in state j, and $\bar{m}^* = \{m_j^*, j \in \mathcal{S}\}$. The expectations in (4.22) and (4.23) are similar to that in (4.14), and details of $E[M_j | \bar{m}^*]$ are given in Appendix A. Augmenting the follow-up study with current-status information from screened but unselected individuals according to this scheme can greatly enhance precision of parameter estimates. In Figure 4.8, we revisit the example from Section 4.4.3 and plot reductions in the variance of $\hat{\beta}_{12}$ achieved when current status information from screening a population in this manner is collected, that is

$$\frac{asvar(\hat{\beta}_{12}) - asvar_2(\hat{\beta}_{12})}{asvar(\hat{\beta}_{12})}$$

where $asvar_2(\hat{\beta}_{12})$ is obtained from (4.21) and $asvar(\hat{\beta}_{12})$ is obtained based on (4.13). In this figure, 50-90% reductions in the variance of $\hat{\beta}_{12}$ are exhibited, depending on the state 3 sampling fraction. The magnitude of the gain in efficiency is driven by the expected sample screened $E[M]$ and its component terms $E[M_j]$, $j \in \mathcal{S}$, examples of which will be given in the next section.

4.5.2 Cost Considerations

The expected cost of a study where cost is incurred for screening individuals in the population builds upon the cost framework in Section 4.4.2. The expected cost of a study with recruitment targets m_j^* for $j \in \mathcal{S}$ is

$$E[C] = \sum_{j \in \mathcal{S}} C_0 E[M_j | m_1^*, m_2^*; \theta] + \sum_{j \in \mathcal{S}} m_j^* E[C_j | Z(A) = j; \theta]$$

where M_j is the number of screened individuals found to be in state $j < K$, C_0 is the cost of screening an individual from the population, C_j is the cost of prospective follow-up for a recruited individual with $Z(A) = j$ and $E[C_j | Z(A) = j; \theta]$ is as in (4.20)

4.5.3 Illustrative Design Setting

Again, we illustrate the effect of the state-specific recruitment targets (with screening) on the design of prevalent cohort studies. The settings mirror those given in Section 4.4.3. In

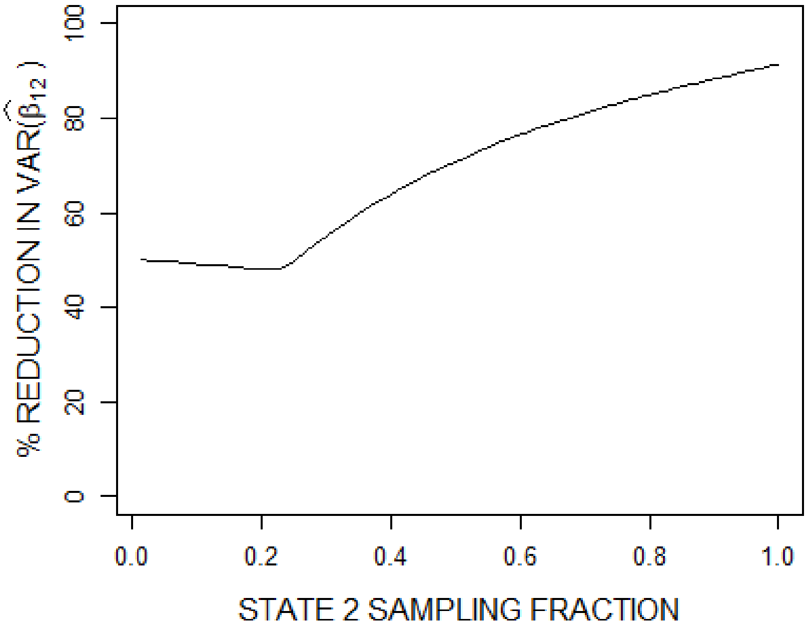


Figure 4.8: Percentage reduction in asymptotic variance of β_{12} by exploiting current status information from the screened population, when state-specific follow-up sample sizes (m_1^*, m_2^*) are chosen to achieve 80% power in the absence of follow-up; $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$, $\beta_{12} = \log 2$, $P(0 \to 1) = 0.03$, $P(1 \to 2) = 0.30$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P(X = 1) = 0.05$

