Design and Analysis of Life History Studies Involving Incomplete Data

by

Fangya Mao

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, 2022

© Fangya Mao 2022

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:	Farouk Nathoo Professor, Canada Research Chair in Biostatistics, Department of Mathematics and Statistics, University of Victoria
Supervisor(s):	Richard J. Cook University Professor, Math Faculty Research Chair, Department of Statistics and Actuarial Science, University of Waterloo
Internal Member:	Jerry Lawless Distinguished Professor Emeritus, Department of Statistics and Actuarial Science, University of Waterloo
Internal Member:	Leilei Zeng Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Ashok Chaurasia Assistant Professor, School of Public Health and Health Sciences, University of Waterloo

Author's Declaration

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

Incomplete life history data can arise in study designs, coarsened observations, missing covariates, and unobserved latent processes. This thesis consists of three different projects developing statistical models and methods to address problems involving such features.

Statistical models which facilitate the exploration of spatial dependence can advance scientific understanding of chronic diseases processes affecting several organ systems or body sites. Motivated by the need to investigate the spatial nature of joint damage in patients with psoriatic arthritis, we develop a multivariate mixture model to characterize latent susceptibility and the progression of joint damage in different locations in Chapter 2. In addition to a large number of joints under consideration and the heterogeneity in risk, the times to joint damage are subject to interval censoring as damage status is only observed at intermittent radiological examination times. We address computational and inferential challenge through use of composite likelihood and two-stage estimation procedures. The key contribution of this chapter is the development of a convenient and general framework for regression modeling to study risk factors for susceptibility to joint damage and the time to damage, as well as spatial dependence of these features.

The design and analysis of two-phase studies have been investigated for biomarker studies involving lifetime data. Two-phase designs aim to guide the efficient selection of a sub-sample of individuals from a phase I cohort to measure some "expensive" markers under budgetary constraints. In a phase I sample information on the response and inexpensive covariates is available for a large cohort, and in phase II, a subsample is selected in which to assay the marker of interest through examination of a biospecimen. The design efficiency is measured in terms of the precision in estimating the effect of the biomarker on some event process (e.g. disease progression) of interest. Chapter 3 considers two-phase designs involving current status observation of the failure process; here individuals are monitored at a single assessment time to determine whether or not they have experienced a failure event of interest. This kind of observation scheme is sometimes desirable in practice as it is more efficient and cost-effective then carrying out multiple assessments. We examine efficient two-phase designs under two analysis methods, namely maximum likelihood and inverse probability weighting. The former tends to be more efficient but requires additional model assumptions involving the nuisance covariate model, while the latter is more robust but yields less efficient estimators since it only analyses data from the phase II subsample. The optimal designs are derived by minimizing the asymptotic variance of the coefficient estimators for the expensive marker. To circumvent the computational challenge in evaluating asymptotic variances at the design stage, we consider

designs involving sub-sampling based on extreme score statistics, extreme observations, or via stratified sub-sampling schemes. The role of the assessment time is highlighted.

Research involving progressive chronic disease processes can be conducted by synthesizing data from different disease registries using different enrolment conditions. In inception cohorts, for example, individuals may be required to not have entered an advanced stage of the disease, while disease registries may focus on individuals who have progressed to a more advanced stage. The former yields left-truncated progression times while the latter yields right-truncated progression times. Chapter 4 considers the development of two-phase designs when the phase I sample contains data pooled from different registries launched to recruit individuals from a common population with different disease-dependent selection criteria. We frame the complex data structure by multistate models and carefully outline model assumptions such as the independence of disease progression and time to death. General likelihood constructions are presented using intensity-based models and we derive a partial likelihood restricted to the failure time of interest under special model assumptions. Both recruitment (phase I) and sub-selection (phase II) biases are accounted for to ensure valid inference. An inverse probability weighting method is also developed to relax or weaken assumptions needed for the likelihood approach. We investigate and compare the performance of various two-phase sampling schemes under each analysis method and provide practical guidance for phase II selection given budgetary constraints.

The contributions of this thesis are reviewed in Chapter 5 where we also mention topics of future research.

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to my supervisor, Dr. Richard J. Cook, one of the wisest people I have had the privilege to work with. I am extremely fortunate and honored to share his profound scientific knowledge and invaluable life wisdom. His exceptional guidance, continuous support, and encouragement have made my Ph.D research much easier and more enjoyable. This thesis would not have been possible without his care and support.

I wish to thank my thesis committee, Drs. Jerry Lawless, Leilei Zeng, Ashok Chaurasia, and Farouk Nathoo for taking their time reviewing this thesis and providing insightful comments and questions. My special thanks go to Drs. Jerry Lawless and Leilei Zeng for sharing inspiring questions, helpful suggestions, and kind support during my Ph.D study.

I am really grateful to the faculty and staff in the Department of Statistics and Actuarial Science at the University of Waterloo for creating a positive and friendly environment for my development as an international graduate student. Special thanks go to Ms. Mary Lou Dufton, Mr. Carlos Mendes, and Mr. Greg Preston for their continuous help in administrative and technical issues. I wish to express my great appreciation to Ms. Ker-Ai Lee for always sharing her valuable computing experiences and helpful suggestions in data analysis with me. I also want to thank Dr. Joel Dubin for supervising my master's essay, encouraging me to further pursue a Ph.D degree in Biostatistics, and for his warm support throughout my time in this department.

To my sincere friends and peers I have met during this journey: Zhaohan Sun, Xinyi Ge, Junhan Fang, Minghui Gao, Marzieh Mussavi Rizi, Wenlin Zhang, Trang Bui, Chi-Kuang Yeh, Yechao Meng, Takaaki Koike, Ce Yang, Bingfeng Xie, Faith Lee, Dongyang Yang, and Jenny Li. I also really appreciate my lovely friends in China: the 504 group (Xingchen Liu, Shujun Guo, Yi Xiong), Yafang He, Siqi Zhang, and my sixteen-years friend Yuhan Wang. Many thanks to them and all friends not listed here for their friendship and joyful accompany.

Last but not the least, I would like to express my heartfelt gratitude to my family, for providing a great source of support and endless love for me. Special thanks go to my parents, Zhiyue Mao and Wanhong Fang, who always believe in me and being there when needed. I am the luckiest to be their daughter.

Dedication

To my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview and introduction of statistical methods

1.1.1 Overview

This thesis is concerned with the development of statistical methods for the design and analysis of life history studies involving incomplete data. Chronic diseases are considered which can be characterized by failure time and more general multistate processes (Klein et al., 2014; Cook and Lawless, 2018). In registry studies involving failure time processes, data are often incomplete because of censoring and truncation (Klein and Moeschberger, 2003). The former may correspond to, for example, right-censored data due to loss to follow-up, or interval-censored data if failure status is only observable at periodic assessment times. Truncation differs from censoring as it relates to inclusion criteria. For example, left truncation arises in settings where individuals are only included if they have not yet developed the event of interest at a recruitment time (Turnbull, 1976). Problems considered include the development of latent variable models to study spatial dependence of susceptibility to disease progression at several body sites. Another theme of this research is the use of two-phase designs in setting involving heavily censored or truncated data.

The first research project is devoted to the analysis of dependent failure time processes under intermittent observation. The model is generalized to accommodate the fact that some processes may not be nonsusceptible to failure event of interest. It is often unknown whether processes not observed to fail are susceptible or not and mixture models can be useful in such settings (Farewell, 1982). In Chapter 2 we develop spatial dependence models for susceptibility and failure times among jointly susceptible processes. The likelihood

is evaluated and composite likelihood methods are proposed to simplify the necessary computing.

Two-phase design problems in life history analysis often arise in biomarker studies where serum samples are collected and stored at study entry for future use. It is usually expensive and inefficient to assay all stored biosamples for one biomarker study and in such cases, two-phase designs offer a cost-effective solution. In phase I inexpensive information on responses and some auxiliary covariates are collected; in phase II a subsample of individuals is selected to measure marker values in biospecimens through an expensive ascertainment process. It is conventional to partition the phase I sample into several exclusive and nonempty strata and to conduct phase II stratified subsampling (Prentice, 1986; Borgan et al., 2000; McIsaac, 2012; Espin-Garcia et al., 2017). More recently, there has been increasing interest in developing asymptotic or approximate optimal designs that aim to achieve the highest precision in estimating the marker effect on some event process of interest (Reilly, 1996; McIsaac and Cook, 2014, 2015; Tao et al., 2020; Chen and Lumley, 2020; Yang et al., 2021). The last two projects of this thesis focus on the cost-effective design of studies aiming to estimate the association between an expensive biomarker and a disease process of interest. Much work on two-phase studies has been carried out for continuous and binary responses, as well as failure processes under right-censoring (e.g. Prentice, 1986; Chen and Lo, 1999; Breslow and Wellner, 2007; Borgan and Samuelsen, 2014; Lawless, 2018; Tao et al., 2020). The current work on two-phase study design considers i) failure time processes under current status observation schemes (Chapter 3), and ii) the analysis of two-stage disease processes when interest lies in estimating the effects of biomarkers on the risk of progression from one disease stage to the next (Chapter 4). In the second setting the use of data from different disease registries creates a phase I sample with sub-samples acquired according to different state-dependent selection processes.

This thesis is written with relatively self-contained chapters with the first section of the chapters devoted to a review of the relevant background literature, followed by an introduction to the notation and methods, reports on the results of simulation studies, and applications to the motivating studies. The last chapter reviews the contributions of this thesis and outlines several topics for future research. In the next section, the relevant statistical models and methods employed for this thesis are briefly reviewed, following which three motivating studies are discussed. The three problems addressed are then briefly described.

1.1.2 Modeling multivariate failure times via Gaussian copulas

With K types of events, we write the failure time vector as $\mathbf{T} = (T_1, T_2, \ldots, T_K)'$, with T_k being a non-negative variable of the time to event k. Given the marginal survival function $\mathcal{F}_k(t; \boldsymbol{\theta}_k) = P(T_k \geqslant t)$, $k = 1, \ldots, K$, the joint survival function $P(\mathbf{T} > \mathbf{t})$ can be specified through a K-dimensional copula function (Sklar, 1959; Joe, 1997) indexed by a vector of dependence parameters $\boldsymbol{\rho}$, such that

$$\mathcal{C}(U_1 \geqslant u_1, \ldots, U_K \geqslant u_K; \boldsymbol{\rho}) = P(U_1 \geqslant u_1, \ldots, U_K \geqslant u_K; \boldsymbol{\rho})$$

where U_k are uniform $[0, 1]$ random variables, $k = 1, \ldots, K$. If we set $U_k = \mathcal{F}_k(T_k; \boldsymbol{\theta}_k)$, then the multivariate survival function for T is

$$P(T_1 > t_1, \ldots, T_K > t_K) = \mathcal{C}(\mathcal{F}_1(t_1; \boldsymbol{\theta}_1), \ldots, \mathcal{F}_K(t_K; \boldsymbol{\theta}_K); \boldsymbol{\rho}). \tag{1.1.1}$$

There are a variety of copula functions available to model the dependence structure among the failure times. For example, with $K = 2$, Chatterjee and Shih (2001) studied the Clayton model (Clayton, 1978), the Frank model (Frank, 1979) and the positive stable model (Hougaard, 1986) in the context of familial disease; Jiang and Cook (2020) utilized the class of Archimedean copulas (Nelsen, 2006) for modeling bivariate interval-censored failure time data with dependent susceptibility. Zhong and Cook (2016) used a Gaussian copula to model multivariate failure times in family studies.

Gaussian copula models accommodate a flexible pairwise association specification through a correlation matrix. It also enjoys many properties similar to the multivariate normal distribution; a most appealing one is the reproducible property in the sense that any subvector retains the same form of distribution as the full vector (Song, 2000). The Gaussian copula function for the multivariate failure time distribution corresponding to (2.2.5) can be expressed as

$$\mathcal{C}(u_1, \ldots, u_K; \boldsymbol{\rho}) = \Phi_K(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_K); \boldsymbol{\rho})$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable and $\Phi_K(\cdot; \boldsymbol{\rho})$ is the cumulative distribution function of a $K \times 1$ multivariate normal random variable with mean zero and $K \times K$ covariance matrix $\Sigma(\boldsymbol{\rho})$ with off-diagonal entries $\rho_{kk'}$. The association between T_k and $T_{k'}$ can be measured by Kendall's tau, given by $\tau_{kk'} = 2 \arcsin(\rho_{kk'})/\pi$, $k, k' \in \{1, \ldots, K\}$. The joint distribution for T is indexed in general by $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\rho}')$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')'$.

1.1.3 Analysis of incomplete covariates in two-phase designs

The idea of two-phase sampling was first introduced by Neyman (1938) as a cost-effective solution to a field survey problem. In a classic two-phase design, one collects relatively cheap information for a cohort of individuals and then randomly draws subsamples from pre-specified strata to ascertain further costly information. This framework is widely used in epidemiological studies when one or more of the risk factors of primary interest are expensive or infeasible to measure on the full cohort due to limited resources (Prentice, 1986; Borgan et al., 2000; Lawless et al., 1999).

We let $\mathbf{X}$ be a $p \times 1$ vector of covariates collected at $t = 0$ and suppose the goal is to evaluate the association between $\mathbf{X}$ and a scalar event time of interest T. We let Y denote the observed version of T. For illustration we take an example of current status data where $Y = (I(T \leqslant A), A)'$ with A denoting the (single) inspection time. We consider the case in which a scalar covariate X_1 in $\mathbf{X}$ is an expensive biomarker and it is cost-prohibitive to measure it for all individuals. Scientific interest often lies in examining covariate effects on the hazard for the event of interest. We partition $\mathbf{X}$ as $(X_1, \mathbf{X}_2')'$ where $\mathbf{X}_2$ represents a vector of discrete inexpensive auxiliary covariates. We write the conditional hazard as

$$h(t|\mathbf{X}) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T \leqslant t + \Delta t | t \leqslant T, \mathbf{X})}{\Delta t}.$$

Let $\boldsymbol{\theta}$ index the distribution function for $T|\mathbf{X}$ and we write $\mathcal{F}(t|\mathbf{X}; \boldsymbol{\theta}) = P(T > t|\mathbf{X})$ as the conditional survival function.

We let $R = I(X_1 \text{ is observed})$ denote the variable indicating that the individual i is selected into the phase II sub-cohort (or sub-sample). With $\boldsymbol{\rho}$ indexing the selection model we write it as

$$P(R = 1|\mathbf{Z}; \boldsymbol{\rho}) = \pi(\mathbf{Z}; \boldsymbol{\rho}),$$

where $\mathbf{Z} = (Y, A, \mathbf{X}_2')'$. For a sample of N observations from independent processes we introduce the subscript i, $i = 1, \ldots, N$. The phase I sample is $\{Y_i, \mathbf{X}_{i2}, i = 1, \ldots, N\}$. If individual i is chosen to measure $\mathbf{X}_i$, we let $R_i = 1$ and let $R_i = 0$ otherwise. Following phase II subsampling we write the available data as $\{Y_i, \mathbf{X}_i^\circ, R_i, i = 1, \ldots, N\}$, where $\mathbf{X}_i^\circ = \mathbf{X}_i$ if $R_i = 1$ and $\mathbf{X}_{i2}$ otherwise. A observed data likelihood (Lawless et al., 1999) is given by

$$\prod_{i=1}^{N} \left[\pi(\mathbf{Z}_i; \boldsymbol{\rho}) P(Y_i|A_i, \mathbf{X}_i; \boldsymbol{\theta}) P(X_{i1}|A_i, \mathbf{X}_{i2})\right]^{R_i} \left[(1 - \pi(\mathbf{Z}_i; \boldsymbol{\rho})) E_{X_1|A,\mathbf{X}_2}[P(Y_i|A_i, X_1, \mathbf{X}_{i2}; \boldsymbol{\theta})]\right]^{1-R_i},$$

where $P(Y|A, \mathbf{X}; \boldsymbol{\theta}) = \mathcal{F}(A|\mathbf{X}; \boldsymbol{\theta})^{I(T>A)}(1 - \mathcal{F}(A|\mathbf{X}; \boldsymbol{\theta}))^{I(T \leqslant A)}$.

A simple alternative to maximum likelihood for dealing with the incomplete covariate data is to restrict the analysis to the phase II subsample in which case inverse probability weighting (Robins et al., 1994) is required. A consistent estimator of $\boldsymbol{\theta}$ is obtainable by using an inverse probability weighted (IPW) estimating function restricted to the phase II data given by

$$\sum_{i=1}^{N} \frac{R_i}{\pi(\mathbf{Z}_i;\boldsymbol{\rho})} \frac{\partial}{\partial \boldsymbol{\theta}} \log P(Y_i|A_i,\mathbf{X}_i;\boldsymbol{\theta}),$$

where the phase II selection probability $\pi(\mathbf{Z};\boldsymbol{\rho})$ is bounded away from zero. Use of IPW estimating function is more robust since it does not require specifications of this nuisance model, but it yields less efficient estimators; it does require specification of the selection model but in the context of two-phase designs this is set by the investigator and so is known.

A variety of other approaches have been developed for dealing with incomplete covariate data which feature different degrees of robustness and efficiency (Robins et al., 1994; Lawless et al., 1999; Chatterjee et al., 2003; Lumley et al., 2011). In Chapters 3 and 4, we focus on inference based on (partial) conditional likelihood and inverse probability weighted (IPW) estimating equations.

1.1.4 Multistate processes

Multistate processes offer an intuitive and appealing framework to model the lifetime dynamics concerning a disease process of interest (Cook and Lawless, 2014, 2018). A multistate model involving $2J$ states can be adopted for progressive disease processes, as show in Figure 1.1, with finite state-space $\mathcal{S} = \{0, 1, \ldots, J-1, D_0, D_1, \ldots, D_{J-1}\}$, where 0 represents the disease-free state and states $1, 2, \ldots, J-1$ represent the worsening disease stages; the set $\mathcal{D} = \{D_0, \ldots, D_{J-1}\}$ contains all absorbing death states, where D_j denotes the death state transiting from state j, $j = 0, 1, \ldots, J-1$.

Omitting covariate information for the time being we let $H(A) = \{Z(u), 0 \leqslant u < A, B\}$ denote the history of disease process over age interval $[0, A)$ for an individual born at time B. We consider Markov intensity of the form

$$\lim_{\Delta a \longrightarrow 0} \frac{P(Z(a+\Delta a^-) = j+1 | Z(a^-) = j, H(a))}{\Delta a} = \lambda_j(a|B), \;\; j = 0, 1, \ldots, J-2$$

for disease progression and

$$\lim_{\Delta a \longrightarrow 0} \frac{P(Z(a+\Delta a^-) = D_j | Z(a^-) = j, H(a))}{\Delta a} = \eta_j(a|B), \;\; j = 0, 1, \ldots, J-1$$

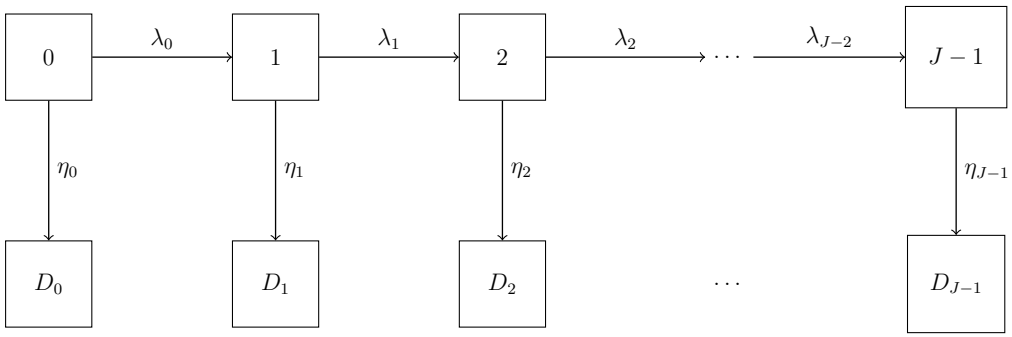


Figure 1.1: A multistate diagram with J disease states and J states representing death from each disease state.

for transitions into death states. We take it as understood in what follows that $\lambda_{J-1}(a|B) = 0$ because $J-1$ is the most advanced disease state. The transition probabilities $P(Z(a_u) = j'|Z(a_l) = j, H(a_l)) = P_{jj'}(a_l, a_u|B)$ can be expressed in terms of the transition intensities. For example, if the disease process is independent of calendar time (i.e. $\lambda_j(a|B) = \lambda_j(a)$ and $\eta_j(a|B) = \eta_j(a)$), then

$$P_{0j}(a_l, a_u) = \int_{a_l}^{a_u} P_{0,j-1}(0,t)\lambda_{j-1}(t)P_{jj}(t, a_u)dt, \text{ for } j = 1, \ldots, J-1,$$

where $P_{jj}(a_l, a_u) = \exp\left(-\int_{a_l}^{a_u}[\lambda_j(t) + \eta_j(t)]dt\right)$.

1.2 Motivating studies

1.2.1 Seroconversion following orthopedic surgery

There is an increasing risk of developing thrombosis in patients undergoing orthopedic surgery (White et al., 1998) and prophylaxis with anithrombotic heparin-based therapies is considered as standard practice. Four multicenter randomized trials were conducted to to compare enoxaprin and fondaparinux for thromboprophylaxis (Eriksson et al., 2001; Lassen et al., 2002; Bauer et al., 2001; Turpie et al., 2002). Patients undergoing orthopedic surgery were randomly assigned to receive one of the antithrombotic drugs (enoxaprin and fondaparinux) and many were seen to develop antibody responses. The development of antibodies may begin any time after surgery. Following recovery from surgery and just prior to discharge from hospital, a blood sample is taken to check the antibody status. This leads to current status data; see Figure 1.2. Interest lies in the identification of genetic markers for the risk of seroconversion. A biobank was collected containing tissue samples

from the four trials. The challenge of how to best select individuals for measurement of the biospecimens under budgetary constraints motivates the work in Chapter 3.

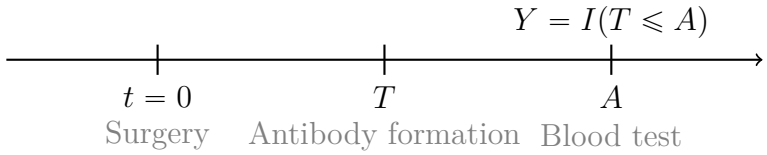


Figure 1.2: Timeline diagram for seroconversion and testing process in thromboprophylaxis trials.

1.2.2 Canadian Longitudinal Study on Aging

To better understand the dynamics of aging in Canadians and hence to improve the health and quality of life as they age, a large, national, 20-year prospective study, the Canadian Longitudinal Studies on Aging (CLSA), was designed to recruit over 50,000 individuals aged 45 to 85 years and follow them every three years until study termination or death (Raina et al., 2009; CLSA, 2015). Upon recruitment, the participants were asked to provide information including the demographic, social, physical/clinical, psychological, economic, and other measures. Additionally, over thirty thousand of 50,000 participants (i.e. the CLSA Comprehensive) are selected to provide in-depth information through periodic physical examinations and biospecimen (blood and urine) collection. It is too expensive to process all biospecimens for all individuals in the CLSA Comprehensive cohort. It is therefore of interest to efficiently select a sub-sample of individuals for biomarker testing when studying the relationship between biomarkers and disease onset or progression - this is both to meet budgetary constraints and to preserve the biospecimens for future studies.

1.2.3 Research programs in psoriasis and psoriatic arthritis

Approximately 2.5% of the North American population have psoriasis (Ps), a chronic immune-mediated dermatological skin disease that causes red raised patches of skin over various body locations including the neck, wrists, lower back, knees, ankles, fingernails and toenails (University Health Network, 2019). Around one third of patients with psoriasis will develop psoriatic arthritis (PsA), a type of inflammatory arthritis with considerable joint pain and stiffness which ultimately decreases functional ability and can even lead to disability due to joint destruction (Gladman et al., 1987, 2005). Such a critical disease condition can significantly impact the quality of life — it can be treated but not cured so understanding risk factors for disease onset and progression are important.

Researchers at the Center for Prognosis Studies in Rheumatic Disease at the University of Toronto created multiple disease registries to study the course of psoriasis and psoriatic arthritis. In particular, the University of Toronto Psoriasis Clinic (UTPC) registry was established in 2006 to enrol patients who have psoriasis but who have not yet developed psoriatic arthritis (Eder et al., 2011). All recruited individuals are assessed carefully upon recruitment and then at scheduled followed-up visits intended to take place every six months according to a standardized protocol. Biospecimens (blood and urine samples) are also collected upon entry to the clinic for future genetic testing.

Another registry, the University of Toronto Psoriatic Arthritis Clinic (UTPAC) registry, was launched much earlier in 1978 (Gladman and Chandran, 2010). Screened patients identified with psoriatic arthritis are recruited in this registry. Participants are monitored prospectively to assess the inflammatory nature of the disease activity and the rates of progression of joint damage. As in the UTPC, patients undergo a detailed clinical assessment to collect retrospective information on disease history (e.g. ages at diagnosis of Ps and PsA) and to provide samples for genetic testing.

These cohort studies motivate two projects in this thesis discussed in Chapters 2 and 4. In general, we consider methodology for assessing the effect of genetic markers on (i) the incidence of psoriatic arthritis among patients with psoriasis; and (ii) the incidence of damage in joints among patients with psoriatic arthritis. Specifically, Chapter 2 considers a framework of spatial dependence modeling of the damage processes in joints over human body among patients with psoriatic arthritis. Chapter 4 consider two-phase study designs under budgetary constraints in genetic marker analysis.

1.3 Outline of the thesis

The outline of this thesis is as follows.

Important scientific insights into chronic diseases affecting several organ systems can be gained from modeling spatial dependence of sites experiencing damage progression. The work in Chapter 2 is motivated by the prospective study of joint damage in patients with psoriatic arthritis using data from the UTPAC registry. We describe models and methods for studying spatial dependence of joint damage in psoriatic arthritis (PsA). Since a large number of joints may remain unaffected even among individuals with a long disease history, spatial dependence is first modelled in latent joint-specific indicators of susceptibility. Among susceptible joints, a Gaussian copula is adopted for dependence modeling of times to damage. Likelihood and composite likelihoods are developed for settings where individuals are under intermittent observation and progression times are subject to type

K interval censoring. Two-stage estimation procedures help mitigate the computational burden arising when a large number of processes (i.e. joints) are under consideration. Simulation studies confirm that the proposed methods provide valid inference, and an application to the motivating data from the University of Toronto Psoriatic Arthritis Clinic yields important scientific insight which can help physicians distinguish PsA from arthritic conditions with different dependence patterns.

Current status data are often encountered in epidemiological or biomedical studies (e.g. the motivating example on seroconversion following anticoagulation therapy described in Section 1.2.1). They arise from an extreme form of interval censoring in which individuals are monitored at a single assessment time to determine whether or not they have experienced a failure event of interest (Sun, 2006). In Chapter 3 we develop methods to guide the efficient selection of individuals for genetic marker testing based on current status data of progression. We consider the design and analysis of two-phase studies aiming to assess the relationship between a genetic marker and the time of disease onset only using the baseline data available at the time of study recruitment. Phase I data is comprised of current status data on the disease onset time and inexpensive covariates, where the design challenge involves the selection of individuals to have their biospecimens assayed following phase II sub-sampling; the role of the assessment time is highlighted. Likelihood and inverse probability-weighted estimating functions are considered for the basis of inference with designs based on sub-sampling individuals with extreme score statistics, extreme observations, or via stratification and sub-sampling with various stratum-specific selection probabilities.

In Chapter 4, attention is directed at the development and study of two-phase designs for pooled lifetime data from multiple disease registries; the registries are considered to have been launched to recruit individuals from a common population with different disease-dependent selection criteria. This work is motivated by the research programs in psoriasis and psoriatic arthritis conducted by researchers in two different clinics in the University of Toronto; see Section 1.2.1) for a brief introduction of the motivated studies. Multistate processes are adopted as they offer an intuitive and appealing framework to model life courses concerning a disease progression of interest. General likelihood constructions are presented under a multistate framework and under some model assumptions we derive partial likelihoods restricted to the parameters of interest. As in Chapter 3, we propose designs based on sub-sampling individuals with extreme score statistics, extreme observations or via stratification and sub-sampling with various stratum-specific selection probabilities. Approaches based on inverse probability weighting are also developed and associated optimal designs are explored on the basis of a pre-specified stratification of the

phase I sample.

Finally, Chapter 5 concludes the thesis with a review and outlines some topics for further research.

Chapter 2

Spatial dependence modeling for interval-censored processes with non-susceptibility

2.1 Introduction

2.1.1 Background

Psoriasis is a skin disease affecting roughly 2.5% of the North American population (Gelfand et al., 2005) with some countries reporting higher prevalence (Karmacharya et al., 2021). Approximately one third of psoriasis patients go on to develop psoriatic arthritis (PsA), a more serious inflammatory musculoskeletal disease involving joint inflammation, pain and damage which can reduce functional ability and quality of life (Gladman et al., 2005). There is considerable heterogeneity in the joint damage process across individuals with PsA, as well as between different joint types within individuals. Some individuals experience rapid joint destruction in many joints, while others may remain damage-free despite a long disease duration (Gladman et al., 1987). Moreover some individuals may experience very rapid joint destruction for a small number of joints, while other joints remain relatively unaffected. Relatively little work has been carried out to characterize joint types and locations most affected by PsA, or the spatial association of affected joints. Insights in this regard will help distinguish PsA from rheumatoid arthritis, osteoarthritis, and other arthritic conditions, and thereby aid in clinical diagnosis (Ruderman and Tambar, 2004).

Helliwell et al. (2000) and Bukhari et al. (2002) used cross-tabulation of damaged joint counts to assess presence of a symmetric pattern in patients with PsA. Cresswell and Farewell (2011) proposed use of generalized linear mixed models for the development of joint damage over consecutive clinic visits among patients in a PsA registry. Chandran et al. (2018) used this framework to investigate ray, row, and symmetric dependence patterns in the hand and foot joints of PsA patients in the same registry. A ray dependence pattern was considered predominant if the strongest association is between joints on the same digit, a row dependence is present if joints in the same distance from the center of the body are most highly associated, and a symmetric dependence pattern is featured if joints in the same location on the opposite side of the body exhibit the highest dependence. These authors found evidence of a symmetric and row dependence in hand and foot joints, but not ray dependence. Importantly Chandran et al. (2018) conducted their investigation based on tests of the null hypotheses of an exchangeable dependence pattern. Previous work has therefore relied on strong assumptions (Helliwell et al., 2000; Bukhari et al., 2002), or has not fully modeled the dependence of the joint damage process (Cresswell and Farewell, 2011; Chandran et al., 2018). We address this challenge here.

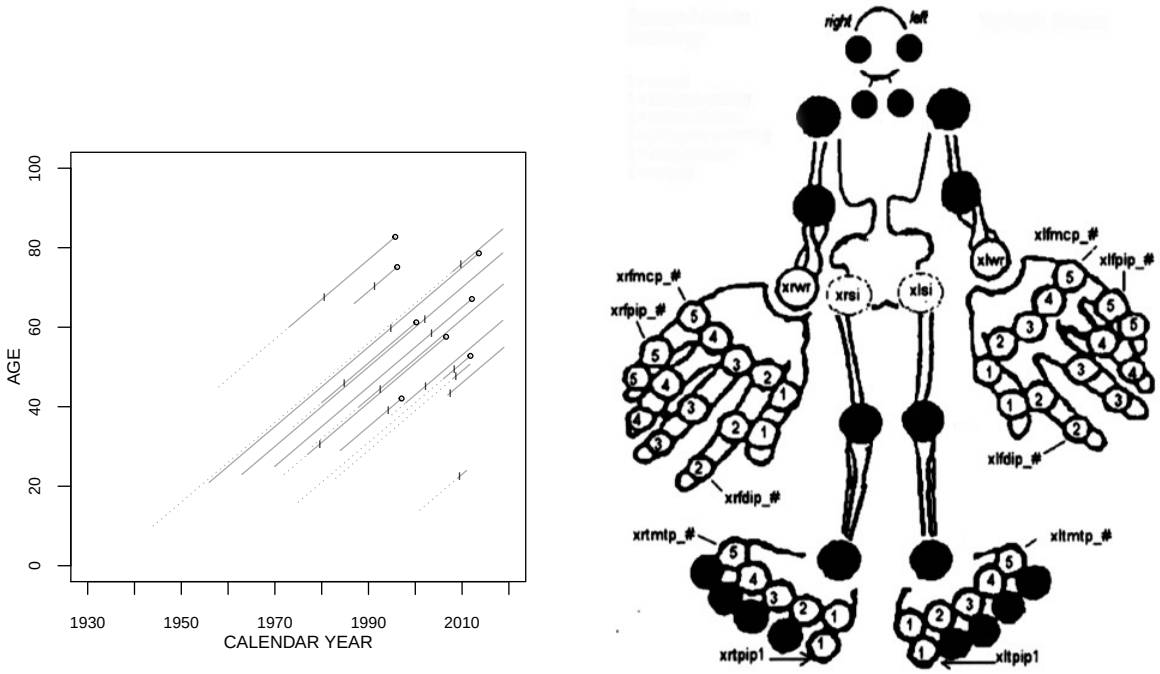
Appropriate models must accommodate a high proportion of joints that do not become damaged despite long follow-up, as well as a flexible dependence structure. Mixture models (Farewell, 1982), often called cure rate models in the failure time context, are useful for addressing the first feature. Random effect models can be adopted to address dependence in correlated failure times with nonsusceptibility (Yau and Ng, 2001; Xiang et al., 2011; Peng and Taylor, 2011), but copula models (Nelsen, 2006) are appealing as they allow separate model specification for marginal processes and dependence structures. Su and Lin (2019) used Archimedean copulas (AC) to model the association for both the susceptibility indicators and failure times with clustered survival data. Chatterjee and Shih (2001) and Jiang and Cook (2020) use models for bivariate binary response for the susceptibility model and copula functions for dependence modeling among the bivariate failure times. The present problem involves a high dimensional response, however, and the dependence structure is of central importance. We therefore adopt a Gaussian copula to model the association between failure times among susceptible joints because it i) accommodates different pairwise associations through specification of a general correlation matrix, ii) provides interpretable measures of pairwise association in failure times of susceptible joints through Kendall's τ, and iii) enjoys many appealing properties like those of the multivariate normal distribution (Song, 2000). To avoid a heavy computational burden from the high dimensional setting, we develop a pairwise composite likelihood (Cox and Reid, 2004; Varin, 2008) and describe an innovative two-stage estimation procedure.

The remainder of this chapter is organized as follows. In sub-Section 2.1.2 we describe the University of Toronto Psoriatic Arthritis Cohort and the disease process of interest motivating this work. In Section 2.2, we define notation and formulate models for the multivariate vector of latent susceptibility indicators and the failure times. These component models are marginal in the sense that dependence parameters are functionally independent of the parameters indexing the marginal models for susceptibility or failure given susceptibility. In Section 2.3, we derive the likelihood and propose estimation and inference procedures based on a pairwise composite likelihood to mitigate the computational burden due to the large number of joints under consideration; this burden is further reduced through the use of a two-stage estimation procedure we describe. Empirical studies carried out in Section 2.4 demonstrate excellent finite sample behaviour of estimators for the marginal and dependence parameters. In Section 2.5 we fit the model to data from the University of Toronto Psoriatic Arthritis Clinic, assess model fit, and describe the important insights gained. Concluding remarks are provided in Section 2.6.

2.1.2 The University of Toronto Psoriatic Arthritis Clinic

The University of Toronto Psoriatic Arthritis Clinic (UTPAC) maintains a registry of PsA patients which was formed in 1978 (Gladman and Chandran, 2010) and is now comprised of almost 2000 patients. The Lexis diagram (Keiding, 2011) of Figure 2.1a depicts information for 15 patients from this PsA registry. The lines emanate from the horizontal axis at birth dates but do not become visible as dotted lines until the onset of psoriasis. These change to solid lines upon the onset of psoriatic arthritis and terminate at loss to follow-up or death, with the latter event denoted by a circle. The vertical hatch mark on each line denotes the recruitment date, at which point blood samples are drawn and stored for future use in genomic and proteomic studies. Recruited individuals are scheduled to attend the clinic annually and to undergo radiographic examination every two years to assess the extent of damage in 28 hand, 12 foot, and 2 sacroiliac (SI) joints; Figure 2.1b contains a homunculus depicting the location of the joints that are assessed. The extent of radiological damage of each joint is measured by the modified Steinbrocker scoring system which assigns a score of 0 for a normal joint, 1 for the presence of soft tissue swelling, 2 if there is evidence of surface erosions, 3 for presence of joint space narrowing, and 4 for the most severe form of damage (Rahman et al., 1998). Since soft tissue swelling is reversible we consider a joint as damaged if a score of 2 or higher was assigned.

Patient characteristics for a subsample of 660 patients providing multiple radiological assessments are summarized in Table 2.1 where it can be seen that about 79.6% of the hand joints were not observed to develop damage even after as much as 40 years with the



(a) **(b)**

Figure 2.1: Illustrative figures showing the data collected at the University of Toronto Psoriatic Arthritis Clinic: (a) A Lexis diagram for a sample of 15 PsA patients from the UTPAC; lines depicts the age of psoriasis onset (start of dotted line), the onset of PsA (where the line becomes solid) and ending upon death (closed circle) or last contact; the vertical hatch mark denotes the age of recruitment to the PsA registry; (b) A homunculus shows the nature of the data collected at the UTPAC.

Table 2.1: Characteristics of 660 patients from the University of Toronto Psoriatic Arthritis Clinic.

No. of patients	660
% of damaged hand joints (damage-free)	20.4 (79.6)
No. female (male)	278 (382)
Mean age at clinic entry (range)	43 (14-86)
Mean age at onset of arthritis (range)	36 (6-86)
Mean no. of visits (range)	4 (1-16)

disease – this motivates formulation of the mixture model in the next section.

2.2 Notation and model formulation

In what follows we discuss the model formulation in terms of joints in PsA. To accommodate the fact that joints in different locations may develop damage at much different rates, we consider J distinct types of joints and let K_j denote the number of type j joints, $j = 1, \ldots, J$; $K = \sum_{j=1}^{J} K_j$ is the total number of joints. To label a particular joint we use a double index (j, k) for joint k of type j, and define a binary variable Z_{jk} such that $Z_{jk} = 1$ if joint (j, k) is susceptible and $Z_{jk} = 0$ otherwise. Let $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jK_j})'$ and $\mathbf{Z} = (\mathbf{Z}_1', \ldots, \mathbf{Z}_J')'$. Let T_{jk} be a nonnegative random variable denoting the time to damage for joint (j, k) with T_{jk} taken as infinite if $Z_{jk} = 0$, we write $\mathbf{T}_j = (T_{j1}, \ldots, T_{jK_j})'$ and $\mathbf{T} = (\mathbf{T}_1', \ldots, \mathbf{T}_J')'$. Covariate information specific to joint (j, k) is denoted by $\mathbf{X}_{jk}$, and we let $\mathbf{X}_j = (\mathbf{X}_{j1}', \ldots, \mathbf{X}_{jK_j}')'$, $j = 1, \ldots, J$, and $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_J')'$.

2.2.1 Dependence modeling of latent susceptibility indicators

The spatial dependence of the vector $\mathbf{Z}$ of latent susceptibility indicators is of primary interest as it reflects the pattern of joint involvement in psoriatic arthritis. Interest also lies in relating genetic and other covariates to the marginal probabilities of susceptibility. To this end, we consider first-order regression models for the marginal mean and second-order models for the pairwise associations (Qaqish and Liang, 1992).

2.2.1.1 Marginal models

Let $\mathbf{X}^{(-j,-k)}$ represent the covariate vector $\mathbf{X}$ excluding $\mathbf{X}_{jk}$ term. We assume $Z_{jk} \perp \mathbf{X}^{(-j,-k)}|\mathbf{X}_{jk}$ and let

$$\pi_{jk} = E(Z_{jk}|\mathbf{X}) = E(Z_{jk}|\mathbf{X}_{jk}), \tag{2.2.1}$$

so π_{jk} denotes the marginal mean of Z_{jk} given $\mathbf{X}$. Let $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK_j})'$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_J')'$. Suppose that $g_1(\cdot)$ is a specified one-one differentiable link function mapping $[0,1]$ onto the real line such that $g_1(\pi_{jk}) = \eta_{j0} + \mathbf{X}_{jk}'\boldsymbol{\eta}_{j1}$, where $\boldsymbol{\eta}_j = (\eta_{j0}, \boldsymbol{\eta}_{j1}')'$ is a vector of joint type-specific coefficients. Since Z_{jk} is binary, the logistic link is natural in this setting, giving the marginal model $g_1(\pi_{jk}) = \log(\pi_{jk}/(1-\pi_{jk}))$.