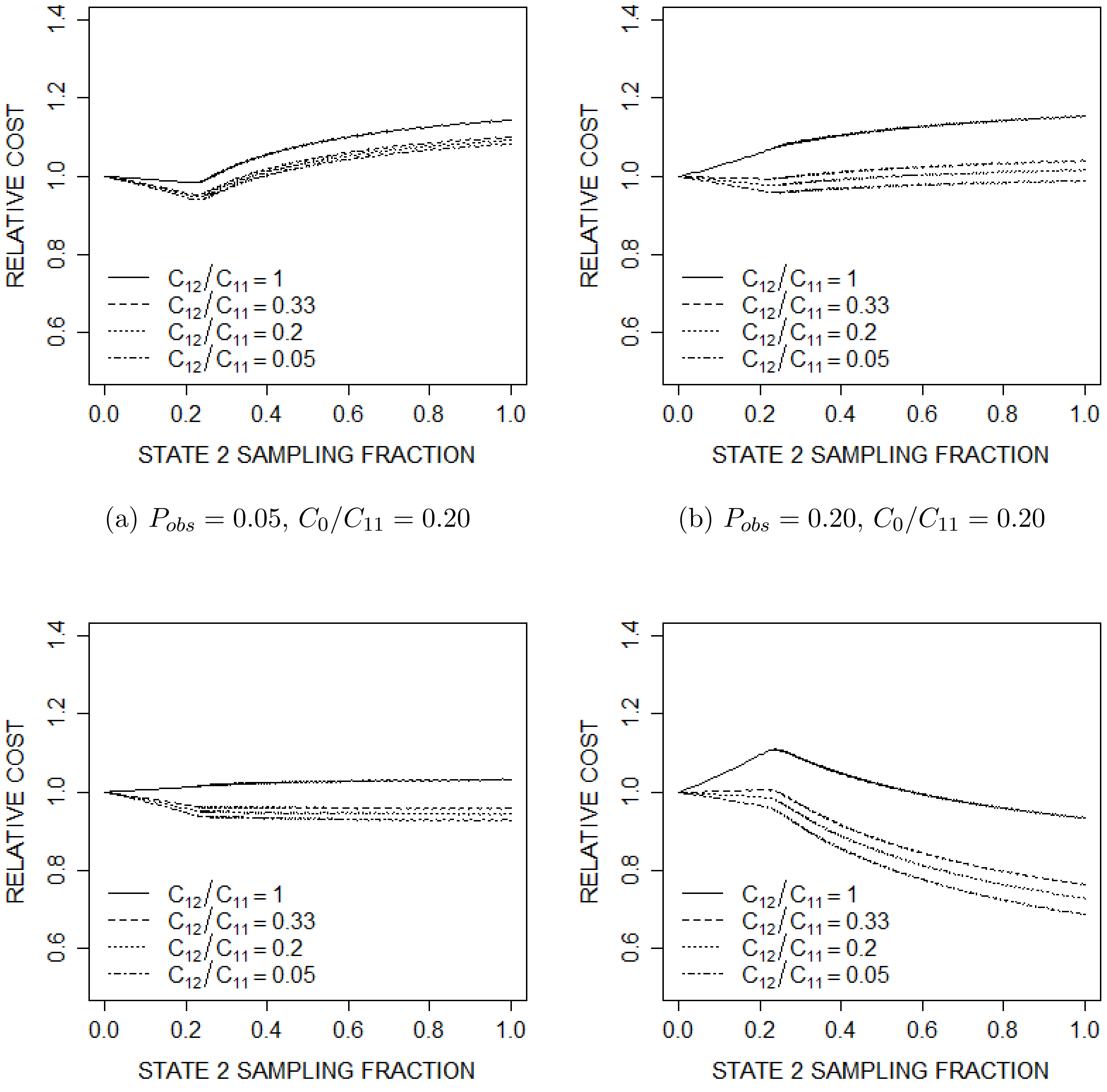
Table 4.1, several designs (m_1^*, m_2^*) are presented, each achieving 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$ with $\alpha = 0.05$ but corresponding to different values of P_{obs}. Across all of these settings, the number of screened individuals, particularly from the 'healthy' state 0, is orders of magnitude larger than the number of individuals selected for prospective follow-up, although certain state 2 sampling fractions $\pi_2 = m_2^*/(m_1^* + m_2^*)$ induce a smaller expected screening sample, for example when $P_{obs} = 0.20$, $(m_1^* = 450, m_2^* = 138)$ admits an almost 50% reduction in the expected screened sample size relative to $(m_1^* = 0, m_2^* = 268)$, although both achieve the same power for β_{12}. Due to the large number of screened individuals in this sampling scheme, the per-individual cost of screening (C_0) and the

	Design		Expected # screened		
P_{obs}	m_1^*	m_2^*	$E[M_0]$	$E[M_1]$	$E[M_2]$
0.05	0	268	84,941	869	268
	600	191	61,716	631	195
	667	0	65,213	667	206
0.10	0	268	84,941	869	268
	550	158	54,365	556	172
	587	0	57,392	587	181
0.20	0	268	84,941	869	268
	450	138	45,573	466	144
	482	0	47,126	482	149

Table 4.1: Designs with current-status augmentation achieving 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$ at a significance level of $\alpha = 0.05$, with $\beta_{12} = \log 2$, $P(0 \to 1) = 0.03$, $P(1 \to 2) = 0.30$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, $P(X = 1) = 0.05$, and $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$ where T_2 is the age of entry into state 2

number of individuals to be screened ($E[M]$) drive the difference in patterns of expected costs seen in Figure 4.9 as compared to Figure 4.6, when inference was based solely on the follow-up sample and no cost was incurred for screening.

Minimum-cost designs, as discussed in the previous section, are better examined in Figure 4.10. When screening is relatively expensive as compared to the cost of follow-up (e.g. top row, with $C_0/C_{11} = 0.20$) designs with $\pi_2 \approx P(Z(A) = 2|Z(0) = 0)/P(Z(A) \in \mathcal{S}|Z(0) = 0)$ are generally most cost effective as they induce smaller screening samples (e.g. smaller $E[M]$). However, as in the previous section we see that as the cost of follow-up in state 2, C_{12}/C_{11}, increases, minimum-cost designs tend to feature smaller π_2 sampling fractions. Also, as the duration of follow-up increases (e.g. as P_{obs} increases), the information that can be gained from prospective follow-up of individuals in state 1 increases and this leads to smaller values π_2 in the corresponding minimum-cost designs.



(a) $P_{obs} = 0.05$, $C_0/C_{11} = 0.20$

(b) $P_{obs} = 0.20$, $C_0/C_{11} = 0.20$

(c) $P_{obs} = 0.05$, $C_0/C_{11} = 0.10$

(d) $P_{obs} = 0.20$, $C_0/C_{11} = 0.10$

Figure 4.9: Expected cost of studies with screening designed with state 2 sampling fraction π_2 and achieving 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$ with $\alpha = 0.05$ and $\beta_{01} = \log 2$, relative to that of the corresponding design with $\pi_2 = 0$, when jointly estimating $\theta_1 = (\alpha_{12}, \beta_{12})$; C_0 is the cost of screening, C_{1k} the cost of one year of follow-up in state k, $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$, $P(0 \to 1) = 0.03$, $P(1 \to 2) = 0.30$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P(X = 1) = 0.05$.

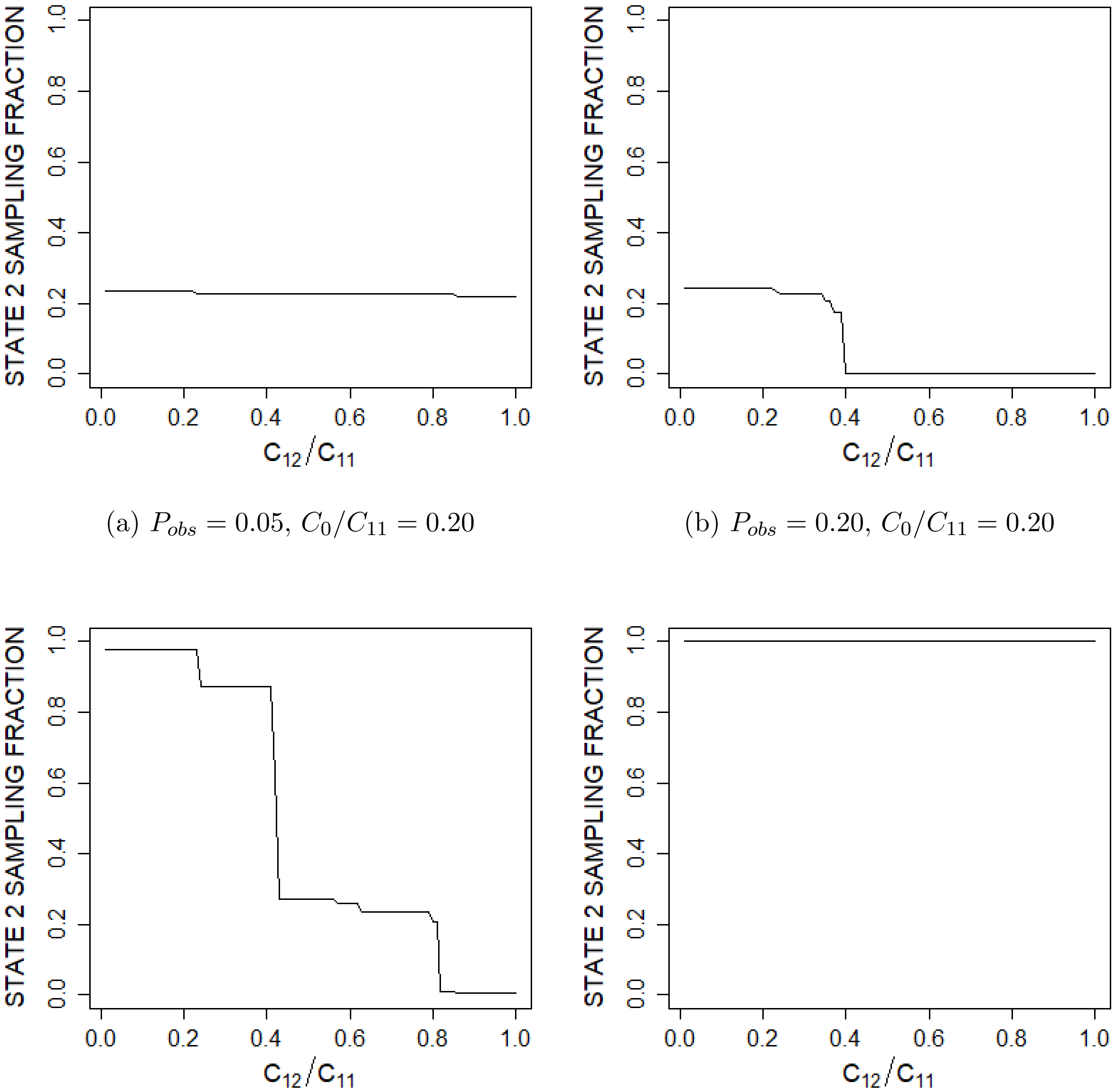
In contrast, when the cost of screening is lower (e.g. $C_0/C_{11} = 0.10$, bottom row of Figure 4.10), a smaller penalty is paid for screening a larger population so minimum-cost designs are achieved with larger π_2 values which induce larger $E[M]$ values.

4.6 Empirical Validation, Estimability, and Auxiliary Data

When fitting a multistate model, care must be taken to ensure the estimability of all parameters. This can become of particular concern when considering piecewise-constant models with several pieces. In practice, if estimability is a problem, either assumptions are made to constrain the parameter space (e.g. by merging states, reducing the number of pieces) or auxiliary data may be used to facilitate estimation.

In the present setting, due to the low rates of psoriasis onset (state 1) and mortality (state 3), especially in the younger age groups, it may be difficult to estimate related parameters for a given dataset. Even when particular interest lies in characterizing progression from psoriasis to psoriatic arthritis, i.e. $\theta_1 = (\alpha_{12}, \beta_{12}$, poor estimability of $\theta_2 = (\alpha_{011}, \alpha_{013}, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ can be problematic. In Table 4.2, we demonstrate the validity of designs based on the Fisher information based on 2,000 simulated datasets with various combinations (m_1^*, m_2^*) chosen to achieve 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq \log 2$ when $\theta_1 = (\alpha_{12}, \beta_{12})$ are estimated jointly and all other parameters in θ_2 are assumed to be known and set to values as specified in Section 4.4.3.