2.2.1.2 Second-order dependence modeling via the odds ratio

Let $\mathcal{S}_2 = \{(j,k,j',k') : (j,k) \prec (j',k'), (j,k), (j',k') \in \mathcal{S}\}$ represent a set of size $K(K-1)/2$ containing all pairwise combinations of elements in the individual joint index set $\mathcal{S}$, where $\prec$ indicates that $j < j'$ or $k < k'$ if $j = j'$. Without loss of generality, we consider two distinct joints labeled (j,k) and (j',k') with $(j,k) \prec (j',k')$ in the following. Let $\mathbf{V}_{jkj'k'} = (\mathbf{X}_{jk}', \mathbf{X}_{j'k'}')'$ and $\mathbf{X}^{(-j,-k,-j',-k')}$ denote the covariate vector $\mathbf{X}$ excluding $\mathbf{X}_{jk}$ and $\mathbf{X}_{j'k'}$. Suppose that

$$(Z_{jk}, Z_{j'k'}) \perp \mathbf{X}^{(-j,-k,-j',-k')}|\mathbf{V}_{jkj'k'}, \tag{2.2.2}$$

then $\omega_{jkj'k'} = E(W_{jkj'k'}|\mathbf{X}) = E(W_{jkj'k'}|\mathbf{V}_{jkj'k'})$ with $W_{jkj'k'} = Z_{jk}Z_{j'k'}$ and the conditional covariance $\text{cov}(Z_{jk}, Z_{j'k'}|\mathbf{V}_{jkj'k'}) = \omega_{jkj'k'} - \pi_{jk}\pi_{j'k'}$. If $\zeta_{jkj'k'}$ denotes the odds ratio characterizing the association between Z_{jk} and $Z_{j'k'}$ given $\mathbf{V}_{jkj'k'}$, then

$$\zeta_{jkj'k'} = \frac{\omega_{jkj'k'}(1 - \pi_{jk} - \pi_{j'k'} + \omega_{jkj'k'})}{(\pi_{jk} - \omega_{jkj'k'})(\pi_{j'k'} - \omega_{jkj'k'})}.$$

Let $g_2(\cdot)$ denote a one-one differentiable function mapping $[0,\infty)$ onto the real line and set

$$g_2^Z(\zeta_{jkj'k'}) = \mathbf{V}_{jkj'k'}'\boldsymbol{\gamma}, \tag{2.2.3}$$

where $\boldsymbol{\gamma}$ is a vector of regression coefficients characterizing the dependence. In the simulation studies of Section 2.4 the log link is used giving $g_2^Z(\zeta_{jkj'k'}) = \log \zeta_{jkj'k'}$. For the conditional distribution of the multivariate binary variable $\mathbf{Z}$ given covariate X, the function $g_1(\cdot)$ in (2.2.1) determines the mean structure and $g_2^Z(\cdot)$ in (2.2.3) characterizes the pairwise dependence structure. We assume higher-order dependence parameters are zero in which case we let $P(\mathbf{Z}|\mathbf{X};\boldsymbol{\varphi})$ denote the conditional distribution for $\mathbf{Z}|\mathbf{X}$ where $\boldsymbol{\varphi} = (\boldsymbol{\eta}', \boldsymbol{\gamma}')'$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_J')'$.

2.2.2 Dependence modeling for failure times in susceptible joints

2.2.2.1 Marginal models

Let $\mathbf{Z}^{(-j,-k)}$ represent the vector of susceptibility indicators excluding the Z_{jk} element. We assume that T_{jk} is independent of the full covariate $\mathbf{X}$ and $\mathbf{Z}^{(-j,-k)}$ given Z_{jk}, so $T_{jk} \perp (\mathbf{X}, \mathbf{Z}^{(-j,-k)})|Z_{jk}$. For a nonsusceptible joint (j,k), the survival probability $P(T_{jk} > t|Z_{jk} = 0) = 1$ for any finite $t \geqslant 0$ since $T_{jk} = \infty$. For a susceptible joint (j,k) the conditional survival probability $P(T_{jk} > t|Z_{jk} = 1)$ is a function of t with range $[0,1]$. If $\mathcal{F}_j(\cdot)$ represents the survivor function for joint (j,k), we have

$$P(T_{jk} > t|\mathbf{Z}, \mathbf{X}) = P(T_{jk} > t|Z_{jk}) = \begin{cases} \mathcal{F}_j(t) & \text{if } Z_{jk} = 1 \\ 1 & \text{if } Z_{jk} = 0, \end{cases} \tag{2.2.4}$$

for any $t \geqslant 0$. The subscript j of $\mathcal{F}_j$ indicates that type-specific marginal models are assumed for $T_{jk}|Z_{jk} = 1$, which includes as a special case, common marginal models for all joints with $\mathcal{F}_1(\cdot) = \cdots = \mathcal{F}_J(\cdot) = \mathcal{F}(\cdot)$. Joint-specific marginal models can be identified by letting $P(T_{jk} > t|Z_{jk} = 1) = \mathcal{F}_{jk}(t)$, but here we focus on joint type-specific survivor functions with the survivor function for type j joints indexed by θ_j and was written $\mathcal{F}_j(t;\theta_j)$.

2.2.2.2 Dependence modeling with the Gaussian copula

Let $\mathcal{S} = \{(j,k) : k = 1,\ldots,K_j, j = 1,\ldots,J\}$ be the set of size K containing indices of all joints for each individual. Furthermore let $m = \sum_j \sum_k Z_{jk}$ be the total number of susceptible joints for an individual ($0 \leqslant m \leqslant K$) and $\bar{\mathcal{S}} = \{(j,k) : Z_{jk} = 1\}$ be the set of size m containing the labels of all susceptible joints where $\bar{\mathcal{S}}$ is the null set if $m = 0$.

We assume $\mathbf{T} \perp \mathbf{X}|\mathbf{Z}$. Note if $m = 0$, then $\mathbf{Z} = \mathbf{0}$ and $P(\mathbf{T} \geqslant \mathbf{t}|\mathbf{Z} = \mathbf{0}) = 1$. If $0 < m \leqslant K$, we let $\bar{\mathbf{T}} = (T_{jk}, (j,k) \in \bar{\mathcal{S}})$ represent the failure time vector $\mathbf{T}$ excluding T_{jk} terms for $(j,k) \notin \bar{\mathcal{S}}$; and $\bar{\mathbf{T}} = \mathbf{T}$ if $m = K$. We therefore have $P(\mathbf{T} \geqslant \mathbf{t}|\mathbf{Z}) = P(\bar{\mathbf{T}} \geqslant \bar{\mathbf{t}}|\mathbf{Z})$ if $0 < m \leqslant K$, or $P(\mathbf{T} \geqslant \mathbf{t}|\mathbf{Z}) = 1$ if $m = 0$, where $\bar{\mathbf{t}}$ is a $m \times 1$ vector with finite nonnegative elements corresponding to the susceptible joints.

If $m \geqslant 2$, the joint survival function $P(\bar{\mathbf{T}} > \bar{\mathbf{t}}|\mathbf{Z})$ can be specified through an m-dimensional copula function (Sklar, 1959; Joe, 1997) indexed by $\boldsymbol{\rho}$, such that

$$\mathcal{C}(U_{jk} \geqslant u_{jk}, (j,k) \in \bar{\mathcal{S}}; \boldsymbol{\rho}) = P(U_{jk} \geqslant u_{jk}, (j,k) \in \bar{\mathcal{S}}; \boldsymbol{\rho})$$

where U_{jk} are uniform $[0,1]$ random variables. If we set $U_{jk} = \mathcal{F}_j(T_{jk};\boldsymbol{\theta}_j)$, then the multivariate survival function for $\bar{\mathbf{T}}|\mathbf{Z}$ is

$$P(T_{jk} > t_{jk}, (j,k) \in \bar{\mathcal{S}}|\mathbf{Z}) = \mathcal{C}(\mathcal{F}_j(t_{jk};\boldsymbol{\theta}_j), (j,k) \in \bar{\mathcal{S}}; \boldsymbol{\rho}). \tag{2.2.5}$$

Gaussian copula functions accommodate a flexible pairwise dependence structure (Nelsen, 2006) and models based on them enjoy many appealing properties – any subvector, for example, retains the same distributional form as the full vector (Song, 2000). We adopt a Gaussian copula to model the multivariate failure times among susceptible joints so that (2.2.5) can be expressed by

$$\mathcal{C}(u_{jk}, (j,k) \in \bar{\mathcal{S}}; \boldsymbol{\rho}) = \Phi_m(\Phi^{-1}(u_{jk}), (j,k) \in \bar{\mathcal{S}}; \boldsymbol{\rho}) \tag{2.2.6}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable and $\Phi_m(\cdot; \boldsymbol{\rho})$ is the cumulative distribution function of a $m \times 1$ multivariate normal random variable with mean zero and $m \times m$ covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\rho})$ with off-diagonal entries $\rho_{jkj'k'}$, $(j,k), (j',k') \in \bar{\mathcal{S}}$. For joints (j,k) and (j',k') and given susceptibility status $(Z_{jk}, Z_{j'k'}) = (1,1)'$, the association between T_{jk} and $T_{j'k'}$ can be measured by Kendall's τ given by $\tau_{jkj'k'} = 2\arcsin(\rho_{jkj'k'})/\pi$, $(j,k), (j',k') \in \bar{\mathcal{S}}$. A second-order model can be proposed to describe the within-cluster association by $g_2^T(\tau_{jkj'k'}) = \mathbf{X}_i'\boldsymbol{\xi}$, where $g_2^T(\cdot)$ is a one-to-one differentiable link function mapping Kendall's τ onto the real line. For example, the Fisher transformation $g_2^T(\tau) = \log((1+\tau)/(1-\tau))$ is a popular choice. The joint distribution for $\mathbf{T}|\mathbf{Z}$ is indexed in general by $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\rho}')'$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_J')'$, but a subset of the parameter in $\boldsymbol{\theta}$ may appear in the joint survival model depending on the realization of $\mathbf{Z}$.

2.3 Methods for estimation and inference

In Appendix A.2 we recast the data and model in terms of counting processes and intensity functions and explicitly state the assumptions required for a conditionally independent visit process (CIVP) defined by Cook and Lawless (2018). Under CIVP assumptions (A.2.1) and (A.2.4) given in Appendix A.2.2, the partial likelihood based on the specified model is of the form

$$L \propto P(\mathbf{T} \in \mathscr{B}|\mathbf{X}) = \sum_{\mathbf{Z} \in \mathcal{Z}} P(\mathbf{T} \in \mathscr{B}|\mathbf{Z}) P(\mathbf{Z}|\mathbf{X}),$$

where $\mathcal{Z}$ is the sample space of $\mathbf{Z}$ and the size of $\mathcal{Z}$ is 2^K; $\mathscr{B} = \prod_{(j,k) \in \mathcal{S}} \mathscr{B}_{jk}$ denotes the censoring region of all joints with $\mathscr{B}_{jk} = (a_{0jk}, a_{1jk}]$ ($0 \leqslant a_{0jk} < a_{1jk} \leqslant \infty$) indicating the interval within which T_{jk} occurs. But we note that when the total number of joints K is large, joint modeling becomes more challenging due to the higher dimensional model. Since the parameter of interest, $\boldsymbol{\psi} = (\boldsymbol{\vartheta}', \boldsymbol{\varphi}')'$, involves at most pairwise associations, we consider a pairwise composite likelihood (Lindsay, 1988; Cox and Reid, 2004) for the settings where q, the length of $\boldsymbol{\psi}$, is less than the sample size N.

2.3.1 The pairwise composite likelihood

Under the pairwise conditional independence visiting conditions (A.2.5) and (A.2.8) given in Appendix A.2.4 and that the observation process is noninformative, the partial likelihood contribution from a pair of joints (j,k) and (j',k') is equivalent to the probability of $(T_{jk}, T_{j'k'})$ falling in the censoring region $\mathscr{B}_{jk} \times \mathscr{B}_{j'k'}$ given covariates $\mathbf{V}_{jkj'k'}$ associated with this pair of joints, that is, $P(T_{jk} \in \mathscr{B}_{jk}, T_{j'k'} \in \mathscr{B}_{j'k'} | \mathbf{V}_{jkj'k'})$. Here we assume $T_{jk}, T_{j'k'} \perp \{\mathbf{X}, \mathbf{Z}^{(-j,-k,-j',-k')}\} | Z_{jk}, Z_{j'k'}$, where $\mathbf{Z}^{(-j,-k,-j',-k')}$ is the $\mathbf{Z}$ vector excluding the Z_{jk} and $Z_{j'k'}$ entries. If $\boldsymbol{\vartheta}_{jkj'k'} = (\boldsymbol{\theta}'_j, \boldsymbol{\theta}'_{j'}, \rho_{jkj'k'})'$ and $\boldsymbol{\varphi}_{jj'} = (\boldsymbol{\eta}'_j, \boldsymbol{\eta}'_{j'}, \boldsymbol{\gamma}')'$, the pairwise composite likelihood contribution arising from $\{\mathscr{B}_i, \mathbf{X}_i\}$ is

$$\mathscr{L}_{i2}(\boldsymbol{\psi}) \propto \prod_{(j,k,j',k') \in \mathcal{S}_2} \mathscr{L}_{ijkj'k'2}, \tag{2.3.1}$$

where $\mathscr{L}_{ijkj'k'2}$, indexed by $\boldsymbol{\psi}_{jkj'k'} = (\boldsymbol{\vartheta}'_{jkj'k'}, \boldsymbol{\varphi}'_{jj'})'$, equals

$$\sum_{Z_{jk}, Z_{j'k'}} P(T_{ijk} \in \mathscr{B}_{ijk}, T_{ij'k'} \in \mathscr{B}_{ij'k'} | Z_{jk}, Z_{j'k'}; \boldsymbol{\vartheta}_{jkj'k'}) P(Z_{jk}, Z_{j'k'} | \mathbf{X}_{ijk}, \mathbf{X}_{ij'k'}; \boldsymbol{\varphi}_{jj'}).$$

Note $P(T_{ijk} \in \mathscr{B}_{ijk}, T_{ij'k'} \in \mathscr{B}_{ij'k'} | Z_{jk}, Z_{j'k'})$ can be expressed in terms of marginal survivor functions $\mathcal{F}_j(\cdot)$, $\mathcal{F}_{j'}(\cdot)$ and a bivariate Gaussian copula function indexed by a dependence parameter $\rho_{jkj'k'}$; see Appendix A.1 for more detailed derivations.

2.3.2 A two-stage estimation algorithm

To estimate $\boldsymbol{\psi}$ we adopt a computationally convenient two-stage estimation procedure for ψ in the spirit of Shih and Louis (1995). We first re-partition $\boldsymbol{\psi} = (\boldsymbol{\psi}'_1, \boldsymbol{\psi}'_2)'$ where $\boldsymbol{\psi}_1 = (\boldsymbol{\theta}', \boldsymbol{\eta}')'$ is the vector of parameters associated with marginal regression models to failure times and susceptibilities, and $\boldsymbol{\psi}_2 = (\boldsymbol{\rho}', \boldsymbol{\gamma}')'$ contains parameters associated with pairwise dependence. Under the working independence assumption needed for two-stage estimation, a stronger CIVP assumption is required than was necessary under the pairwise conditional independence conditions; see (A.2.10) and (A.2.11) of Appendix A.2.4. Under this stronger CIVP assumption we derive the working independence composite likelihood for $\boldsymbol{\psi}_1$ arising from intermittent assessments:

$$\mathscr{L}_{i1}(\boldsymbol{\psi}_1) \propto \prod_{(j,k) \in \mathcal{S}} P(T_{ijk} \in \mathscr{B}_{ijk} | \mathbf{X}_{ijk}; \boldsymbol{\psi}_{1j}) \tag{2.3.2}$$

for estimation in $\boldsymbol{\psi}_1 = (\boldsymbol{\psi}'_{11}, \ldots, \boldsymbol{\psi}'_{1J})'$ with $\boldsymbol{\psi}_{1j} = (\boldsymbol{\theta}'_j, \boldsymbol{\eta}'_j)'$, where

$$P(T_{ijk} \in \mathscr{B}_{ijk} | \mathbf{X}_{ijk}; \boldsymbol{\psi}_{1j}) = \sum_z P(T_{ijk} \in \mathscr{B}_{ijk} | Z_{ijk} = z; \boldsymbol{\theta}_j) \cdot P(Z_{ijk} = z | \mathbf{X}_{ijk}; \boldsymbol{\eta}_j).$$

Stage I involves constructing and maximizing the marginal composite likelihood $\mathscr{L}_1(\boldsymbol{\psi}_1) = \prod_{i=1}^N \mathscr{L}_{i1}(\boldsymbol{\psi}_1)$ is labeled as stage I. The corresponding score vector is $\mathscr{S}_1(\boldsymbol{\psi}_1) = \sum_{i=1}^N \mathscr{S}_{i1}(\boldsymbol{\psi}_1)$, where $\mathscr{S}_{i1}(\boldsymbol{\psi}_1) = \partial \log \mathscr{L}_{i1}(\boldsymbol{\psi}_1)/\partial \boldsymbol{\psi}_1$ and let $\mathscr{I}_{i1} = -\partial \mathscr{S}_{i1}(\boldsymbol{\psi}_1)/\partial \boldsymbol{\psi}_1'$. We let $\breve{\boldsymbol{\psi}}_1$ denote the value of $\boldsymbol{\psi}$ solving the composite score equation $\mathscr{S}_1(\boldsymbol{\psi}_1) = 0$. Under suitable regularity conditions (Varin et al., 2011),

$$\sqrt{N}(\breve{\boldsymbol{\psi}}_1 - \boldsymbol{\psi}_1) \xrightarrow{d} \text{MVN}(\mathbf{0}, \bar{\mathcal{A}}_{11}^{-1}\bar{\mathcal{B}}_{11}\bar{\mathcal{A}}_{11}^{-1}) \tag{2.3.3}$$

where $\bar{\mathcal{A}}_{11} = E[\mathscr{I}_{i1}(\boldsymbol{\psi}_1)]$ and $\bar{\mathcal{B}}_{11} = E\{\mathscr{S}_{i1}(\boldsymbol{\psi}_1)\mathscr{S}_{i1}'(\boldsymbol{\psi}_1)\}$. In practice, these matrices can be estimated empirically by $\hat{\bar{\mathcal{A}}}_{11}(\breve{\boldsymbol{\psi}}_1) = N^{-1}\sum_{i=1}^N \mathscr{I}_{i1}(\breve{\boldsymbol{\psi}}_1)$ and $\hat{\bar{\mathcal{B}}}_{11}(\breve{\boldsymbol{\psi}}_1) = N^{-1}\sum_{i=1}^N \mathscr{S}_{i1}(\breve{\boldsymbol{\psi}}_1)\mathscr{S}_{i1}'(\breve{\boldsymbol{\psi}}_1)$. The `nlm` function in `R` can be used for the optimization; by specifying `hessian=TRUE` in this function, $\hat{\bar{\mathcal{A}}}_{11}(\breve{\boldsymbol{\psi}}_1)$ is specified through the Hessian matrix obtained by finite differencing of the observed data log-likelihood function.

In stage II, the estimation for $\boldsymbol{\psi}_2$ is carried out by maximizing $\mathscr{L}_2(\breve{\boldsymbol{\psi}}_1, \boldsymbol{\psi}_2) = \prod_{i=1}^N \mathscr{L}_{i2}(\breve{\boldsymbol{\psi}}_1, \boldsymbol{\psi}_2)$, where $\mathscr{L}_{i2}(\breve{\boldsymbol{\psi}}_1, \boldsymbol{\psi}_2)$ is given in (2.3.1) but with $\boldsymbol{\psi}_1$ set at $\breve{\boldsymbol{\psi}}_1$. If $\mathscr{S}_{i22}(\boldsymbol{\psi}_2; \boldsymbol{\psi}_1) = \partial \log \mathscr{L}_{i2}(\boldsymbol{\psi})/\partial \boldsymbol{\psi}$, we set $\mathscr{S}_{22}(\boldsymbol{\psi}_2; \breve{\boldsymbol{\psi}}_1) = \sum_{i=1}^N \mathscr{S}_{i22}(\boldsymbol{\psi}_2; \breve{\boldsymbol{\psi}}_1) = 0$ and obtain an estimator of $\boldsymbol{\psi}_2$, denoted by $\breve{\boldsymbol{\psi}}_2$. We then let $\breve{\boldsymbol{\psi}} = (\breve{\boldsymbol{\psi}}_1, \breve{\boldsymbol{\psi}}_2)$ denote the estimator of $\boldsymbol{\psi}$ obtained from this two-stage procedure. Under suitable regularity conditions (Boos and Stefanski, 2013),

$$\sqrt{N}(\breve{\boldsymbol{\psi}}_2 - \boldsymbol{\psi}_2) \xrightarrow{d} \text{MVN}(\mathbf{0}, \bar{\mathcal{A}}_{22}^{-1}\left[\bar{\mathcal{B}}_{22} - \bar{\mathcal{A}}_{21}\bar{\mathcal{A}}_{11}\bar{\mathcal{A}}_{21}'\right]\bar{\mathcal{A}}_{22}^{-1}),$$

where $\bar{\mathcal{A}}_{21} = E[-\partial \mathscr{S}_{i2}(\boldsymbol{\psi})/\partial \boldsymbol{\psi}_1']$, $\bar{\mathcal{A}}_{22} = E[-\partial \mathscr{S}_{i2}(\boldsymbol{\psi})/\partial \boldsymbol{\psi}_2']$, and $\bar{\mathcal{B}}_{22} = E[\mathscr{S}_{i22}(\boldsymbol{\psi})\mathscr{S}_{i22}'(\boldsymbol{\psi})]$. The empirical estimates of these expectations are

$$\hat{\bar{\mathcal{A}}}_{21}(\breve{\boldsymbol{\psi}}) = -N^{-1}\sum_{i=1}^N \frac{\partial \mathscr{S}_{i2}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}_1'}\Big|_{\boldsymbol{\psi}=\breve{\boldsymbol{\psi}}}, \hat{\bar{\mathcal{A}}}_{22}(\breve{\boldsymbol{\psi}}) = -N^{-1}\sum_{i=1}^N \frac{\partial \mathscr{S}_{i2}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}_2'}\Big|_{\boldsymbol{\psi}=\breve{\boldsymbol{\psi}}},$$
$$\text{and } \hat{\bar{\mathcal{B}}}_{22}(\breve{\boldsymbol{\psi}}) = N^{-1}\sum_{i=1}^N \mathscr{S}_{i22}(\breve{\boldsymbol{\psi}})\mathscr{S}_{i22}'(\breve{\boldsymbol{\psi}}).$$

Again, the general-purpose optimization function such as `nlm` or `optim` in R can be applied to facilitate estimation and computation of the corresponding standard error.

2.4 Simulation studies

Two sets of simulation studies are conducted to examine the finite sample properties of the two-stage pairwise composite likelihood estimator for the complex spatially correlated interval-censored data. In the first set of studies, we considered three joint types

$(J = 3)$ comprised of $(K_1, K_2, K_3) = (2, 2, 2)$ or $(2, 6, 6)$ joints for each type. For illustrative purpose a 2×1 binary covariate vector $\mathbf{X} = (X_1, X_2)'$ is considered and X_1 and X_2 are generated by independent Bernoulli distributions with successful probability p_1 and p_2, respectively. We set $p_1 = 0.45$ and $p_2 = 0.05$ or 0.1. The susceptibility indicator $\mathbf{Z}$ is simulated from a joint distribution in which d-order dependencies are zero $(d \geqslant 3)$ and the marginal models for Z_{jk} conditional on $\mathbf{X}$ is given by $P(Z_{jk} = 1|\mathbf{X}) = \exp(\eta_{j0} + \eta_1 X_1 + \eta_2 X_2)/(1 + \exp(\eta_{j0} + \eta_1 X_1 + \eta_2 X_2))$, where η_{j0} is a type-specific intercept and $(\eta_1, \eta_2)'$ is a vector of regression coefficients. When $(K_1, K_2, K_3) = (2, 2, 2)$, the marginal susceptibilities $(\pi_1, \pi_2, \pi_3)' = (0.2, 0.15, 0.15)'$; when $(K_1, K_2, K_3) = (2, 6, 6)$, the marginal susceptibilities within each joint type $(\pi_1, \pi_2, \pi_3)' = (0.2, 0.05, 0.05)'$, where $\pi_j = P(Z_{jk} = 1)$, $j = 1, 2, 3$. By setting $\eta_1 = -0.2$ and $\eta_2 = 0$ or 0.1, we then can solve η_{j0}. Moreover, we consider type-specific pairwise association between Z_{jk} and $Z_{j'k'}$ parameterized by odds ratios $\zeta_{jkj'k'} = \zeta_{jj'} = \exp(\gamma_{jj'})$. We set $\zeta_{jj} = 1.2$, $j = 1, 2, 3$ and $\zeta_{jj'} = 1.05$, $j \neq j'$. For the nonsusceptible joints (i.e. $Z_{jk} = 0$), we set $T_{jk} = \infty$; and for those susceptibles, we generate $\bar{\mathbf{T}}$ from the multivariate survival function of the form (2.2.5), where we consider a piecewise constant hazard function for the marginal model $P(T_{jk} \geqslant t|Z_{jk} = 1) = \exp(-\alpha_j t)$, and type-specific pairwise association between failure times of $T_{jk}|Z_{jk} = 1$ and $T_{j'k'}|Z_{j'k'} = 1$ is measured by Kendall's τ with $\tau_{jkj'k'} = \tau_{jj'}$. We set $\tau_{jj} = 0.15$, $j = 1, 2, 3$ and $\tau_{jj'} = 0.05$, $j \neq j'$. We set $P(T_{1k} \leqslant 1|Z_{1k} = 1) = 0.9$, $P(T_{2k} \leqslant 1|Z_{2k} = 1) = 0.85$ and $P(T_{3k} \leqslant 1|Z_{3k} = 1) = 0.85$ and solve α_j. To generate the censoring intervals, we simulate individual clinical visits by using a Poisson process with rate λ over $(0, 1]$ and we set $\lambda = 5$. The results for the first set of studies are summarized in Table 2.2 and Table A.1. We find that empirical biases (BIAS) of all estimates are negligible for the parameter of interest and decrease as the sample size N (and) or the total number of joints K increases, as expected. The empirical coverage probabilities (ECP) are close to the nominal 95% level, which indicates a good agreement between the empirical (ESE) and model-based standard errors (ASE).

In the second set of simulation studies, we set up marginal models and pairwise associations for the susceptibility indicator while left the d-order dependencies unspecified $(d \geqslant 3)$. We obtain a joint distribution of $\mathbf{Z}$ via the iterative proportional fitting procedure by calling `ObtainMultBinaryDist` function from `R` package `mipfp` (Barthelemy and Suesse, 2018). For simple illustration, we considered two joint types (J=2) comprised of $(K_1, K_2) = (2, 2)$ or $(4, 4)$ joints for each type. We set $p_1 = 0.5$ and $p_2 = 0.05$ or 0.1. By setting $(\pi_1, \pi_2)' = (0.65, 0.75)'$ and $(\eta_1, \eta_2)' = (0.45, 0.5)'$, we can solve for η_{j0}. Two-piece piecewise constant hazard models are considered for marginal failure times $P(T_{jk} \geqslant t|Z_{jk} = 1) = \exp(-(\alpha_{j1} I(t < b_j) + \alpha_{j2} I(t \geqslant b_j)))$, where $\boldsymbol{\alpha}_j = (\alpha_{j1}, \alpha_{j2})'$ is a 2×1 vector of constant hazards with a cutpoint $b_j = 0.55$. We set $P(T_{1k} \leqslant 1|Z_{1k} = 1) = 0.85$,

Table 2.2: Empirical properties of two-stage composite likelihood estimates based on one thousand simulated samples of size $N = 6000$ with $J = 3$ and zero d-order dependencies ($d \geqslant 3$); two-piece piecewise constant proportional hazards models for $T_{jk}|Z_{jk} = 1$ without covariates and the logistic models for $Z_{jk}|\mathbf{X}_{jk}$ in stage I.

	STAGE I										STAGE II									
	$(K_1, K_2, K_3) = (2, 2, 2)$					$(K_1, K_2, K_3) = (2, 6, 6)$						$(K_1, K_2, K_3) = (2, 2, 2)$					$(K_1, K_2, K_3) = (2, 6, 6)$				
PARA	TRUE	BIAS	ESE	ASE	ECP	TRUE	BIAS	ESE	ASE	ECP	PARA	TRUE	BIAS	ESE	ASE	ECP	TRUE	BIAS	ESE	ASE	ECP
(a) $p_2 = 0.05, \eta_2 = 0.00$																					
$\log \alpha_{11}$	0.834	−0.008	0.083	0.088	0.932	×	−0.007	0.084	0.083	0.935	$\rho^{\dagger}_{11}$	0.386	0.012	0.228	0.233	0.952	×	0.005	0.227	0.238	0.941
$\log \alpha_{12}$	0.834	−0.007	0.272	0.287	0.910	×	−0.002	0.276	0.269	0.941	$\rho^{\dagger}_{12}$	0.124	−0.003	0.132	0.127	0.955	×	−0.007	0.133	0.133	0.957
$\log \alpha_{21}$	0.640	−0.015	0.129	0.126	0.940	×	−0.015	0.128	0.124	0.935	$\rho^{\dagger}_{13}$	0.124	−0.011	0.131	0.129	0.946	×	−0.008	0.132	0.130	0.958
$\log \alpha_{22}$	0.640	−0.005	0.346	0.333	0.929	×	−0.012	0.340	0.336	0.932	$\rho^{\dagger}_{22}$	0.386	0.005	0.342	0.368	0.931	×	0.010	0.262	0.262	0.934
$\log \alpha_{31}$	0.640	−0.018	0.129	0.126	0.942	×	−0.016	0.129	0.132	0.928	$\rho^{\dagger}_{23}$	0.124	−0.006	0.159	0.160	0.955	×	−0.009	0.161	0.154	0.961
$\log \alpha_{32}$	0.640	−0.024	0.340	0.332	0.928	×	−0.008	0.342	0.346	0.922	$\rho^{\dagger}_{33}$	0.386	0.007	0.342	0.347	0.947	×	0.010	0.258	0.265	0.944
η_{10}	−1.299	0.009	0.066	0.068	0.939	×	0.007	0.067	0.065	0.948	γ_{11}	0.182	−0.013	0.102	0.100	0.960	×	−0.009	0.102	0.102	0.955
η_{20}	−1.648	0.013	0.103	0.102	0.939	×	0.012	0.091	0.090	0.936	γ_{12}	0.049	−0.009	0.061	0.061	0.949	×	−0.011	0.056	0.056	0.949
η_{30}	−1.648	0.016	0.103	0.102	0.940	×	0.011	0.092	0.093	0.938	γ_{13}	0.049	−0.008	0.061	0.060	0.942	×	−0.008	0.056	0.057	0.944
η_1	−0.200	0.000	0.035	0.035	0.945	×	0.000	0.034	0.034	0.954	γ_{22}	0.182	−0.014	0.143	0.142	0.952	×	−0.015	0.089	0.088	0.952
η_2	0.000	−0.002	0.077	0.076	0.958	×	−0.003	0.074	0.072	0.958	γ_{23}	0.049	−0.008	0.072	0.073	0.939	×	−0.011	0.061	0.060	0.960
											γ_{33}	0.182	−0.010	0.143	0.147	0.945	×	−0.018	0.090	0.090	0.955
(b) $p_2 = 0.05, \eta_2 = 0.10$																					
$\log \alpha_{11}$	0.834	−0.008	0.084	0.086	0.931	×	−0.007	0.084	0.082	0.931	$\rho^{\dagger}_{11}$	0.386	−0.001	0.228	0.233	0.946	×	0.003	0.226	0.232	0.945
$\log \alpha_{12}$	0.834	−0.007	0.275	0.285	0.916	×	−0.004	0.274	0.268	0.934	$\rho^{\dagger}_{12}$	0.124	−0.003	0.132	0.129	0.961	×	−0.005	0.133	0.136	0.941
$\log \alpha_{21}$	0.640	−0.016	0.128	0.130	0.931	×	−0.022	0.130	0.126	0.939	$\rho^{\dagger}_{13}$	0.124	−0.010	0.131	0.132	0.944	×	−0.005	0.132	0.128	0.958
$\log \alpha_{22}$	0.640	−0.009	0.341	0.341	0.928	×	−0.025	0.345	0.335	0.927	$\rho^{\dagger}_{22}$	0.386	0.015	0.348	0.364	0.937	×	0.005	0.260	0.260	0.947
$\log \alpha_{31}$	0.640	−0.020	0.129	0.127	0.938	×	−0.016	0.129	0.130	0.936	$\rho^{\dagger}_{23}$	0.124	−0.004	0.158	0.159	0.955	×	−0.010	0.160	0.155	0.955
$\log \alpha_{32}$	0.640	−0.029	0.340	0.334	0.935	×	−0.011	0.344	0.344	0.929	$\rho^{\dagger}_{33}$	0.386	0.008	0.345	0.344	0.953	×	0.000	0.259	0.266	0.950
η_{10}	−1.304	0.008	0.067	0.068	0.93	×	0.007	0.066	0.066	0.947	γ_{11}	0.182	−0.011	0.102	0.102	0.951	×	−0.009	0.102	0.098	0.960
η_{20}	−1.653	0.015	0.102	0.106	0.932	-2.864	0.016	0.093	0.091	0.945	γ_{12}	0.049	−0.008	0.061	0.061	0.945	×	−0.012	0.056	0.055	0.962
η_{30}	−1.653	0.018	0.103	0.104	0.939	-2.864	0.012	0.092	0.092	0.934	γ_{13}	0.049	−0.009	0.061	0.060	0.951	×	−0.008	0.056	0.056	0.950
η_1	−0.200	0.000	0.035	0.035	0.947	×	0.000	0.034	0.034	0.948	γ_{22}	0.182	−0.014	0.145	0.145	0.952	×	−0.015	0.089	0.090	0.953
η_2	0.100	−0.001	0.071	0.070	0.951	×	−0.002	0.067	0.066	0.950	γ_{23}	0.049	−0.009	0.072	0.073	0.937	×	−0.010	0.062	0.060	0.957
											γ_{33}	0.182	−0.012	0.144	0.146	0.948	×	−0.019	0.090	0.090	0.950

Note. PARA, parameter; TRUE, true value of the parameter; ESE, sample standard deviation; ASE, model-based standard error; × means the value is the same as that specified in the "TRUE" column when $(K_1, K_2, K_3) = (2, 2, 2)$; $\rho^{\dagger}_{jj'} = \tan(.5\pi\rho_{jj'})$, $\gamma_{jj'} = \log \zeta_{jj'}$, $j \leqslant j'; j, j' \in \{1, 2\}$.

$P(T_{2k} \leqslant 1 | Z_{2k} = 1) = 0.9$, and $\alpha_{11} = \alpha_{12}$, $\alpha_{22} = 1.1\alpha_{21}$. For pairwise associations among susceptibilities and failure times of suceptibles, we set $(\zeta_{11}, \zeta_{12}, \zeta_{22})' = (1.15, 1.1, 1.05)'$ and $(\tau_{11}, \tau_{12}, \tau_{22})' = (0.1, 0.15, 0)'$. The results for the second set of studies are summarized in Table 2.3. When either the total number of joints K or the visit rate λ increases, the estimating efficiency in parameters is improved with smaller bias and reduced standard errors as expected.

2.5 Hand joint damage in psoriatic arthritis

We apply the proposed method to the data on 28 hand joints (14 in each hand) in 660 patients from the University of Toronto Psoriatic Arthritis Clinic. Some of the Human Leukocyte Antigen (HLA) markers have been identified as important risk factors for the development of the disease (Gladman and Farewell, 1995). For the marginal (joint-level) model for susceptibility we therefore consider gender (female versus male), age at the diagnosis of PsA, and a set of HLA markers that have previously been shown to be associated with PsA including HLA-A2, B13, B27, B37, B38, B39, Cw6, DR4, DR7 (Gladman and Farewell, 1995).

Patients with an earlier age of disease onset had a higher risk of susceptibility to damage (OR = exp(0.035) = 1.04; 95% CI: 1.02-1.05, $p < 0.001$). Among the HLA markers, individuals who are HLA B37 positive had a lower risk of susceptibility to damage in hand joints (OR = exp(−0.874) = 0.42; 95% CI: 0.18-0.97, $p = 0.022$), while those with HLA B39 positive appeared to be more susceptible to damage (OR = exp(1.399) = 4.05; 95% CI: 1.23-13.39, $p = 0.042$).

Plots of the cumulative probability of damage in the left (left panel (a)) and right hand joints (right panel (b)) are provided in Figure 2.2 based on models with piecewise constant hazards with two pieces arising from a single cutpoint at 11.47 years post-onset (the median of finite value of a_{1jk}), while empirically averaging over the latent susceptibility indicator and covariates. Overlaid on these plots are nonparametric Turnbull estimates of the marginal failure time distributions from the interval-censored times to damage, obtained by the `ic_np` function in the `R` package `icenReg` (version 2.0.5) (Turnbull, 1976; Anderson-Bergman, 2017) under a working indepedence assumption. The plots show that the marginal features of the fitted model are well calibrated to the raw data as represented by the nonparametric estimates.

Interest lies in modeling the spatial dependence in joint involvement among patients with PsA in order to assess the similarities or differences with the patterns of other arthritic conditions (Chandran et al., 2018). As mentioned in Section 2.1.2 we explore the strength of

Table 2.3: Empirical properties of two-stage composite likelihood estimates based on one thousand simulated samples of size $N = 2000$ with $J = 2$ and unspecified d-order dependencies ($d \geqslant 3$); two-piece piecewise constant proportional hazards models for $T_{jk}|Z_{jk} = 1$ without covariates and the logistic models for $Z_{jk}|\mathbf{X}_{jk}$ are fitted in stage I.