In order to jointly estimate $\theta = (\theta_1, \theta_2)$, it may be helpful to employ auxiliary data. For example, if mortality rates are assumed to be unrelated to the state occupied, γ may be estimated from national population-level death records. To estimate the age-group specific intensities for psoriasis onset, cross-sectional survey data from studies such as that by



(a) $P_{obs} = 0.05$, $C_0/C_{11} = 0.20$

(b) $P_{obs} = 0.20$, $C_0/C_{11} = 0.20$

(c) $P_{obs} = 0.05$, $C_0/C_{11} = 0.10$

(d) $P_{obs} = 0.20$, $C_0/C_{11} = 0.10$

Figure 4.10: Optimal sampling fraction π_2 for minimum-cost designs of studies with screening achieving 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$ with significance level $\alpha = 0.05$; when jointly estimating $\theta_1 = (\alpha_{12}, \beta_{12})$; C_0 is the cost of screening, C_{1k} the cost of one year of follow-up in state k, $P_{obs} = P(T_2 < A + \tau | Z(A) = 1)$, $\beta_{12} = \log 2$, $P(0 \rightarrow 1) = 0.03$, $P(1 \rightarrow 2) = 0.30$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P(X = 1) = 0.05$.

(m_1^*, m_2^*)								
DESIGN	PARAMETER	TRUE	BIAS	ESE	AVSE	ASE	ECP(%)	POW(%)
				Follow-up only				
$(292, 0)$	$\log \lambda_{12}$	-4.818	-0.020	0.202	0.200	0.196	95.0	100
	β_{12}	0.693	0.007	0.259	0.252	0.247	95.3	81.0
$(280, 120)$	$\log \lambda_{12}$	-4.818	-0.015	0.210	0.201	0.198	94.2	100
	β_{12}	0.693	0.005	0.261	0.251	0.247	94.6	80.1
$(265, 260)$	$\log \lambda_{12}$	-4.818	-0.010	0.210	0.203	0.201	95.0	100
	β_{12}	0.693	-0.010	0.258	0.250	0.247	94.8	80.6
				Follow-up + current data augmentation				
$(139, 0)$	$\log \lambda_{12}$	-4.818	-0.020	0.199	0.198	0.194	95.4	100
	β_{12}	0.693	0.013	0.252	0.251	0.247	94.8	81.5
$(0, 80)$	$\log \lambda_{12}$	-4.818	-0.001	0.197	0.195	0.194	95.0	100
	β_{12}	0.693	0.005	0.243	0.248	0.246	95.2	82.8

Table 4.2: Empirical validation (based on $2,000$ simulated datasets) of designs (m_1^*, m_2^*) with 80% power for testing $H_0 : \beta_{12} = 0$ vs $H_A : \beta_{12} \neq 0$ at a significance level of $\alpha = 0.05$ when jointly estimating $\theta_1 = (\alpha_{12}, \beta_{12})$, with $\beta_{12} = \log 2$, $P(0 \to 1) = 0.03$, $P(1 \to 2) = 0.30$, $P(X = 1) = 0.50$, $e^{\alpha_{013}}/e^{\alpha_{011}} = 2$, and $P_{obs} = P(T_2 < A + \tau | Z(A) = 1) = 0.20$, where T_2 is the age of entry into state 2 and intensities for mortality are as in Section 3.2; ESE is the empirical standard error, AVSE is the average standard error, ASE is the asympotic standard error from the Fisher information, ECP(%) is the empirical coverage probability expressed as a percentage with nominal level of 95%, and POW(%) is the empirical power expressed as a percentage.

Gelfand et al. [2005] focusing on the United States and the Multinational Assessment of Psoriasis and Psoriatic Arthritis (MAPP) survey [Lebwohl et al., 2014] involving screening in North America and Europe, could be exploited.

4.7 Discussion

We have developed a framework for the design of prevalent cohort studies with state-dependent recruitment, where the primary interest is to estimate the effect of a covariate on the occurrence of a non-terminal disease progression event, while accounting for mortality. Our designs are based on the asymptotic Fisher information; the resulting variances are shown to agree well with finite-sample variance estimates (see Table 4.2). We compare the value of recruiting individuals from different states for prospective follow-up. When interest lies in estimating a covariate effect on the $j \to k$ transition, recruiting from states j and k is analogous to left and right truncation (with respect to the event of interest), although here mortality is taken into account as well. The impact of augmenting this with current-status information from screening the population on study design is also considered. Throughout, attention is focused on identifying minimum-cost designs achieving given power, reflecting the differing costs of follow-up in different states and recruitment for screening.

In this chapter, we considered the setting where retrospective transition times are available for individuals selected for follow-up, while transitions observed prospectively are subject to administrative censoring. In many settings, while it may be reasonable to assume the above for some events (e.g. death), it is more appropriate to assume interval-censoring for non-terminal events, due to the intermittent nature of assessments. Accommodating loss-to-follow-up is straightfoward if dropout is non-informative, and if a large degree of loss-to-follow-up is anticipated tracing studies may be considered to mitigate the resulting loss of information [Moon et al., 2018]. It is much more challenging to account for dependent missingness and loss-to-follow-up at the design stage, accounting for this at the design stage would require assumptions about the missingness process.

Here we assumed individuals are recruited from a population screen, which induces a distribution for the age at accrual given the state at accrual $A|Z(A)$. If individuals are obtained in a different manner, for example through direct referrals from physicians or from existing registry data, the distribution of A would be different. Further, we assumed that the value of X was available for all screened individuals; this is plausible for demographic variables, but less so for expensive genetic markers. In the latter case, the current-status information would be $\{A, Z(A)\}$ rather than $\{A, Z(A), X\}$. Estimation could still be performed based on an EM algorithm, exploiting the distribution of $X|(A, Z(A))$, and the Fisher information derived for the corresponding observed-data likelihood.

Appendix A: Expectation of screened sample sizes M_j

Consider the 4-state illness-death model from Section 4.5, and suppose $m_0^* = 0$ to reflect the assumption that individuals in the healthy state are not considered for prospective follow-up. Given targets (m_1^*, m_2^*) for recruitment from states 1 and 2, the numbers M_j screened from states $j \in \mathcal{S} = \{0, 1, 2\}$ are random. The marginal expectations of these counts, $E[M_j | \bar{m}^*]$ are required for the Fisher information in (4.21), where $\bar{m}^* = (m_0^*, m_1^*, m_2^*)$ here. Assuming a stationary birth process in the population, $p_j = P(Z(A) = j | Z(A) \in \mathcal{S})$ is the prevalence of j (for $j = 0, 1, 2$) in the alive population. The probability π_h that quota h is satisfied last (for $h = 1, 2$) is

$$\pi_1 = \sum_{\ell=0}^{\infty} P(M_2 \geq m_2^* | M - m_1^* = \ell) P(M - m_1^* = \ell)$$
$$\pi_2 = \sum_{\ell=0}^{\infty} P(M_1 \geq m_1^* | M - m_2^* = \ell) P(M - m_2^* = \ell).$$

where $(M_j | M - m_h^*) \sim BIN(M - m_h^*, p_j/(1 - p_h))$ for $j \neq h \in \{1, 2\}$ and $(M - m_h^*) \sim NBIN(m_h^*, p_h)$ for $h \in \{1, 2\}$. We then consider the marginal distribution of each of the M_j given quota h is satisfied last, for $j \neq h \in \{0, 1, 2\}$,

$$P(M_j = k | \text{quota } h \text{ satisfied last}) = \left[\sum_{\ell=0}^{\infty} P(M_j = k | M - m_h^* = \ell + k) P(M - m_h^* = \ell + k) \right] \Big/ \pi_h$$

for $k \geq m_j^*$, where $m^* = \sum_{h=1}^{2} m_h^*$ and $M = \sum_{j \in \mathcal{S}} M_j$. Finally, we write the desired marginal (unconditional) expectations

$$E[M_1] = \pi_1 m_1^* + \pi_2 E[\, M_1 \mid \text{quota 2 is satisfied last }]$$
$$E[M_2] = \pi_2 m_2^* + \pi_1 E[\, M_2 \mid \text{quota 1 is satisfied last }]$$
$$E[M_0] = \pi_1 E[M_0 | \text{ quota 1 is satisfied last }] + \pi_2 E[M_0 | \text{ quota 2 is satisfied last }]$$

where

$$E[M_j|\text{ quota } h \text{ is satisfied last}] = \sum_{k=0}^{\infty} kP(M_j = k|\text{ quota } h \text{ is satisfied last })$$

Chapter 5

Conclusion and Future Work

In this thesis, several of aspects of study designs were considered in the context of longitudinal studies to model the progression of chronic diseases via multistate processes. A summary of key contributions is presented in Section 5.1 and an outline of future research directions is given in Section 5.2.

5.1 Key Contributions

In Chapter 2, we look at impact of loss-to-follow-up in prospective cohort studies with intermittent assessments on efficiency. We establish a framework for cost-effective selection of a subset of lost individuals for tracing to recover information, where this selection is informed by the data collected prior to dropout and considering the expected information to be gained by tracing. We then demonstrate that meaningful gains in efficiency (in the order of up to 60% in the scenarios considered) can be achieved relative to simple random selection for tracing, at no additional cost; the implications of gains of this magnitude on the conduct of potential tracing studies in large cohort studies such as the CLSA have the potential to be quite large.

In Chapter 3, we take a rigorous approach to investigate the relationship between the frequency of assessments and sample size in design of prospective longitudinal studies with intermittent assessments. We derive the asymptotic variance in the presence of misclassification, and this serves as basis for accounting for misclassification of disease status at the design stage. Our theoretical results are supported with several examples to illustrate the interplay between design factors (e.g. frequency of intermittent assessments and sample size), the disease process, and the misclassification process and identify minimum-cost designs to achieve a given level of power.

In Chapter 4, we consider the impact of prevalent cohort sampling in the context of multistate processes. We evaluate the relative value of state-dependent sampling in the failure-time setting, incorporating a measure of cost to account for the differential ease of obtaining prospective and retrospective information in practice. This idea is extended to the multistate framework by considering state-dependent sampling schemes where the number of individuals to be recruited from each state may be pre-specified, and current status information from screening the population to obtain these samples may be available as well. As in the previous chapter, we identify minimum-cost designs under each of these schemes which achieve a desired level of power.