	$(K_1, K_2) = (2, 2)$										$(K_1, K_2) = (4, 4)$									
	$p_2 = 0.05$					$p_2 = 0.10$					$p_2 = 0.05$					$p_2 = 0.10$				
PARA	TRUE	BIAS	ESE	ASE	ECP	TRUE	BIAS	ESE	ASE	ECP	TRUE	BIAS	ESE	ASE	ECP	TRUE	BIAS	ESE	ASE	ECP
(a) Visit rate $\lambda = 10$																				
$\log \alpha_{11}$	0.640	−0.008	0.078	0.079	0.939	×	−0.011	0.079	0.078	0.932	0.640	−0.006	0.060	0.058	0.940	×	−0.004	0.058	0.058	0.938
$\log \alpha_{12}$	0.640	−0.014	0.197	0.200	0.930	×	−0.022	0.202	0.199	0.921	0.640	−0.008	0.146	0.144	0.938	×	−0.003	0.145	0.143	0.941
$\log \alpha_{21}$	0.790	−0.007	0.056	0.057	0.956	×	−0.006	0.053	0.056	0.961	0.790	−0.004	0.043	0.044	0.953	×	−0.004	0.043	0.044	0.947
$\log \alpha_{22}$	0.885	−0.013	0.170	0.168	0.937	×	−0.010	0.159	0.166	0.945	0.885	−0.011	0.117	0.121	0.957	×	−0.009	0.116	0.120	0.958
η_{10}	0.379	0.021	0.139	0.139	0.952	0.355	0.027	0.142	0.139	0.949	0.379	0.013	0.104	0.102	0.955	0.355	0.008	0.100	0.099	0.949
η_{20}	0.864	0.020	0.129	0.128	0.954	0.842	0.014	0.121	0.125	0.965	0.864	0.013	0.091	0.092	0.965	0.842	0.011	0.089	0.091	0.959
η_1	0.450	0.009	0.083	0.084	0.966	×	0.014	0.079	0.084	0.974	0.450	0.006	0.063	0.066	0.970	×	0.007	0.065	0.066	0.949
η_2	0.500	0.032	0.224	0.220	0.956	×	0.021	0.160	0.158	0.968	0.500	0.018	0.175	0.173	0.954	×	0.011	0.128	0.124	0.949
$\rho^{\dagger}_{11}$	0.251	−0.006	0.103	0.101	0.948	×	−0.006	0.099	0.101	0.950	0.251	−0.005	0.050	0.051	0.953	×	−0.004	0.050	0.051	0.951
$\rho^{\dagger}_{12}$	0.000	−0.006	0.055	0.052	0.936	×	−0.006	0.055	0.052	0.941	0.000	−0.005	0.037	0.035	0.923	×	−0.002	0.036	0.035	0.940
$\rho^{\dagger}_{22}$	0.386	0.001	0.097	0.097	0.943	×	−0.001	0.100	0.097	0.936	0.386	0.002	0.054	0.052	0.938	×	<0.001	0.054	0.051	0.935
γ_{11}	0.140	0.015	0.198	0.178	0.928	×	0.027	0.204	0.180	0.935	0.140	0.009	0.105	0.091	0.925	×	0.007	0.098	0.091	0.940
γ_{12}	0.049	0.006	0.112	0.107	0.938	×	0.009	0.114	0.107	0.938	0.049	0.002	0.070	0.067	0.946	×	−0.001	0.068	0.067	0.948
γ_{22}	0.095	−0.031	0.206	0.183	0.927	×	−0.038	0.208	0.182	0.938	0.095	−0.024	0.118	0.093	0.896	×	−0.017	0.109	0.092	0.916
(b) Visit rate $\lambda = 20$																				
$\log \alpha_{11}$	0.640	−0.006	0.065	0.067	0.948	×	−0.005	0.065	0.067	0.949	0.640	−0.003	0.052	0.051	94.4	×	−0.001	0.051	0.050	0.954
$\log \alpha_{12}$	0.640	−0.009	0.167	0.167	0.939	×	−0.005	0.162	0.167	0.949	0.640	−0.004	0.125	0.121	0.933	×	0.001	0.122	0.121	0.948
$\log \alpha_{21}$	0.790	−0.003	0.047	0.049	0.965	×	−0.003	0.046	0.049	0.971	0.790	−0.002	0.038	0.038	0.953	×	−0.001	0.037	0.038	0.950
$\log \alpha_{22}$	0.885	−0.001	0.134	0.140	0.958	×	−0.004	0.135	0.138	0.956	0.885	−0.002	0.097	0.101	0.958	×	<0.001	0.097	0.100	0.951
η_{10}	0.379	0.014	0.117	0.118	0.952	0.355	0.012	0.114	0.117	0.955	0.379	0.008	0.089	0.087	0.950	0.355	0.005	0.088	0.086	0.948
η_{20}	0.864	0.008	0.109	0.107	0.967	0.842	0.009	0.108	0.106	0.959	0.864	0.007	0.079	0.079	0.950	0.842	0.006	0.076	0.078	0.953
η_1	0.450	0.007	0.077	0.078	0.960	×	0.004	0.078	0.078	0.948	0.450	0.003	0.062	0.063	0.955	×	0.003	0.061	0.063	0.961
η_2	0.500	0.013	0.205	0.205	0.957	×	0.006	0.144	0.146	0.956	0.500	0.014	0.171	0.165	0.948	×	0.007	0.121	0.118	0.943
$\rho^{\dagger}_{11}$	0.251	−0.001	0.093	0.092	0.954	×	−0.001	0.091	0.092	0.955	0.251	−0.002	0.046	0.047	0.953	×	−0.003	0.047	0.047	0.937
$\rho^{\dagger}_{12}$	0.000	−0.005	0.049	0.048	0.949	×	−0.003	0.048	0.048	0.947	0.000	−0.002	0.034	0.033	0.936	×	−0.003	0.035	0.033	0.941
$\rho^{\dagger}_{22}$	0.386	−0.001	0.087	0.085	0.946	×	−0.001	0.084	0.085	0.959	0.386	0.001	0.048	0.046	0.946	×	0.001	0.047	0.046	0.939
γ_{11}	0.140	0.002	0.171	0.163	0.945	×	0.007	0.171	0.163	0.945	0.140	0.004	0.094	0.085	0.930	×	0.003	0.093	0.085	0.929
γ_{12}	0.049	0.005	0.100	0.099	0.956	×	−0.001	0.102	0.098	0.942	0.049	0.000	0.063	0.063	0.949	×	0.002	0.064	0.063	0.951
γ_{22}	0.095	−0.021	0.176	0.168	0.950	×	−0.028	0.187	0.168	0.925	0.095	−0.015	0.097	0.086	0.926	×	−0.008	0.095	0.086	0.931

Note. PARA, parameter; TRUE, true value of the parameter; $\rho^{\dagger}_{jj'} = \tan(.5\pi\rho_{jj'})$, $\gamma_{jj'} = \log \zeta_{jj'}$, $j \leqslant j'; j, j' \in \{1, 2\}$; ESE, sample standard deviation; ASE, model-based standard error; × means the value is the same as that specified in the first complete "TRUE" column under each setting of (K_1, K_2).

Table 2.4: Regression coefficient estimates from second-order dependence models in stage II for susceptibility and failure times are given joint susceptibility based on data from the University of Toronto Psoriatic Arthritis Clinic.

	Susceptibility			Failure Time		
	EST	ASE	p-VAL	EST	ASE	p-VAL
Intercept	2.637	0.182	<0.001	2.074	0.092	<0.001
I_{ray}	0.139	0.077	0.074	0.048	0.074	0.256
I_{row}	0.692	0.526	0.188	0.119	0.145	0.205
I_{sym}	1.143	0.341	<0.001	−0.071	0.134	0.298

the symmetric, ray, or row-type associations for susceptibility and the time to joint damage using second-order models for the latent susceptibility indicators and copula functions for the time to damage among susceptible joints. We adopt the two-stage estimation approach using a composite pairwise likelihood. For each pair of hand joints we define three binary indicators with $I_{\text{ray}} = 1$ if two joints are on the same digit, $I_{\text{row}} = 1$ if the two joints are at the same knuckle location of different digits, and I_{sym} indicating whether the two joints are in the same location of the opposite hands; note that we have $I_{\text{ray}} = 1$ if $I_{\text{sym}} = 1$. Table 2.4 displays estimates of the parameters in the second-order models given by $g_2^Z(\zeta_{jkj'k'}) = \kappa_0 + \boldsymbol{\kappa}_1'\mathbf{W}$ and $g_2^T(\tau_{jkj'k'}) = \iota_0 + \boldsymbol{\iota}_1'\mathbf{W}$ for susceptibility indicators and failure times, respectively, where $\mathbf{W} = (I_{\text{ray}}, I_{\text{row}}, I_{\text{sym}})'$, $g_2^Z(\cdot)$ and $g_2^T(\cdot)$ are one-to-one differentiable link functions mapping to the real line; here we consider $g_2^Z(x) = \log(x)$, and $g_2^T(x) = \tan(.5\pi \sin(.5\pi x))$. Left column of the table contains the results of using a logarithm link function for the odds ratio (OR) of susceptibility for a pair of hand joints. By incorporating the estimated intercept, it is apparent that any pair of hand joints appear to be naturally and significantly associated ($\log \text{OR} = 2.64$; 95% CI: $2.28 - 2.99$). While susceptibility features neither ray-type ($\log \text{OR} = 2.78$; 95% CI: $2.47 - 3.08$) nor row-type association ($\log \text{OR} = 3.33$; 95% CI: $2.48 - 4.18$) beyond this within-individual association, there is significant evidence for a stronger symmetric dependence in susceptibility with $\log \text{OR} = 4.47$ (95% CI: $3.99 - 4.95$). Table 2.4 also reports on the findings regarding the pairwise association in the failure times among susceptible joints, parameterized by a second-order model based on Kendall's τ. There is strong evidence for the existence of natural pairwise association in failure times ($\tau = 0.472$, 95% CI: $0.43 - 0.51$), but little suggestion of anything beyond an exchangeable dependence pattern for the failure times among susceptible joints.

To assess the adequacy of second-order models we plot the estimated concordance func-

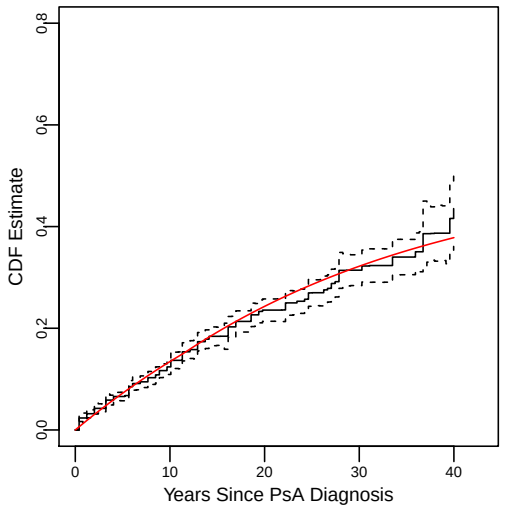
tions $P(\max(T_{jk}, T_{j'k'}) \leqslant t)$ and the estimated bivariate cumulative probabilities $P(T_{jk} \leqslant t_1, T_{j'k'} \leqslant t_2)$ based on the fitted models along with their nonparametric estimates. Plots for symmetric pairs of joints are displayed in Figure 2.2 and Figures A.2-A.4 for joints with other pairwise associations given in Appendix A.4. Specifically, the bivariate nonparametric estimates of $P(T_{jk} \leqslant t_1, T_{j'k'} \leqslant t_2)$ are obtained by using `ICNPMLE` function in the `R` package `icenReg` (version 1.2.8) (Anderson-Bergman, 2017). Again there is good agreement between the parametric and nonparametric estimates which suggests the model is fitting these aspects of the data reasonably well.

2.6 Discussion

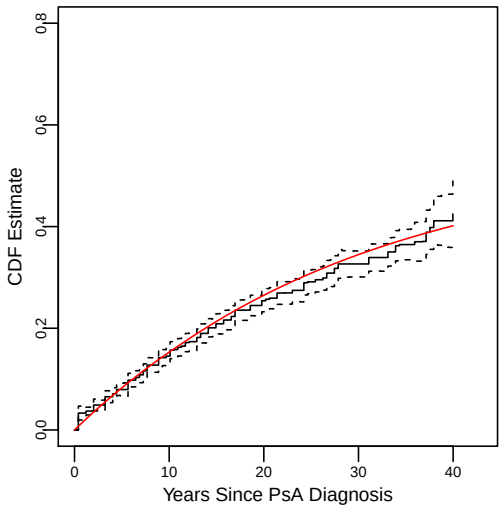
We propose a flexible and general framework for modeling multivariate interval-censored failure times which accommodates nonsusceptibility and flexible spatial dependence modelling. We therefore extend the approaches for modeling dependent interval-censored failure times with latent nonsusceptibility indicators, develop composite likelihood and two-stage estimation methods to facilitate analyses involving large cluster sizes, and provide software for implementation of these methods. A strength of this model is that it enables dependence modeling for susceptibility, and for failure times given joint susceptibility, based on parameters which are functionally independent of the parameters characterizing the marginal processes; marginal models for correlated binary data are used for the susceptibility indicators and a Gaussian copula is used for dependence modeling of the failure times among susceptible joints. These yield pairwise dependence models for susceptibility in terms of odds ratios for pairwise association in failure times among susceptible joints in terms of Kendall's τ. Application of these methods has lead to important scientific insights into the nature of the spatial dependence structure of joint damage in PsA.

We use a composite likelihood as our objective function, a weighted product of likelihood functions corresponding to lower dimensional events (Lindsay, 1988; Varin, 2008). Working independence and pairwise composite likelihoods are specified and a two-stage estimation procedure is adopted for computational efficiency and robustness in the sense of Varin et al. (2011). We implicitly set all weights equal so that they can be ignored but Varin et al. (2011) notes that unequal weights can be chosen to improve efficiency. We consider settings in which the dimension of the parameter vector of interest, q, is fixed and less than the sample size N. The results of the simulation studies show that the estimators have small finite sample bias and the empirical standard errors closely track the robust standard errors.

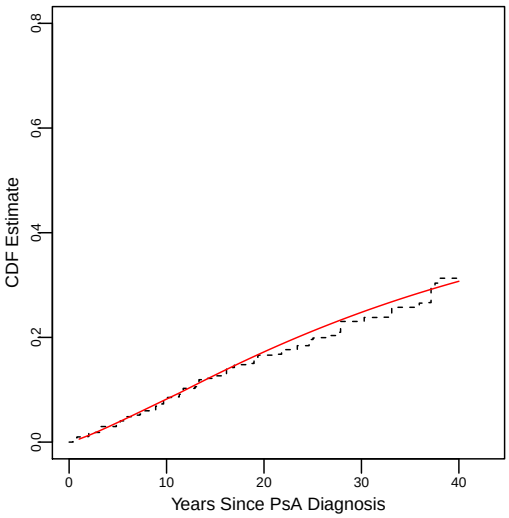
There was excellent agreement between the parametric and nonparametric estimates



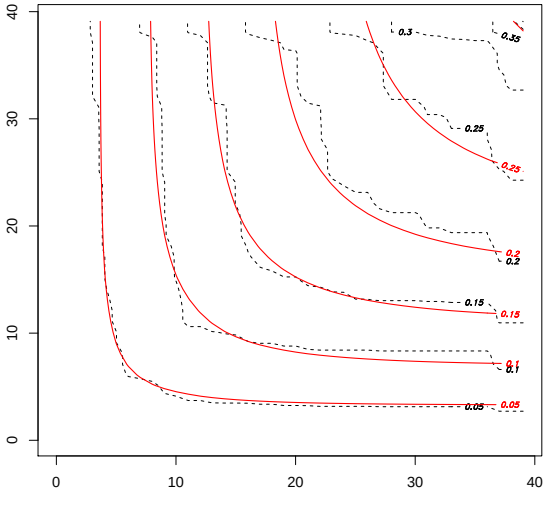
(a) Left thumb or finger joints



(b) Right thumb or finger joints



(c) Marginal CDF of $\max(T_{jk}, T_{j'k'})$



(d) Bivariate CDF of $(T_{jk}, T_{j'k'})$

Figure 2.2: Estimation of the marginal probability of damage in the left (a) and right (b) thumb or finger joints based on the fitted two-piece piecewise constant hazard (smooth line) and a nonparametric Turnbull estimate (nonsmooth line) along with 95% CI band (dashed line) using a resampling technique; and the estimated marginal concordance function (c) and bivariate cumulative probability (d) of damage in symmetric pairs based on the fitted second-order models for susceptibility and failure times of the susceptibles (solid line) and nonparametric estimates (dashed line).

of the marginal and bivariate failure time distributions suggesting that the model provides reasonable fit to the data and inferences can be drawn. The fitted model gave strong evidence of a symmetric dependence pattern in the susceptibility indicators for hand joints in PsA. These finding can provide insights which can help in the clinical diagnosis of PsA. The model accommodates regression on covariates for the marginal latent susceptibility indicators which can be used to investigate the relation between HLA markers and susceptibility for joint damage; if interest lies in assessing whether covariates might affect the hazard of damage among susceptible joints it would be straightforward to extend the model to incorporate this. In applications estimability challenges can arise with such expanded models when the same covariates appear in the susceptibility and failure time models (Li et al., 2001; Hanin and Huang, 2014) but we do not pursue that here. Finally we note that our dependence modeling was confined to the spatial association, but the framework we describe for the susceptibility indicators allows regression modeling of covariate effects on the dependence parameters.

For illustration, the proposed method was applied to data on hand joints in patients from the UTPAC to gain insights into the nature of the spatial dependence in joint involvement among PsA patients (Chandran et al., 2018); extending the model to incorporate data on different types of joints (e.g. foot joints or spinal joints) is relatively straightforward. The methods can also be applied to disease processes in which multiple organ systems can be affected. Examples of these include systemic lupus erythematosus, another autoimmune disease, in which patients experience disease activity in the heart, lung, kidney, central nervous system, and many other locations. Likewise in diabetes individuals may experience vascular disease, nephropathy and retinopathy. Characterizing which kinds of patients are at risk of the different types of complications, and which complications tend to occur together can help optimize patient care.

Chapter 3

Two-phase designs with current status data

3.1 Introduction

Current status data arise when interest lies in a time to an event but when the failure status of each individual in a sample is determined at an assessment time resulting in either right- or left-censored observations (Groeneboom and Wellner, 1992; Sun, 2006). Such data are often encountered in demography, epidemiology and biomedical studies. We consider a cohort study where interest lies in modeling the relationship between a biomarker and a failure time under current status observation. In many large cohort studies blood or tissue samples are collected and stored in a biobank upon recruitment, but it is often too costly to assay all biospecimens to measure the biomarker in all individuals. The phase I data is comprised of the current status observations and inexpensive covariates that are readily available. The challenge in two-phase designs is to develop statistically efficient and cost-effective selection strategies to create a phase II subsample of individuals in which the biomarkers are complete. For excellent surveys of the literature on outcome-dependent sampling see Lawless et al. (1999) and Ding et al. (2017). Much work on two-phase designs has been carried out for the failure time setting accommodating right-censoring (Prentice, 1986; Chen and Lo, 1999; Breslow and Wellner, 2007; Lawless, 2018; Tao et al., 2020). Relatively little work has been carried out for two-phase designs involving interval-censored data, but recent developments include Li et al. (2008), Zhou et al. (2017) and Zhou et al. (2020). For the more extreme form of current status data, the relevant literature has mostly been confined to the development of analytical methods for dealing with incomplete covariates (Li and Nan, 2011; Wen and Lin, 2011; Wen and

Chen, 2013). In this paper we consider design issues of two-phase studies involving current status observation of a failure time, in the phase I sample. A key feature of current status data is the time of assessment, which is neither a response nor an auxiliary covariate – it provides key information which aids in the interpretation of the status indicator. That is, individuals who are assessed very early and are found to be event-free, and individuals who are assessed very late and are found to have failed, do not convey much information about the covariate effect of interest. We explore two-phase design problems with a view to investigating this intuition under proportional hazards modeling. Specifically we develop efficient phase II sub-sampling strategies based on current status data in the likelihood and estimating function frameworks. The former yields more efficient estimators but relies on modeling assumptions regarding the nuisance biomarker distribution given auxiliary covariates. Correct specification of this model is challenging when the auxiliary covariates include continuous variables or are high dimensional. Inverse probability weighting does not require modeling the biomarker distribution and so is more robust, but it yields less efficient estimators.

The remainder of this chapter is organized as follows. We introduce notation and model assumptions for the two analysis frameworks in Section 3.2. In Section 3.3 we present the information for the maximum likelihood estimator of interest exploiting a framework developed by Tao et al. (2020) for the current status setting; simulation studies are also reported on which investigate various two-phase designs using maximum likelihood. Section 3.4 gives the optimal sampling rule under inverse probability weighted estimating equations via Neyman allocation and related influential functions in the sense of Breslow and Wellner (2007) and Chen and Lumley (2020). In Section 3.5 additional simulation studies are reported to demonstrate the robustness of using piecewise baseline hazards to approximate continuous distributions. Section 3.6 presents two illustrative applications of the proposed two-phase designs and estimators and concluding remarks and topics for further research are mentioned in Section 3.7.

3.2 Analysis methods

3.2.1 Notation and model assumptions

Let T denote a random failure time of interest corresponding to, say, age of disease onset, and let $\mathbf{X}$ denote a $p \times 1$ covariate vector. Under a current status observation scheme the failure status of each individual is assessed at an individual-specific assessment time A giving data (Y, A) where $Y = I(T \leqslant A)$ and $I(\cdot)$ is the indicator function. Suppose that

a sample of N independent individuals provides observations $\{(Y_i, A_i, \mathbf{X}_i), i = 1, \ldots, N\}$. We consider the partition $\mathbf{X} = (X_1, \mathbf{X}_2')'$ where $\mathbf{X}_2$ is a $(p-1) \times 1$ vector of covariates which are inexpensive to observe and complete, and X_1 records the value of a biomarker of interest which is costly to measure. Biospecimens are stored in a biobank so measurements of X_1 can be made for a subsample of individuals. We let $\mathbf{Z}_i = (Y_i, A_i, \mathbf{X}_{i2}')'$ and let $\{\mathbf{Z}_i, i = 1, \ldots, N\}$ denote the data in the phase I sample. Let $R = I(X_1 \text{ is observed})$ and $\mathbf{X}^\circ$ denote the observed vector so that $\mathbf{X}^\circ = \mathbf{X}$ if $R = 1$ and $\mathbf{X}^\circ = \mathbf{X}_2$ if $R = 0$, respectively. In the design of two-phase studies, whether X_1 is observed or not is governed by a selection model indexed by ρ and given by

$$P(R = 1|Y, A, \mathbf{X}) = P(R = 1|\mathbf{Z}) = \pi(\mathbf{Z}; \rho), \tag{3.2.1}$$

where $X_1 \perp R|\mathbf{Z}$; X_1 is therefore missing at random (MAR) (Little and Rubin, 2002). We also assume $T \perp A|\mathbf{X}_2$.

Primary interest resides in evaluating the effect of an expensive covariate X_1 on the failure process through the proportional hazards (PH) model

$$h(t|\mathbf{X}; \boldsymbol{\theta}) = h(t; \boldsymbol{\alpha}) \exp(\beta_1 X_1 + \boldsymbol{\beta}_2' \mathbf{X}_2), \tag{3.2.2}$$

where $h(\cdot; \boldsymbol{\alpha})$ is the baseline hazard function indexed by $\boldsymbol{\alpha}$, $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2')'$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$. We also let $H(t|\mathbf{X}; \boldsymbol{\theta}) = \int_0^t h(s|\mathbf{X}; \boldsymbol{\theta}) ds$ and $\mathcal{F}(t|\mathbf{X}; \boldsymbol{\theta}) = 1 - F(t|\mathbf{X}; \boldsymbol{\theta}) = \exp(-H(t|\mathbf{X}; \boldsymbol{\theta}))$. We adopt a flexible weakly parametric piecewise constant (PWC) form for the baseline hazard, which approximates a reasonably wide range of distributional shapes (Friedman, 1982) and has proven popular for many applications (Grenfell and Anderson, 1985; Diamond et al., 1986; Cook et al., 2008; Cook and Tolusso, 2009). If $0 = b_0 < b_1 < \cdots < b_{K-1} < b_K = \infty$ denote a set of prespecified cut points, we set $h(t; \boldsymbol{\alpha}) = \sum_{k=1}^K \exp(\alpha_k) I(t \in [b_{k-1}, b_k))$ to approximate the shape of the baseline hazard function $h(\cdot; \boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)'$. Guidance on specification of the cut-points for the PWC hazard model for current status data is given by Zhan (1999).

In sub-Sections 3.2.2 and 3.2.3 we describe estimation and inference for β_1 based on maximum likelihood (Lawless et al., 1999) and inverse probability weighted estimating functions (Robins et al., 1994) respectively.

3.2.2 Maximum likelihood

Under the condition that the joint distribution of $(A, \mathbf{X}_2')'$ is non-informative for $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\eta}')'$, the observed likelihood $L(\boldsymbol{\vartheta})$ based on $\{Y_i, A_i, \mathbf{X}_i^\circ, R_i, i = 1, 2, \ldots, N\}$ propor-

tional to

$$\prod_{i=1}^{N}\Big\{\int\left[1-\mathcal{F}(A_i|X_1,\mathbf{X}_{i2};\theta)\right]^{Y_i}\mathcal{F}(A_i|X_1,\mathbf{X}_{i2};\theta)^{1-Y_i}dG(X_1|\mathbf{X}_{i2};\eta)\Big\}^{1-R_i} \times\Big\{\left[1-\mathcal{F}(A_i|\mathbf{X}_i;\theta)\right]^{Y_i}\mathcal{F}(A_i|\mathbf{X}_i;\theta)^{1-Y_i}dG(\mathbf{X}_{i1}|X_{i2};\boldsymbol{\eta})\Big\}^{R_i}, \tag{3.2.3}$$

where $G(X_1|\mathbf{X}_2;\boldsymbol{\eta})$ is the conditional distribution function for $X_1|\mathbf{X}_2$. The observed data score vector $S(\boldsymbol{\vartheta})=\partial\log L(\boldsymbol{\vartheta})/\partial\boldsymbol{\theta}$ can be written as

$$S(\boldsymbol{\vartheta})=\sum_{i=1}^{N}\left\{R_i\begin{pmatrix}\mathscr{S}_1(Y_i|A_i,\mathbf{X}_i;\boldsymbol{\theta})\\ \mathscr{S}_2(X_{i1}|\mathbf{X}_{i2};\boldsymbol{\eta})\end{pmatrix}+(1-R_i)\begin{pmatrix}E\{\mathscr{S}_1(Y_i|A_i,\mathbf{X}_i;\boldsymbol{\theta})|\mathbf{Z}_i;\boldsymbol{\vartheta}\}\\ E\{\mathscr{S}_2(X_{i1}|\mathbf{X}_{i2};\eta)|\mathbf{Z}_i;\boldsymbol{\vartheta}\}\end{pmatrix}\right\},$$

where $\mathscr{S}_1(Y_i|A_i,\mathbf{X}_i;\boldsymbol{\theta})=\partial\log P(Y_i|A_i,\mathbf{X}_i;\boldsymbol{\theta})/\partial\theta=Q(\mathbf{Z}_i,X_{i1};\boldsymbol{\vartheta})\partial\log\mathcal{F}(A_i|\mathbf{X}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ with $Q(\mathbf{Z}_i,X_{i1};\boldsymbol{\vartheta})=-F^{-1}(A_i|\mathbf{X}_i;\boldsymbol{\theta})\left[Y_i-F(A_i|\mathbf{X}_i;\boldsymbol{\theta})\right]$ and $\mathscr{S}_2(X_{i1}|\mathbf{X}_{i2};\boldsymbol{\eta})=\partial\log G(X_{i1}|\mathbf{X}_{i2};\boldsymbol{\eta})/\partial\boldsymbol{\eta}$. The solution to $S(\boldsymbol{\vartheta})=\mathbf{0}$ is the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}=(\hat{\boldsymbol{\theta}}',\hat{\boldsymbol{\eta}}')'$ and under suitable regularity conditions (Boos and Stefanski, 2013), $\sqrt{N}(\hat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta})\xrightarrow{d}\text{MVN}(0,\mathcal{I}(\boldsymbol{\Omega};\boldsymbol{\rho}))$, where $\mathcal{I}(\boldsymbol{\Omega};\boldsymbol{\rho})=E_{\text{R}\mathbf{Z}\text{X}_1}\{-\partial S(\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}'\}/N=E_{\text{R}\mathbf{Z}\text{X}_1}\{S(\boldsymbol{\vartheta})S'(\boldsymbol{\vartheta})\}/N$ is the expected information and $\boldsymbol{\Omega}$ denotes a parameter vector including $\boldsymbol{\vartheta}$ and other nuisance parameters indexing the distribution of $(A,\mathbf{X}_2')'$. In the analysis of data the observed information matrix $I(\boldsymbol{\vartheta})=-\partial S(\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}'$ evaluated at $\hat{\boldsymbol{\vartheta}}$, is used for inference, however the functional dependence of $\mathcal{I}(\boldsymbol{\Omega};\boldsymbol{\rho})$ on $\boldsymbol{\rho}$ is the basis for the optimal design described in later sections.

3.2.3 Inverse probability weighted estimating functions

Restricting attention to individuals with complete data is appealing when the nuisance covariate distribution $G(X_1|\mathbf{X}_2;\boldsymbol{\eta})$ is challenging to specify, provided inverse probability weights are used to address the possibility that those with complete data represent a biased sub-sample (Robins et al., 1994).

In the present setting, provided the phase II selection probability $\pi(\mathbf{Z};\boldsymbol{\rho})$ in (3.2.1) is bounded away from zero, the solution to the inverse probability weighted (IPW) estimating equation

$$U_1(\boldsymbol{\theta};\boldsymbol{\rho})=\sum_{i=1}^{N}\frac{R_i}{\pi(\mathbf{Z}_i;\boldsymbol{\rho})}\mathscr{S}_1(Y_i|A_i,\mathbf{X}_i;\boldsymbol{\theta})=\mathbf{0} \tag{3.2.4}$$

is consistent and asymptotically normally distributed. Joint estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ is known to enhance efficiency of estimation for $\boldsymbol{\theta}$ (Robins et al., 1994). A consistent estimator of $\boldsymbol{\rho}$ is obtained by solving the score equation

$$U_2(\boldsymbol{\rho})=\sum_{i=1}^{N}\{R_i-\pi(\mathbf{Z}_i;\boldsymbol{\rho})\}/\{\pi(\mathbf{Z}_i;\boldsymbol{\rho})(1-\pi(\mathbf{Z}_i;\boldsymbol{\rho}))\}\left[\partial\pi(\mathbf{Z}_i;\boldsymbol{\rho})/\partial\boldsymbol{\rho}\right]=\mathbf{0}.$$

Letting $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\rho}')'$, we write the full set of estimating functions as $U(\boldsymbol{\phi}) = (U_1'(\boldsymbol{\theta}; \boldsymbol{\rho}), U_2'(\boldsymbol{\rho}))'$ and denote the solution to $U(\boldsymbol{\phi}) = \mathbf{0}$ as $\tilde{\boldsymbol{\phi}} = (\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\rho}}')'$ where we use "$\sim$" to distinguish these estimators from maximum likelihood estimators of Section 3.2.2. Under suitable regularity conditions (Robins et al., 1995), $\sqrt{N}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{d} \text{MVN}(0, \Sigma(\boldsymbol{\Omega}, \boldsymbol{\rho}))$, where $\Sigma(\boldsymbol{\Omega}, \boldsymbol{\rho}) = \mathcal{A}^{-1}(\boldsymbol{\Omega}, \boldsymbol{\rho})\mathcal{B}(\boldsymbol{\Omega}, \boldsymbol{\rho})\mathcal{A}^{-1}(\boldsymbol{\Omega}, \boldsymbol{\rho})$ is of robust sandwich form with $\mathcal{A}(\boldsymbol{\Omega}, \boldsymbol{\rho}) = E_{\text{R}\mathbf{Z}\text{X}_1}\{-\partial U(\boldsymbol{\phi})/\partial \boldsymbol{\phi}'\}/N$ and $\mathcal{B}(\boldsymbol{\Omega}, \boldsymbol{\rho}) = E_{\text{R}\mathbf{Z}\text{X}_1}\{U(\boldsymbol{\phi})U'(\boldsymbol{\phi})\}/N$. We consider phase II sampling schemes based on this framework for analysis in Section 3.4.

3.3 Design and extreme current status data

In this section, we discuss two-phase designs for estimators of β_1 via maximum likelihood. Following the work of Derkach (2014) and Tao et al. (2020), we first present the information for $\hat{\beta}_1$ and then optimal designs under the setting where $\beta_1 = o(1)$. This leads to a theoretical finding supporting the notion that subjects with "extreme current status responses" should be preferentially sampled in phase II as they are more "informative" regarding β_1.

3.3.1 Phase II selection based on extreme current status data

Following Tao et al. (2020) we let $\mu = \beta_1 X_1 + \boldsymbol{\beta}_2' \mathbf{X}_2$ denote the linear predictor and note that observed score function $S_{\beta_1} = \partial \log L(\boldsymbol{\vartheta})/\partial \beta_1$ for β_1 the parameter of prime interest, can be written as

$$S_{\beta_1} = RM_\mu X_1 + (1-R)\frac{\int M_\mu X_1 (1-\mathcal{F}(A|\mathbf{X}))^Y \mathcal{F}(A|\mathbf{X})^{1-Y} dG(X_1|\mathbf{X}_2)}{\int (1-\mathcal{F}(A|\mathbf{X}))^Y \mathcal{F}(A|\mathbf{X})^{1-Y} dG(X_1|\mathbf{X}_2)},$$

where $M_\mu = \partial \log P(Y|A, \mathbf{X})/\partial \mu$ may be viewed as a score-type residual with $E(M_\mu) = 0$; see Appendix B.1 for details. If $\boldsymbol{\theta}_\circ = (\boldsymbol{\beta}_2', \boldsymbol{\alpha}')'$, $\boldsymbol{\theta} = (\beta_1, \boldsymbol{\theta}_\circ')'$ and

$$\mathcal{I}_{\beta_1\beta_1} = -E\left(\frac{\partial^2 \log L(\boldsymbol{\vartheta})}{\partial \beta_1^2}\right), \ \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ'} = -E\left(\frac{\partial^2 \log L(\boldsymbol{\vartheta})}{\partial \beta_1 \partial \boldsymbol{\theta}_\circ'}\right), \ \mathcal{I}_{\beta_1\boldsymbol{\eta}'} = -E\left(\frac{\partial^2 \log L(\boldsymbol{\vartheta})}{\partial \beta_1 \partial \boldsymbol{\eta}'}\right),$$
$$\mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ} = -E\left(\frac{\partial^2 \log L}{\partial \boldsymbol{\theta}_\circ \partial \boldsymbol{\theta}_\circ'}\right), \ \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\eta}} = -E\left(\frac{\partial^2 \log L(\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\eta}'}\right), \text{ and } \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} = -E\left(\frac{\partial^2 \log L(\boldsymbol{\vartheta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}\right).$$

Moreover $N^{-1/2}S_{\beta_1}$ and $N^{-1/2}(\hat{\beta}_1 - \beta_1)$ converge in distribution to normal random variables with zero means and variances $V_{\beta_1} = \mathcal{I}_{\beta_1\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}^{-1}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} - \mathcal{I}_{\beta_1\eta}\mathcal{I}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\beta_1}$ and $V_{\beta_1}^{-1}$ respectively.

In general to compute V_{β_1} expectations are required with respect to all random variables and hence it is not analytically tractable (Bickel et al., 1993; Robins et al., 1995). To

address this, Tao et al. (2020) showed that V_{β_1} can be decomposed as

$$E\left\{R\text{Var}[M_\mu|R=1,A,X_2]\text{Var}(X_1|\mathbf{X}_2)\right\} \tag{3.3.1}$$

and a non-negative component independent of the phase II sub-sampling rules under $\beta_1 = o(1)$; see Appendix B.1 for a sketch of the derivation.

Tao et al. (2020) considered right-censored failure time setting in which case hence M_μ is a martingale-type residual. Following their argument, an asymptotic optimal (or approximate optimal when β_1 differs from zero) sampling rule for our current status setting involves selecting subjects with the largest and smallest values of $M_\mu \text{Var}(X_1|\mathbf{X}_2)^{1/2}$, subject to the constraint

$$P(R=1) = E[P(R=1|\mathbf{Z};\boldsymbol{\rho})] = \gamma, \tag{3.3.2}$$

where γ $(0 < \gamma \leqslant 1)$ is the fixed marginal probability of selecting an individual in the phase II sample.

For right-censored data under the PH models, the computation of martingale residuals is easy using the `coxph` function (Therneau, 2020) in `R` (R Core Team, 2020), but for current-status data, further simplifications are warranted. We first note that for specified Y, M_μ is a monotone function of the inspection time A. When $Y = 0$, $Q(\mathbf{Z}, X_1) = 1$ so $M_\mu = \log \mathcal{F}(A|\mathbf{X})$ is a nonpositive decreasing function in A. When $Y = 1$, $Q(\mathbf{Z}, X_1) = -\mathcal{F}(A|\mathbf{X})/F(A|\mathbf{X}) \leqslant 0$ and $\partial M_\mu/\partial A = h(A|\mathbf{X};\boldsymbol{\theta})Q(\mathbf{Z}, X_1)\left[1 + \log \mathcal{F}(A|\mathbf{X})/F(A|\mathbf{X})\right] \leqslant 0$ so M_μ is again a nonnegative decreasing function in A. Hence, extreme values of M_μ are associated with extreme current status responses corresponding to $Y = 1$ with a small assessment time A (i.e., large positive residuals) and $Y = 0$ with a large assessment time A (i.e., small positive residuals).

The mathematical relation between the design efficiency and extreme current status responses offers a new perspective. In the recent literature extreme outcome sampling schemes have been reported to have good properties in time-to-event settings and genetic studies (Lawless, 2018; Zhou et al., 2020; Espin-Garcia et al., 2017). The essence behind such sampling schemes can be easily understood and justified in simple linear regression but it is less obvious with hazard-based models and censored time-to-event data.

3.3.2 Residual and extreme response dependent sampling

The budgetary constraint (3.3.2) is equivalent to imposing a constraint on the size of the phase II sub-sample which we denote by n with $n \leqslant N$. Here we describe some two-phase sampling schemes based on the derivation in Section 3.3.1. We refer to the asymptotically

optimal designs aiming to minimize (3.3.1) when $\beta_1 = o(1)$ as TAO-OPT, since it exploits the framework of Tao et al. (2020).

TAO-OPT selects m_1 individuals with the highest and $m_0 = n - m_1$ with the lowest values of $M_\mu \text{Var}(X_1|\mathbf{X}_2)^{1/2}$, where m_0 and $m_1 = n - m_0$ are determined by maximizing (3.3.1). In particular, the score-type residuals M_μ are estimated based on fitting a PH model to phase I data conditioning on X_2 alone and with a PWC hazard function. Full implementation of TAO-OPT requires computing $\text{Var}(X_1|\mathbf{X}_2)$, which is not possible with phase I data alone. To address this we propose several practical designs to approximate TAO-OPT, including extreme residual dependent sampling (EXT-M_μ), extreme response dependent sampling (EXT-(A, Y)) and a two-stage adaptive design TAO-OPTA. We describe these in what follows.

Extreme Residual Sampling (EXT-M_μ), under the assumption $X_1 \perp \mathbf{X}_2$ (i.e. treating $\text{Var}(X_1|\mathbf{X}_2)$ as constant), selects m_1 individuals with the highest and $m_0 = n - m_1$ with the lowest values of M_μ, where m_0 and $m_1 = n - m_0$ are determined to maximize $E\{R\text{Var}(M_\mu|R = 1, \mathbf{X}_2)\}$. Again, the residuals M_μ are estimated using phase I data only based on a model only conditioning on $\mathbf{X}_2$.

Extreme Response Dependent Sampling (EXT-(A, Y)) selects $\min(n/2, Y_\cdot)$ individuals who had experienced events (i.e. $Y_i = 1$) with the smallest assessment times A and $n - \min(n/2, Y_\cdot)$ subjects who are censored (i.e. $Y_i = 0$) with the largest A, where $Y_\cdot = \sum_{i=1}^{N} Y_i$.

Adaptive Extreme Residual Sampling (TAO-OPTA) includes a two-stage adaptive procedure in which a preliminary simple random sample of size n_a ($< n$) is chosen to acquire information on X_1 and estimate $\boldsymbol{\eta}$ in the covariate model $G(X_1|\mathbf{X}_2; \boldsymbol{\eta})$ and $\text{Var}(X_1|\mathbf{X}_2)$. We then select $n_b = n - n_a$ individuals to maximize (3.3.1). Larger values of n_a result in higher precision in estimating $\text{Var}(X_1|\mathbf{X}_2)$ but less efficiency gains in the overall design.

3.3.3 Simulation studies

Here we report on simulation studies conducted to investigate the three extreme response dependent sampling schemes described in Section 3.3.2 under maximum likelihood. We compare their performance to stratified sampling schemes including balanced, extreme, optimal and adaptive optimal stratified sampling schemes. We first describe how these are implemented.

3.3.3.1 Two-phase stratified designs

Phase I samples are often stratified to provide a framework for conventional two-phase designs. Here we define strata based on phase I data $\{A_i, Y_i, i = 1, \ldots, N\}$ or $\{\mathbf{Z}_i' =$

$(A_i, Y_i, \mathbf{X}'_{i2})', i = 1, \ldots, N\}$ to implement phase II stratified sub-sampling schemes. In particular we consider discretizing the observed assessment time A into three strata based on the empirical percentiles $(c_1, c_2)'$ $(c_1 < c_2)$. This renders a new categorical variable $\bar{A}$ with a sample space $\{1, 2, 3\}$ and a total number of strata as $J = 6$ or 12 (when a binary $\mathbf{X}_2$ is used). Let $\bar{Z}$ denote a categorical variable indicating the resultant strata and if we let J represent the total number of strata, then $\bar{Z} \in \{1, 2, \ldots, J\}$. Following this stratification the selection model (3.2.1) used for the phase II sub-sampling scheme can be defined by

$$\pi(\mathbf{Z}; \boldsymbol{\rho}) = \sum_{j=1}^{J} \rho_j I(\bar{Z} = j), \tag{3.3.3}$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_J)'$ is a $J \times 1$ vector of the stratum-specific selection probabilities.

Balanced Stratified Sampling (BAL) samples subjects evenly across the pre-specified phase I strata with selection probabilities $\rho_j = n_j/N_j$, where $n_j = \min(n/J + m_j, N_j)$ and m_j is set by iterative balanced sampling among the rest of non-exhausted strata until $\sum_j n_j = n$ is satisfied.

Extreme Stratified Sampling (EXT-$(\bar{A}, Y)$) oversamples subjects from strata labeling $(\bar{A}, Y) = (1, 1)$ and $(\bar{A}, Y) = (3, 0)$, and to avoid a near-zero selection probability in some strata, we set the lower bound of ρ_j to be 0.05 for each j; for example, $n_j = \min(0.95n/4 + m_j, N_j)$ when $J = 12$, or $\min(0.95n/2 + m_j, N_j)$ when $J = 6$, if stratum j is an "extreme" stratum; and $\min(m_j + 0.05N_j, N_j)$, otherwise.

Optimal Stratified Sampling (OPT) schemes are derived by minimizing the asymptotic variance of β_1 estimators based on the selection model (3.3.3). Though infeasible in practice, they can be conducted as a benchmark in simulation studies.

Adaptive Optimal Stratified Sampling (OPTA) proposed by McIsaac and Cook (2015) splits the phase II selection into phase IIa and IIb. In phase IIa a sample of size $n_a (< n)$ is selected to measure X_1 and related parameters are then estimated to guide with the design components for an optimal selection in phase IIb. Note that though OPTA does not require knowledge of the true parameters it relies on correct specification of the form of all distributions. There is a trade-off between the precise estimation in phase IIa and the approximated optimal selection in phase IIb. The larger the phase IIa sample, the more precise the estimation; however the final selection may be more distant away from the optimal design. We set $n_a/n = 0.5$ and carry out balanced phase IIa sub-sampling throughout the simulation studies unless otherwise specified.