5.2 Future Work

5.2.1 Misclassification in Assessment of State Occupancy

While the implications of misclassified disease assessments were discussed in chapter 3 in the context of inception cohort studies with intermittent observation of disease status, misclassification arises in other settings as well. For example, researchers at the Toronto Centre for Prognosis Studies in Rheumatic Disease have developed a telephone screening tool, the Toronto Psoriatic Arthritis Screen (ToPAS [Gladman et al., 2009] and ToPAS-II [Tom et al., 2015]) questionnaires, to identify potential cases of psoriatic arthritis among individuals with psoriasis, as an alternative to the gold-standard Classification Criteria for Psoriatic Arthritis (CASPAR) which requires evaluation by a rheumatologist [Taylor et al., 2006]. Similar diagnostic tools have been developed by other groups: the Psoriatic Arthritis Screening and Evaluation (PASE) [Husni et al., 2007], the Psoriasis Epidemiology Screening Tool (PEST) [Ibrahim et al., 2009], and the Early Arthritis for Psoriatic Patients (EARP) [Tinazzi et al., 2012]. When these four tools were compared in an independent cross-sectional sample of 302 individuals with psoriasis, the sensitivities ranged from 44-91% and specificities from 88-97% [Mishra et al., 2017], relative to CASPAR. The key value of these diagnostic tools is their ease of use, as they can be administered easily and quickly (e.g. via telephone). Such tools are well suited for use in a tracing study as the barriers to obtaining a response are far lower [Tom et al., 2015]. In the present work it was assumed that responses obtained through tracing were not subject to misclassification, but allowing for misclassification at this stage would capture a more realistic perspective on tracing in practice.

Accounting for misclassification can be done in a number of ways, for example (i) using published values of the misclassification rates and treating them as fixed, (ii) using a two-stage estimation procedure where estimates of misclassification rates are obtained in stage one (using validation data) and large sample results enable the derivation of the limiting variance of estimators of the disease process parameters in stage two, or (iii) jointly modelling the disease process and misclassification process using validation data. In Chapter 3, approach (i) was adopted, although in practice ignoring the variability in misclassification estimates leads to underestimating the variance of the disease process parameters. Incorporating approaches (ii) or (iii) to account for misclassification at the design stage is an area warranting future development [Yi, 2016].

5.2.2 Respondent-driven Assessment Times

Throughout this thesis, we considered designing studies where the timing of planned assessments was known in advance and we assumed that participants adhered to this schedule. While this assumption may be reasonable in highly controlled settings such as clinical trials, it is often untenable in prospective longitudinal studies carried out over an extended period of time. A number of challenges arise when the timing of assessments is driven by participants. Some work has been done to analyze longitudinal data with irregular assessment times, Pullenayegum and Lim [2016] present a thorough review of a broad range of existing methods and proposes some guidance for study design. In short, the authors suggest attempting to minimize deviations from the protocol in assessment times, and barring this, suggest carefully recording the reasons for these deviations when possible. Given this information, it may be possible to account for potential irregularities in the visit process at the design stage, by positing models for some degree of deviations from the scheduled visit protocol; this remains an area for future work. Alternatively, tracing studies also offer an avenue for recovering information when visit times deviate significantly from the initial protocol, and may aid in understanding of mechanisms of dropout [Farewell et al., 2003].

5.2.3 Incorporating Auxiliary Data

When fitting multistate models, particularly if adopting piecewise constant intensity models with several pieces, the number of parameters to estimate can be large. In some situations, it may be necessary to augment data from a cohort study with additional sources of auxiliary data to facilitate estimation. For example, external population-level mortality data may be necessary to fit illness-death models when the duration of follow-up does not admit observation of enough deaths. As another example, the illustrations in Sections 4.4.3 and 4.5.3 focused on a four-state illness-death model with prospective follow-up of individuals recruited from the diseased states $\{1, 2\}$ only, not from the healthy state 0. In this setting, estimation of the transition intensity into state 1 may be problematic, as all instances of entry into state 1 in the sample are right-truncated; auxiliary data can serve as a useful tool to augment the sample data and facilitate estimation of such transition intensities. Augmenting the standard likelihoods as described in Chapters 2, 3, and 4 with auxiliary data warrants future development.

References

P. S. Albert and C. Hendricks Brown. The design of a panel study under an alternating Poisson process assumption. *Biometrics*, 47(3):921–932, 1991.

P. Armitage and T. Colton. *Encyclopedia of Biostatistics*. Wiley, 2005.

M. Asgharian, C. M'Lan, and D. Wolfson. Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association*, 97(457): 201–209, 2002.

S. Blumenthal. Proportional sampling in life length studies. *Technometrics*, 9(2):205–218, 1967.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press, 2006.

CLSA. Canadian Longitudinal Study on Aging (CLSA). https://www.clsa-elcv.ca. Accessed: 2018-06-19.

L. M. Collins and J. W. Graham. The effect of the timing and spacing of observations in longitudinal studies of tobacco and other drug use: Temporal design considerations. *Drug and Alcohol Dependence*, 68:85–96, 2002.

R. J. Cook. Information and efficiency consideration in planning studies based on two-state markov processes. *Journal of Statistical Research*, 34:161–178, 2000.

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

D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. Wiley, New York, first edition, 1965.

E. Demidenko. Sample size determination for logistic regression revisited. *Statistics in Medicine*, 26(18):3385–3397, 2007.

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

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. *New England Journal of Medicine*, 1993(329):977–986, 1993.

S. W. Duffy, I. D. Nagtegaal, M. Wallis, F. H. Cafferty, N. Houssami, J. Warwick, P. C. Allgood, O. Kearins, N. Tappenden, E. O'Sullivan, et al. Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. *American Journal of Epidemiology*, 168(1):98–104, 2008.

Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy: ETDRS report number 12. *Ophthalmology*, 98(5):823–833, 1991.

L. Eder, V. Chandran, H. Shen, R. J. Cook, S. Shanmugarajah, C. F. Rosen, and D. D. Gladman. Incidence of arthritis in a prospective cohort of psoriasis patients. *Arthritis Care & Research*, 63(4):619–622, 2011a.

L. Eder, T. Law, V. Chandran, S. Shanmugarajah, H. Shen, C. F. Rosen, R. J. Cook, and D. D. Gladman. Association between environmental factors and onset of psoriatic arthritis in patients with psoriasis. *Arthritis Care & Research*, 63(8):1091–1097, 2011b.

A. F. Emery and A. V. Nenarokomov. Optimal experiment design. *Measurement Science and Technology*, 9(6):864–876, 1998.

V. T. Farewell, J. F. Lawless, D. D. Gladman, and M. B. Urowitz. Tracing studies and analysis of the effect of loss to follow-up on mortality estimation from patient registry data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):445–456, 2003.

D. Farzanfar, A. Abumuamar, J. Kim, E. Sirotich, Y. Wang, and E. Pullenayegum. Longitudinal studies that use data collected as part of usual care risk reporting biased results: a systematic review. *BMC Medical Research Methodology*, 17(1):133, 2017.

W. A. Fuller. *Measurement Error Models.* John Wiley & Sons, 1987.

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

J. M. Gelfand, A. B. Troxel, J. D. Lewis, S. K. Kurd, D. B. Shin, X. Wang, D. J. Margolis, and B. L. Strom. The risk of mortality in patients with psoriasis: results from a population-based study. *Archives of Dermatology*, 143(12):1493–1499, 2007.

D. Gladman. Mortality in psoriatic arthritis. *Clinical & Experimental Rheumatology*, 26 (5):S62, 2008.

D. Gladman, R. Shuckett, M. Russell, J. Thorne, and R. Schachter. Psoriatic arthritis (psa)-an analysis of 220 patients. *QJM: An International Journal of Medicine*, 62(2): 127–141, 1987.

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

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

D. D. Gladman, C. T. Schentag, B. D. Tom, V. Chandran, J. Brockbank, C. Rosen, and V. T. Farewell. Development and initial validation of a screening questionnaire for psoriatic arthritis: the Toronto psoriatic arthritis screen (ToPAS). *Annals of the Rheumatic Diseases*, 68(4):497–501, 2009.

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

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

M. E. Husni, K. H. Meyer, D. S. Cohen, E. Mody, and A. A. Qureshi. The pase questionnaire: pilot-testing a psoriatic arthritis screening and evaluation tool. *Journal of the American Academy of Dermatology*, 57(4):581–587, 2007.

W.-T. Hwang and R. Brookmeyer. Design of panel studies for disease progression with multiple stages. *Lifetime Data Analysis*, 9(3):261–274, 2003.

G. Ibrahim, M. Buch, C. Lawson, R. Waxman, and P. Helliwell. Evaluation of an existing screening tool for psoriatic arthritis in people with psoriasis and the development of a new instrument: the psoriasis epidemiology screening tool (pest) questionnaire. *Clinical & Experimental Rheumatology*, 27(3):469, 2009.

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

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

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, 2011.

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

N. Keiding, C. Holst, and A. Green. Retrospective estimation of diabetes incjdence from information in a prevalent population and historical mortality. *American Journal of Epidemiology*, 130(3):588–600, 1989.

S. Kobayashi, F. Sata, S. Sasaki, T. S. Braimoh, A. Araki, C. Miyashita, H. Goudarzi, S. Kobayashi, and R. Kishi. Combined effects of ahr, cyp1a1, and xrcc1 genotypes and prenatal maternal smoking on infant birth size: Biomarker assessment in the Hokkaido study. *Reproductive Toxicology*, 65:295–306, 2016.

M. Kosorok and W. Chao. Further details on the analysis of longitudinal ordinal response data in continuous time. *Techincal Report*, 92, 1995.

M. R. Kosorok and W.-H. Chao. The analysis of longitudinal ordinal response data in continuous time. *Journal of the American Statistical Association*, 91(434):807–817, 1996.

N. Kreiger, A. Tenenhouse, L. Joseph, T. Mackenzie, S. Poliquin, J. P. Brown, J. C. Prior, and R. S. Rittmaster. Research notes: The Canadian multicentre osteoporosis study (CaMos): Background, rationale, methods. *Canadian Journal on Aging/La Revue canadienne du vieillissement*, 18(03):376–387, 1999.

R. Lakshman, F. Whittle, W. Hardeman, M. Suhrcke, E. Wilson, S. Griffin, and K. K. Ong. Effectiveness of a behavioural intervention to prevent excessive weight gain during infancy (the baby milk trial): study protocol for a randomised controlled trial. *Trials*, 16(1):1, 2015.

T. Lancaster. *The Econometric Analysis of Transition Data*. Cambridge University Press, 1992.

J. M. Lange and V. N. Minin. Fitting and interpreting continuous-time latent Markov models for panel data. *Statistics in Medicine*, 32(26):4581–4595, 2013.

J. F. Lawless and N. N. Rad. Estimation and assessment of Markov multistate models with intermittent observations on individuals. *Lifetime Data Analysis*, 21(2):160–179, 2015.

M. G. Lebwohl, H. Bachelez, J. Barker, G. Girolomoni, A. Kavanaugh, R. G. Langley, C. F. Paul, L. Puig, K. Reich, and P. C. van de Kerkhof. Patient perspectives in the management of psoriasis: results from the population-based multinational assessment of psoriasis and psoriatic arthritis survey. *Journal of the American Academy of Dermatology*, 70(5):871–881, 2014.

R. J. Little and D. Rubin. Analysis with missing data, 1987.

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

J. Ma, L. Thabane, J. Beyene, and P. Raina. Power analysis for population-based longitudinal studies investigating gene-environment interactions in chronic diseases: A simulation study. *PLoS ONE*, 11(2):e0149940, 2016.