3.3.3.2 Data generation

For each individual we generate X_2 from a Bernoulli distribution with $P_2 = P(X_2 = 1)$ and simulate $X_1|X_2$ from $G(x_1|x_2)$ where $P(X_1 = 1|X_2; \boldsymbol{\eta}) = \exp(\eta_0 + \eta_1 X_2)/\{1 + \exp(\eta_0 + \eta_1 X_2)\}$ and $\boldsymbol{\eta} = (\eta_0, \eta_1)'$ is set by specifying the marginal probabilities $P_1 = P(X_1 = 1)$, $P_2 = P(X_2 = 1)$ and the odds ratio $\varphi = P(X_1 = 1, X_2 = 1)P(X_1 = 0, X_2 = 0)/[P(X_1 = 1, X_2 = 0)P(X_1 = 0, X_2 = 1)]$. Given $\mathbf{X} = (X_1, X_2)'$, T is taken to be Weibull distributed with cumulative hazard function $H(t|X; \boldsymbol{\theta}) = (\lambda t)^\kappa \exp(\beta_1 X_1 + \beta_2 X_2)$, where $\boldsymbol{\theta} = (\lambda, \kappa, \beta_1, \beta_2)'$. We let $\tau = 1$ denote the maximum inspection time (or the study end) and generate the assessment time as $A = \min(A^\dagger, \tau)$ where $A^\dagger \sim \text{GAM}(\mu, \psi)$ with $\mu = E[A^\dagger]$ and $\psi = \text{Var}(A^\dagger)$. We set the probability of failure by τ as $q_1 = P(T \leqslant \tau)$ and the observed event rate $q_2 = P(Y = 1)$. Without loss of generality, we set $\tau = 1$.

Let $(c_1, c_2)'$ denote the empirical tertiles of $\{A_i, i = 1, \ldots, N\}$. We consider the setting where there is a high probability of failure by the maximum inspection time by setting $q_1 = 0.6$ and set the probability of a positive failure status to $q_2 = 0.3$; we then stratify the phase I data based on $(Y, \bar{A}, X_2)'$ yielding twelve strata. We also consider the setting in which the probability of failure by the maximum inspection time is low with $q_1 = 0.1$, and specify the assessment time distribution here so that $q_2 = 0.05$; in order to avoid near-empty strata in this setting we stratify simply on the current status responses $(\bar{A}, Y)'$. We consider settings where $(\boldsymbol{\beta}', \varphi, \psi)'$ take on different values: for the scenarios where $X_1 \perp X_2$ we set $\varphi = 1$ and when $X_1 \not\perp X_2$ we set $\varphi = 2$. Moreover, we consider $\beta_1 = -0.2, 0$ or 0.2, $\beta_2 = -0.2, 0$ or 0.2, and $\psi = 0.1, 0.2$, or 0.3. The phase I sample size is given by $N = 2000$ and we consider phase II sub-sample sizes with $n = 300$ or 600.

3.3.3.3 Simulation results

To obtain the maximum likelihood estimators of $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\eta}')'$ described in sub-Section 3.2.2, the general-purpose optimizer `optim` in `R` (R Core Team, 2020) is used with the "limited memory Broyden-Fletcher-Goldfarb-Shanno" (L-BFGS-B) algorithm. Table 3.1 summarizes the simulation results for the estimated log hazard ratio $\hat{\beta}_1$. The empirical bias (BIAS) of all estimates are relatively small and the corresponding coverage probabilities are all close to the nominal 95% level, a reflection of the good agreement between the empirical (ESE) and analytical standard errors (ASE). Among the stratified sampling schemes, OPT resulted in the smallest standard error for β_1 estimators as expected, and OPTA yielded a better approximation to the OPT design than the alternative stratified sampling designs (i.e. SRS, BAL and EXT-$(\bar{A}, Y)$). The derivation of the OPT and OPTA designs, however, are based on correct specification of the parametric models for $(A, X_2)'$

and the residual or response-dependent sampling schemes do not have this requirement. TAO-OPT is the most efficient design among all designs, though only feasible in simulation studies. As its practical substitute, EXT-M_μ and EXT-(A, Y) provide better approximations to TAO-OPT than TAO-OPTA with $n_a = 0.5n$. Table B.1 in Appendix B.3 presents different choices for the proportions of the phase IIa samples n_a/n for TAO-OPTA and their comparison to EXT-M_μ and EXT-(A, Y). TAO-OPTA performs better when n_a is smaller, but a sufficient phase IIa sample is necessary to make reliable estimates for $\text{Var}(X_1|X_2)$. When $\text{Var}(X_1|X_2)$ is constant (i.e. $\varphi = 1$), EXT-(A, Y), EXT-M_μ and TAO-OPT appear to be equally efficient or the difference is trivial; while $\text{Var}(X_1|X_2)$ is not constant (as $\varphi = 2$) the efficiency gains of TAO-OPT comparing to EXT-(A, Y) and EXT-M_μ is more substantial however the magnitude seems to be negligible especially when the event rate is low. These results are consistent with the theoretical derivations in Section 3.3.1.

3.4 Influence functions and Neyman allocation

In this section we provide optimal sampling rules for the inverse probability weighted estimator discussed in Section 3.2.3. A closed-form expression of the optimal design (OPT) for IPW estimators is possible under stratification of phase I data and selection model (3.3.3) (Reilly and Pepe, 1995; McIsaac and Cook, 2015; Chen and Lumley, 2020).

Subject to a constraint on phase II sample size n ($\leqslant N$) the optimal design for IPW estimators follows Neyman allocation (Neyman, 1938) to minimize the sampling variance of the estimator of interest, expressable as a summation of i.i.d. random variables (Breslow et al., 2009; Chen and Lumley, 2020). Specifically the IPW estimator $\tilde{\beta}_1$ can be written as

$$\sqrt{N}(\tilde{\beta}_1 - \beta_1) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\pi(\mathbf{Z}_i; \boldsymbol{\rho})} \Delta_i(\beta_1) + o_p(1),$$

where $\Delta_i(\beta_1)$ is an influence measure of the observation from subject i to the estimation in β_1 (Sasieni, 1993). Chen and Lumley (2020) note that the optimal allocation is

$$n_j \propto n N_j \sigma_j \tag{3.4.1}$$

with $\sum_{j=1}^{N} n_j = n$ and $\sigma_j = \text{Var}(\Delta(\beta_1)|\bar{Z} = j)^{1/2}$.

The influence function $\Delta_i(\beta_1)$ can be approximated by a case influence `dfbeta` statistic (Therneau and Grambsch, 2000; Klein and Moeschberger, 2003) $\tilde{\beta}_1 - \tilde{\beta}_{1(i)}$, where $\tilde{\beta}_{1(i)}$ represents the IPW estimator of β_1 when observation i is removed. For computational convenience we use its first-order expansion proposed in Cain and Lange (1984) to approximate the case influence $\tilde{\beta}_1 - \tilde{\beta}_{1(i)}$. We first rewrite the IPW estimator of $\boldsymbol{\theta}$ as a function of

a weight w_i so the weighted score equation in (3.2.4) becomes $U_1(\boldsymbol{\theta}) = U_1(\tilde{\boldsymbol{\theta}}(w_i), w_i) = 0$ such that $\tilde{\boldsymbol{\theta}}(1) = \tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}(0) = \tilde{\boldsymbol{\theta}}_{(i)}$. Using the first-order Taylor series expansion of $\tilde{\boldsymbol{\theta}}(w_i)$ at $w_i = 1$ and taking derivatives on the both sides of the weighted score equation, we have $\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_{(i)} \approx \left[-\partial U_1/\partial \tilde{\boldsymbol{\theta}}\right]^{-1} \partial U_1/\partial w_i$. Then $\tilde{\beta}_1 - \tilde{\beta}_{1(i)}$ is approximated by the first element of the vector $\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_{(i)}$ (Cain and Lange, 1984) which is denoted by $\tilde{\Delta}_i(\beta_1)$.

The derivation of OPT and two-stage OPTA using inverse probability weighting rely on correct specification of all distribution forms and the nuisance integration. Neyman allocation based on the approximate influence measure $\tilde{\Delta}_i(\beta_1)$ is expected to yield a similar design efficiency of OPT while alleviating these constraints. However the computation of $\{\tilde{\Delta}_i(\beta_1), i = 1, \ldots, N\}$ requires the full cohort information on $(Y, A, \mathbf{X}')'$. In a similar spirit to the two-stage OPTA design of Section 3.3.3.1, we conduct a two-stage adaptive Neyman allocation of $\{\tilde{\Delta}_i(\beta_1), i = 1, \ldots, N\}$ with BAL in phase IIa and approximated Neyman allocation in phase IIb (NEYA). Specifically we proceed as follows. First a phase IIa subsample of size n_a is created using a BAL sub-sampling scheme and an estimate of σ_j in (3.4.1) is obtained, denoted by $\tilde{\sigma}_j$, using the data from the phase IIa sample. Second a phase IIb subsample is obtained using approximate Neyman allocation based on estimates from the phase IIa. Hence NEYA only requires the specification of $P(Y|A, \mathbf{X})$, which can be expressed in terms of the hazard model of interest in (3.2.2). See Appendix B.2 for further details about how NEY and NEYA were implemented here.

We next conduct a set of simulation studies to investigate the performance of three suboptimal sampling schemes including SRS, BAL, and EXT-$(\bar{A}, Y)$, two optimal designs OPT and NEY, and two adaptive optimal designs OPTA and NEYA. Again, the "limited memory Broyden-Fletcher-Goldfarb-Shanno" (L-BFGS-B) algorithm is used to obtain the estimators of θ using the general-purpose optimizer `optim` in `R` (R Core Team, 2020). The simulation results summarized in Table 3.2 show that OPT and NEY are the most efficient and as expected have similar performances. In general OPTA and NEYA offer a good approximation to OPT or NEY than alternative suboptimal designs. However, when the event is rare and the phase II sample size is small, BAL performs better than NEYA. This may be because NEYA requires sufficient selection in phase IIa to ensure reasonable estimates for σ_j.

3.5 A study of robustness and practical issues

The simulation studies thus far have involved correct specification of the PH models at both the design and analysis stage (i.e. $\kappa = 1$). Here we suppose the baseline cumulative hazard function is of the form $(\lambda t)^\kappa$ with $\kappa = 1.25$ in order to investigate the impact of

misspecification in design and/or analysis stages. In the models adopted we use PWC baseline hazards with cutpoints chosen to correspond to the empirical quantiles of the assessment times $\{A_i, i = 1, \ldots, N\}$. In particular, a piecewise constant model with four pieces ($\approx N^{1/5.5}$) set the cutpoints as the 25th, 50th, and 75th percentiles of A_is; and a PWC model with ten pieces ($\approx N^{1/3.3}$) set the cutpoints as the 10th, 20th, ..., and 90th percentiles of the A_i. Table 3.3 presents the simulation results of six practical sampling schemes (three under each analysis method). The sandwich variance (Boos and Stefanski, 2013) is used here. The results indicated that the PWC approximations used in the design and analysis stages yields a satisfactory approximation to the true model.

3.6 Illustrative applications

3.6.1 Diabetes in patients with psoriatic arthritis

Here we applied the proposed sampling schemes and inference procedures to a current status data set on diabetes from the University of Toronto Psoriatic Arthritis (PsA) clinic study (Gladman and Chandran, 2011). Patients with PsA were referred to this clinic since 1978. Demographical information such as gender, education level, lifestyle and so on were collected at first clinical visit as well as the development of diabetes. Biosamples were also collected and stored in the biobank at study entry for future studies. It is of great interest to study the risk of developing diabetes and assessing important human leukocyte antigens (HLA) markers such as DR4 in patients with PsA. The ascertainment of HLA markers through biosample testing tends to be expensive for a large cohort. A two-phase design would offer a cost-effective solution to this problem. Here we used a sample of 1021 patients (43.2% females, 22.6% DR4 positive) for analysis. For illustrative purposes we treated gender as an "inexpensive" covariate and the marker DR4 as an "expensive" covariate.

Suppose that the cohort of 1021 patients is the phase I sample and the information on the marker DR4 was missing. As the event rate is substantially low (around 6.8%), we stratified the phase I sample into six exclusive and nonempty strata on the status of diabetes and the discretized inspection time (i.e. age at the first visit to the clinic). We discretized the inspection times by its empirical 33th and 66th percentiles. In particular, the 33th and 66th percentiles of the age at inspection in patients are 37 and 50, respectively. We applied the practical sampling schemes including BAL, EXT-$(\bar{A}, Y)$, NEYA, TAO-OPTA, EXT-M_μ and EXT-(A, Y), to select $n = 200$ or 400 patients to collect information on DR4. We implemented these designs for practical purposes since they do not require additional

model assumptions other than the PH model of primary interest and (or) the nuisance covariate model.

Regarding the number of pieces used for the PWC model, we considered the integers around $1021^{1/5} \approx 4$, $1021^{1/4} \approx 6$ and $1021^{1/3} \approx 10$. The cutpoints b_ks are chosen such that each $(b_{k-1}, b_k]$ contains roughly equal number of inspection times. The estimation results are very similar and Table 3.4 summarizes the regression results using a piecewise constant hazard model with six pieces. One can see that (i) the regression coefficients estimates under the proposed two-phase designs are all pretty close to that under the full data analysis and indicate that DR4 positive is significantly associated with higher risk of developing diabetes; (ii) most of the designs select all patients with diabetes however the selection varies among subjects without diabetes; (iii) under the same allocation rules, ML estimators are generally more efficient then IPW estimators; (iv) different n_a renders different efficiency for adaptive designs. One needs to keep n_a large enough to obtain reasonable estimates for σ_j for NEYA, though estimates for $\text{Var}(X_1|X_2)$ require smaller n_a; and (v) the proposed TAO-OPTA, EXT-(A, Y) and EXT-M_μ obtained smaller standard error when estimating the DR4 marker effect, which is consistent with our theoretical derivation and simulation results.

3.6.2 Seroconversion after prophylactic anticoagulation therapy

Here we consider an application of the proposed methodology to data from a study of surgical patients at risk of developing a serological response following surgery. Specifically interest lies in understanding non-drug factors associated with the development of an antibody response after surgery. Following recovery from surgery and just prior to discharge from hospital, a blood sample is taken to assess seroconversion status leading to current status data. Here we focus on a sample of several thousand patients obtained by pooling data form several surgical studies, of which 3663 were female (59.94%), 5367 underwent planned orthopedic surgery, 1370 had surgery to repair a fracture. Sixty-three percent of the individuals were classified as overweighted or obese, by having a body-mass index (BMI) more than 25 kg/m^2.

Genetic samples were collected in these studies but they are not available to us so we use for illustration we consider BMI as the "expensive" covariate measured only in a phase II sub-sample. The frequency of antibody formation is rather low ($183/6111 \approx 2.97\%$) so to avoid empty strata we discretize the timing of the blood-test using the empirical 33rd and 66th percentiles (i.e. 5.58 and 6.84 days respectively) and adopted six strata defined by antibody status and the categorical variable defined by the timing of the blood tests;

the seroconversion rates during the periods $[0, 5.58]$, $(5.58, 6.84]$ and $(6.84, 10]$ are 17/2017, 63/2019, and 103/2075, respectively.

We adopt piecewise constant baseline hazards with two (or four) pieces, using the empirical median (or the 25th, 50th, and 75th percentiles) of the blood-test time as the cut-point(s). Figure 3.1 displays the nonparametric and parametric estimates of the marginal cumulative distribution function based on the full phase I sample; there is a good agreement among these estimates suggesting that both PWC models provide reasonable fit to the marginal distribution. To reflect various budgetary constraints, we consider the phase II sample size n ranging from 366 ($= 183 \times 2$) to 2400. Table 3.5 summarizes the regression results using the PWC PH model with two pieces under these two-phase designs (including SRS, BAL, NEYA, TAO-OPTA, EXT-(A, Y), and EXT-M_μ) with $n = 1000$ and the full data analysis. Based on analysis of the full phase I sample (possible since here we have access to all of the data for the BMI variable) individuals designated as overweight/obesity (i.e. $I(\text{BMI} \geqslant 25)$) have a significantly elevated hazard for the formation of antibodies (HR= $\exp(0.462) = 1.59$; 95% CI: 1.14, 2.23). For the analyses based on the phase II sub-sampling schemes we note that the stratified sub-sampling schemes ensure budgetary constraints are met, but the precision of the resulting estimators is greatly reduced. The point estimates fluctuate about the estimate from the full sample analysis but the confidence intervals are considerably wider and none of the stratified designs yield evidence of a BMI effect. The proposed residual-based sampling designs (i.e. TAO-OPTA and EXT-M_μ) along with EXT-(A, Y) all yield phase II samples that give maximum likelihood estimates with smaller standard errors and confidence intervals that are narrow enough to exclude the null value. For the TAO-OPTA design we find HR= $\exp(0.392) = 1.48$ (95% CI: 1.06, 2.08) when $n_a = 200$, for the EXT-M_μ design we obtain HR= $\exp(0.435) = 1.54$ (95% CI: 1.11, 2.16), and for the EXT-(A, Y) design we obtain HR= $\exp(0.416) = 1.51$ (95% CI: 1.08, 2.12).

3.7 Discussion

The key to cost-effective and efficient two-phase design is the identification of the types of individuals who are collectively the most informative about the parameter of interest. Often this is done through stratification of a phase I sample and carrying out stratified phase II sub-sampling (Ding et al., 2017; Zhou et al., 2018; Lawless, 2018). This requires a strategy for stratification of the phase I sample and it can be unclear what the best strategy is in a given situation. Tao et al. (2020) provided such guidance in developing optimal two-phase designs using maximum likelihood approach for a wide range of senarios; we adapt this

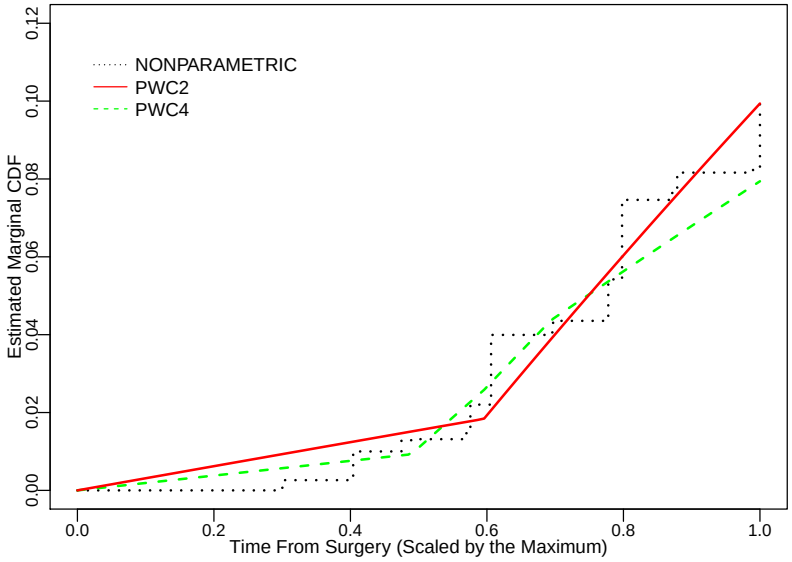


Figure 3.1: Nonparametric and parametric (piecewise constant hazard model with two pieces (PWC2) and with four pieces (PWC4)) estimates of the marginal cumulative distribution function for the time to seroconversion in orthopedic surgery.

work for the setting of current status data. We provide a practical and efficient response-dependent design (i.e., EXT-(A, Y)) which does not require either phase I stratification or parametric assumptions. Through extensive simulation studies we demonstrated the merits of EXT-(A, Y) by comparing to both optimal and common suboptimal designs. This derivation can also provide some insights to efficient sampling in settings of general type K censoring. But we need to note that the score-type residual M_μ is a function of both a current status response (A, Y) and a linear predictor $\mu = \beta_1 X_1 + \boldsymbol{\beta}_2' \mathbf{X}_2$. It is rather straightforward to show that under a proportional hazards model, M_μ is a nonpositive decreasing function of μ when $Y = 0$, and is nonnegative increasing in μ, when $Y = 1$. Hence, when μ is extreme it may influence the order of M_μ. In such cases, the efficiency advantage of EXT-(A, Y) may be limited so residual-based designs are preferred.

We also study an important alternative design approach involving inverse probability weighted estimating functions. As opposed to the extreme sampling fashion favoured under ML, balanced allocation of phase II samples is preferred under inverse probability weighting. Neyman allocation with related influence functions has been reported to approximate the optimal stratum-specific selection well (Breslow et al., 2009; Chen and Lumley, 2020). Our simulation results are consistent with the established findings.

The expensive covariate X_1 we considered and examined in the derivation and simulation studies is scalar. For multivariate covariate X_1, the aimed design efficiency may no longer depend on one entry in the covariance matrix but a sub-block. Then the optimal criteria we could consider include, trace ("A-optimal"), determinant ("D-optimal") or eigenvalues of the block of variance matrix corresponding to β_1 estimator. When multivariate X_1 comes into consideration, modeling the relationship between X_1 and $\mathbf{X}_2$ could get much more complex. Additional challenge arises in the settings where the phase I auxiliary covariate vector $\mathbf{X}_2$ contains continuous components and correlated with the expensive covariate X_1. Zeng and Lin (2014) proposed an approximation of the conditional density function $P(X_1|\mathbf{X}_2)$ by kernel smoothing in two-phase cohort studies. Generalization to the case allowing some components of $\mathbf{X}$ to be time-dependent is generally impractical under the current status observational scheme we considered here. If necessary additional collection of data is warranted and the maximum likelihood approach can be very challenging due to the difficulties in estimating $G(X_1|\mathbf{X}_2;\boldsymbol{\eta})$. The approach based on inverse weighted estimating equations alleviates this burden but at the price of lower efficiency (Reilly and Pepe, 1995; Robins et al., 1995; Lawless et al., 1999).

We considered proportional hazards model since it is probably the most commonly used approach to evaluate covariate effects in regression analysis for time-to-event data. Alternative models are needed to provide practical interpretation given specific context, or for settings where the proportional hazard assumption does not hold. In fact our key derivations here can be extended to other transformation models such as the proportional odds, probit, the additive hazard models and so forth. We demonstrated our methods using a piecewise constant function to approximate the true model and advocated its convenience and robustness carrying forward the essence of a nonparametric method. Important questions arise in practical use of PWC models include: how many pieces are appropriate to estimate the unknown failure time distribution, and where are the locations. For example, Goodman et al. (2011) proposed a data-driven method to determine the number and the placement of change points for PWC model based on a Wald-type test. Piecewise polynomials, wavelets, B-splines or any other flexible functions can also be employed.

In the simulation study and (or) the illustrative analysis, we didn't require positive selection probability for every study subject in some designs including OPT, TAO-OPT, EXT-(A, Y) and EXT-M_μ under maximum likelihood. It may raise identifiability issues when, for instance, the interaction effect of X_1 and X_2 is of interest. Hence, additional constraints may need to be adopted in the practical use of these designs. Another limitation of our development is that we required a conditional independence assumption between the inspection time A and $(T, X_1)'$ given the inexpensive covariate X_2, which may not be

hold in many settings (Rossini and Tsiatis, 1996; Chen et al., 2012). For these occasions a careful modeling of $P(A|T, \mathbf{X})$ to ensure valid inference is necessitated and extensions of our methods warrants further exploration.

Table 3.1: Simulation results of the estimated log hazard ratio in X_1 under maximum likelihood based on 1000 samples of size $N = 2000$; $\psi = 0.2, n = 300$.

			Stratified Designs										Non-stratified Designs							
			SRS		BAL		EXT-$(\bar{A}, Y)$		OPT		OPTA		EXT-(A, Y)		EXT-M_μ		TAO-OPT		TAO-OPTA	
(q_1, q_2)	(β_1, β_2)	n	ESE/ESD	ECP	ESE/ESD	ECP	ESE/ESD	ECP	ESE/ESD	ECP	ESE/ESD	ECP	ESE/ESD	ECP	ESE/ESD	ECP	ESE/ESD	ECP	ESE/ESD	ECP
(a) $X_1 \perp X_2$ (i.e. constant Var$(X_1\|X_2)$)																				
(0.6, 0.3)	(-0.2, -0.2)	300	0.29/0.29	0.95	0.26/0.25	0.95	0.24/0.22	0.94	0.21/0.19	0.94	0.22/0.21	0.94	0.18/0.18	0.95	0.18/0.17	0.95	0.18/0.17	0.95	0.21/0.20	0.95
		600	0.20/0.20	0.96	0.17/0.17	0.95	0.17/0.17	0.95	0.15/0.15	0.95	0.15/0.15	0.95	0.14/0.14	0.94	0.14/0.13	0.94	0.14/0.13	0.94	0.15/0.15	0.95
	(-0.2, 0.2)	300	0.29/0.29	0.95	0.26/0.25	0.94	0.23/0.22	0.94	0.20/0.19	0.94	0.21/0.21	0.94	0.18/0.18	0.95	0.18/0.17	0.95	0.18/0.17	0.95	0.21/0.21	0.94
		600	0.20/0.20	0.95	0.17/0.17	0.96	0.17/0.16	0.94	0.16/0.15	0.93	0.16/0.15	0.94	0.14/0.14	0.94	0.14/0.13	0.94	0.14/0.13	0.94	0.15/0.15	0.95
	(0.2, -0.2)	300	0.27/0.26	0.95	0.23/0.23	0.95	0.20/0.20	0.94	0.18/0.17	0.94	0.20/0.19	0.94	0.16/0.16	0.94	0.16/0.16	0.94	0.16/0.16	0.95	0.19/0.18	0.94
		600	0.18/0.18	0.96	0.16/0.16	0.95	0.15/0.15	0.94	0.14/0.14	0.94	0.14/0.14	0.94	0.13/0.12	0.94	0.13/0.12	0.94	0.13/0.12	0.94	0.14/0.14	0.94
	(0, 0)	300	0.27/0.27	0.95	0.25/0.24	0.95	0.21/0.21	0.95	0.19/0.19	0.95	0.20/0.20	0.95	0.17/0.16	0.94	0.17/0.16	0.94	0.17/0.16	0.94	0.20/0.20	0.95
		600	0.19/0.19	0.95	0.17/0.17	0.94	0.16/0.15	0.94	0.15/0.14	0.93	0.15/0.14	0.94	0.14/0.13	0.93	0.14/0.13	0.94	0.14/0.13	0.93	0.15/0.14	0.95
(0.1, 0.05)	(-0.2, -0.2)	300	3.94/28.51	0.93	0.36/0.34	0.95	0.43/0.41	0.94	0.31/0.31	0.95	0.34/0.32	0.95	0.32/0.31	0.95	0.31/0.31	0.95	0.31/0.31	0.95	0.36/0.34	0.95
		600	1.42/2.84	0.96	0.30/0.29	0.94	0.30/0.29	0.95	0.29/0.28	0.95	0.29/0.29	0.95	0.29/0.28	0.95	0.29/0.28	0.95	0.29/0.28	0.95	0.29/0.28	0.95
	(-0.2, 0.2)	300	3.81/23.73	0.93	0.36/0.34	0.95	0.42/0.41	0.96	0.31/0.31	0.96	0.33/0.32	0.94	0.32/0.31	0.94	0.31/0.31	0.96	0.31/0.31	0.96	0.35/0.34	0.95
		600	1.09/2.44	0.97	0.29/0.29	0.94	0.29/0.29	0.95	0.28/0.28	0.96	0.29/0.29	0.95	0.28/0.28	0.96	0.28/0.28	0.95	0.28/0.28	0.95	0.29/0.28	0.95
	(0.2, -0.2)	300	2.85/14.43	0.95	0.32/0.32	0.95	0.39/0.37	0.95	0.29/0.29	0.95	0.31/0.30	0.95	0.30/0.29	0.95	0.29/0.29	0.95	0.29/0.28	0.94	0.32/0.31	0.95
		600	0.96/1.85	0.96	0.27/0.26	0.95	0.26/0.26	0.95	0.25/0.25	0.95	0.26/0.26	0.96	0.25/0.25	0.95	0.25/0.25	0.96	0.25/0.25	0.96	0.25/0.25	0.95
	(0, 0)	300	3.28/15.46	0.94	0.33/0.32	0.95	0.40/0.38	0.95	0.30/0.30	0.95	0.31/0.31	0.94	0.30/0.29	0.95	0.29/0.29	0.95	0.29/0.29	0.95	0.34/0.33	0.96
		600	1.07/1.77	0.96	0.28/0.27	0.95	0.27/0.27	0.95	0.27/0.26	0.95	0.27/0.27	0.96	0.26/0.26	0.96	0.27/0.26	0.95	0.27/0.26	0.96	0.27/0.27	0.96
(b) $X_1 \not\perp X_2$ (i.e. non-constant Var$(X_1\|X_2)$)																				
(0.6, 0.3)	(-0.2, -0.2)	300	0.30/0.29	0.95	0.25/0.25	0.95	0.23/0.23	0.96	0.19/0.19	0.95	0.20/0.21	0.95	0.18/0.18	0.96	0.18/0.18	0.96	0.17/0.18	0.96	0.21/0.21	0.96
		600	0.2/0.2	0.96	0.17/0.17	0.95	0.17/0.17	0.96	0.15/0.16	0.96	0.15/0.15	0.95	0.14/0.14	0.95	0.14/0.14	0.95	0.14/0.14	0.96	0.15/0.15	0.95
	(-0.2, 0.2)	300	0.28/0.29	0.96	0.25/0.25	0.96	0.22/0.22	0.95	0.18/0.18	0.94	0.19/0.19	0.95	0.17/0.17	0.96	0.17/0.17	0.96	0.17/0.17	0.95	0.20/0.19	0.96
		600	0.2/0.2	0.96	0.17/0.17	0.95	0.16/0.16	0.95	0.15/0.15	0.95	0.15/0.15	0.95	0.14/0.14	0.95	0.14/0.13	0.95	0.14/0.13	0.95	0.15/0.15	0.95
	(0.2, -0.2)	300	0.27/0.27	0.96	0.24/0.23	0.94	0.20/0.20	0.96	0.17/0.18	0.96	0.19/0.19	0.95	0.16/0.16	0.96	0.16/0.16	0.96	0.16/0.16	0.95	0.19/0.19	0.95
		600	0.18/0.19	0.96	0.16/0.16	0.95	0.15/0.15	0.95	0.14/0.14	0.95	0.14/0.14	0.95	0.13/0.12	0.95	0.13/0.12	0.95	0.13/0.12	0.95	0.14/0.14	0.95
	(0, 0)	300	0.27/0.28	0.96	0.24/0.24	0.95	0.22/0.21	0.95	0.18/0.18	0.95	0.19/0.19	0.96	0.17/0.17	0.95	0.17/0.17	0.94	0.17/0.16	0.94	0.19/0.19	0.96
		600	0.19/0.19	0.96	0.17/0.17	0.94	0.16/0.15	0.95	0.14/0.14	0.96	0.15/0.14	0.94	0.13/0.13	0.95	0.13/0.13	0.94	0.13/0.13	0.94	0.15/0.14	0.95
(0.1, 0.05)	(-0.2, -0.2)	300	4.19/29.16	0.92	0.36/0.35	0.95	0.44/0.42	0.95	0.32/0.32	0.95	0.34/0.33	0.95	0.33/0.31	0.94	0.32/0.31	0.95	0.32/0.31	0.95	0.35/0.35	0.95
		600	1.34/2.21	0.97	0.31/0.30	0.94	0.31/0.30	0.95	0.30/0.29	0.95	0.30/0.29	0.94	0.30/0.29	0.95	0.30/0.29	0.95	0.30/0.29	0.95	0.30/0.29	0.94
	(-0.2, 0.2)	300	3.70/25.95	0.94	0.35/0.34	0.94	0.42/0.41	0.95	0.32/0.31	0.95	0.33/0.32	0.94	0.32/0.31	0.95	0.31/0.31	0.95	0.31/0.31	0.95	0.35/0.34	0.96
		600	1.11/0.53	0.97	0.30/0.29	0.95	0.31/0.29	0.94	0.29/0.28	0.94	0.30/0.29	0.95	0.29/0.28	0.94	0.29/0.28	0.94	0.29/0.28	0.94	0.30/0.28	0.94
	(0.2, -0.2)	300	2.85/14.43	0.95	0.32/0.32	0.95	0.39/0.37	0.95	0.29/0.29	0.95	0.31/0.30	0.95	0.30/0.29	0.95	0.29/0.29	0.95	0.29/0.28	0.94	0.32/0.31	0.95
		600	0.64/0.66	0.96	0.28/0.27	0.95	0.28/0.27	0.95	0.26/0.26	0.95	0.27/0.26	0.95	0.26/0.25	0.95	0.27/0.25	0.95	0.26/0.25	0.95	0.27/0.26	0.94
	(0, 0)	300	3.30/18.37	0.94	0.34/0.33	0.96	0.40/0.39	0.95	0.31/0.30	0.94	0.33/0.31	0.93	0.31/0.3	0.93	0.31/0.30	0.95	0.31/0.29	0.95	0.34/0.33	0.94
		600	0.86/0.50	0.97	0.29/0.28	0.94	0.29/0.28	0.94	0.28/0.27	0.94	0.29/0.27	0.94	0.28/0.27	0.94	0.28/0.27	0.94	0.28/0.27	0.94	0.29/0.27	0.94

Note. OPTA conducts balanced sampling and TAO-OPTA implements simple random sampling in phase II a with $n_a = 0.5n$; When $X_1 \not\perp X_2$, the odds ratio between X_1 and X_2 is set by $\varphi = 2$.

Table 3.2: Simulation results of the estimated log hazard ratio in X_1 following oracle and practical Neyman allocations under inverse probability weighting based on 500 samples of size $N = 2000$; $\psi = 0.2$.

			OPT		NEY		OPTA		NEYA		SRS		BAL		EXT-$(\bar{A}, Y)$	
(q_1, q_2)	(β_1, β_2)	φ	ESD/ASE	ECP	ESD/ASE	ECP	ESD/ASE	ECP	ESD/ASE	ECP	ESD/ASE	ECP	ESD/ASE	ECP	ESD/ASE	ECP
(a) Phase II sample size $n = 300$ ($n_a = 240, n_b = 60$)																
(0.6, 0.3)	(-0.2, -0.2)	1	0.24/0.24	0.95	0.24/0.24	0.95	0.26/0.26	0.95	0.26/0.25	0.94	0.29/0.29	0.95	0.28/0.27	0.94	0.35/0.33	0.94
		2	0.24/0.24	0.96	0.23/0.24	0.96	0.26/0.26	0.96	0.26/0.26	0.94	0.31/0.29	0.94	0.28/0.27	0.94	0.35/0.33	0.95
	(-0.2, 0.2)	1	0.24/0.24	0.95	0.24/0.24	0.94	0.26/0.26	0.95	0.27/0.25	0.94	0.30/0.29	0.94	0.28/0.27	0.94	0.37/0.33	0.94
		2	0.23/0.24	0.95	0.24/0.24	0.96	0.25/0.25	0.96	0.25/0.25	0.95	0.29/0.29	0.94	0.28/0.26	0.94	0.34/0.32	0.94
	(0.2, -0.2)	1	0.23/0.22	0.95	0.23/0.22	0.95	0.24/0.24	0.94	0.25/0.24	0.93	0.27/0.26	0.93	0.25/0.25	0.95	0.31/0.29	0.94
		2	0.23/0.23	0.95	0.22/0.22	0.95	0.24/0.24	0.96	0.25/0.24	0.95	0.28/0.27	0.94	0.26/0.25	0.94	0.31/0.29	0.95
	(0, 0)	1	0.23/0.23	0.94	0.23/0.23	0.95	0.26/0.25	0.95	0.26/0.24	0.94	0.28/0.27	0.95	0.27/0.26	0.95	0.35/0.31	0.93
		2	0.24/0.23	0.95	0.23/0.23	0.95	0.25/0.25	0.95	0.26/0.24	0.93	0.29/0.28	0.95	0.27/0.26	0.94	0.34/0.31	0.95
(0.1, 0.05)	(-0.2, -0.2)	1	0.32/0.32	0.96	0.33/0.33	0.96	0.37/0.35	0.95	0.38/0.37	0.95	4.56/0.65	0.85	0.38/0.36	0.94	0.48/0.44	0.94
		2	0.33/0.33	0.96	0.33/0.33	0.97	0.36/0.36	0.95	0.38/0.38	0.96	4.97/0.66	0.84	0.38/0.36	0.95	0.49/0.45	0.93
	(-0.2, 0.2)	1	0.31/0.32	0.97	0.32/0.33	0.97	0.35/0.35	0.95	0.38/0.37	0.96	4.40/0.66	0.86	0.37/0.36	0.95	0.45/0.44	0.96
		2	0.33/0.32	0.96	0.33/0.33	0.96	0.35/0.35	0.96	0.38/0.37	0.94	4.38/0.66	0.88	0.37/0.36	0.95	0.47/0.44	0.94
	(0.2, -0.2)	1	0.29/0.30	0.96	0.30/0.30	0.96	0.33/0.32	0.94	0.35/0.34	0.94	3.02/0.62	0.90	0.33/0.33	0.95	0.40/0.39	0.95
		2	0.31/0.30	0.95	0.32/0.31	0.95	0.33/0.33	0.95	0.34/0.35	0.95	3.48/0.63	0.90	0.34/0.33	0.95	0.43/0.40	0.95
	(0, 0)	1	0.31/0.31	0.96	0.32/0.31	0.95	0.33/0.33	0.95	0.36/0.35	0.95	3.87/0.64	0.89	0.35/0.34	0.95	0.44/0.41	0.95
		2	0.33/0.31	0.95	0.34/0.32	0.94	0.34/0.34	0.94	0.37/0.36	0.95	3.93/0.65	0.89	0.36/0.35	0.95	0.44/0.42	0.96
(b) Phase II sample size $n = 600$ ($n_a = 300, n_b = 300$)																
(0.6, 0.3)	(-0.2, -0.2)	1	0.17/0.17	0.95	0.16/0.17	0.95	0.17/0.17	0.96	0.18/0.17	0.94	0.20/0.20	0.95	0.18/0.19	0.95	0.23/0.22	0.94
		2	0.17/0.17	0.96	0.16/0.17	0.96	0.17/0.17	0.96	0.17/0.18	0.96	0.20/0.20	0.95	0.19/0.19	0.96	0.22/0.22	0.95
	(-0.2, 0.2)	1	0.16/0.17	0.96	0.17/0.17	0.95	0.18/0.17	0.94	0.20/0.17	0.94	0.20/0.20	0.94	0.19/0.19	0.96	0.22/0.21	0.94
		2	0.16/0.17	0.96	0.16/0.17	0.96	0.17/0.17	0.95	0.19/0.17	0.94	0.20/0.20	0.96	0.19/0.18	0.95	0.21/0.20	0.94
	(0.2, -0.2)	1	0.16/0.16	0.94	0.16/0.16	0.95	0.16/0.16	0.95	0.17/0.16	0.93	0.18/0.18	0.96	0.17/0.17	0.95	0.19/0.19	0.94
		2	0.16/0.16	0.95	0.16/0.16	0.95	0.16/0.16	0.97	0.18/0.16	0.94	0.19/0.19	0.95	0.17/0.17	0.95	0.20/0.19	0.96
	(0, 0)	1	0.17/0.16	0.93	0.17/0.16	0.94	0.17/0.17	0.94	0.18/0.17	0.93	0.19/0.19	0.94	0.19/0.18	0.94	0.21/0.20	0.94
		2	0.17/0.16	0.94	0.16/0.16	0.96	0.17/0.17	0.96	0.17/0.17	0.95	0.19/0.19	0.96	0.18/0.18	0.96	0.20/0.20	0.94
(0.1, 0.05)	(-0.2, -0.2)	1	0.29/0.29	0.95	0.3/0.29	0.96	0.30/0.29	0.96	0.30/0.29	0.96	1.61/0.52	0.94	0.31/0.30	0.94	0.31/0.30	0.95
		2	0.30/0.29	0.94	0.31/0.30	0.95	0.31/0.30	0.94	0.30/0.30	0.95	1.52/0.53	0.96	0.32/0.30	0.94	0.32/0.30	0.95
	(-0.2, 0.2)	1	0.29/0.29	0.95	0.30/0.29	0.95	0.29/0.29	0.95	0.29/0.29	0.95	1.27/0.52	0.96	0.30/0.30	0.95	0.30/0.30	0.96
		2	0.30/0.29	0.95	0.30/0.29	0.96	0.31/0.29	0.94	0.31/0.29	0.94	1.22/0.52	0.96	0.31/0.30	0.94	0.31/0.30	0.95
	(0.2, -0.2)	1	0.26/0.26	0.95	0.26/0.26	0.95	0.26/0.26	0.96	0.26/0.26	0.96	0.95/0.45	0.94	0.27/0.27	0.95	0.27/0.27	0.95
		2	0.27/0.26	0.94	0.28/0.27	0.95	0.27/0.27	0.96	0.27/0.27	0.95	0.74/0.47	0.95	0.29/0.27	0.94	0.28/0.27	0.94
	(0, 0)	1	0.28/0.27	0.94	0.29/0.28	0.95	0.27/0.27	0.96	0.28/0.28	0.96	1.20/0.48	0.95	0.28/0.28	0.95	0.28/0.28	0.96
		2	0.29/0.28	0.94	0.30/0.28	0.94	0.29/0.28	0.95	0.29/0.28	0.95	0.88/0.49	0.96	0.30/0.28	0.95	0.30/0.28	0.95

Note. OPTA and NEYA implement balanced sampling in phase IIa; φ: the odds ratio between X_1 and X_2.

Table 3.3: Simulation results of the estimated log hazard ratio in X_1 with a phase IIa selection based on 200 samples of size 2000; $\kappa = 1.25$, $\psi = 0.2$, $q_1 = 0.6, q_2 = 0.3$, $\beta_1 = -0.2, \beta_2 = -0.2$.

			Maximum Likelihood									Inverse Prob. Weighting								
			EXT-(A,Y)			OPTA			TAO-OPTA			BAL			OPTA			NEYA		
Design	Analysis	φ	BIAS	ESE	ECP	BIAS	ESE	ECP	BIAS	ESE	ECP	BIAS	ESE	ECP	BIAS	ESE	ECP	BIAS	ESE	ECP
(a) Phase II sample size $n = 300$ ($n_a = 240, n_b = 60$)																				
WEI	WEI	1	-0.018	0.177	0.945	0.004	0.213	0.945	-0.029	0.248	0.940	-0.023	0.265	0.975	-0.014	0.249	0.955	-0.013	0.262	0.945
		2	-0.012	0.173	0.970	0.006	0.207	0.975	-0.018	0.260	0.945	-0.025	0.270	0.935	< 0.001	0.251	0.945	-0.007	0.255	0.945
	PWC_4	1	-0.017	0.176	0.945	0.003	0.212	0.950	-0.030	0.247	0.945	-0.025	0.269	0.950	-0.017	0.250	0.965	-0.016	0.264	0.950
		2	-0.012	0.173	0.970	0.006	0.207	0.970	-0.017	0.258	0.950	-0.027	0.274	0.945	-0.004	0.252	0.950	-0.011	0.255	0.935
	PWC_{10}	1	-0.018	0.177	0.945	0.003	0.212	0.950	-0.030	0.247	0.945	-0.026	0.272	0.955	-0.018	0.252	0.955	-0.017	0.266	0.950
		2	-0.012	0.174	0.970	0.006	0.207	0.975	-0.017	0.258	0.955	-0.027	0.277	0.935	-0.007	0.254	0.950	-0.013	0.258	0.955
PWC_4	WEI	1	-0.018	0.177	0.945	-0.009	0.210	0.945	-0.029	0.248	0.940	-0.023	0.265	0.975	-0.008	0.261	0.945	-0.013	0.254	0.950
		2	-0.012	0.173	0.970	0.004	0.208	0.975	-0.016	0.259	0.945	-0.025	0.270	0.935	-0.006	0.256	0.955	-0.012	0.269	0.935
	PWC_4	1	-0.017	0.176	0.945	-0.010	0.209	0.950	-0.029	0.247	0.945	-0.025	0.269	0.950	-0.011	0.260	0.950	-0.015	0.255	0.955
		2	-0.012	0.173	0.970	0.004	0.208	0.970	-0.015	0.257	0.950	-0.027	0.274	0.945	-0.010	0.256	0.950	-0.017	0.269	0.935
	PWC_{10}	1	-0.018	0.177	0.945	-0.010	0.209	0.950	-0.030	0.247	0.945	-0.026	0.272	0.955	-0.012	0.264	0.945	-0.017	0.257	0.955
		2	-0.012	0.174	0.970	0.004	0.208	0.975	-0.016	0.258	0.955	-0.027	0.277	0.935	-0.013	0.259	0.955	-0.017	0.272	0.940
(b) Phase II sample size $n = 600$ ($n_a = 300, n_b = 300$)																				
WEI	WEI	1	-0.016	0.145	0.935	-0.007	0.159	0.915	-0.022	0.158	0.960	-0.017	0.187	0.940	-0.009	0.168	0.975	-0.004	0.171	0.955
		2	-0.015	0.137	0.925	-0.003	0.155	0.915	-0.018	0.150	0.945	-0.008	0.191	0.955	-0.017	0.172	0.955	-0.023	0.182	0.935
	PWC_4	1	-0.016	0.145	0.935	-0.008	0.160	0.915	-0.023	0.158	0.960	-0.017	0.188	0.945	-0.011	0.169	0.975	-0.006	0.171	0.950
		2	-0.016	0.137	0.920	-0.003	0.155	0.915	-0.018	0.149	0.945	-0.010	0.190	0.955	-0.019	0.174	0.950	-0.024	0.183	0.940
	PWC_{10}	1	-0.017	0.145	0.935	-0.008	0.160	0.910	-0.023	0.158	0.955	-0.018	0.189	0.940	-0.010	0.170	0.975	-0.006	0.171	0.950
		2	-0.016	0.137	0.920	-0.003	0.156	0.920	-0.019	0.149	0.945	-0.009	0.191	0.955	-0.019	0.176	0.945	-0.025	0.186	0.935
PWC_4	WEI	1	-0.016	0.145	0.935	-0.009	0.158	0.930	-0.021	0.158	0.950	-0.017	0.187	0.940	-0.005	0.169	0.965	-0.002	0.176	0.955
		2	-0.015	0.137	0.925	-0.006	0.155	0.930	-0.017	0.150	0.950	-0.008	0.191	0.955	-0.017	0.174	0.950	-0.022	0.179	0.950
	PWC_4	1	-0.016	0.145	0.935	-0.010	0.159	0.925	-0.022	0.157	0.950	-0.017	0.188	0.945	-0.007	0.169	0.960	-0.004	0.177	0.955
		2	-0.016	0.137	0.920	-0.006	0.156	0.930	-0.017	0.150	0.950	-0.010	0.190	0.955	-0.019	0.175	0.960	-0.023	0.180	0.955
	PWC_{10}	1	-0.017	0.145	0.935	-0.011	0.158	0.930	-0.022	0.158	0.950	-0.018	0.189	0.940	-0.007	0.170	0.960	-0.004	0.177	0.960
		2	-0.016	0.137	0.920	-0.006	0.156	0.935	-0.018	0.150	0.950	-0.009	0.191	0.955	-0.019	0.177	0.955	-0.024	0.182	0.945

Note. OPTA, NEYA: balanced sampling is conducted in phase IIa; TAO-OPTA: simple random sampling is conducted in phase IIa.

Table 3.4: Regression analysis fitting a piecewise constant hazard model with six pieces to diabetes data from the University of Toronto Psoriatic Arthritis clinic study.

			Maximum Likelihood		Inverse. Prob. Weighting		Number of Selected with $(\bar{A}, Y)$					
Design	n	n_a/n	DR4	Gender	DR4	Gender	(1,1)	(2,1)	(3,1)	(1,0)	(2,0)	(3,0)
FULL	1021		0.680(0.252)	-0.159(0.247)	0.680 (0.252)	-0.159(0.246)	9	16	44	328	321	303
Stratified Designs												
BAL	200		0.727(0.325)	-0.163(0.252)	0.713(0.348)	-0.164(0.327)	9	16	43	44	44	44
	400		0.714(0.276)	-0.161(0.249)	0.698(0.283)	-0.158(0.273)	9	16	44	110	110	110
EXT-$(\bar{A}, Y)$	200		0.752(0.333)	-0.166(0.253)	0.713(0.358)	-0.175(0.340)	9	16	33	48	48	47
	400		0.714(0.276)	-0.161(0.249)	0.699(0.282)	-0.157(0.271)	9	16	44	111	111	110
NEYA	200	2/4			0.686(0.373)	-0.166(0.359)	8	9	24	33	49	76
		3/4			0.717(0.360)	-0.173(0.347)	9	16	25	39	48	62
	400	1/4			0.687(0.281)	-0.167(0.271)	8	14	42	53	101	182
		2/4			0.698(0.276)	-0.164(0.267)	9	16	42	68	105	160
		3/4			0.692(0.277)	-0.158(0.266)	9	16	44	87	108	135
Non-stratified Designs												
TAO-OPTA	200	1/4	0.667(0.307)	-0.149(0.251)			9	16	42	16	16	100
		2/4	0.789(0.330)	-0.170(0.253)			9	16	28	32	31	84
		3/4	0.740(0.396)	-0.167(0.255)			5	9	20	48	47	69
	400	1/4	0.691(0.261)	-0.153(0.246)			9	16	44	32	32	267
		2/4	0.659(0.266)	-0.143(0.248)			9	16	44	64	63	204
		3/4	0.719(0.275)	-0.163(0.249)			9	16	42	96	94	143
EXT-M_μ	200		0.633(0.298)	-0.137(0.249)			9	16	44	0	0	131
	400		0.632(0.256)	-0.144(0.244)			9	16	44	0	28	303
EXT-(A, Y)	200		0.598(0.295)	-0.137(0.248)			9	16	44	0	0	131
	400		0.634(0.256)	-0.148(0.246)			9	16	44	0	28	303

Note. Estimates of regression coefficients and standard errors (in brackets) under the BAL, EXT-$(\bar{A}, Y)$, NEYA, TAO-OPTA, and EXT-(A, Y) are averaged over 1000 replicates; TAO-OPTA implements simple random sampling and NEYA conducts balanced sampling in phase IIa.

Table 3.5: Regression analysis fitting a piecewise constant hazard model with two pieces to data from 6111 patients received orthopetic surgery with the phase II sample size $n = 1000$. Estimated log hazard ratios for overweight/obesity (i.e. BMI $\geqslant 25$) and the corresponding 95% confidence intervals in round brackets are presented.

			Stratified Designs					
	Full		SRS		BAL		NEYA	
Framework	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
Likelihood	0.46	(0.13, 0.80)	0.48	(-0.35, 1.31)	0.33	(-0.02, 0.67)	-	-
IPW[1]	0.46	(0.12, 0.81)	0.54	(-0.41, 1.48)	0.47	(-0.05, 0.98)	0.47	(-0.04, 0.98)
							0.47	(-0.04, 0.98)
			Non-Stratified Designs					
	Full		TAO-OPTA[2]		EXT-M_μ		EXT-(A, Y)	
Likelihood	0.46	(0.13, 0.80)	0.39	(0.06, 0.73)	0.44	(0.10, 0.77)	0.42	(0.08, 0.75)
			0.37	(0.04, 0.71)				

[1] for NEYA the first row is based on an internal pilot of $n_a = 400$ while the second row is for $n_a = 800$;
[2] first row is for $n_a = 200$ and second row is for $n_a = 400$.

Chapter 4

Response-dependent subsampling involving multiple disease registries

4.1 Introduction

With the development and accumulation of modern epidemiological research, multiple large prevalent cohorts may have been formed to collect data from patients recruited in different clinical stages of chronic disease processes. In time to event analyses, the increasing availability of different data sources has led to increasing interest in statistical analysis exploiting data from a combination of cohorts (Wang, 1999; Copas and Farewell, 2001; Saarela et al., 2009; Wolfson et al., 2019). However, cost-effective design of biomarker studies based on combined registry data has not received much attention. Often budgetary issues arise in this context as it is too expensive and inefficient to assess the stored biosamples for every enrolled patient. Much research has been directed at two-phase designs for failure time data subject to censoring; see Prentice (1986); Borgan and Samuelsen (2014); McIsaac and Cook (2015); Ding et al. (2017); Lawless (2018); Tao et al. (2020) among others. In this article, we address the development of two-phase designs and the associated analysis for pooled prevalent cohort data under likelihood and inverse probability weighting. Two types of selection bias arise in this context from i) recruitment procedures implemented by each disease registry and ii) response-dependent sub-sampling for the combined registries. To accommodate such a complex data structure and various observation patterns, multistate models are employed as they offer an intuitive and appealing framework to model the lifetime dynamics concerning a disease progression of interest (Cook and Lawless, 2014, 2018).

To motivate the methodology, we consider two research programs conducted by University of Toronto in psoriasis and psoriatic arthritis; see Section 1.2.3. Psoriasis (Ps) and psoriatic arthritis (PsA) are often deemed as chronologically-ordered diseases or disorder conditions and two research programs, led separately by the University of Toronto Psoriasis Clinic (UTPC) and the University of Toronto Psoriatic Arthritis Clinic (UTPAC), have been conducted to investigate these two chronologically-ordered events, respectively. In particular, UTPC established a registry of patents with Ps in 2006 (Eder et al., 2011); another disease registry maintained by UTPAC was launched much earlier in 1977 by enrolling patients in PsA (Gladman and Chandran, 2011). Upon recruitment, patients went through a detailed interview on disease history and clinical examination, provided blood or urine samples for genetic testing, and then followed prospectively according to a standardized protocol. Biomarker studies in identifying human leukocyte antigens (HLA) associated with the risk of PsA development in Ps patients have been received a lot of attention (Rahman and Elder, 2005; Eder et al., 2012; Chandran, 2013; Wu and Cook, 2018). With the availability of biosamples from both UTPC and UTPAC registries, we frame the the pooled registry data as the phase I sample and aim to provide some guidance in cost-effective selection of a subcohort for the ascertainment of the expensive biomarker information. The UTPC registry provides Ps-prevalent cohort data and prospective incidence of PsA onsets and the UTPAC registry provide PsA-prevalent cohort data where retrospective information on the development of PsA from Ps is available.

Multistate models have been employed extensively for Ps and PsA studies (Chandran et al., 2010; Farewell and Su, 2011; Cook and Lawless, 2018; Zeng et al., 2020) and their utilization on life history processes subject to biased observation schemes dates back to Keiding (1991) and Commenges (2002). In this chapter we begin by considering a homogeneous population (Commenges, 1999) following a six-state data model involving two disease stages (i.e. Ps and PsA) as in Cook and Lawless (2018) and Zeng et al. (2020). The state space contains $\{0, 1, 2, D_0, D_1, D_2\}$, where state 0 represents disease-free and alive, states 1 and 2 represent the disease state with state 1 representing an initial phase and state 2 a more advanced disease state, respectively. States D_j represent an absorbing state of death from state j, $j = 0, 1, 2$; see Figure 4.1. In the later sections we will consider likelihood-based methods and discuss model assumptions for desirable simplifications based on this general six-state model.

The organization of this chapter is as follows. In Section 4.2, we introduce the notation, observed data and model formulations. Likelihood construction upon two assumptions on the intensity models is presented for six-state processes. In Section 4.3, the estimation and inference procedure base on a simplified partial likelihood is developed. Extreme residual

and response dependent designs are derived in Section 4.4 and their performance and comparison to other commonly-used two-phase designs are investigated by a comprehensive simulation studies conducted in Section 4.5. Section 4.6 returns to a general setting to relax the nondifferential mortality assumption and discusses the analysis and design problems in a more general setting; The proposed methods are examined in another set of simulation studies in Section 4.7 and applied to the motivating example in Section 4.8. This chapter ends with a discussion of the inverse probability weighting approach in Section 4.9 and concluding remarks in Section 4.10.

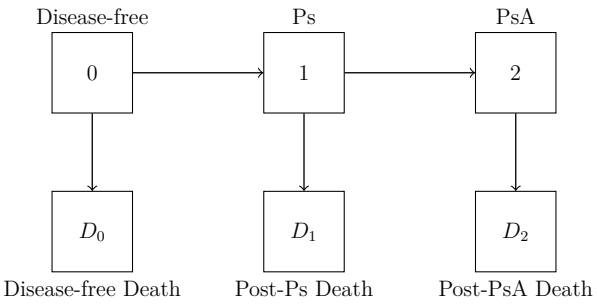


Figure 4.1: A state space diagram for a six-state two-stage disease process.

4.2 Notation, data, and model

4.2.1 Disease progression under a six-state process model

We consider a birth cohort defined as individuals born in a window of calendar time $\mathscr{B} = (B_L, B_R]$. For a generic individual we let B denote the birth date so that at calendar time $e = B + a$ they are age a, $a > 0$. Let A_D denote the age of death (or the survival time) and $A_j < A_D$ the age of a $j-1 \to j$ transition if one occurs, $j = 1, 2$; if the transition does not occur before death we set $A_j = \infty$. Let $Z(a)$ denote the state occupied at age a with $P(Z(0) = 0) = 1$. Let $H(a) = \{Z(u), 0 < u \leqslant a, Z(0) = 0, B\}$ denote the life history up to age a and $\bar{H}(a) = \{H(a), \mathbf{V}\}$ be the expanded history, where $\mathbf{V}$ denotes covariates reflecting demographic characteristics and possibly baseline marker information. In our motivating example, primary interest lies in studying the effect of a specific new marker on the Ps to PsA (i.e. $1 \to 2$) transition.

4.2.2 Two-phase designs with biased phase I samples

Different disease registries will typically impose different disease-related selection conditions, leading to the formation of biased samples from the target population. Analysis

of data from a combination of registries requires a careful treatment to these recruitment rules. We next describe a somewhat idealized recruitment procedure to characterize selection into the UTPAC and UTPC respectively. In reality individuals are recruited over different windows of calendar time and according to a variety of poorly understood and documented processes. Some individuals are recruited by particular screening efforts but we assume here that the cross-sectional sampling scheme we describe next is the method employed for all recruited individuals for convenience. For further simplification we assume that the recruitment process occurs at a particular date E_0.

4.2.2.1 The University of Toronto Psoriatic Arthritis Cohort

The UTPAC registry began to enrol PsA patients in 1977 (Gladman and Chandran, 2011). In what follows we will refer to this registry as *Registry 2* since individuals were recruited in state 2 of the six-state model described by Figure 4.1. For each individual enrolled in *Registry 2*, retrospective life history data are recorded at the time of recruitment including date of birth B, age at onset of the initial disease A_1, age at onset of the advanced disease A_2; the age at recruitment is $A_0 = E_0 - B$. We suppose that recruited individuals take part in a follow-up study of C years duration, conducted to learn about the disease course. We consider C as an administrative or completely random censoring time that is independent and noninformative for the process of interest. Let $A_C = A_0 + C$ denote the age at possible censoring and $A^\dagger = A_D \wedge A_C$ denote the time on study, where $a \wedge b$ denotes the minimum of a and b. Let $\delta_1 = I(A_1 \leqslant A_D)$, $\delta_2 = I(A_2 \leqslant A^\dagger)$ and $\delta_D = I(A_D \leqslant A^\dagger)$ indicate the observation of the $0 \to 1$ and $1 \to 2$, to transitions and death, respectively, where $I(\cdot)$ is the indicator function. Those recruited into *Registry 2* provide information on $\{Z_0 = 2, \bar{H}(A^\dagger)\}$, where $Z_0 = Z(A_0)$ and we observe $A_1 < A_2 < \infty$ (retrospectively) so $\delta_1 = 1$, $\delta_2 = 1$.

Often recruitment to a registry or cohort may depend on age, gender and other demographic variables. The Canadian Longitudinal Study in Aging (Raina et al., 2009), for example, used stratified sampling so survey weights may be used in analyses. We assume here, however, that the variables governing the complex survey design are contained in $\mathbf{V}$ and so survey weighting is not needed or discussed in what follows. Given interest lies in the process following entry to state 1, we consider the likelihood contribution from an individual recruited in *Registry 2* based on

$$P(H(A^\dagger)|Z_0 = 2, \bar{H}(A_1)) = \frac{P(H(A^\dagger), Z_0 = 2|\bar{H}(A_1))}{P(Z_0 = 2|\bar{H}(A_1))}. \tag{4.2.1}$$

the conditional probability of observing a disease evolution given that an individual recruited in state 2 and $\bar{H}(A_1)$. Since E_0 is fixed, here and in what follows we suppress the

dependence on $A_0 = E_0 - B$ in the conditions for simplicity. A broad class of recruitment rules can be accommodated by such a likelihood construction - the recruitment procedure can depend on the state at recruitment, birth time (or age at recruitment), onset of the initial disease, and covariates $\mathbf{V}$.

4.2.2.2 The University of Toronto Psoriasis Cohort

The UTPC registry was established in 2006 to recruit patients with psoriasis (Eder et al., 2011). The goal of establishing this cohort was to observe incident cases of PsA and to study risk factors for its development. In what follows we will refer to this registry as *Registry 1* since individuals were recruited in state 1 of the six-state model described by Figure 4.1. Similar to the data collection procedure assumed for *Registry 2*, recruited individuals undergo an interview to retrospectively record their life history prior to recruitment, including their dates of birth B and their ages at onset of the initial disease A_1. Upon recruitment, a prospective follow-up study of length C is conducted to monitor the disease development. An individual in *Registry 1* therefore provides data on $\{Z_0 = 1, \bar{H}(A^\dagger)\}$. Given $\bar{H}(A_0)$, the conditional probability $P(H(A^\dagger)|Z_0 = 1, \bar{H}(A_1))$ equals

$$\frac{P(H(A^\dagger), Z_0 = 1|\bar{H}(A_1))}{P(Z_0 = 1|\bar{H}(A_1))},$$

where the numerator is the conditional probability of observing a disease path given $\bar{H}(A_1)$, and the denominator adjusts for the recruitment bias. It is evident that the recruitment process can depend on the state at recruitment, birth time (or age at recruitment), onset of the initial disease, and covariates $\mathbf{V}$.

4.2.2.3 Pooled registries, incomplete covariate and two-phase design

We now consider two-phase design problems involving a phase I sample as obtained by pooling data from *Registry 1* and *Registry 2*. To proceed further, we idealize the recruitment of individuals to the two registries by assuming that, by screening the birth cohort at E_0, a sample of N_1 individuals in state 1 is recruited into *Registry 1* and a sample of N_2 individuals in state 2 is enrolled in *Registry 2*. A schematic of possible life history paths prior to and following recruitment to *Registry 1* and *Registry 2* is depicted in Figure 4.2.

Let X denote a fixed biomarker of interest which is only observed when a biospecimen collected upon recruitment to the respective registry is assayed, giving $(X, \mathbf{V}')'$ is the complete covariate vector. Under budgetary constraints, one cannot measure X for the whole (combined) cohort. An important issue is how to efficiently select a phase II

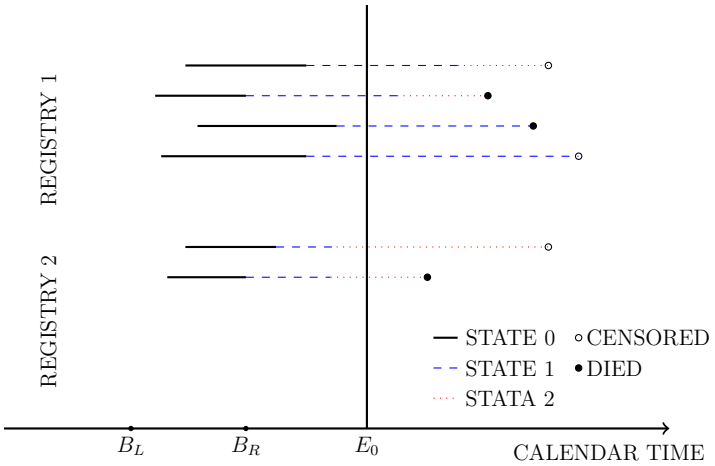


Figure 4.2: A schematic of possible life history paths prior to and following recruitment to *Registry 1* and *Registry 2*.

subcohortsample for the ascertainment of X. Two-phase designs offer a natural framework to construct a solution. Let $\Delta = I(X \text{ is observed})$ indicate whether the marker value of interest is available. Here X is missing at random as the selection model, designated by investigators, is based on observed information (Little and Rubin, 2002). The general selection model can be expressed as

$$P(\Delta = 1|\bar{H}(A^\dagger), X) = \sum_{k=1}^{2} I(Z_0 = k)P(\Delta = 1|Z_0 = k, \bar{H}(A^\dagger)) \tag{4.2.2}$$

under the budgetary constraint

$$P(\Delta = 1) = \tau, \tag{4.2.3}$$

where $0 < \tau \leqslant 1$. Let $\mathbf{X}^\circ = (X, \mathbf{V}')'$ if $\Delta = 1$ and otherwise $\mathbf{X}^\circ = \mathbf{V}$. The Lexis diagram (Keiding, 2011) of Figure 4.3 depicts information for two individuals recruited in *Registry 1* and *Registry 2*, respectively. The horizontal axis represents the calendar time indicating the calendar window $(B_L, B_R]$ of the target birth cohort, birth dates B_1 and B_2 for the two individuals, and the date at screening and recruitment E_0. The vertical axes measure time since birth (age); the left axis reflects the age of events for an individual that could be recruited to *Registry 1* and right axis has the corresponding information for an individual that could be recruited to *Registry 2*. Information on transitions over $H(A^\dagger)$ can be re-expressed by the transition indicators $(\delta_1, \delta_2, \delta_D)'$ and the realized sojourn times in states j, denoted by S_j, $j = 0, 1, 2$, where $S_0 = A_1 \wedge A_D$; $S_1 = (A_2 \wedge A^\dagger) - A_1$ if $\delta_1 = 1$, or 0 if $\delta_1 = 0$; and $S_2 = A^\dagger - (A_2 \wedge A^\dagger)$. The realized sojourn times in states j, S_j (j=0, 1, 2), and the time elapse T_0 from A_1 to age at recruitment A_0 are also labeled. Note that

$S_2 = 0$ for the individual born at B_1 as the $1 \to 2$ disease progression did not occur during the follow-up.

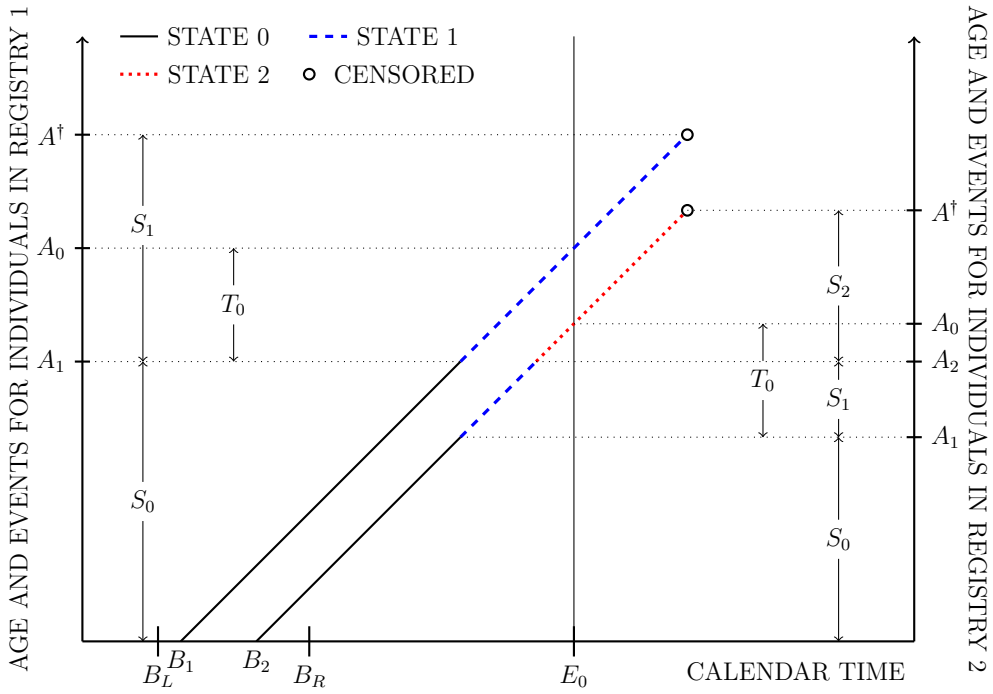


Figure 4.3: A Lexis diagram of two diseased individuals recruited into *Registry 1* (left) and *Registry 2* (right), with horizontal axis shown the calendar times and the vertical axis presented corresponding age; and realized sojourn times S_j in state j (j=0,1,2) are also labeled.

Hence, *Registry 1* provides observations $\{Z_{i0} = 1, H_i(A_i^\dagger), \mathbf{X}_i^\circ, \Delta_i, i = 1, \ldots, N_1\} = \{Z_{i0} = 1, B_i, A_{i1}, S_{i1}, S_{i2}, \delta_{i1} = 1, \delta_{i2}, \delta_{iD}, \mathbf{X}_i^\circ, \Delta_i, i = 1, \ldots, N_1\}$ and *Registry 2* contains data on $\{Z_{i0} = 2, H_i(A_i^\dagger), \mathbf{X}_i^\circ, \Delta_i, i = N_1 + 1, \ldots, N\} = \{Z_{i0} = 2, B_i, A_{i1}, S_{i1}, S_{i2}, \delta_{i1} = 1, \delta_{i2} = 1, \delta_{iD}, \mathbf{X}_i^\circ, \Delta_i, i = N_1 + 1, \ldots, N\}$, where $N = N_1 + N_2$ is the combined sample size. The likelihood contribution from the combined registry data is proportional to

$$\prod_{i=1}^{N} \prod_{k=1}^{2} \left\{ \frac{[P(H_i(A_i^\dagger), X_i | \bar{H}_i(A_{i1}))]^{\Delta_i} [P(H_i(A_i^\dagger) | \bar{H}_i(A_{i1}))]^{1-\Delta_i}}{P(Z_{i0} = k | \bar{H}_i(A_{i1}))} \right\}^{I(Z_{i0}=k)}. \tag{4.2.4}$$

Note that identifiability problems may arise in attempting to maximize this likelihood without external information; for example, mortality rates among disease-free individuals will not be estimable – see Appendix C.1. To address this, we next outline several assumptions on intensity models under maximum likelihood. These assumptions are subsequently relaxed in Section 4.6 and 4.9.

4.2.3 Model, assumption, and likelihood

Let $\lambda_j(a|\bar{H}(a^-), X)$, $j = 0, 1$, and $\gamma_j(a|\bar{H}(a^-), X)$, $j = 0, 1, 2$, denote the general intensity functions for the $j \to j+1$, and $j \to D_j$ transitions. We consider the following assumptions:

Assumption 1. *Given the history up to A_1 and phase I covariates $\mathbf{V}$, the biomarker X is not related to the onset disease process, that is, $\lambda_0(a|X, \bar{H}(a^-)) = \lambda_0(a|\bar{H}(a^-))$*

Assumption 2. *Given the history up to A_1 and phase I covariates $\mathbf{V}$, the biomarker X is not associated with death (i.e. $\gamma_j(a|X, \bar{H}(a^-)) = \gamma_j(a|\bar{H}(a^-))$, $j = 0, 1, 2$).*

The conditional independence of the death intensities from the biomarker X and the death processes might be appropriate in many settings. For example, candidate HLA markers for a disease process may not generally be associated with death. However, the conditional independence between X and the initial disease progression is somewhat a restrictive assumption. It is more reasonable if we assume that $\mathbf{V}$ contains known relevant markers for the $0 \to 1$ and $1 \to 2$ transitions and consider X as a candidate new marker for the $1 \to 2$ transition. More generally, however, Assumption 1 can be relaxed if inverse probability weighting (IPW) is applied to score contributions from the subsample with complete observation on X. We discuss IPW in Section 4.9.

If $\lambda_1(t|A_1, \mathbf{V}, X)$ is the intensity for disease progression (i.e. transition from state 1 to state 2), we assume it takes a multiplicative form of

$$\lambda_1(t|A_1, \mathbf{V}, X; \boldsymbol{\theta}) = \lambda(t; \boldsymbol{\alpha}) \exp(\beta_1 X + \boldsymbol{\beta}_2' \mathbf{V} + \beta_3 \log A_1), \tag{4.2.5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ with $\boldsymbol{\alpha}$ indexing the hazard function $\lambda(\cdot)$ and $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2', \beta_3)'$ representing the regression coefficients. This model accommodates a relation between the age of onset A_1 of the initial disease condition and the risk of a $1 \to 2$ transition. We assume that there is no trend in the covariate distribution of $X|\mathbf{V}$ over calendar time (i.e. $X \perp B|\mathbf{V}$) and let $\boldsymbol{\eta}$ index the covariate model $P(X|\mathbf{V}; \boldsymbol{\eta})$.

Let $\gamma_j(a|\bar{H}(a^-)) = \gamma_j(a|\mathbf{V}; \boldsymbol{\psi}_j)$ denote $j \to D_j$ transition intensities, $j = 1, 2$, and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \boldsymbol{\psi}_2')'$. Let

$$\mathscr{L}_{\boldsymbol{\theta}} = \mathcal{F}_1(S_1|A_1, \mathbf{V}, X; \boldsymbol{\theta}) \lambda_1^{\delta_2}(S_1|A_1, \mathbf{V}, X; \boldsymbol{\theta}) \tag{4.2.6}$$

correspond to the likelihood contribution relevant to the disease history for the $1 \to 2$ transition with $\mathcal{F}_1(S_1|A_1, \mathbf{V}, X; \boldsymbol{\theta}) = \exp(-\int_0^{S_1} \lambda_1(t|A_1, \mathbf{V}, X; \boldsymbol{\theta}) dt)$, $\mathscr{L}_{\boldsymbol{\eta}} = P(X|\mathbf{V}; \boldsymbol{\eta})$, and

$$\mathscr{L}_{\boldsymbol{\psi}} = \mathcal{G}_1(S_1|A_1, \mathbf{V}; \boldsymbol{\psi}_1)[\gamma_1(A_2^\dagger|\mathbf{V}; \boldsymbol{\psi}_1)]^{\delta_D(1-\delta_2)}[\mathcal{G}_2(S_2|A_1, S_1, \mathbf{V}; \boldsymbol{\psi}_2)\gamma_2^{\delta_D}(A^\dagger|\mathbf{V}; \boldsymbol{\psi}_2)]^{\delta_2} \tag{4.2.7}$$

with

$$\mathcal{G}_1(t|A_1,\mathbf{V};\boldsymbol{\psi}_1)=\exp(-\int_{A_1}^{A_1+t}\gamma_1(a|\mathbf{V};\boldsymbol{\psi}_1)da)$$

and

$$\mathcal{G}_2(t|A_1,S_1,\mathbf{V};\boldsymbol{\psi}_2)=\exp(-\int_{A_2}^{A_2+t}\gamma_2(a|\mathbf{V};\boldsymbol{\psi}_1)da).$$

Given $\{Z_0,\bar{H}(A_1)\}$, the likelihood contribution of a recruited individual in the combined prevalent cohort in (4.2.4) becomes

$$\prod_{j=1,2}\left[\left(\frac{\mathscr{L}_{\boldsymbol{\theta}}\mathscr{L}_{\boldsymbol{\eta}}\mathscr{L}_{\boldsymbol{\psi}}}{P(Z_0=j|A_1,\delta_1=1,\mathbf{V})}\right)^{\Delta}\times\left(\frac{E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}};\boldsymbol{\eta}]\mathscr{L}_{\boldsymbol{\psi}}}{P(Z_0=j|A_1,\delta_1=1,\mathbf{V})}\right)^{1-\Delta}\right]^{I(Z_0=j)}, \quad (4.2.8)$$

where the denominator adjusts for the recruitment bias in each disease registry. If $T_0=A_0-A_1$, then $P(Z_0=1|A_1,\delta_1=1,\mathbf{V};\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\psi}_1)$ is expressed as

$$E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1,\mathbf{V},X;\boldsymbol{\theta});\boldsymbol{\eta}]\times\mathcal{G}_1(T_0|A_1,\mathbf{V};\boldsymbol{\psi}_1) \quad (4.2.9)$$

and $P(Z_0=2|A_1,\delta_1=1,\mathbf{V};\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\psi})$ involves a double integral given by

$$\int_0^{T_0}E_{X|\mathbf{V}}[\mathcal{F}_1(t|A_1,\mathbf{V},X;\boldsymbol{\theta});\boldsymbol{\eta}]\times\mathcal{G}_1(t|A_1,\mathbf{V};\boldsymbol{\psi}_1)\mathcal{G}_2(T_0|A_1,S_1=t,\mathbf{V};\boldsymbol{\psi}_2)dt. \quad (4.2.10)$$

4.3 Partial likelihood with independent mortality

In many settings disease processes affect quality of life but have relatively little impact on mortality. In such cases, we can make the following further assumption.

Assumption 3. *Mortality is independent of disease status, that is,* $\gamma_1(a|\mathbf{V})=\gamma_2(a|\mathbf{V})=\gamma(a|\mathbf{V})$.

This condition holds typically for relatively benign diseases and may be reasonable for our motivated psoriasis and psoriatic arthritis example. Assumptions 1-3 will be assumed in what follows unless otherwise indicated. These assumptions are subsequently weakened using inverse probability weighting.