J. Mehtälä, K. Auranen, and S. Kulathinal. Optimal designs for epidemiologic longitudinal studies with binary outcomes. *Statistical Methods in Medical Research*, 24(6):803–818, 2015.

I. Michalek, B. Loring, and S. John. A systematic review of worldwide epidemiology of psoriasis. *Journal of the European Academy of Dermatology and Venereology*, 31(2): 205–212, 2017.

S. Mishra, H. Kancharla, S. Dogra, and A. Sharma. Comparison of four validated psoriatic arthritis screening tools in diagnosing psoriatic arthritis in patients with psoriasis (compaq study). *British Journal of Dermatology*, 176(3):765–770, 2017.

N. C. Moon, L. Zeng, and R. J. Cook. Tracing studies in cohorts with attrition: Selection models for efficient sampling. *Statistics in Medicine*, 2018.

A. L. Moskowitz, J. L. Krull, K. A. Trickey, and B. F. Chorpita. Quality vs. quantity: Assessing behavior change over time. *Journal of Psychopathology and Behavioral Assessment*, pages 1–20, 2017.

J. R. Nesselroade. Interindividual differences in intraindividual change. In L. M. Collins and J. L. Horn, editors, *Best methods for the analysis of change: Recent advances, unanswered questions, future directions*, chapter 6. 1991.

A. Ogdie and P. Weiss. The epidemiology of psoriatic arthritis. *Rheumatic Disease Clinics*, 41(4):545–568, 2015.

A. Ogdie, K. Haynes, A. B. Troxel, T. J. Love, S. Hennessy, H. Choi, and J. M. Gelfand. Risk of mortality in patients with psoriatic arthritis, rheumatoid arthritis and psoriasis: a longitudinal cohort study. *Annals of the rheumatic diseases*, 73(1):149–153, 2014.

E. M. Pullenayegum and L. S. Lim. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research*, 25(6):2992–3014, 2016.

N. N. Rad. *Multistate Models for Biomarker Processes*. PhD thesis, University of Waterloo, 2014.

P. Raina, C. Wolfson, S. Kirkland, and L. Griffith. The Canadian Longitudinal Study on Aging (CLSA) report on health and aging in Canada. 2018.

P. S. Raina, C. Wolfson, S. A. Kirkland, L. E. Griffith, M. Oremus, C. Patterson, H. Tuokko, M. Penning, C. M. Balion, D. Hogan, et al. The Canadian longitudinal study on aging (CLSA). *Canadian Journal on Aging/La Revue canadienne du vieillissement*, 28(03):221–229, 2009.

E. Riboli. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Annals of Oncology*, 3(10):783–791, 1992.

S. M. Ross. *Introduction to Probability Models*. Academic press, 2014.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

H. Shen and R. J. Cook. Regression with incomplete covariates and left-truncated time-to-event data. *Statistics in Medicine*, 32(6):1004–1015, 2013.

Y. Shen, J. Ning, and J. Qin. Nonparametric and semiparametric regression estimation for length-biased survival data. *Lifetime Data Analysis*, 23(1):3–24, 2017.

R. Simon. Length biased sampling in etiologic studies. *American Journal of Epidemiology*, 111(4):444–452, 1980.

J. D. Singer and J. B. Willett. Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin*, 110(2):268, 1991.

Statistics Canada. Deaths and mortality rates, by age group and sex: Canada, provinces and territories, annual. *CANSIM (database)*, 2018.

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

W. Taylor, D. Gladman, P. Helliwell, A. Marchesoni, P. Mease, and H. Mielants. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. *Arthritis & Rheumatology*, 54(8):2665–2673, 2006.

A. C. Timmons and K. J. Preacher. The importance of temporal design: How do measurement intervals affect the accuracy and efficiency of parameter estimates in longitudinal research? *Multivariate Behavioral Research*, 50(1):41–55, 2015.

I. Tinazzi, S. Adami, E. M. Zanolin, C. Caimmi, S. Confente, G. Girolomoni, P. Gisondi, D. Biasi, and D. McGonagle. The early psoriatic arthritis screening questionnaire: a simple and fast method for the identification of arthritis in patients with psoriasis. *Rheumatology*, 51(11):2058–2063, 2012.

A. C. Titman, G. A. Lancaster, K. Carmichael, and D. Scutt. Accounting for bias due to a non-ignorable tracing mechanism in a retrospective breast cancer cohort study. *Statistics in Medicine*, 30(4):324–334, 2011.

B. D. Tom, V. Chandran, V. T. Farewell, C. F. Rosen, and D. D. Gladman. Validation of the Toronto psoriatic arthritis screen version 2 (ToPAS 2). *The Journal of Rheumatology*, 42(5):841–846, 2015.

H. Tuokko, L. E. Griffith, M. Simard, and V. Taler. Cognitive measures in the Canadian Longitudinal Study on Aging. *The Clinical Neuropsychologist*, 31(1):233–250, 2017.

B. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

M.-C. Wang. Gap time bias in incident and prevalent cohorts. *Statistica Sinica*, 9(4): 999–1010, 1999.

C. Wolfson, D. B. Wolfson, M. Asgharian, C. E. M'Lan, T. Østbye, K. Rockwood, and D. Hogan. A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344(15):1111–1116, 2001.

G. Y. Yi. *Statistical Analysis with Measurement Error or Misclassification*. Springer, 2016.

L. Zeng and R. Cook. Cross-sectional analysis of multistate disease processes from population screening. *In Preparation*, 2018.

L. Zeng, R. J. Cook, and K.-A. Lee. Design of cancer trials based on progression-free survival with intermittent assessment. *Statistics in Medicine*, 37(12):1947–1959, 2018.