Under Assumptions 1-3, we can construct a simplified partial likelihood without specifying the nuisance mortality models. If $\mathcal{G}(a|\mathbf{V};\boldsymbol{\psi})=\exp(-\int_0^a\gamma(u|\mathbf{V};\boldsymbol{\psi})du)$, we have

$$P(Z_0=1|A_1,\delta_1=1,\mathbf{V},X)=\mathcal{F}_1(T_0|A_1,\mathbf{V},X;\boldsymbol{\theta})\times\frac{\mathcal{G}(A_0|\mathbf{V};\boldsymbol{\psi})}{\mathcal{G}(A_1|\mathbf{V};\boldsymbol{\psi})} \quad (4.3.1)$$

and $P(Z_0 = 2|A_1, \delta_1 = 1, \mathbf{V}, X)$ can be factored as

$$[1 - \mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta})] \times \frac{\mathcal{G}(A_0|\mathbf{V}; \boldsymbol{\psi})}{\mathcal{G}(A_1|\mathbf{V}; \boldsymbol{\psi})}. \tag{4.3.2}$$

The likelihood (4.2.8) then becomes

$$\begin{aligned}&\left[\frac{\{\mathscr{L}_{\boldsymbol{\theta}}\mathscr{L}_{\boldsymbol{\eta}}\}^{\Delta}(E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}])^{1-\Delta}}{E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]}\right]^{I(Z_0=1)} \left[\frac{\{\mathscr{L}_{\boldsymbol{\theta}}\mathscr{L}_{\boldsymbol{\eta}}\}^{\Delta}(E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}])^{1-\Delta}}{1 - E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]}\right]^{I(Z_0=2)} \\ &\times \left[\frac{\mathscr{L}_{\boldsymbol{\psi}}}{\mathcal{G}(A_0|\mathbf{V}; \boldsymbol{\psi})/\mathcal{G}(A_1|\mathbf{V}; \boldsymbol{\psi})}\right]^{I(Z_0\in\{1,2\})}.\end{aligned} \tag{4.3.3}$$

If $\boldsymbol{\psi}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\eta}')'$ are functionally independent, we can maximize $\sum_{i=1}^{N} \log L_{ip}(\boldsymbol{\vartheta})$ to estimate $\boldsymbol{\vartheta}$, where $\log L_{ip}(\boldsymbol{\vartheta})$ is a log partial likelihood corresponding to the first line of (4.3.3), given by

$$\begin{aligned}I(Z_{i0} \in \{1,2\}) \big[\Delta_i(\log \mathscr{L}_{i\boldsymbol{\theta}} + \log \mathscr{L}_{i\boldsymbol{\eta}}) + (1 - \Delta_i) \log E_{X|\mathbf{V}_i}[\mathscr{L}_{i\boldsymbol{\theta}}; \boldsymbol{\eta}]\big] \\ -I(Z_{i0} = 1) \log E_{X|\mathbf{V}_i}[\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}] \\ -I(Z_{i0} = 2) \log(1 - E_{X|\mathbf{V}_i}[\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}]).\end{aligned} \tag{4.3.4}$$

Taking the first derivative of $\log L_{ip}(\boldsymbol{\vartheta})$ with respect to $\boldsymbol{\vartheta}$ we write the partial score vector as $S_{ip}(\boldsymbol{\vartheta}) = \partial \log L_{ip}(\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}$ is given by

$$\begin{aligned}&\Delta I(Z_{i0} \in \{1,2\}) \begin{pmatrix}\mathscr{S}_{i\boldsymbol{\theta}} \\ \mathscr{S}_{i\boldsymbol{\eta}}\end{pmatrix} + (1 - \Delta_i)\frac{I(Z_{i0} \in \{1,2\})}{E_{X|\mathbf{V}_i}[\mathscr{L}_{i\boldsymbol{\theta}}; \boldsymbol{\eta}]} \begin{pmatrix}E_{X|\mathbf{V}_i}[\mathscr{S}_{i\boldsymbol{\theta}}\mathscr{L}_{i\boldsymbol{\theta}}; \boldsymbol{\eta}] \\ E_{X|\mathbf{V}_i}[\mathscr{S}_{i\boldsymbol{\eta}}\mathscr{L}_{i\boldsymbol{\theta}}; \boldsymbol{\eta}]\end{pmatrix} \\ &-\frac{I(Z_{i0} = 1)}{E_{X|\mathbf{V}_i}[\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}]} \begin{pmatrix}E_{X|\mathbf{V}_i}[\partial\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta})/\partial\boldsymbol{\theta}; \boldsymbol{\eta}] \\ E_{X|\mathbf{V}_i}[\mathscr{S}_{i\boldsymbol{\eta}}\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}]\end{pmatrix} \\ &+\frac{I(Z_{i0} = 2)}{(1 - E_{X|\mathbf{V}_i}[\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}])} \begin{pmatrix}E_{X|\mathbf{V}_i}[\partial\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta})/\partial\boldsymbol{\theta}; \boldsymbol{\eta}] \\ E_{X|\mathbf{V}_i}[\mathscr{S}_{\boldsymbol{\eta}}\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}]\end{pmatrix},\end{aligned} \tag{4.3.5}$$

where $\mathscr{S}_{i\boldsymbol{\theta}} = \partial \log \mathscr{L}_{i\boldsymbol{\theta}}/\partial\boldsymbol{\theta}$ and $\mathscr{S}_{i\boldsymbol{\eta}} = \partial \log \mathscr{L}_{i\boldsymbol{\eta}}/\partial\boldsymbol{\eta}$. The observed information matrix is then $I_p(\boldsymbol{\vartheta}) = -N^{-1}\sum_{i=1}^{N} S_{ip}(\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta} = I_{p1}(\boldsymbol{\vartheta}) + I_{p2}(\boldsymbol{\vartheta})$, where

$$I_{p1}(\boldsymbol{\vartheta}) = -N^{-1}\sum_{i=1}^{N}\left\{\Delta_i I(Z_{i0} \in \{1,2\})\frac{\partial}{\partial\boldsymbol{\vartheta}'}\begin{pmatrix}\mathscr{S}_{i\boldsymbol{\theta}} \\ \mathscr{S}_{i\boldsymbol{\eta}}\end{pmatrix} + (1 - \Delta_i)\frac{\partial}{\partial\boldsymbol{\vartheta}'}\left[\frac{I(Z_0\in\{1,2\})}{E_{X|\mathbf{V}_i}[\mathscr{L}_{i\boldsymbol{\theta}}; \boldsymbol{\eta}]}\begin{pmatrix}E_{X|\mathbf{V}_i}[\mathscr{S}_{i\boldsymbol{\theta}}\mathscr{L}_{i\boldsymbol{\theta}}; \boldsymbol{\eta}] \\ E_{X|\mathbf{V}_i}[\mathscr{S}_{i\boldsymbol{\eta}}\mathscr{L}_{i\boldsymbol{\theta}}; \boldsymbol{\eta}]\end{pmatrix}\right]\right\}, \tag{4.3.6}$$

and $I_{p2}(\boldsymbol{\vartheta})$ is

$$\begin{aligned}&N^{-1}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\vartheta}'}\left[\frac{I(Z_{i0} = 1)}{E_{X|\mathbf{V}_i}[\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}]}\begin{pmatrix}E_{X|\mathbf{V}_i}[\partial\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta})/\partial\boldsymbol{\theta}; \boldsymbol{\eta}] \\ E_{X|\mathbf{V}_i}[\mathscr{S}_{\boldsymbol{\eta}}\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}]\end{pmatrix}\right] \\ &-\frac{\partial}{\partial\boldsymbol{\vartheta}'}\left[\frac{I(Z_{i0} = 2)}{(1 - E_{X|\mathbf{V}_i}[\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}])}\begin{pmatrix}E_{X|\mathbf{V}_i}[\partial\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta})/\partial\boldsymbol{\theta}; \boldsymbol{\eta}] \\ E_{X|\mathbf{V}_i}[\mathscr{S}_{\boldsymbol{\eta}}\mathcal{F}_1(T_{i0}|A_{i1}, \mathbf{V}_i, X_i; \boldsymbol{\theta}); \boldsymbol{\eta}]\end{pmatrix}\right].\end{aligned}$$

We let $\hat{\boldsymbol{\vartheta}}$ be the estimator of $\boldsymbol{\vartheta}$ solving the overall score equation. For parametric models, under standard regularity conditions (Boos and Stefanski, 2013), $\sqrt{N}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ is asymptotically normal with mean 0 and asymptotic variance $\mathcal{I}(\boldsymbol{\vartheta}; \boldsymbol{\Omega}, \boldsymbol{\rho})^{-1}$, where $\boldsymbol{\Omega}$ represents the full parameter vector, $\boldsymbol{\rho}$ indexes the phase II selection model, and $\mathcal{I}(\boldsymbol{\vartheta}; \boldsymbol{\Omega}, \boldsymbol{\rho}) = E[I_p(\boldsymbol{\vartheta}); \boldsymbol{\Omega}, \rho]$ is the expected information. Note that

$$\mathcal{I}(\boldsymbol{\vartheta}; \boldsymbol{\Omega}, \boldsymbol{\rho}) = E[I_{p1}(\boldsymbol{\vartheta}); \boldsymbol{\Omega}, \boldsymbol{\rho}] + E[I_{p2}(\boldsymbol{\vartheta}); \boldsymbol{\Omega}] \tag{4.3.7}$$

suggesting that the phase II selection models will only affect the first term.

4.4 Response-dependent phase II sub-sampling

4.4.1 Review of residual-based designs for right-censored data

We here consider the weighted residual-dependent designs of Tao et al. (2020) in the context of our combined prevalent cohort data. The motivation for considering this stems intuitively from the form of $I_{p1}(\boldsymbol{\vartheta})$ in (4.3.6) and the decomposition of the expected information $\mathcal{I}(\boldsymbol{\vartheta}; \boldsymbol{\Omega}, \boldsymbol{\rho})$ in (4.3.7). The matrix $I_{p1}(\boldsymbol{\vartheta})$ takes similar form to the information matrix in conventional failure time analysis. For standard right-censored failure time data under proportional hazards models, Tao et al. (2020) proved that when the incompletely observed covariate is not a strong predictor for the event process of interest (i.e. $\beta_1 = o(1)$) a residual-based two-phase design can be asymptotically optimal for estimating β_1. In such designs, a preliminary proportional hazards model – one not adjusting for the (incompletely observed) expensive covariate – is fitted to the phase I data, and martingale-type residuals are computed based on this preliminary model and associated estimates.

4.4.2 Residual-based designs for combined cohort data

Death from state 1 is a competing event for the time to enter state 2; the gap time between entry times to state 1 and 2 is defined by the sojourn time in state 1 with the occurrence of disease progression. Standard time-to-event analysis as well as the development of optimal designs do not directly apply to the combined prevalent cohort data. Moreover, in *Registry 1*, the sojourn time in state 1 is left-truncated by $T_0 = A_0 - A_1$, while in *Registry 2*, individuals are recruited with PsA so the sojourn time in state 1 is exactly known but right-truncated by T_0.

The expected information matrix $\mathcal{I}(\boldsymbol{\vartheta}; \boldsymbol{\Omega}, \boldsymbol{\rho})$ defined in (4.3.7) is expressible as

$$\begin{pmatrix} \mathcal{I}_{\beta_1\beta_1} & \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ} & \mathcal{I}_{\beta_1\boldsymbol{\eta}} \\ \mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} & \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ} & \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\eta}} \\ \mathcal{I}_{\boldsymbol{\eta}\beta_1} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\theta}_\circ} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} \end{pmatrix}, \tag{4.4.1}$$

where $\boldsymbol{\theta}_\circ = (\boldsymbol{\beta}_2', \beta_3, \boldsymbol{\alpha}')'$ and $\mathcal{I}_{\mathbf{ab}} = -E[\partial^2 \log L_p(\boldsymbol{\vartheta})/\partial \mathbf{a} \partial \mathbf{b}']$. We note that $N^{-1/2}(\hat{\beta}_1 - \beta_1)$ converges in distribution to a normal random variable with mean zero and variance $V_{\beta_1}^{-1}$ with

$$V_{\beta_1} = \left[\mathcal{I}_{\beta_1\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}^{-1}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathcal{I}_{\boldsymbol{\eta}\beta_1}\right]. \tag{4.4.2}$$

If we let $\mu = \beta_1 X + \boldsymbol{\beta}_2' \mathbf{V} + \beta_3 \log A_1$ denote the linear predictor of interest, then $\lambda_1(t|S_0, X, \mathbf{V}; \boldsymbol{\theta}) = \lambda(t; \boldsymbol{\alpha}) \exp(\mu)$ and $M_\mu(\boldsymbol{\theta}) = \delta_2 - \int_0^{S_1} \lambda_1(t|A_1, X, \mathbf{V}; \boldsymbol{\theta})$. When $\beta_1 = o(1)$, V_{β_1} can be decomposed as

$$E_{\mathbf{W}}\Big\{ I(Z_0 \in \{1,2\}) E[\Delta|\mathbf{W}] \mathrm{Var}[M_\mu|\mathbf{W}, \Delta = 1] \mathrm{Var}(X|\mathbf{V}) \Big\}. \tag{4.4.3}$$

and a non-negative component independent of the phase II sub-sampling rules, where $\mathbf{W} = (B, A_1, A^\dagger, \delta_1, \delta_D, \mathbf{V}')'$; see Appendix C.3 for a sketch of the derivation.

Following the argument of Tao et al. (2020), an $o(\beta_1)$ optimal sampling rule involves selecting subjects with the largest and smallest values of $M_\mu \mathrm{SD}(X|\mathbf{V})$ with $\mathrm{SD}(X|\mathbf{V}) = \mathrm{Var}(X|\mathbf{V})^{1/2}$, subject to the budgetary constraint defined in (4.2.3). In the design stage, we first fit a preliminary model

$$\lambda_1^*(t|A_1, \mathbf{V}) = \lambda(t; \boldsymbol{\alpha}_*) \exp(\boldsymbol{\xi}_2' \mathbf{V} + \xi_3 \log A_1) \tag{4.4.4}$$

using the phase I data only. When $\beta_1 = 0$, the preliminary model (4.4.4) coincides with the true model (4.2.5) and the estimator $(0, \hat{\boldsymbol{\alpha}}_*', \hat{\boldsymbol{\xi}}_2', \hat{\boldsymbol{\xi}}_3)'$ approaches to the true $\boldsymbol{\theta}$. By evaluating M_μ at $\boldsymbol{\theta} = (0, \hat{\boldsymbol{\alpha}}_*', \hat{\boldsymbol{\xi}}_2', \hat{\xi}_3)'$, we obtain the estimated "residual" $\hat{M}_\mu$. As in Tao et al. (2020), we sample individuals with extreme values of $\hat{M}_\mu \mathrm{SD}(X|\mathbf{V})$ and call such designs the weighted residual dependent designs (WRES-M_μ). WRES-M_μ samples m_1 subjects who experienced disease progression and had the largest values of $M_\mu \mathrm{SD}(X|\mathbf{V})$ and $m_0 = n - m_1$ subjects who has the smallest values of $M_\mu \mathrm{SD}(X|\mathbf{V})$ where m_1 is determined by maximizing (4.4.3). However the evaluation of $\mathrm{SD}(X|\mathbf{V})$ is not possible based on the phase I sample. Hence we can either i) use an adaptive procedure; or ii) consider an (unweighted) residual dependent sampling scheme (i.e. by treating $\mathrm{SD}(X|\mathbf{V})$ as a constant).

We next describe the details of the sub-sampling based on three practical alternatives to WRES-M_μ, including adaptive WRES-M_μ (AWRES-M_μ), unweighted residual dependent sampling (RES-M_μ), and extreme response dependent sampling (EXT-(δ_2, S_1)). To reflect a budgetary constraint we set the size of phase II sample as n ($\leqslant N$).

4.4.2.1 AWRES-M_μ

Note that $\mathrm{SD}(X|\mathbf{V})$ is inestimable base on phase I data, so it is natural to conduct a two-step adaptive design (McIsaac and Cook, 2015). Here we consider a phase IIa subsample of size n_a ($< n$) selected through simple random sampling where the expensive covariate X is measured in selected individuals. Then based on this subsample we obtain an estimate of $\boldsymbol{\eta}$, denoted by $\hat{\boldsymbol{\eta}}_a$, and evaluate $\mathrm{SD}(X|\mathbf{V};\boldsymbol{\eta})$ at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}_a$ for the entire phase I sample. This enables us to implement WRES-M_μ to guide the selection of the rest of the phase I sample (excluding the phase IIa subsample already chosen) to construct a phase IIb sample of size $n_b = n - n_a$.

4.4.2.2 RES-M_μ

Unweighted residual dependent sampling (RES-M_μ) samples m_1 subjects who had experienced disease progression and the largest values of M_μ and $m_0 = n - m_1$ subjects who have the smallest values of M_μ, where m_1 is determined to maximize (4.4.3) given a constant $\mathrm{Var}(X|\mathbf{V})$.

4.4.2.3 EXT-(δ_2, S_1)

While controlling for other variables, M_μ is a non-increasing function of S_1. Hence extreme values of M_μ connects to extreme responses (i.e., $\delta_2 = 1$ with small S_1 and $\delta_2 = 0$ with large S_1). As a practical alternative, we consider sampling subjects with extreme responses, which does not rely on the fit of preliminary models. Under extreme response dependent sampling (EXT-(δ_2, S_1)), we select $\min\{n/2, \sum_{i=1}^N \delta_{i2}\}$ subjects who had experienced the disease progression with the shortest observed sojourn times and $n - \min\{n/2, \sum_{i=1}^N \delta_{i2}\}$ individuals who had the longest observed sojourn time in state 1 without disease progression.

4.5 Simulation studies

4.5.1 Data generation

In what follows we describe data generation for a single individual but the process is repeated to create the entire sample. Consider a stationary birth process (i.e. no trend over calendar time in the transition intensities or covariate distributions) with $B \sim \mathrm{Unif}(0, 100)$ and set the recruitment date $E_0 = 100$. For simplicity, we consider both X and V as

Bernoulli and assume that $(X, V) \perp B$. We generate V from a Bernoulli distribution with success probability $P(V = 1) = 0.5$ and then simulate $X|V$ from a logistic model with

$$P(X|V; \boldsymbol{\eta}) = \frac{\exp(\eta_0 + \eta_1 V)}{1 + \exp(\eta_0 + \eta_1 V)}, \tag{4.5.1}$$

where $\boldsymbol{\eta} = (\eta_0, \eta_1)'$. By setting the marginal probability of X, $P(X = 1) = 0.1$, and the odds ratio e^{η_1}, we solve for η_0.

We generate possible ages of disease onset and death from a six-state model of Figure 4.1 under proportional intensities with constant baseline intensities $\lambda_0(a|V) = \lambda$ for the $0 \to 1$ transition, $\lambda_1(t|A_1, X, V) = \alpha \exp(\beta_1 X + \beta_2 V + \beta_3 \log A_1)$ for the $1 \to 2$ transition, and $\gamma(a|V) = \gamma$ for the $j \to D_j$ transitions, $j = 0, 1, 2$. Psoriasis is a common skin disease may affect $2 - 3\%$ of the general population and about 30% psoriasis patients develop psoriatic arthritis (Gladman et al., 2005; University Health Network, 2019). The probability that an individual occupies state j at recruitment time is given by $P(Z_0 = j)$ and we set $P(Z_0 = 0) = 0.5$, $P(Z_0 = 1) = 0.01$ and $P(Z_0 = 2) = 0.01$ to solve λ, α and γ.

Random censoring times C are generated from an exponential withdrawal time to give $P(\delta_2 = 1|Z_0 = 1) = r \times 100\%$ with $r = 0.1$ or 0.3, leading to different observed frequency of events in *Registry 1*. We set the size of the birth cohort $\mathcal{N} = 10^6$, the sizes of two registries $N_1 = N_2 = 1000$, and the size of phase II subsample $n = 600$.

4.5.2 Efficiency comparisons

In this section we conduct simulation studies to evaluate the empirical efficiency of WRES-M_μ, AWRES-M_μ, RES-M_μ, and EXT-(δ_2, S_1) for estimating β_1 under the partial likelihood given in (4.3.3). To understand their merits, we compare their performance to other commonly used designs including simple random sampling (SRS) and two-phase stratified designs such as the balanced stratified sampling (BAL).

4.5.2.1 Two-phase stratified designs

For two-phase stratified designs, different stratification strategies are implemented on the phase I data leading to different allocation of phase II subsamples. We let $\bar{S}_1 \in \{1, 2, 3\}$ is a categorical variable generated by the continuous variable S_1 discretized by a vector of cutpoints $(c_1, c_2)'$ corresponding to the empirical tertiles of $\{S_{i1}, i = 1, \ldots, N\}$ and $\bar{S}_1 = k$ if $S_1 \in [c_{k-1}, c_k)$, $k = 1, 2, 3$. We consider five phase I stratification schemes based on (1) Z_0; (2) δ_2; (3) Z_0, δ_2; (4) $\delta_2, \bar{S}_1$; and (5) $Z_0, \delta_2, \bar{S}_1$. The total number of strata then range from 2 to 9. For ease of reference, let Y denote the label of the phase I strata with $Y \in \{1, \ldots, J\}$

and J is the total number of phase I strata. For example, when the phase I data is stratified on Z_0, we obtain two strata with $Y = 1$ corresponding to $Z_0 = 1$ and $Y = 2$ corresponding to $Z_0 = 2$, respectively. Let $M_j = \sum_{i=1}^{N} I(Y_i = j)$ and $n_j = \sum_{i=1}^{N} \Delta_i I(Y_i = j)$.

Balanced sampling (BAL) aims to achieve approximately equal numbers per stratum following phase II selection. Under BAL, we select $n_j = \min\{n/J, M_j\} + n_j^\circ$ subjects from stratum j, where n_j° represents the number of individuals additionally selected from stratum j and is set by iterative balanced sampling among the rest of non-exhausted strata until $\sum_j n_j = n$ is satisfied.

When the phase I data is stratified on $(\delta_2, \bar{S}_1)$ and $J = 6$, an extreme stratified sampling scheme, EXT-$(\delta_2, \bar{S}_1)$, is implemented to oversample subjects from the strata with $(\delta_2, \bar{S}_1) \in \{(1, 1), (0, 3)\}$; these strata are deemed to be "extreme" strata as they contains subjects who have unusual or "extreme" responses. To avoid extremely small selection probability in some strata, we set the lower bound of n_j to be $0.05n$ for each j. This gives $n_j = \min(0.95n/2+n_j^\circ, M_j)$, if stratum j is one of the "extreme" strata; and $\min(0.05N_j+n_j^\circ, M_j)$, otherwise. Again, n_j° is obtained by iterative balanced sampling among the rest of non-exhausted strata until $\sum_j n_j = n$ is satisfied.

4.5.2.2 Results

The "limited memory Broyden-Fletcher-Goldfarb-Shanno" (L-BFGS-B) algorithm is used to obtain the maximum likelihood estimators of $\boldsymbol{\theta}$ using the general-purpose optimizer `optim` in `R` (R Core Team, 2020). Tables 4.1 and 4.2 summarize the estimation results for the log hazard ratios $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ under WRES-M_μ, AWRES-M_μ with $n_a/n = 0.2$, RES-M_μ, and EXT-(δ_2, S_1). A full data analysis (FULL) is conducted as a benchmark, which is possible in simulation studies but infeasible in practice. Under partial likelihood analysis of (4.3.3) the average estimates are all close to the true value of β_1 (not shown). Good agreement is found between empirical standard deviations (ESE) and the average model-based standard errors (ASE). And the empirical coverage probabilities is close to the nominal 95% level. Figure 4.4 and Figure 4.5 show the relative efficiency of five BAL designs (i.e. BAL-Z_0, BAL-δ_2, BAL-(Z_0, δ_2), BAL-$(\delta_2, \bar{S}_1)$, BAL-$(Z_0, \delta_2, \bar{S}_1)$), EXT-$\delta_2, S_1$, RES-$M_\mu$, and WRES-$M_\mu$ comparing to SRS with the phase II sample size of $n = 600$, as SRS is probably the most commonly-used design in practice due to its convenience and simplicity. The relative efficiency (ARE) was defined as 100 times the mean asymptotic variance of $\hat{\beta}_1$ under the SRS design divided by that under each alternative design. For ease of presentation, instead of reporting AREs of all five BAL designs in Figure 4.4 and Figure 4.5, we only present their highest ARE in each setting of $\boldsymbol{\beta}$.

The general conclusion is that residual-dependent designs yield estimators with the highest precision for estimating β_1. In the presented settings RES-M_μ and AWRES-M_μ have similar performances and approximate WRES-M_μ well. In particular, RES-M_μ achieves the efficiency level comparable to that of WRES-M_μ especially when X and V are non-correlated, as expected. As AWRES-M_μ is implemented through a two-stage adaptive procedure where larger values of n_a lead to more accurate estimates of $\text{Var}(X|V)$ but less efficiency gains in the overall design. Hence AWRES-M_μ has better performance with a smaller phase IIa subsample (not shown) but a sufficient sample size n_a is necessary to make reliable estimates for $Var(X|V)$. EXT-$(\delta_2, \bar{S}_1)$ generally outerperformed BAL designs, but for the settings where the effects of auxiliary covariates (i.e. V and $\log A_1$) are substantially large and the observed frequency of events in *Registry 1* is moderate ($r = 0.3$).

4.5.3 A sensitivity study: differential mortality

In many settings, the nondifferential mortality assumption (Assumption 3 in Section 4.3) may be implausible. Here we investigate the impact of violations to this assumption on estimation based on the partial likelihood approach under different designs, we consider

$$\frac{\gamma_2(a|V)}{\gamma_1(a|V)} = \exp(\nu)$$

and conduct a set of simulation studies with $\nu = \log 1.2$, or $\log 1.5$.

A full data analysis is implemented to investigate the impact of violation of Assumption 3 on the estimation and inference on $\boldsymbol{\vartheta}$ using partial likelihood under Assumption 1-3. The results obtained from 1000 replications are summarized in Table C.1 in Appendix C.2. We observe bias in the parameter indexing the baseline hazard but minimal bias on covariate coefficients. The robust sandwich variances (Boos and Stefanski, 2013) are calculated and the corresponding standard erros agree with the empirical standard deviations well, which leads to empirical coverage probabilities close to the nominal 95% level.

We then investigate the performance of the two-phase designs implemented in Section 4.5.2. Table 4.3 presents the finite sample properties for β_1 under these two-phase designs, where the sandwich variance is used. Again the empirical bias of all estimates are relatively small and the corresponding coverage probabilities are all close to the nominal 95% level. The precision of estimators from those designs is similar to the previous settings where $\exp(\nu) = 1$. In particular, residual-dependent designs achieved higher precision comparing to the alternative designs.

4.6 Likelihood with differential mortality

We return to the likelihood construction in (4.2.8) for occasions where the independent mortality assumption (i.e. Assumption 3) is violated because $\gamma_1(a|V;\boldsymbol{\psi}_1) \neq \gamma_2(a|V;\boldsymbol{\psi}_2)$. Under Assumptions 1-2, the log likelihood conditional on $\{Z_{i0}, \bar{H}_i(A_{i1}), i = 1,\ldots,N_1, N_1+1,\ldots,N\}$ is given by

$$\begin{aligned}\ell_c(\boldsymbol{\varphi}) = \sum_{i=1}^{N_1} I(Z_{i0}=1)\Bigg[&\Big\{\Delta_i\left(\log \mathscr{L}_{i\boldsymbol{\theta}} + \log \mathscr{L}_{i\boldsymbol{\eta}}\right) + (1-\Delta_i)\log E_{X|V_i}[\mathscr{L}_{i\boldsymbol{\theta}};\boldsymbol{\eta}] + \log \mathscr{L}_{i\boldsymbol{\psi}}\Big\} \\ &-\Big\{\log E_{X|V_i}[\mathcal{F}_1(T_{i0}|A_{i1},V_i,X;\boldsymbol{\theta});\boldsymbol{\eta}] - \log \mathcal{G}_1(T_{i0}|A_{i1},V_i;\boldsymbol{\psi}_1)\Big\}\Bigg] \\ + \sum_{i=N_1+1}^{N} I(Z_{i0}=2)\Bigg[&\Big\{\Delta_i\left(\log \mathscr{L}_{i\boldsymbol{\theta}} + \log \mathscr{L}_{i\boldsymbol{\eta}}\right) + (1-\Delta_i)\log E_{X|V_i}[\mathscr{L}_{i\boldsymbol{\theta}};\boldsymbol{\eta}] + \log \mathscr{L}_{i\boldsymbol{\psi}}\Big\} \\ &-\log\int_0^{T_{i0}} E_{X|V_i}[\mathcal{F}_1(t|A_{i1},V_i,X;\boldsymbol{\theta});\boldsymbol{\eta}]\mathcal{G}_1(t|A_{i1},V_i;\boldsymbol{\psi}_1)\mathcal{G}_2(T_{i0}|A_{i1},S_1=t,V_i;\boldsymbol{\psi}_2)dt\Bigg],\end{aligned} \tag{4.6.1}$$

where $\boldsymbol{\varphi} = (\beta_1, \boldsymbol{\theta}_\circ', \boldsymbol{\eta}', \boldsymbol{\psi}')'$. Denote the score vector by $S_c(\boldsymbol{\varphi}) = \partial\ell_c(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}$, and let the expected information matrix $\mathcal{I}_c = -E[\partial S_c(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}'] = E[S_c(\boldsymbol{\varphi})S_c'(\boldsymbol{\varphi})]$ be given by

$$\begin{pmatrix} \mathcal{I}_{\beta_1\beta_1} & \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ} & \mathcal{I}_{\beta_1\boldsymbol{\eta}} & \mathcal{I}_{\beta_1\boldsymbol{\psi}} \\ \mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} & \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ} & \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\eta}} & \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\psi}} \\ \mathcal{I}_{\boldsymbol{\eta}\beta_1} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\theta}_\circ} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\psi}} \\ \mathcal{I}_{\boldsymbol{\psi}\beta_1} & \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\theta}_\circ} & \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\eta}} & \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\psi}} \end{pmatrix}, \tag{4.6.2}$$

where $\mathcal{I}_{ab} = -E[\partial^2\ell_c(\boldsymbol{\varphi})/\partial a\partial b']$.

Let $\breve{\boldsymbol{\varphi}}$ denote the estimator of $\boldsymbol{\varphi}$ defined by solving $S_c(\boldsymbol{\varphi}) = \mathbf{0}$. The asymptotic variance of the maximum likelihood estimator $\breve{\beta}_1$ is given by

$$\left[\mathcal{I}_{\beta_1\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}^{-1}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\psi}}\mathcal{I}^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}\mathcal{I}_{\boldsymbol{\psi}\beta_1}\right]^{-1}, \tag{4.6.3}$$

the inverse of which can be again decomposed as (4.4.3) and a non-negative component independent of the phase II designs. This result follows the fact that $\mathcal{I}_{\beta_1\boldsymbol{\psi}}\mathcal{I}^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}\mathcal{I}_{\boldsymbol{\psi}\beta_1}$ does not depend on Δ and a similar derivation outlined in Appendix C.3. Hence the weighted residual dependent designs (WRES-M_μ) described in Section 4.4 still achieves an $o(\beta_1)$-optimal efficiency in this setting with differential mortality. But we note here that such a decomposition is not hold in general when the intensities of transitions to death depends on the biomarker X.

4.7 Simulation studies

Here we report a second set of simulations designed to evaluate the empirical performance of four non-stratified designs proposed in Section 4.4, simple random sampling and six stratified designs described in Section 4.5.2 under the general model for differential mortality. For data generation, we use the same procedure described in Section 4.5.1 but let $\lambda_1(t|A_1, X, V) = (\alpha t)^{1.1} \exp(\beta_1 X + \beta_2 V + \beta_3 \log A_1)$ for the $1 \to 2$ transition, and $\gamma_1(a|V) = (\gamma a)^{1.1}$ and $\gamma_2(a|V) = 1.5\gamma_1(a|V)$ for the $j \to D_j$ transitions, $j = 1, 2$, respectively. We consider $\beta_1 \in \{0.0, 0.3, 0.7\}$, $\beta_2 = 0$, $\beta_3 = 1$, and an administrative censoring time such that $P(\delta_2 = 1|Z_0 = 1) = 0.1$.

In practice, the form of the baseline intensities for the $1 \to 2$ and $j \to D_j$ transitions ($j = 1, 2$) are unknown. Hence we adopt piecewise constant functions to approximate these baseline intensities. In particular, two-piece piecewise constant functions are used with the empirical medians of $\{S_{i1} : \delta_{i2} = 1, i = 1, \ldots, N\}$ and $\{A_i^\dagger : \delta_{iD} = 1, i = 1, \ldots, N\}$ as the corresponding cutpoints, respectively. The results obtained from 1000 replications suggest that the empirical bias of all estimates for $\boldsymbol{\beta}$ are relatively small and that the corresponding empirical standard errors agree with the analytical standard errors well, where the sandwich variance (Boos and Stefanski, 2013) is used. Table 4.4 summarizes the estimated log hazard ratio of X_1 from eleven phase II sub-sampling schemes. The residual-based designs (i.e. RES-M_μ, AWRES-M_μ and WRES-M_μ) appear to be much more efficient than the conventional stratified designs in terms of estimating precision in β_1.

4.8 Markers for psoriatic arthritis in psoriasis

In this section, we illustrate the use of the proposed designs to data determined by pooling samples from the UTPC and UTPAC disease registries. The primary interest is in covariate effects relating important HLA markers to the development of PsA in patients with Ps. Some genetic studies have reported that HLA B27 is associated only with PsA and not with psoriasis (Gladman et al., 1986; Gladman and Farewell, 1995; Hebert et al., 2012; Chandran, 2013). We consider a phase I sample comprised of 670 Ps patients from the UTPC and 1146 PsA patients from the UTPAC, who provide complete information on HLA B27 and dates of birth, recruitment, and Ps onset; see Table 4.5 for a relevant information summary. For illustrative purposes, we treat the marker HLA B27 as an expensive covariate where its measurement is subject to a budgetary constraint, reflected by a restricted phase II sample size $n \leqslant 1816$. For the Ps to PsA transition, we consider the following intensity

model

$$\lambda_1(t|X, S_0; \boldsymbol{\theta}) = \sum_{j=1}^{4} \exp(\alpha_j) I(t \in [b_{j-1}, b_j)) \exp(\beta_1 X + \beta_2 I(S_0 \in [18, 40) + \beta_3 I(S_0 \in [40, \infty)),$$

where b_j are the cutpoints with $(b_0, b_1, b_2, b_3, b_4)' = (0, 10, 20, 40, \infty)'$ and X is the binary HLA B27. The nuisance covariate model is just $P(X; \boldsymbol{\eta}) = \exp(\eta_0)/(1 + \exp(\eta_0))$. Hence the weighted residual-based designs (i.e. WRES-M_μ, AWRES-M_μ) are not considered. In addition, we assume that the death transitions from Ps and PsA are governed by piecewise constant intensities with common cutpoints $(0, 60, 80, \infty)$ but different parameters. Therefore we use the likelihood described in Section 4.6 for the estimation and inference of parameters of interest.

A full data analysis (FULL) in which the known X values are used is conducted as a benchmark. We consider nine different two-phase designs incldluing, simple random sampling (SRS), five different balanced sampling schemes (BAL-Z_0, BAL-δ_2, BAL-Z_0, δ_2, BAL-$(_2, \bar{S}_1)$, BAL-$(Z_0, \delta_2, \bar{S}_1)$), extreme stratified sampling (EXT-$(\delta_2, \bar{S}_1)$), extreme response dependent sampling (EXT-(δ_2, S_1)), extreme residual dependent sampling (RES-M_μ). The results are summarized in Table 4.6. All regression estimates are close to that using the full data. The standard error estimates of HLA B27 under RES-M_μ and EXT-(δ_2, S_1) are smaller than that under other two-phase designs. Among the two-phase stratified designs, the standard error estimates of HLA B27 under BAL designs fluctuates with different choices of phase I stratification strategy while EXT-$(\delta_2, \bar{S}_1)$ had the best performance and achieved comparable efficiency level to the extreme response or residual-dependent designs as n increases. These observations are consistent with the simulation results.

4.9 Two-phase designs via inverse probability weighting

In this section we briefly outline an alternative approach based on inverse probability of observing weights.

4.9.1 Inverse probability weighting

An important alternative to maximum likelihood estimation is to use inverse probability weighted (IPW) complete data score equations. The IPW estimating equation is viewed as an approximation to the score functions under a full data analysis, which is typically

not feasible. Hence, information from a subject who gives complete data represents several potentially missing subjects. As the IPW method is restricted to individuals in the phase II subsample, it avoids the need to model the nuisance covariate assumption. More importantly, the construction of the IPW estimating functions completely relax the conditions Assumptions 1-2 and requires a weaker version of Assumption 3 to use the weighted partial score function. Therefore, the IPW method is more robust but less efficient than the partial likelihood method.

Given $\{Z_0, A_0, A_1, \delta_1, \mathbf{V}, X\}$, the weighted score function is

$$\frac{\Delta}{\pi(\boldsymbol{\rho})}\Big\{I(Z_0=1)\frac{\partial}{\partial\vartheta}\log P(S_1,S_2,\delta_2,\delta_D|Z_0=1,A_0,A_1,\delta_1=1,\mathbf{V},X) + I(Z_0=2)\frac{\partial}{\partial\vartheta}\log P(S_1,S_2,\delta_2=1,\delta_D|Z_0=2,A_0,A_1,\delta_1=1,\mathbf{V},X)\Big\}.$$

Note that $\pi(\boldsymbol{\rho})$ must be bounded away from zero.

The nondifferential mortality assumption required for the IPW approach is weaker than Aassumption 3 in the sense that X can be related to the mortality rates. Under this assumption, a consistent estimator of $\boldsymbol{\varphi} = (\boldsymbol{\theta}', \boldsymbol{\rho}')'$ is obtained by setting

$$U(\boldsymbol{\varphi}) = (U_1'(\boldsymbol{\theta};\boldsymbol{\rho}), U_2'(\boldsymbol{\rho}))' = 0 \tag{4.9.1}$$

where

$$U_1(\boldsymbol{\theta};\boldsymbol{\rho}) = \sum_{i=1}^{N}\frac{\Delta_i}{\pi_i(\hat{\boldsymbol{\rho}})}\Big\{I(Z_{i0}\in\{1,2\})\mathscr{S}_{\boldsymbol{\theta}}(\delta_{i2},S_{i1}|A_{i1},\mathbf{V}_i,X_i;\boldsymbol{\theta}) - I(Z_{i0}=1)\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\mathcal{F}_1(T_{i0}|A_{i1},\mathbf{V}_i,X_i;\boldsymbol{\theta})\right] - I(Z_{i0}=2)\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log(1-\mathcal{F}_1(T_{i0}|A_{i1},\mathbf{V}_i,X_i;\boldsymbol{\theta}))\right]\Big\},$$

with $\mathscr{S}_{\boldsymbol{\theta}}(\delta_{i2},S_{i1}|A_{i1},\mathbf{V}_i,X_i;\boldsymbol{\theta})$ defined as $\mathscr{S}_{\boldsymbol{\theta}}$ in (4.3.5), and

$$U_2(\boldsymbol{\rho}) = \sum_{i=1}^{N}\frac{\Delta_i-\pi_i(\boldsymbol{\rho})}{\pi_i(\boldsymbol{\rho})(1-\pi_i(\boldsymbol{\rho}))}\left[\frac{\partial\pi_i(\boldsymbol{\rho})}{\partial\boldsymbol{\rho}}\right].$$

Let $\tilde{\boldsymbol{\varphi}} = (\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\rho}}')'$ denote the solution to (4.9.1) with "$\sim$" distinguishing these estimators from maximum likelihood estimators of Section 4.3. Under suitable regularity conditions (Robins et al., 1994), $\sqrt{N}(\tilde{\boldsymbol{\varphi}}-\boldsymbol{\varphi})$ is asymptotically normal with mean zero and the variance

$$\Sigma(\boldsymbol{\Omega},\boldsymbol{\rho}) = \mathcal{A}^{-1}(\boldsymbol{\Omega},\boldsymbol{\rho})\mathcal{B}(\boldsymbol{\Omega},\boldsymbol{\rho})\mathcal{A}^{-1}(\boldsymbol{\Omega},\boldsymbol{\rho}), \tag{4.9.2}$$

where $\mathcal{A} = E[-\partial U(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}';\boldsymbol{\Omega},\boldsymbol{\rho}]/N$ and $\mathcal{B} = E[U(\boldsymbol{\varphi})U'(\boldsymbol{\varphi});\boldsymbol{\Omega},\boldsymbol{\rho}]/N$.

4.9.2 Phase II selection: prospective vs retrospective information

The efficiency of a design rule on Δ is not only associated with the likelihood contribution $\mathscr{L}_{\boldsymbol{\theta}}$ (or $\mathscr{S}_{\boldsymbol{\theta}}$) of the disease history for the $1 \to 2$ transition but also related to the truncation types and levels; we use the term "truncation" here to indicate that T_0 measures the left- or right- truncation times of the sojourn times in state 1 for those recruited in two registries, respectively. Heuristically, the phase II selection under IPW would have different preferences from that under ML in terms of the selection between two prevalent cohorts.

To see this, we conduct a set of simulation studies under a simple setting where $\mathbf{V}$ is independent of X and not adjusted in the primary intensity model for the $0 \to 1$ transition, i.e., $\beta_2 = 0$ in and $\eta_1 = 0$ in (4.5.1). The data generation follows the same procedure described in Section 4.5.1. Given the phase I sample pooled from two disease registries, as an illustration we consider the phase I strata defined by the state at recruitment Z_0 and the status of the event of interest δ_2. Then the phase I sample is categorized into three classes $(Z_0, \delta_2) \in \{(1,1),(1,0),(2,1)\}$. Let $N_{jk} = \sum_{i=1}^{N} I(Z_{i0} = j) I(\delta_{i2} = k)$ and $n_{jk} = \sum_{i=1}^{N} \Delta_i I(Z_{i0} = j) I(\delta_{i2} = k)$ denote the sizes of strata in phase I and II samples for $(j,k) = (1,0),(1,1),(2,1)$. A simple class of response-dependent phase II selection model is written as

$$\pi(\boldsymbol{\rho}) = \rho_{00} I(Z_0 = 1) I(\delta_2 = 0) + \sum_{j=1}^{2} \rho_{j1} I(Z_0 = j) I(\delta_2 = 1), \tag{4.9.3}$$

where $\boldsymbol{\rho} = (\rho_{10}, \rho_{11}, \rho_{21})'$ is a 3×1 vector of stratum-specific selection probabilities and $\rho_{jk} = n_{jk}/N_{jk}$ is determined by design.

Figures 4.6 and 4.7 show contour plots of asymptotic standard errors for β_1 estimates under various combinations of (n_{10}, n_{11}, n_{21}) and the phase II subsample size $n = n_{10} + n_{11} + n_{21} = 600$. As can be seen ML and IPW the precision of estimators varies according to phase II allocations in different ways. When the number of events in *Registry 1* (i.e. N_{11}) is small ($r = P(\delta_2 = 1 | Z_0 = 1) = 0.1$), efficiency gains in the ML estimator $\hat{\beta}_1$ first increases as n_{21} (the number of events from *Registry 2*) increases, but then decreases after n_{21} reaches around 40% of $n - n_{11}$ for a specified n_{11} (the number of events selected from *Registry 1*); while increasing n_{21} leads to a higher precision for the IPW estimator $\tilde{\beta}_1$. When $r = 0.3$, for a relative small n_{11} (e.g. $\leqslant 0.1n$), larger n_{21} first results in efficiency improvement then causes efficiency deduction for $\hat{\beta}_1$, however, an increment in n_{21} has much smaller influences on efficiency changes for $\tilde{\beta}_1$. Moreover, when n_{11} is relatively large (e.g. $\geqslant 0.3n$), increasing n_{21} may enlarge the efficiency loss in $\hat{\beta}_1$ while improve estimating precision in $\tilde{\beta}_1$.

It is apparent in Figure 4.8 that different approaches to estimation lead to different optimal selection models. Figure 4.8 presents the optimal selection probabilities for each of three strata defined by (Z_0, δ_2) versus various constraints on the size of the phase II subsample under ML and that under IPW. Optimal stratified selection probabilities are determined by minimizing $\mathcal{I}^{-1}(\boldsymbol{\Omega}, \boldsymbol{\rho})_{[1,1]}$ defined by (4.3.7) under ML or $\Sigma(\boldsymbol{\Omega}, \boldsymbol{\rho})_{[1,1]}$ in (4.9.2) under IPW, while respecting the budgetary constraints. Note that the optimal designs under IPW assign the highest selection probability for stratum $(Z_0, \delta_2) = (1, 1)$ (i.e. $\rho_{11} = n_{11}/N_{11}$) among all three strata, however, ρ_{11} in optimal designs under ML remains rather small.

4.10 Discussion and topics of future research

In this chapter, we investigate two-phase design problems with combined data from two prevalent cohorts providing prospective and retrospective disease progression information, respectively. We start from a six-state model to describe both the disease processes of interest and different observational schemes. We focused on the likelihood method for analysis and carefully discussed the assumptions required on intensity models. Two types of selection bias, recruitment bias in phase I and selection bias in phase II, are adjusted to ensure the validness of the analysis. When adjusting for the selection bias, the design efficiency of phase II allocation appears to be independent of recruitment rules but only related to the likelihood contribution $\mathscr{L}_{\boldsymbol{\theta}}$. Subjects with some unusual observation, including responses and (or) covariates, are usually deemed to be more informative about the event process of interest. This perception has motivated the development of, for example, case-cohort designs (Prentice, 1986) and outcome-dependent sampling (Ding et al., 2017; Lawless, 2018; Zhou et al., 2020). The weighted residual dependent sampling proposed in Tao et al. (2020) is also targeting on those subjects who are unusual in terms of both responses and expensive covariates. We observed that the extension of Tao et al. (2020) in our setting still appears to be highly efficient relative to simple random sampling and two-phase stratified designs in both simulation studies and the data application. And the unweighted or weighted residual, M_μ or $M_\mu \mathrm{SD}(X|\mathbf{V})$, can be understood as a summary measure of the "unusual" extent of an observation in this context.

There are several limitations of our proposed approach. The first limitation is that we focused on the proportional hazard model with specified baseline hazards due to its popularity in many applications, although it covers a wide range of settings and the weakly parametric piecewise constant hazard function provides flexibility. The development of general semiparametric transformation methods is of interest. Secondly, we assume that precise disease history is available upon recruitment. However, this is rather unrealistic

in many prevalent cohort studies due to, for example, recall bias. In practice, the disease (both initial and subsequent) onset times are usually subject to various types of censoring, such as left censoring, interval censoring and so on. Generalizations to accommodate those settings is a direction of future studies.

Moreover, we developed and compared our designs when X is univariate. However the extension of our proposed designs to multivariate covariates setting can be rather straightforward. But multivariate X can bring substantial complexity and challenge in specifying the nuisance covariate model $P(X|\mathbf{V};\boldsymbol{\eta})$. Zeng and Lin (2014) proposed a semiparametric approach to deal with multivariate X through Kernel estimation, however their method is limited to the settings where X is not large in dimension. To this end, an inverse probability weighting (IPW) method briefly discussed in Section 4.9, which does not require specifying the covariate model, may be desirable to pursue (Robins et al., 1994).

Table 4.1: Simulation results based on 1000 simulated samples with $N_1 = N_2 = 1000$ and $n = 600$; $100P(\delta_2 = 1|Z_0 = 1) = 10$: ESE is the empirical standard deviation, ASE is the average model-based standard error and ECP is empirical coverage probability.

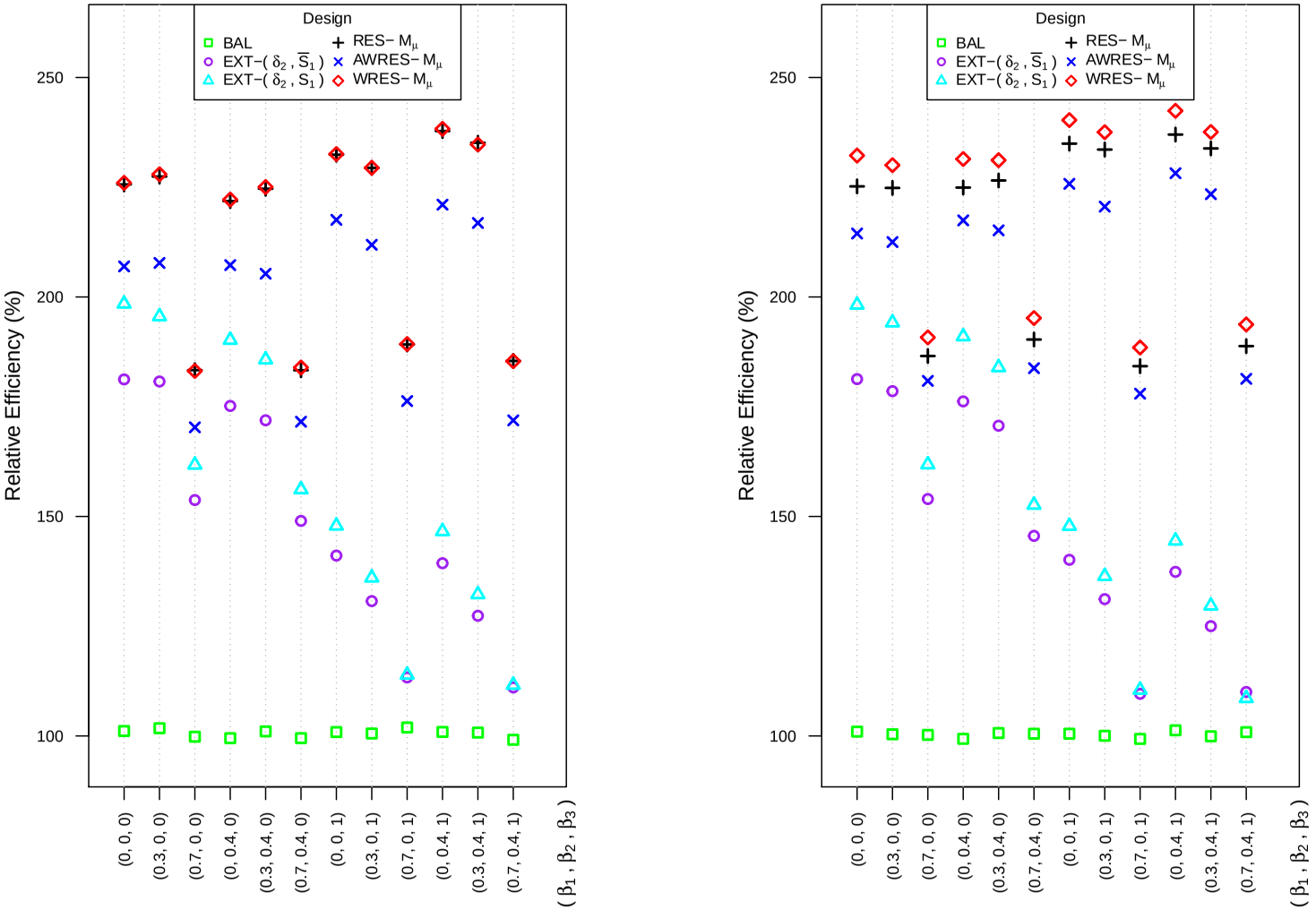
		β_1			β_2			β_3			Phase II Allocation	
$(\beta_1, \beta_2, \beta_3)$	Design	ASE	ESE	ECP	ASE	ESE	ECP	ASE	ESE	ECP	*Registry 1*	*Registry 2*
						(a) $X \perp V$						
(0, 0, 0)	FULL	0.10	0.10	0.94	0.13	0.13	0.96	0.05	0.06	0.93	1000	1000
	WRES-M_μ	0.12	0.13	0.95	0.13	0.13	0.96	0.05	0.06	0.93	299	301
	AWRES-M_μ	0.13	0.14	0.94	0.13	0.13	0.96	0.05	0.06	0.93	299	301
	RES-M_μ*	0.12	0.13	0.95	0.13	0.13	0.96	0.05	0.06	0.94	299	301
	EXT-(δ_2, S_1)*	0.13	0.14	0.95	0.13	0.13	0.96	0.05	0.06	0.94	325	275
(0.7, 0, 0)	FULL	0.09	0.10	0.94	0.13	0.13	0.94	0.05	0.05	0.94	1000	1000
	WRES-M_μ	0.12	0.13	0.94	0.13	0.14	0.95	0.05	0.06	0.94	299	301
	AWRES-M_μ	0.13	0.13	0.94	0.13	0.14	0.95	0.05	0.06	0.94	299	301
	RES-M_μ	0.12	0.12	0.95	0.13	0.14	0.94	0.05	0.06	0.94	299	301
	EXT-(δ_2, S_1)	0.13	0.13	0.95	0.13	0.14	0.94	0.05	0.06	0.94	326	274
(0.3, 0.4, 1)	FULL	0.10	0.10	0.95	0.12	0.12	0.96	0.09	0.10	0.95	1000	1000
	WRES-M_μ	0.11	0.12	0.95	0.12	0.12	0.95	0.09	0.10	0.95	272	328
	AWRES-M_μ	0.12	0.12	0.95	0.12	0.12	0.95	0.09	0.10	0.95	278	322
	RES-M_μ	0.11	0.12	0.95	0.12	0.12	0.95	0.09	0.10	0.94	272	328
	EXT-(δ_2, S_1)	0.15	0.16	0.95	0.12	0.12	0.95	0.09	0.10	0.95	328	272
						(b) $X \not\perp V$						
(0, 0, 0)	FULL	0.10	0.10	0.95	0.13	0.13	0.96	0.05	0.06	0.93	1000	1000
	WRES-M_μ	0.12	0.13	0.94	0.13	0.13	0.96	0.05	0.06	0.93	299	301
	AWRES-M_μ	0.13	0.13	0.95	0.13	0.13	0.96	0.05	0.06	0.94	299	301
	RES-M_μ	0.12	0.13	0.96	0.13	0.13	0.96	0.05	0.06	0.94	299	301
	EXT-(δ_2, S_1)*	0.13	0.14	0.95	0.13	0.13	0.96	0.05	0.06	0.93	324	276
(0.7, 0, 0)	FULL	0.09	0.10	0.94	0.13	0.13	0.95	0.05	0.05	0.96	1000	1000
	WRES-M_μ	0.12	0.12	0.95	0.13	0.14	0.94	0.05	0.05	0.95	298	302
	AWRES-M_μ	0.12	0.12	0.96	0.13	0.13	0.94	0.05	0.05	0.95	299	301
	RES-M_μ	0.12	0.13	0.95	0.13	0.13	0.94	0.05	0.05	0.95	299	301
	EXT-(δ_2, S_1)*	0.13	0.14	0.94	0.13	0.14	0.94	0.05	0.05	0.95	324	276
(0.3, 0.4, 1)	FULL	0.10	0.10	0.95	0.12	0.12	0.96	0.09	0.09	0.94	1000	1000
	WRES-M_μ	0.11	0.11	0.95	0.12	0.12	0.95	0.09	0.09	0.94	268	332
	AWRES-M_μ	0.12	0.12	0.95	0.12	0.12	0.95	0.09	0.09	0.94	275	325
	RES-M_μ	0.11	0.11	0.95	0.12	0.12	0.95	0.09	0.10	0.94	272	328
	EXT-(δ_2, S_1)	0.15	0.15	0.95	0.12	0.12	0.95	0.09	0.10	0.94	330	270

Note. $*$: the number of failed simulations is greater than 10 (within the first 1000 replications); When $X \not\perp V$, OR(X, V) is set as 2; AWRES-M_μ selects a phase IIa subsample of size $n_a = 0.2n$.

Table 4.2: Simulation results based on 1000 simulated samples with $N_1 = N_2 = 1000$ and $n = 600$; $100P(\delta_2 = 1|Z_0 = 1) = 30$.

		β_1			β_2			β_3			Phase II Allocation	
$(\beta_1, \beta_2, \beta_3)$	Design	ASE	ESE	ECP	ASE	ESE	ECP	ASE	ESE	ECP	*Registry 1*	*Registry 2*
						(a) $X \perp V$						
(0, 0, 0)	FULL	0.09	0.10	0.94	0.10	0.09	0.96	0.04	0.04	0.95	1000	1000
	WRES-M_μ	0.11	0.12	0.94	0.10	0.09	0.96	0.04	0.04	0.95	308	292
	AWRES-M_μ	0.12	0.12	0.93	0.10	0.09	0.96	0.04	0.04	0.95	306	294
	RES-M_μ	0.11	0.12	0.94	0.10	0.09	0.96	0.04	0.04	0.95	308	292
	EXT-(δ_2, S_1)	0.12	0.12	0.95	0.10	0.09	0.95	0.04	0.04	0.95	353	247
(0.7, 0, 0)	FULL	0.09	0.09	0.94	0.09	0.10	0.93	0.04	0.04	0.95	1000	1000
	WRES-M_μ*	0.11	0.12	0.95	0.10	0.10	0.94	0.04	0.04	0.96	308	292
	AWRES-M_μ	0.12	0.13	0.94	0.10	0.10	0.94	0.04	0.04	0.96	306	294
	RES-M_μ	0.11	0.12	0.95	0.10	0.10	0.94	0.04	0.04	0.95	308	292
	EXT-(δ_2, S_1)*	0.12	0.12	0.94	0.10	0.10	0.93	0.04	0.04	0.96	355	245
(0.3, 0.4, 1)	FULL	0.09	0.09	0.95	0.09	0.09	0.94	0.06	0.06	0.95	1000	1000
	WRES-M_μ	0.11	0.11	0.95	0.09	0.09	0.94	0.06	0.06	0.95	286	314
	AWRES-M_μ	0.11	0.11	0.95	0.09	0.09	0.94	0.06	0.06	0.94	289	311
	RES-M_μ	0.11	0.11	0.95	0.09	0.09	0.94	0.06	0.06	0.95	286	314
	EXT-(δ_2, S_1)	0.15	0.15	0.95	0.09	0.09	0.93	0.06	0.06	0.94	359	241
						(b) $X \not\perp V$						
(0, 0, 0)	FULL	0.09	0.10	0.95	0.10	0.09	0.96	0.04	0.04	0.95	1000	1000
	WRES-M_μ	0.11	0.12	0.94	0.10	0.09	0.96	0.04	0.04	0.95	309	291
	AWRES-M_μ	0.12	0.12	0.95	0.10	0.09	0.96	0.04	0.04	0.95	306	294
	RES-M_μ	0.11	0.12	0.95	0.10	0.09	0.96	0.04	0.04	0.95	308	292
	EXT-(δ_2, S_1)	0.12	0.12	0.95	0.10	0.09	0.96	0.04	0.04	0.95	353	247
(0.7, 0, 0)	FULL	0.09	0.09	0.95	0.10	0.09	0.95	0.04	0.04	0.95	1000	1000
	WRES-M_μ	0.11	0.11	0.96	0.10	0.10	0.95	0.04	0.04	0.95	308	292
	AWRES-M_μ*	0.12	0.12	0.95	0.10	0.10	0.95	0.04	0.04	0.95	306	294
	RES-M_μ	0.11	0.12	0.95	0.10	0.10	0.95	0.04	0.04	0.96	308	292
	EXT-(δ_2, S_1)	0.12	0.13	0.96	0.10	0.10	0.95	0.04	0.04	0.96	352	248
(0.3, 0.4, 1)	FULL	0.09	0.09	0.94	0.09	0.09	0.95	0.06	0.06	0.94	1000	1000
	WRES-M_μ	0.10	0.11	0.95	0.09	0.09	0.95	0.06	0.06	0.94	280	320
	AWRES-M_μ	0.11	0.11	0.93	0.09	0.09	0.95	0.06	0.06	0.94	285	315
	RES-M_μ	0.10	0.11	0.95	0.09	0.09	0.95	0.06	0.06	0.94	286	314
	EXT-(δ_2, S_1)	0.15	0.16	0.94	0.09	0.09	0.95	0.06	0.06	0.94	359	241

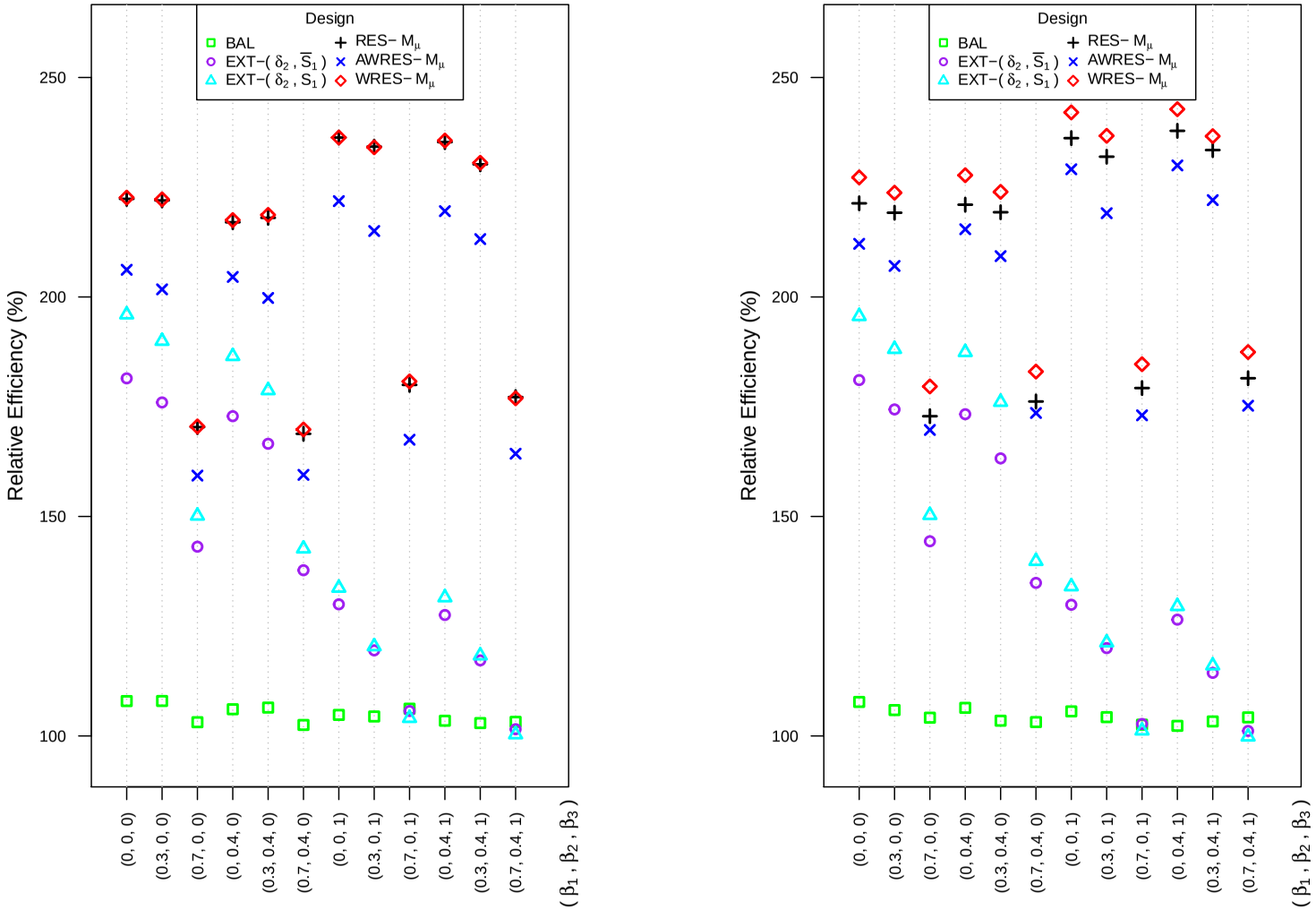
Note. $*$: the number of failed simulations is greater than 10 (within the first 1000 replications); When $X \not\perp V$, $\text{OR}(X, V)$ is set as 2; AWRES-M_μ selects a phase IIa subsample of size $n_a = 0.2n$.



(a) $\text{OR}(X, V) = 1, 100P(\delta_2 = 1|Z_0 = 1) = 10$

(b) $\text{OR}(X, V) = 2, 100P(\delta_2 = 1|Z_0 = 1) = 10$

Figure 4.4: Relative efficiency of other two-phase designs to SRS. The relative efficiency is computed as the mean asymptotic variance of $\hat{\beta}_1$ from SRS divided by that under each of the alternative designs. BAL represents the most efficient one of the five BAL designs in each setting of β. AWRES-M_μ selects a phase IIa subsample of size $n_a = 0.2n$; $N_1 = N_2 = 1000$, $n = 600$.



(a) $\text{OR}(X, V) = 1, 100P(\delta_2 = 1|Z_0 = 1) = 30$

(b) $\text{OR}(X, V) = 2, 100P(\delta_2 = 1|Z_0 = 1) = 30$

Figure 4.5: Relative efficiency of other two-phase designs to SRS. The relative efficiency is computed as the mean asymptotic variance of $\hat{\beta}_1$ from SRS divided by that under each of the alternative designs. BAL represents the most efficient one of the five BAL designs in each setting of β. AWRES-M_μ selects a phase IIa subsample of size $n_a = 0.2n$; $N_1 = N_2 = 1000$, $n = 600$.

Table 4.3: Comparison of two-phase designs in terms of estimated log hazard ratios $\hat{\beta}_1$ under violation of Assumption 3: BIAS is 100×the empirical bias, ESE is the empirical standard deviation, ASE is the average of the robust sandwich standard error and ECP is empirical coverage probability, nsim = 500 with $N_1 = N_2 = 1000$ and $n = 600$; $\text{OR}(X, V) = 2$.

		$100P(\delta_2 = 1\|Z_0 = 1) = 10$								$100P(\delta_2 = 1\|Z_0 = 1) = 30$							
		$\boldsymbol{\beta} = (0,0,0)'$				$\boldsymbol{\beta} = (0.7,0.4,1)'$				$\boldsymbol{\beta} = (0,0,0)'$				$\boldsymbol{\beta} = (0.7,0.4,1)'$			
exp(ν)	Design	BIAS	ASE	ESE	ECP	BIAS	ASE	ESE	ECP	BIAS	ASE	ESE	ECP	BIAS	ASE	ESE	ECP
1.2	SRS	0.35	0.18	0.19	0.95	1.81	0.16	0.17	0.92	0.35	0.16	0.17	0.95	0.01	0.15	0.15	0.93
	BAL-Z_0	1.03	0.18	0.17	0.95	1.83	0.16	0.17	0.94	1.22	0.16	0.16	0.95	-0.13	0.15	0.15	0.95
	BAL-δ_2	-0.15	0.18	0.18	0.96	2.30	0.16	0.17	0.94	0.81	0.16	0.16	0.94	-0.61	0.15	0.15	0.94
	BAL-(Z_0, δ_2)	0.71	0.18	0.19	0.94	0.87	0.16	0.16	0.95	-0.50	0.17	0.17	0.94	-0.61	0.15	0.14	0.96
	BAL-$(\delta_2, S_1)^*$	-0.71	0.19	0.20	0.94	1.81	0.17	0.17	0.93	0.80	0.17	0.18	0.94	-0.90	0.15	0.15	0.94
	BAL-(Z_0, δ_2, S_1)	1.33	0.19	0.21	0.93	1.18	0.17	0.18	0.92	0.83	0.18	0.19	0.93	-0.35	0.15	0.15	0.95
	EXT-$(\delta_2, \bar{S}_1)$	0.41	0.14	0.14	0.95	1.57	0.16	0.17	0.93	0.77	0.12	0.12	0.95	-0.57	0.15	0.15	0.95
	EXT-$(\delta_2, S_1)^*$	0.60	0.13	0.14	0.93	2.07	0.16	0.16	0.94	0.46	0.12	0.12	0.95	0.07	0.15	0.15	0.95
	RES-M_μ	1.15	0.12	0.13	0.95	2.82	0.12	0.13	0.93	1.16	0.11	0.11	0.93	1.78	0.11	0.12	0.93
	AWRES-M_μ	0.80	0.12	0.13	0.94	2.50	0.12	0.13	0.94	1.09	0.11	0.12	0.94	1.50	0.11	0.12	0.95
	WRES-M_μ	0.94	0.12	0.13	0.94	2.82	0.12	0.12	0.95	1.17	0.11	0.12	0.93	1.55	0.11	0.12	0.94
1.5	SRS	0.94	0.17	0.17	0.95	1.76	0.16	0.17	0.93	0.62	0.15	0.14	0.95	-1.67	0.15	0.15	0.92
	BAL-$Z_0{}^*$	1.07	0.17	0.18	0.92	2.75	0.16	0.16	0.92	0.49	0.15	0.16	0.93	-0.50	0.14	0.15	0.94
	BAL-δ_2	0.28	0.17	0.18	0.95	1.99	0.16	0.16	0.94	0.32	0.15	0.15	0.93	-1.02	0.15	0.16	0.93
	BAL-(Z_0, δ_2)	-0.17	0.18	0.18	0.95	2.18	0.16	0.16	0.93	0.05	0.15	0.16	0.94	-1.17	0.14	0.15	0.93
	BAL-(δ_2, S_1)	1.19	0.18	0.17	0.95	2.01	0.17	0.19	0.90	0.23	0.16	0.16	0.93	-1.04	0.15	0.15	0.94
	BAL-(Z_0, δ_2, S_1)	1.17	0.18	0.18	0.94	2.08	0.16	0.17	0.93	-0.69	0.16	0.16	0.94	-1.00	0.15	0.15	0.94
	EXT-$(\delta_2, \bar{S}_1)$	-0.45	0.13	0.14	0.94	3.30	0.16	0.16	0.95	-0.08	0.12	0.12	0.93	0.29	0.15	0.15	0.93
	EXT-(δ_2, S_1)	0.29	0.12	0.13	0.95	3.48	0.16	0.16	0.93	0.48	0.11	0.12	0.95	-0.75	0.15	0.16	0.93
	RES-$M_\mu{}^*$	0.15	0.12	0.12	0.95	4.40	0.12	0.13	0.94	-0.36	0.11	0.11	0.95	2.12	0.11	0.11	0.95
	AWRES-M_μ	0.27	0.12	0.12	0.94	4.41	0.12	0.14	0.91	0.16	0.11	0.11	0.95	2.17	0.11	0.12	0.92
	WRES-M_μ	0.02	0.12	0.12	0.95	4.23	0.12	0.13	0.94	-0.26	0.10	0.11	0.94	2.12	0.11	0.11	0.95

Note. $*$: the number of failed simulations is greater than 5 (within the first 500 replications) for $\boldsymbol{\beta} = (0,0,0)'$ and $100P(\delta_2 = 1|Z_0 = 1) = 10$.

Table 4.4: Comparison of two-phase designs in terms of estimated log hazard ratios $\hat{\beta}_1$ under models accommodating differential mortality: BIAS is 100×the empirical bias, ESE is the empirical standard deviation, ASE is the average robust standard error and ECP is 100×the empirical coverage probability, nsim = 500 with $N_1 = N_2 = 1000$ and $n = 600$; $\text{OR}(X, V) = 2$, $\beta_2 = 0$, $\beta_3 = 1$, $100P(\delta_2 = 1|Z_0 = 1) = 10$.

	$\beta_1 = 0.0$				$\beta_1 = 0.3$				$\beta_1 = 0.7$			
DESIGN	BIAS	ASE	ESE	ECP	BIAS	ASE	ESE	ECP	BIAS	ASE	ESE	ECP
SRS	0.07	0.17	0.17	95.5	-0.75	0.17	0.16	93.1	0.81	0.16	0.16	93.8
BAL-Z_0	0.67	0.18	0.17	94.1	<0.01	0.16	0.16	95.1	1.50	0.16	0.16	93.2
BAL-(δ_2)	0.27	0.18	0.17	93.7	1.08	0.16	0.16	93.9	1.92	0.16	0.16	93.3
BAL-(Z_0, δ_2)	1.03	0.18	0.17	93.0	0.59	0.17	0.16	93.5	1.64	0.16	0.16	93.9
BAL-(δ_2, S_1)	0.25	0.18	0.18	93.6	0.94	0.18	0.17	93.9	1.09	0.17	0.16	93.3
BAL-(Z_0, δ_2, S_1)	0.97	0.17	0.18	95.4	0.67	0.17	0.17	94.8	1.63	0.17	0.16	93.4
EXT-$(\delta_2, \bar{S}_1)$	1.08	0.15	0.15	95.5	1.74	0.15	0.15	94.2	1.43	0.16	0.15	94.2
EXT-(δ_2, S_1)	1.04	0.14	0.14	95.4	1.24	0.15	0.15	94.0	1.89	0.16	0.16	94.0
RES-M_μ	0.31	0.10	0.10	94.6	0.15	0.11	0.11	95.1	2.09	0.12	0.12	94.9
AWRES-M_μ	0.48	0.11	0.10	95.0	0.36	0.11	0.10	95.0	1.79	0.12	0.11	93.5
WRES-M_μ	0.17	0.11	0.11	94.9	-0.05	0.11	0.11	95.1	1.38	0.12	0.12	95.8

Note. An internal pilot of $n_a = 200$ chosen by SRS is used for AWRES-M_μ.

Table 4.5: Summary of analysis data from UTPC and UTPAC as of July 2019.

	UTPC	UTPAC	UTPC + UTPAC
[a]No. of patients	670	1146	1816
No. of Ps → PsA conversions	59	1146	1205
No. of Ps → Death conversions	9	0	9
No. of PsA → Death conversions	1	109	110
No. HLA B27 (positive)	24	194	218
No. Female	295	490	785
[b]Age at recruitment (year)	16, 35, 46, 57, 82	14, 34, 43, 53, 87	
Age at Ps onset (year)	0, 18, 28, 42, 76	0, 18, 26, 38, 84	
Length of follow-up (month)	2, 100, 124, 146, 228	4, 100, 180, 273, 498	

[a] discrete variables are summarized with frequency;

[b] continuous variables are summarized with five quantiles at the minimum, lower quartile, median, upper quartile and the maximum.

Table 4.6: Estimates of parameters associated with the hazard model for the Ps to PsA transition, with average robust standard errors in parentheses, using the combined registry data from UTPC and UTPAC as the phase I sample.

	Covariate			PWC log Hazard				Phase II Allocation	
Design	B27	Age at Ps Onset		α_1	α_2	α_3	α_4	UTPC	UTPAC
		$[18, 40)$	$[40, \infty)$	$[0, 10)$	$[10, 20)$	$[20, 40)$	$[40, \infty)$		
FULL	1.46(0.17)	0.94(0.30)	1.78(0.36)	-6.96(0.28)	-6.26(0.28)	-5.53(0.27)	-4.18(0.26)	670	1146
Phase II Sample Size $n = 600$									
SRS	1.65(0.31)	0.92(0.32)	1.80(0.38)	-6.98(0.29)	-6.27(0.29)	-5.52(0.28)	-4.19(0.27)	222	378
BAL-Z_0	1.53(0.28)	0.93(0.32)	1.78(0.38)	-6.97(0.30)	-6.27(0.29)	-5.52(0.28)	-4.20(0.27)	300	300
BAL-δ_2	1.49(0.27)	0.95(0.32)	1.78(0.38)	-6.99(0.30)	-6.28(0.29)	-5.54(0.28)	-4.21(0.27)	315	285
BAL-(Z_0, δ_2)	1.50(0.29)	0.91(0.31)	1.76(0.38)	-6.94(0.29)	-6.24(0.29)	-5.50(0.28)	-4.17(0.27)	330	270
BAL-$(\delta_2, \bar{S}_1)$	1.67(0.30)	0.91(0.31)	1.81(0.38)	-6.97(0.29)	-6.27(0.29)	-5.52(0.28)	-4.18(0.27)	270	330
BAL-$(Z_0, \delta_2, \bar{S}_1)$	1.66(0.32)	0.89(0.31)	1.80(0.38)	-6.95(0.29)	-6.24(0.28)	-5.49(0.28)	-4.15(0.27)	285	315
EXT-$(\delta_2, \bar{S}_1)$	1.28(0.22)	0.95(0.32)	1.71(0.38)	-6.98(0.30)	-6.27(0.29)	-5.53(0.29)	-4.22(0.27)	290	310
EXT-(δ_2, S_1)	1.28(0.22)	1.01(0.32)	1.69(0.38)	-7.00(0.30)	-6.3(0.30)	-5.55(0.29)	-4.25(0.27)	304	296
EXT-M_μ	1.29(0.20)	0.99(0.32)	1.75(0.38)	-7.02(0.30)	-6.32(0.29)	-5.57(0.28)	-4.24(0.27)	300	300
Phase II Sample Size $n = 900$									
SRS	1.54(0.24)	0.92(0.31)	1.78(0.38)	-6.96(0.29)	-6.26(0.29)	-5.52(0.28)	-4.18(0.27)	333	567
BAL-Z_0	1.44(0.22)	0.93(0.31)	1.76(0.37)	-6.96(0.29)	-6.26(0.29)	-5.52(0.28)	-4.19(0.27)	450	450
BAL-δ_2	1.43(0.21)	0.96(0.31)	1.78(0.37)	-6.99(0.29)	-6.29(0.29)	-5.54(0.28)	-4.21(0.27)	471	429
BAL-(Z_0, δ_2)	1.42(0.22)	0.92(0.31)	1.75(0.37)	-6.95(0.29)	-6.25(0.29)	-5.50(0.28)	-4.18(0.27)	480	420
BAL-$(\delta_2, \bar{S}_1)$	1.59(0.24)	0.90(0.31)	1.80(0.37)	-6.96(0.29)	-6.26(0.28)	-5.51(0.27)	-4.17(0.26)	402	498
BAL-$(Z_0, \delta_2, \bar{S}_1)$	1.59(0.25)	0.89(0.30)	1.80(0.37)	-6.95(0.28)	-6.25(0.28)	-5.50(0.27)	-4.16(0.26)	402	498
EXT-$(\delta_2, \bar{S}_1)$	1.26(0.18)	0.97(0.31)	1.70(0.38)	-6.98(0.29)	-6.28(0.29)	-5.53(0.28)	-4.22(0.27)	411	489
EXT-(δ_2, S_1)	1.34(0.18)	0.99(0.32)	1.73(0.38)	-7.02(0.30)	-6.32(0.29)	-5.58(0.28)	-4.25(0.27)	454	446
EXT-M_μ	1.41(0.18)	1.00(0.32)	1.78(0.38)	-7.04(0.30)	-6.34(0.29)	-5.6(0.28)	-4.26(0.27)	450	450

Note. Estimates and corresponding standard errors in parenthesis are averaged over 1000 replications

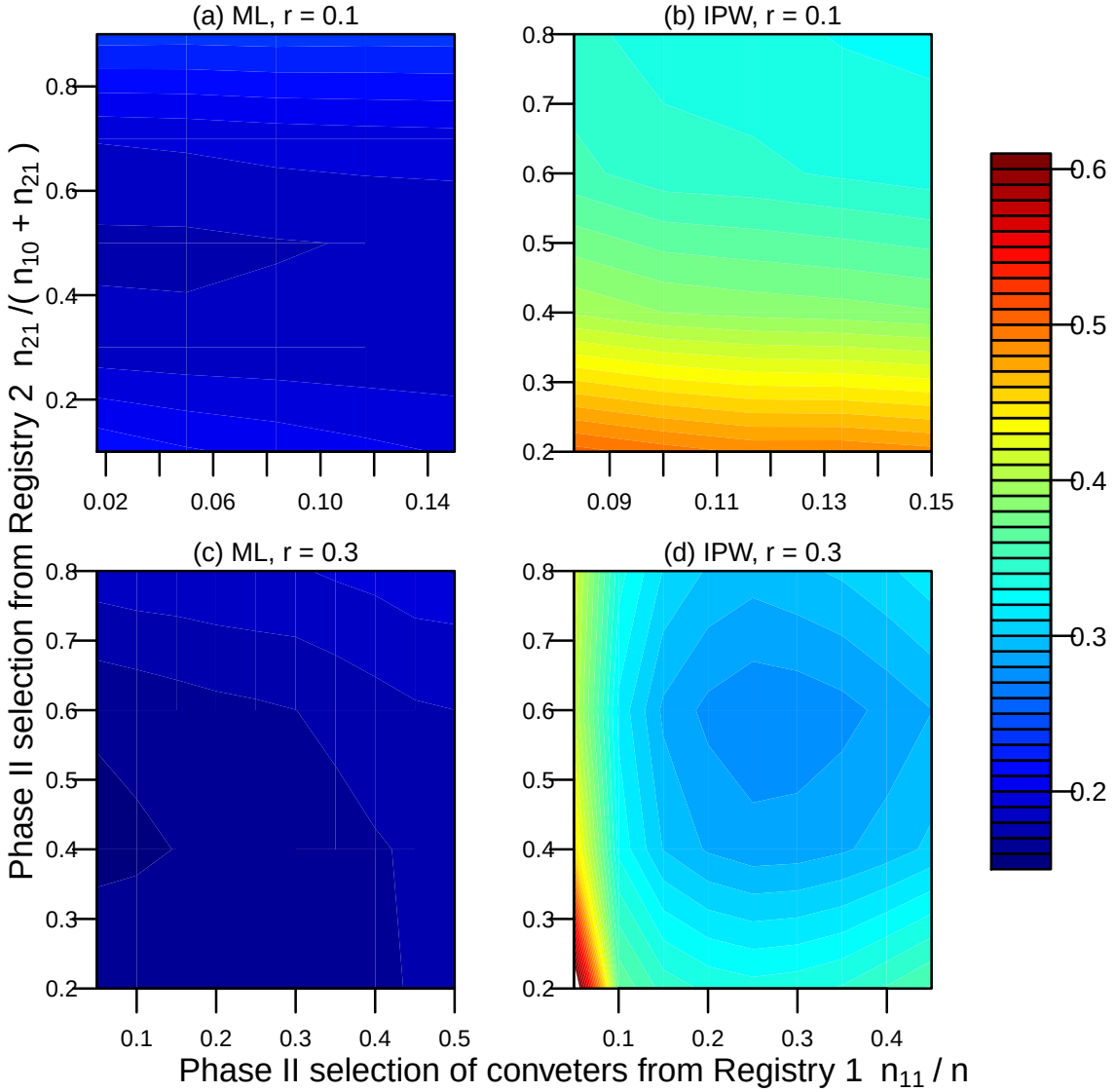


Figure 4.6: Analytical standard error (ASE) of the estimated log hazard ratio for expensive covariate X under maximum likelihood ($\hat{\beta}_1$: (a) and (c)) and inverse probability weighting ($\tilde{\beta}_1$: (b) and (d)), based on 500 samples of $N = 2000$ ($N_1 = N_2 = 1000$) in phase I and $n = 600$ in phase II; $(\beta_1, \beta_2, \beta_3) = (0.2, 0, 0.2)$, $X \perp V$; $n_{11} = \sum_{i=1}^{N_1} \Delta_i \delta_{i2}$ and $n_{21} = \sum_{i=N_1+1}^{N} \Delta_i$; $P(\delta_2 = 1 | Z_0 = 1) = r$.

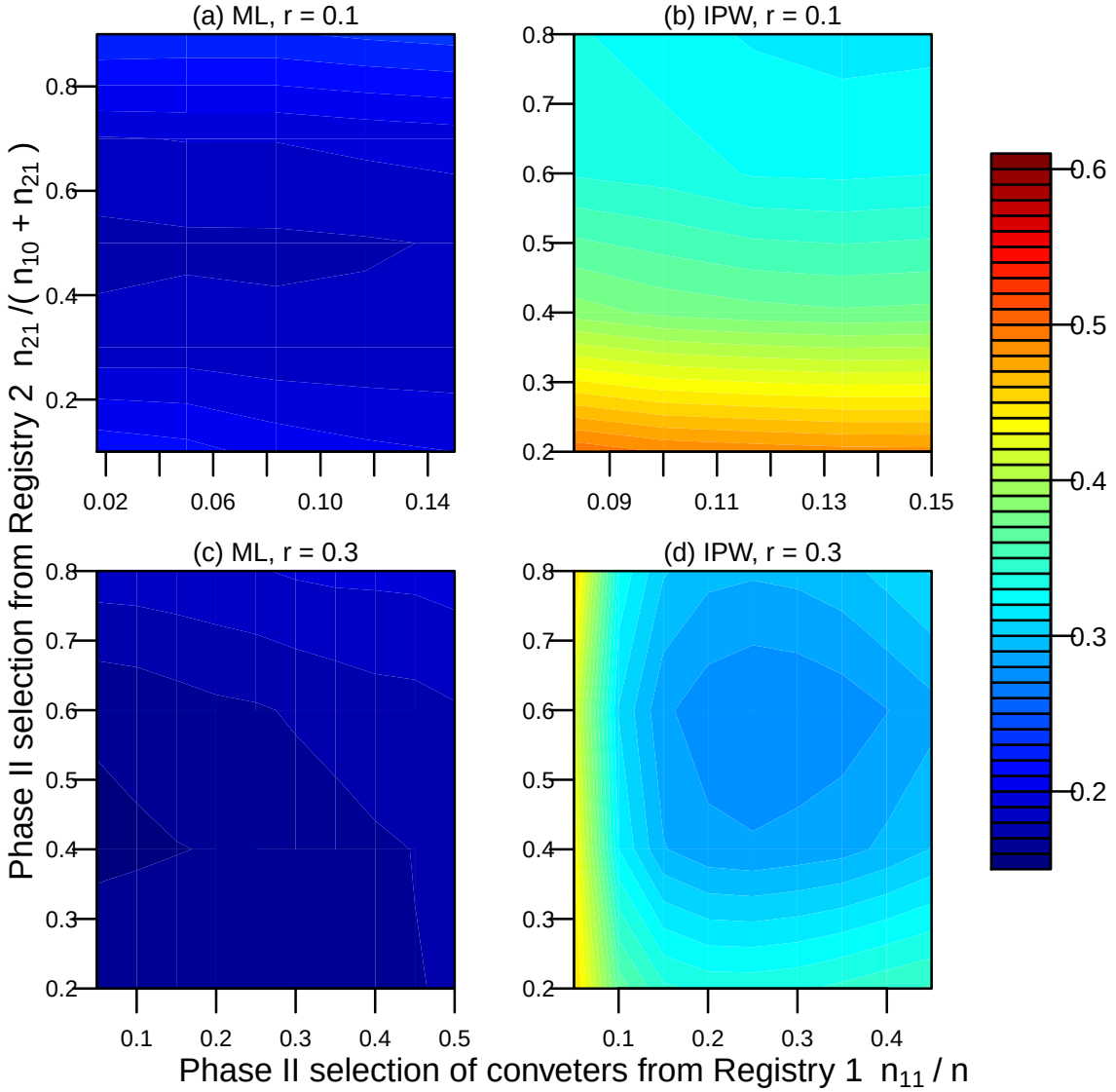
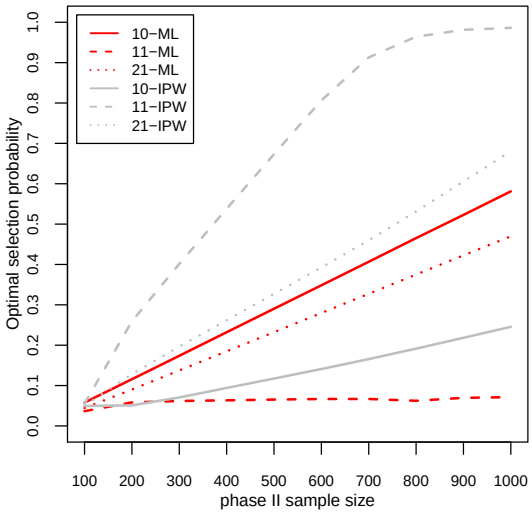
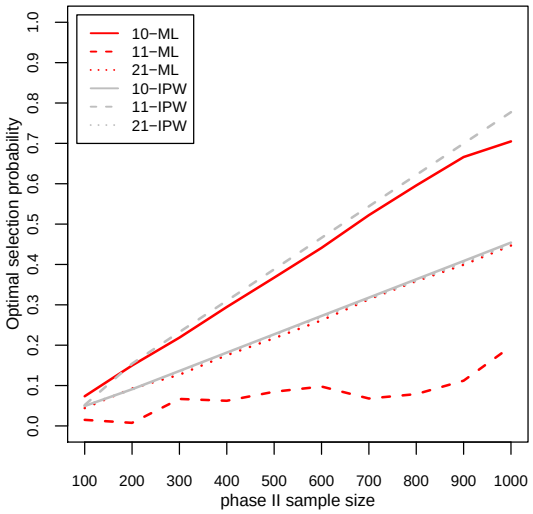


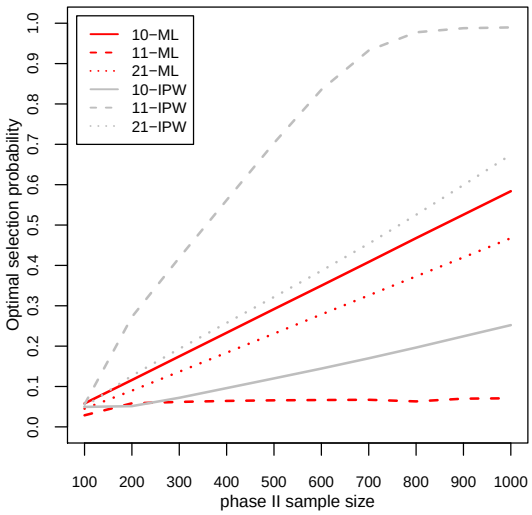
Figure 4.7: Analytical standard error (ASE) of the estimated log hazard ratio for expensive covariate X_1 under maximum likelihood ($\hat{\beta}_1$: (a) and (c)) and inverse probability weighting ($\tilde{\beta}_1$: (b) and (d)), based on 1000 samples of $N = 2000$ ($N_1 = N_2 = 1000$) in phase I and $n = 600$ in phase II; $(\beta_1, \beta_2, \beta_3) = (0.2, 0, 0)$, $X \perp V$; $n_{11} = \sum_{i=1}^{N_1} \Delta_i \delta_{i2}$ and $n_{21} = \sum_{i=N_1+1}^{N} \Delta_i$; $P(\delta_2 = 1 | Z_0 = 1) = r$.



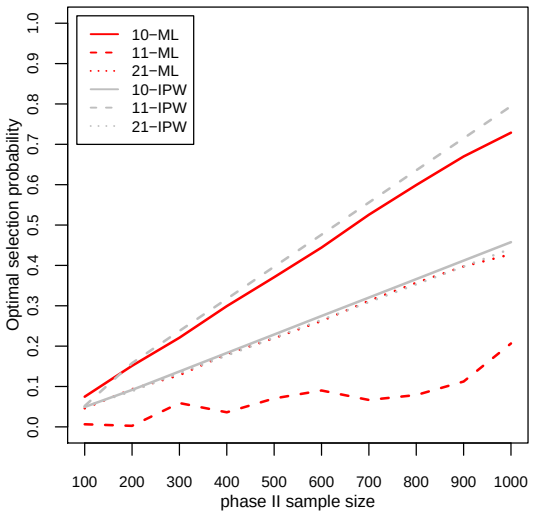
(a) $\beta_3 = 0.0, 100P(\delta_2 = 1|Z_0 = 1) = 10$



(b) $\beta_3 = 0.0, 100P(\delta_2 = 1|Z_0 = 1) = 30$



(c) $\beta_3 = 0.2, 100P(\delta_2 = 1|Z_0 = 1) = 10$



(d) $\beta_3 = 0.2, 100P(\delta_2 = 1|Z_0 = 1) = 30$

Figure 4.8: Optimal stratum-specific selection under maximum likelihood and inverse probability weighting, averaging over 100 phase I samples of $N = 2000$ ($N_1 = N_2 = 1000$); Phase I sample is stratified on (Z_0, δ_2) hence consists of 3 strata: $(1, 1)$, $(2, 1)$ and $(1, 0)$; $(\beta_1, \beta_2, \beta_3) = (0.2, 0, 0)$, $X \perp V$.

Chapter 5

Review and Future Work

5.1 Overview

This thesis has presented new statistical methods addressing challenging problems in the design and analysis of life history studies. In Chapter 2 we concentrate on spatial dependence modeling of multivariate failure time processes subject to nonsusceptibility and intermittent observation. Chapters 3 and 4 deal with two-phase design problems motivated by biomarker studies in life history analysis.

The developments in Chapter 2 are motivated by the need to study patterns of joint damage data in patients with psoriatic arthritis enrolled in University of Toronto Psoriatic Arthritis (PsA) Clinic. To accommodate the fact that most joints do not go on to develop damage, we introduce a binary susceptibility indicator for each joint. To model the spatial dependence patterns of interest, we consider two types of spatial dependencies, one among the susceptibility indicators and the other for failure times given the joint susceptibility. Overall we propose a flexible framework that allows for separate specifications of marginal susceptibility models and damage processes at the joint level, and two types of spatial dependence structures. We adopt a Gaussian copula to describe the dependence structure of failure times and obtain interpretable measures of pairwise associations through Kendall's τ. Composite likelihood is utilized for computational efficiency and robustness in the sense of Varin et al. (2011). Important insights were gained in the nature of the spatial dependence to help in the understanding and diagnosis of psoriatic arthritis.

The next two chapters consider quite different problems related to the design of two-phase studies for incomplete life history data. In Chapter 3 we study two-phase designs with current status data. Maximum likelihood and inverse probability weighting are considered

to deal with missing covariates arising from two-phase designs. Under maximum likelihood, we adapt a small $o(\beta_1)$ optimal design proposed by Tao et al. (2020) to the current status setting and propose three alternative designs for practical implementation. In particular, we note the mathematical relation between score residuals and extreme current status responses, which provides a new perspective to understand the good property of extreme outcome dependent sampling in the current status setting.

In Chapter 4 we consider two-phase design problems when a phase I sample is formed by pooling data from multiple disease registries. Multistate models are adopted to describe the data structure and accommodate various observational patterns including truncation and censoring. Both recruitment (phase I) and selection (phase II) biases are addressed to ensure valid inference. We consider several assumptions regarding the associated intensity functions to construct partial likelihoods for estimation and inference purposes. Under the likelihood framework we develop extreme residual and extreme response dependent designs to improve efficiency over standard stratified designs in the spirit of Tao et al. (2020). We also discuss an alternative inverse probability weighting approach to relax those assumptions and explore related efficient designs.

In the following sections we outline several topics for future research.

5.2 Future research on Chapter 2

5.2.1 Generalized score tests for spatially dependent interval-censored processes

In Chapter 2 we develop a flexible framework to allow regression analysis in 1) susceptibility to damage at the joint level; 2) the hazard for the time-to-damage at joint level; and 3) the spatial dependence structure of the susceptibilities and of the failure time processes of the susceptible joints. Scientific interests may especially reside in screening and identifying the genetic risk factors related to the marginal susceptibilities. Wald and likelihood ratio tests require fitting a full model under the alternative hypothesis, however, score tests only require model fitting under the null. Here we discuss about generalized score test based on the pairwise composite likelihood, for the identification of genetic risk factors in the marginal susceptibility model.

Let $\mathbf{G}$ be a $p_2 \times 1$ vector of genetic markers and $\bar{\mathbf{X}}_{jk} = (\mathbf{X}'_{jk}, \mathbf{G}')'$ represent the extended covariate vector. Let $\bar{\mathbf{X}} = (\bar{\mathbf{X}}'_{jk}; j = 1, \ldots, J, k = 1, \ldots, K)'$ and $\bar{\mathbf{X}}^{(-j,-k)}$ represents the full covariate vector excluding the $\bar{\mathbf{X}}_{jk}$ term. We assume that $Z_{jk} \perp \bar{\mathbf{X}}^{(-j,-k)} | \bar{\mathbf{X}}_{jk}$ and the

marginal model for $Z_{jk}|\bar{\mathbf{X}}_{jk}$ is given by

$$P(Z_{jk}=1|\bar{\mathbf{X}}_{jk};\boldsymbol{\eta}_j)=\frac{\exp(\eta_{j0}+\mathbf{X}'_{jk}\boldsymbol{\eta}_{j1}+\mathbf{G}'\boldsymbol{\eta}_{j2})}{1+\exp(\eta_{j0}+\mathbf{X}'_{jk}\boldsymbol{\eta}_{j1}+\mathbf{G}'\boldsymbol{\eta}_{j2})},$$

where η_{j0} is the baseline joint-type-specific effect, $\boldsymbol{\eta}_{j1}$ is a vector of covariate effect of $\mathbf{X}_{jk}$ and $\boldsymbol{\eta}_{j2}$ is a $p_2\times 1$ parameter vector of genetic effect on the susceptibility of joints of type j, $j=1,\ldots,J$. A common genetic effect across joint types is identified when $\boldsymbol{\eta}_{j2}=\boldsymbol{\eta}_{j'2}$, $j\neq j'=1,\ldots,J$.

We let $\boldsymbol{\eta}_2$ be a $p\times 1$ vector containing all $\boldsymbol{\eta}_{j2}$ with $p=J\cdot r$ if the genetic effect is joint-type-specific and then $\boldsymbol{\eta}_2=(\boldsymbol{\eta}'_{12},\ldots,\boldsymbol{\eta}'_{J2})'$; and $p=r$ if the genetic effect is common across joint types. We consider the null hypothesis of no genetic effect as $H_0:\boldsymbol{\eta}_2=\mathbf{0}$ and the alternative $H_1:\boldsymbol{\eta}_2\neq\mathbf{0}$. Two versions of pseudo-score statistics are naturally constructed and have the usual asymptotic χ^2_p distribution where the degree of freedom p (Molenberghs and Verbeke, 2006).

5.2.1.1 An independence composite generalized score test

We partition $\boldsymbol{\psi}_1=(\boldsymbol{\psi}'_{1\circ},\boldsymbol{\eta}'_2)'$, where $\boldsymbol{\psi}_{1\circ}=(\theta',\boldsymbol{\eta}'_0,\boldsymbol{\eta}'_1)'$ is a $q_1\times 1$ subvector of $\boldsymbol{\psi}_1$ excluding $\boldsymbol{\eta}_2$. Let $U(\boldsymbol{\psi}_1)=(U'_{\boldsymbol{\psi}_{1\circ}},U'_{\boldsymbol{\eta}_2})'$, derived from the first derivative of the log of pairwise composite likelihood $\sum_i\sum_{(j,k)}\log\mathscr{L}_{ijk}$ with respect to $\boldsymbol{\psi}_1=(\boldsymbol{\psi}'_{1\circ},\boldsymbol{\eta}'_2)$, where

$$U_{\boldsymbol{\psi}_{1\circ}}=\frac{\partial}{\partial\boldsymbol{\psi}_{1\circ}}\sum_{i,j,k}\log\mathscr{L}_{ijk}$$

and

$$U_{\boldsymbol{\eta}_2}=\frac{\partial}{\partial\boldsymbol{\eta}_2}\sum_{i,j,k}\log\mathscr{L}_{ijk},$$

where the summation is over $k=1,\ldots,K_j$, $j=1,\ldots,J$, and $i=1,\ldots,N$.

To derive a generalized score test statistic, we follow the guidance of Boos (1992). First we need to find the asymptotic covariance matrix of $U_{\boldsymbol{\eta}_2}$. Without model misspecification, we can derive a score statistic under H_0 as

$$\begin{aligned}\mathcal{T}_U=&\left(U'_{\boldsymbol{\psi}_{1\circ}},\mathbf{0}_{p\times p}\right)\begin{bmatrix}I_{\boldsymbol{\psi}_{1\circ}\boldsymbol{\psi}_{1\circ}} & I_{\boldsymbol{\psi}_{1\circ}\boldsymbol{\eta}_2}\\ I_{\boldsymbol{\eta}_2\boldsymbol{\psi}_{1\circ}} & I_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2}\end{bmatrix}^{-1}\begin{pmatrix}U_{\boldsymbol{\psi}_{1\circ}}\\ \mathbf{0}_{p\times p}\end{pmatrix}\\ =&U'_{\boldsymbol{\psi}_{1\circ}}A_U(\boldsymbol{\psi}_1)^{-1}U_{\boldsymbol{\psi}_{1\circ}},\end{aligned}$$

where $A_U(\boldsymbol{\psi}_1)=I_{\boldsymbol{\psi}_{1\circ}\boldsymbol{\psi}_{1\circ}}-I_{\boldsymbol{\psi}_{1\circ}\boldsymbol{\eta}_2}I^{-1}_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2}I_{\boldsymbol{\eta}_2\boldsymbol{\psi}_{1\circ}}$ with $I_{\boldsymbol{\psi}_{1\circ}\boldsymbol{\psi}_{1\circ}}=-N^{-1}\partial S_{\boldsymbol{\psi}_{1\circ}}/\partial\boldsymbol{\psi}'_{1\circ}$, $I_{\boldsymbol{\psi}_{1\circ}\boldsymbol{\eta}_2}=I'_{\boldsymbol{\eta}_2\boldsymbol{\psi}_{1\circ}}=-N^{-1}\partial U_{\boldsymbol{\psi}_{1\circ}}/\partial\boldsymbol{\eta}'_2$, and $I_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2}=-N^{-1}\partial U_{\boldsymbol{\eta}_2}/\partial\boldsymbol{\eta}'_2$. And under suitable regularity conditions, $\mathcal{T}_U\xrightarrow{d}\chi^2_p$.

To add additional robustness, we construct a generalized score test $\mathcal{T}_{GU}$ statistic following Boos(1992)'s approach, where $A_U(\boldsymbol{\psi}_1)$ is replaced by $\Sigma_U(\boldsymbol{\psi}_1) = C_U B_U(\boldsymbol{\psi}_1)^{-1} C_U'$ of a sandwich form with

$$B_U(\boldsymbol{\psi}_1) = N^{-1} \sum_{i=1}^{N} [U_i(\boldsymbol{\psi}_1) U_i(\boldsymbol{\psi}_1)]$$

and

$$C_U = (\mathbf{I}_{q_1 \times q_1}, -I_{\boldsymbol{\psi}_{1\circ}\boldsymbol{\eta}_2} I_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2}^{-1}),$$

where $\mathbf{I}_{q_1 \times q_1}$ is a $q_1 \times q_1$ identity matrix.

5.2.1.2 A pairwise composite generalized score test

As an alternative to the two-stage estimation procedure, a simultaneous procedure can theoretically improve some estimation efficiency for the stage I parameters $\boldsymbol{\psi}_1$ in a two-stage estimation procedure. Based on pairwise composite likelihood, another pseudo-score test statistic can be derived as follows.

For ease of exhibition, we change the order of the elements in $\boldsymbol{\psi}$ such that $\boldsymbol{\psi} = (\boldsymbol{\psi}_\circ', \boldsymbol{\eta}_2')'$, where $\boldsymbol{\psi}_\circ = (\theta', \rho', \gamma', \boldsymbol{\eta}_0', \boldsymbol{\eta}_1')'$ is a $q_2 \times 1$ subvector of $\boldsymbol{\psi}$ excluding $\boldsymbol{\eta}_2$. Let $S(\boldsymbol{\psi}) = (S_{\boldsymbol{\psi}_\circ}', S_{\boldsymbol{\eta}_2}')'$, derived from the first derivative of the log of pairwise composite likelihood

$$\sum_{i=1}^{N} \sum_{\substack{(j,k),(j',k') \in \bar{\mathcal{S}} \\ (j',k') \succ (j,k)}} \log \mathscr{L}_{ijkj'k'}$$

with respect to $\boldsymbol{\psi} = (\boldsymbol{\psi}_\circ', \boldsymbol{\eta}_2')$, where

$$S_{\boldsymbol{\psi}_\circ} = \frac{\partial}{\partial \boldsymbol{\psi}_\circ} \sum_{i=1}^{N} \sum_{\substack{(j,k),(j',k') \in \bar{\mathcal{S}} \\ (j',k') \succ (j,k)}} \log \mathscr{L}_{ijkj'k'},$$

and

$$S_{\boldsymbol{\eta}_2} = \frac{\partial}{\partial \boldsymbol{\eta}_2} \sum_{i=1}^{N} \sum_{\substack{(j,k),(j',k') \in \bar{\mathcal{S}} \\ (j',k') \succ (j,k)}} \log \mathscr{L}_{ijkj'k'}.$$

Again, we follow the guidance of Boos (1992) to derive a generalized score test statistic. First we need to find the asymptotic covariance matrix of $S_2(\boldsymbol{\psi})$. Without model misspecification, we can derive a score statistic under H_0 as

$$\begin{aligned} \mathcal{T}_S =& \left(S_{\boldsymbol{\psi}_\circ}', \mathbf{0}_{p\times p}\right) \begin{bmatrix} I_{\boldsymbol{\psi}_\circ\boldsymbol{\psi}_\circ} & I_{\boldsymbol{\psi}_\circ\boldsymbol{\eta}_2} \\ I_{\boldsymbol{\eta}_2\boldsymbol{\psi}_\circ} & I_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2} \end{bmatrix}^{-1} \begin{pmatrix} S_{\boldsymbol{\psi}_\circ} \\ \mathbf{0}_{p\times p} \end{pmatrix} \\ =& S_{\boldsymbol{\psi}_\circ}' A_S(\boldsymbol{\psi})^{-1} S_{\boldsymbol{\psi}_\circ}, \end{aligned}$$

where $A_S(\boldsymbol{\psi}) = I_{\psi_\circ\psi_\circ} - I_{\psi_\circ\boldsymbol{\eta}_2} I^{-1}_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2} I_{\boldsymbol{\eta}_2\psi_\circ}$ with $I_{\psi_\circ\psi_\circ} = -N^{-1}\partial S_{\psi_\circ}/\partial \boldsymbol{\psi}'_\circ$, $I_{\psi_\circ\boldsymbol{\eta}_2} = I'_{\boldsymbol{\eta}_2\psi_\circ} = -N^{-1}\partial S_{\psi_\circ}/\partial \boldsymbol{\eta}'_2$, and $I_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2} = -N^{-1}\partial S_{\boldsymbol{\eta}_2}/\partial \boldsymbol{\eta}'_2$. And under suitable regularity conditions, we have $\mathcal{T}_S \xrightarrow{d} \chi^2_p$.

Again, to provide additional robustness, one can construct a generalized score test $\mathcal{T}_{GS}$ statistic following Boos (1992), where $A_S(\boldsymbol{\psi})$ is replaced by $\Sigma_S(\boldsymbol{\psi}) = C_S B_S(\boldsymbol{\psi})^{-1} C'_S$ of a sandwich form with

$$B_S(\boldsymbol{\psi}) = N^{-1} \sum_{i=1}^{N} [S_i(\boldsymbol{\psi}) S'_i(\boldsymbol{\psi})]$$

and

$$C_S = (\mathbf{I}_{q_2 \times q_2}, -I_{\psi_\circ\boldsymbol{\eta}_2} I^{-1}_{\boldsymbol{\eta}_2\boldsymbol{\eta}_2}).$$

5.2.2 Weighted second-order estimating equations for spatial dependent interval-censored processes with nonsusceptibility

The composite likelihood approach employed in Chapter 2 was adopted to avoid the need to work with the full likelihood. Another strategy is to consider second-order generalized estimating equations (GEE2) as these offer a convenient framework for estimation and inference for regression analysis of marginal responses and pairwise associations (Prentice and Zhao, 1991). In the spirit of Jiang and Cook (2020), we consider an adaptation of GEE2 to the present setting involving a latent vector of spatially correlated binary susceptibility indicators $\mathbf{Z}_i$ for individual i in sample with $i = 1, \ldots, N$. Here, we present weighted second-order estimating functions that are constructed as a weighted sum of estimating functions one would use if $\mathbf{Z}_i$ were observed. The weights are dependent on distributional assumptions and specific models for both the failure times and the vector of susceptibility indicators.

Recall that $\boldsymbol{\psi} = (\boldsymbol{\varphi}', \boldsymbol{\vartheta}')'$ denotes the set of all parameters with $\boldsymbol{\varphi} = (\boldsymbol{\eta}', \boldsymbol{\gamma}')'$ the set of regression coefficients associated with the first- and second-order models for $\mathbf{Z}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\rho}')'$ for $\mathbf{T}$, defined based on all pairs of joints. We consider an estimating function of the form

$$U_1(\boldsymbol{\varphi}; \boldsymbol{\psi}) = \sum_{i=1}^{N} \sum_{(j,k,j',k') \in \mathcal{S}_2} U_{ijkj'k'1}(\boldsymbol{\varphi}; \boldsymbol{\psi})$$

for a sample of size N, where $\mathcal{S}_2 = \{(j,k,j',k') : (j,k) \succ (j',k'), k = 1, \ldots, K_j, j = 1, \ldots, J\}$ is an index set of size $K(K-1)/2$ containing all pairwise combinations of the indices of individual joints, with $\succ$ indicating that $j < j'$ or $k < k'$ if $j = j'$. The specific

constructions are of the form

$$U_{ijkj'k'1}(\boldsymbol{\varphi};\boldsymbol{\psi}) = \sum_{\mathbf{z}_{ijkj'k'}\in\mathcal{Z}_2} \varrho_1(\mathbf{z}_{ijkj'k'};\boldsymbol{\psi}) \left[H'_{ijkj'k'1}(\boldsymbol{\varphi}) D^{-1}_{ijkj'k'1}(\boldsymbol{\varphi}) \begin{pmatrix} \mathbf{Z}_{ijkj'k'} - \boldsymbol{\pi}_{ijkj'k'} \\ W_{ijkj'k'} - \omega_{ijkj'k'} \end{pmatrix} \right], \tag{5.2.1}$$

where $\mathbf{Z}_{ijkj'k'} = (Z_{ijk}, Z_{ij'k'})'$ is the pair of indicators for joints (j,k) and (j',k') and $W_{ijkj'k'} = Z_{ijk}Z_{ij'k'}$ is the product. Recall $\pi_{ijk} = E[Z_{ijk}|\mathbf{X}_{ijk}]$ and let $\boldsymbol{\pi}_{ijkj'k'} = (\pi_{ijk}, \pi_{ij'k'})'$. Moreover, if $\mathbf{V}_{ijkj'k'} = (\mathbf{X}_{ijk}, \mathbf{X}_{ij'k'})'$ then let $\omega_{ijkj'k'} = E[W_{ijkj'k'}|\mathbf{V}_{ijkj'k'}]$. The derivative matrix is given by

$$H_{ijkj'k'1}(\boldsymbol{\varphi}) = \begin{pmatrix} \partial\boldsymbol{\pi}_{ijkj'k'}/\partial\boldsymbol{\eta}' & \mathbf{0} \\ \partial\omega_{ijkj'k'}/\partial\boldsymbol{\eta}' & \partial\omega_{ijkj'k'}/\partial\boldsymbol{\gamma}' \end{pmatrix},$$

and

$$D_{ijkj'k'1}(\boldsymbol{\varphi}) = \begin{pmatrix} \mathrm{cov}(\mathbf{Z}_{ijkj'k'}|\mathbf{V}_{ijkj'k'}) & \mathrm{cov}(\mathbf{Z}_{ijkj'k'}, W_{ijkj'k'}|\mathbf{V}_{ijkj'k'}) \\ \mathrm{cov}(W_{ijkj'k'}, \mathbf{Z}_{ijkj'k'}|\mathbf{V}_{ijkj'k'}) & \mathrm{var}(W_{ijkj'k'}|\mathbf{V}_{ijkj'k'}) \end{pmatrix},$$

which may be viewed as a complete data 3×3 covariance matrix. The term in square brackets in (5.2.1) is derived as in Prentice and Zhao (1991) for correlated binary data. Let

$$\varrho_1(\mathbf{z}_{ijkj'k'};\boldsymbol{\psi}) = P(\mathbf{Z}_{ijkj'k'} = \mathbf{z}_{ijkj'k'}|O_{ijkj'k'};\boldsymbol{\psi}) \tag{5.2.2}$$

be the conditional expectations of $\mathbf{Z}_{ijkj'k'}$ given the observed data $O_{ijkj'k'} = \{\mathscr{B}_{ijk}, \mathscr{B}_{ij'k'}, \mathbf{V}_{ijkj'k'}\}$. This estimating function has the same spirit of the observed data score in missing data problems that can be reformulated as the conditional expectations of the complete data score. To compute (5.2.2) however we require estimators of all elements of $\boldsymbol{\psi}$, so to address this we specify a second set of estimating functions.

Let $a_{i0} = 0 < a_{i1} < \ldots < a_{ir_i}$ be the r_i assessment times defining $r_i + 1$ intervals $\mathcal{A}_{il} = [a_{i,l-1}, a_{il})$, $l = 1,\ldots,r_{i+1}$ where $a_{i,r_i+1} = \infty$, for individual i, $i = 1,\ldots,N$. For the failure time process, let $\mathbf{Y}_{ijk} = (Y_{ijk,1},\ldots,Y_{ijk,r_i+1})'$ be an $(r_i+1)\times 1$ vector with elements $Y_{ijk,l} = N_{ijk}(a_{il}) - N_{ijk}(a_{i,l-1})$ indicating that the failure time occurred in $\mathcal{A}_{il}$ for joint (j,k), where $\{N_{ijk}(s), 0 < s\}$ is the counting process for failure of joint (j,k), $l = 1,\ldots,r_i+1$. Let $\boldsymbol{\delta}_{ijk}$ be an $(r_i+1)\times 1$ vector with elements $\delta_{ijk,l} = E[Y_{ijk,l}|Z_{ijk};\boldsymbol{\theta}_j]$.

The set of estimating functions for $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\rho}')$ with $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1,\ldots,\boldsymbol{\theta}'_J)'$ is constructed as

$$U_2(\boldsymbol{\vartheta};\boldsymbol{\psi}) = \sum_{i=1}^{N} \sum_{(j,k,j',k')\in\mathcal{S}_2} U_{ijkj'k'2}(\boldsymbol{\vartheta};\boldsymbol{\psi})$$

with elements

$$U_{ijkj'k'2}(\boldsymbol{\vartheta};\boldsymbol{\psi}) = \sum_{\mathbf{z}_{ijkj'k'}\in\mathcal{Z}_2} \varrho_1(\mathbf{z}_{ijkj'k'};\boldsymbol{\psi}) \left[H'_{ijkj'k'2}(\boldsymbol{\vartheta})\Delta_{ijkj'k'}D^{-1}_{ijkj'k'2}(\boldsymbol{\vartheta}) \begin{pmatrix} \mathbf{Y}_{ijkj'k'} - \boldsymbol{\delta}_{ijkj'k'} \\ M_{ijkj'l'} - \sigma_{ijkj'k'} \end{pmatrix} \right],$$

where $\mathbf{Y}_{ijkj'k'} = (\mathbf{Y}'_{ijk}, \mathbf{Y}'_{ij'k'})'$, $\boldsymbol{\delta}_{ijkj'k'} = (\boldsymbol{\delta}'_{ijk}, \boldsymbol{\delta}'_{ij'k'})'$ and $M_{ijkj'k'} = \mathbf{Y}_{ijk} \otimes \mathbf{Y}_{ij'k'} = (Y_{ijk,1}\mathbf{Y}'_{ij'k'}, \ldots, Y_{ijk,r_i+1}\mathbf{Y}'_{ij'k'})'$ is a $(r_i+1)^2 \times 1$ vector with $\otimes$ denoting the Kronecker product. The parameterized matrices are the derivative matrix

$$H_{ijkj'k'2}(\boldsymbol{\vartheta}) = \begin{pmatrix} \partial\boldsymbol{\delta}_{ijkj'k'}/\partial\boldsymbol{\theta}' & \mathbf{0} \\ \partial\sigma_{ijkj'k'}/\partial\boldsymbol{\theta}' & \partial\sigma_{ijkj'k'}/\partial\boldsymbol{\rho}' \end{pmatrix}, \tag{5.2.3}$$

and the $(r_i+1)(r_i+3) \times (r_i+1)(r_i+3)$ covariance matrix

$$D_{ijkj'k'2}(\boldsymbol{\vartheta}) = \begin{pmatrix} \mathrm{cov}(\mathbf{Y}_{ijkj'k'}|\mathbf{Z}_{ijkj'k'}) & \mathrm{cov}(\mathbf{Y}_{ijkj'k'}, M_{ijkj'k'}|\mathbf{Z}_{ijkj'k'}) \\ \mathrm{cov}(M_{ijkj'k'}, \mathbf{Y}'_{ijkj'k'}|\mathbf{Z}_{ijkj'k'}) & \mathrm{var}(M_{ijkj'k'}|\mathbf{Z}_{ijkj'k'}) \end{pmatrix}.$$

We let

$$\Delta_{ijkj'k'} = \begin{pmatrix} Z_{ijk}\mathbf{I}_{r_i+1} & \mathbf{0}_{(r_i+1)\times(r_i+1)} & \mathbf{0}_{(r_i+1)\times(r_i+1)^2} \\ \mathbf{0}_{(r_i+1)\times(r_i+1)} & Z_{ij'k'}\mathbf{I}_{r_i+1} & \mathbf{0}_{(r_i+1)\times(r_i+1)^2} \\ \mathbf{0}_{(r_i+1)^2\times(r_i+1)} & \mathbf{0}_{(r_i+1)^2\times(r_i+1)} & \mathbf{Z}_{ijkj'k'}\mathbf{I}_{(r_i+1)^2} \end{pmatrix}, \tag{5.2.4}$$

where $\mathbf{I}_r$ is an identity matrix of rank r; the off-diagonal entries of $\Delta_{ijkj'k'}$ are zero.

Let $U_i(\boldsymbol{\psi}) = (U'_{i1}(\boldsymbol{\varphi};\boldsymbol{\psi}), U'_{i2}(\boldsymbol{\vartheta};\boldsymbol{\psi}))'$ denote the estimating functions, we then can obtain an estimate for $\boldsymbol{\psi}$ by setting $U(\boldsymbol{\psi}) = \sum_{i=1}^{N} U_i(\boldsymbol{\psi}) = 0$ and we denote it as $\tilde{\boldsymbol{\psi}}$. Under correct specification of the conditional moments, $U(\boldsymbol{\psi})$ is an unbiased estimating function for $\boldsymbol{\psi}$ and under suitable conditions (Boos and Stefanski, 2013),

$$\sqrt{N}(\tilde{\boldsymbol{\psi}} - \boldsymbol{\psi}) \to \mathrm{MVN}(\mathbf{0}, \bar{\mathbb{A}}^{-1}\bar{\mathbb{B}}\left[\bar{\mathbb{A}}^{-1}\right]'),$$

where $\bar{\mathbb{A}} = E[-\partial U_i(\boldsymbol{\psi})/\partial\boldsymbol{\psi}']$ and $\bar{\mathbb{B}} = E[U_i(\boldsymbol{\psi})U'_i(\boldsymbol{\psi})]$.

5.3 Future research on Chapter 3 and 4

5.3.1 Two-phase designs with cross-sectional samples

In large-scale epidemiological and public health studies, it is resource-intensive and impractical to recruit a random sample in the hope of forming an inception cohort. Other sampling schemes often used include cross-sectional sampling of individuals from the population (Van Es et al., 2000), but it is prohibitively expensive to process biosamples for an entire cohort constructed in this way thus motivates the development and study of two-phase designs (Neyman, 1938). The phase I sample we consider here involves individuals sampled cross-sectionally from a population and followed until the end of study or death. Biosamples are collected and stored in biobanks at the study entry for future studies.

5.3.1.1 A four-state illness-death model and the phase I sample

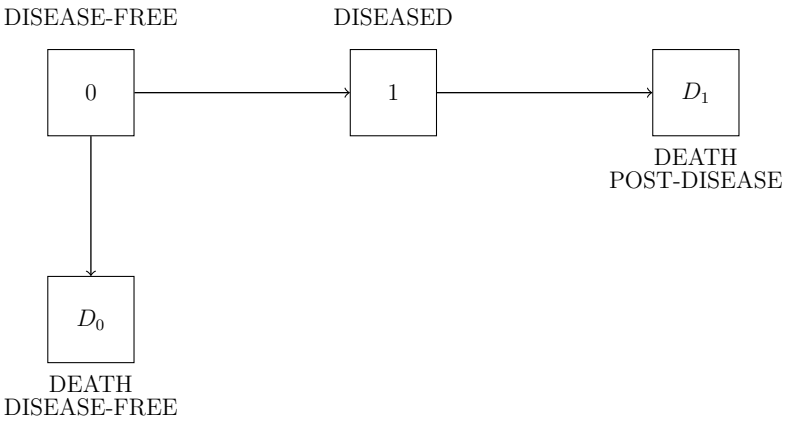


Figure 5.1: A state space diagram for a four-state illness-death model.

We consider a birth cohort of individuals with each generating an irreversible four-state illness-death model involving an initial healthy state 0, a state representing the disease of interest (state 1) and states D_j, representing death from states j, $j = 0, 1$, respectively; see Figure 5.1. Suppose that N individuals are sampled from a birth cohort of size $\mathcal{N}$ ($\mathcal{N} > N$) defined by a window of calendar time $\mathscr{B} = (B_L, B_R]$ at recruitment date S_0 ($S_0 \geqslant B_R$). Let Δ_1 be an indicator of recruitment. For a recruited individual born at time B, we denote the age at recruitment by $A_0 = S_0 - B$. Suppose that information on the life path prior to A_0, denoted by $\mathcal{Z}(A_0) = \{Z(a), 0 < a < A_0\}$, is available; in the context of the illness-death process, when it is required that selected individuals be alive, this would be the disease status and A_1, the age of disease onset among individuals with $Z(A_0) = 1$. Suppose follow-up is planned for C_A years to acquire information on the disease course. Let C_R denote a random time from recruitment to withdrawal and $C = \min\{C_A, C_R\}$. For simplicity we assume that the censoring process is non-informative and independent of the disease process. In terms of age, the prospective follow-up period of an individual is $(A_0, A^\dagger)$, where $A^\dagger = \min\{A_D, A_0 + C\}$ and $\delta_D = I(A^\dagger = A_D)$ indicates that death was observed. To accommodate censoring we let $\bar{\mathcal{Z}}(\infty) = \{Z(a), 0 < a < A^\dagger, A^\dagger, \delta_D\}$ denote the observed life history process where the overbar "$-$" denotes the fact that follow-up may be incomplete due to censoring, and the argument ∞ reflects that we are considering all information that is ultimately available. The Lexis diagram in Figure 5.2 gives a graphical representation indicating the life course of an individual recruited from the birth cohort in state 1 and observed to make a $1 \rightarrow D_1$ transition during follow-up.

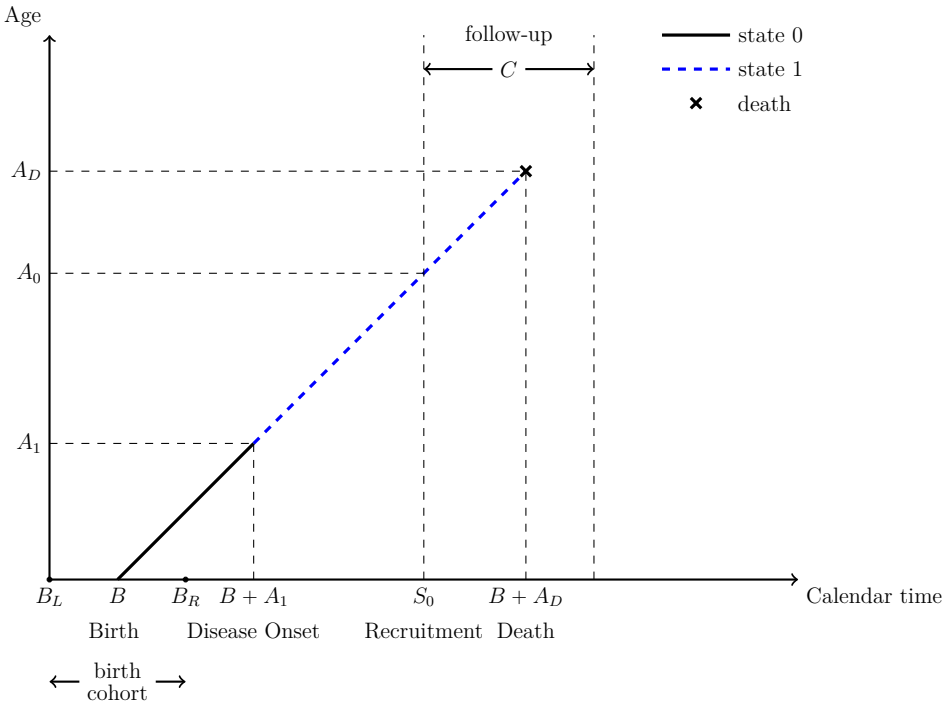


Figure 5.2: A Lexis diagram of a generic individual selected at state 1 with the horizontal axis depicting the calendar times of birth, disease onset, and death; the ages of disease onset are represented on the vertical axis; the administrative follow-up time is also identified.

In counting process notation, let $N_{jj'}(t)$ indicate a $j \to j'$ transition occurred over $(0, t]$ and $dN_{jj'}(t) = I(\text{a } j \to j' \text{ transition occurs at time } t)$. And let $\bar{Y}(t) = I(0 < t \leqslant A^{\dagger})$, $d\bar{N}_{jj'}(t) = \bar{Y}(t)dN_{jj'}(t)$ and $\bar{N}_{jj'}(t) = \int_0^t d\bar{N}_{jj'}(s)$ denote the corresponding counting process for $j \to j'$ transitions. Then $\bar{N}_{\cdot D}(\infty) = \sum_{j=0}^{1} \bar{N}_{jD_j}(\infty) = 1$ if the individual was observed to die and $\bar{N}_{01}(\infty) = 1$ if they made the $0 \to 1$ transition (i.e. developed the disease) while under retrospective or prospective observation.

We consider a partition of $\mathbf{X} = (X_1, \mathbf{X}_2')'$, where X_1 denotes a fixed biomarker which may be ascertained through assay of the stored biological specimens, and $\mathbf{X}_2$ represents the inexpensive auxiliary covariates. This will be the case in many biomarker studies which may rely on biobanks of stored serum to obtain proteomic or genomic data. We assume that $X_1 \perp B|\mathbf{X}_2$ and denote the conditional density for $X_1|\mathbf{X}_2$ as $P(X_1|\mathbf{X}_2)$. Hence, the phase I data contain information on the recruitment age A_0 (or equivalently on the birth date B), the auxiliary covariate information $\mathbf{X}_2$, and the observed life process $\{Z(a), 0 < a < A^{\dagger}\}$.

5.3.1.2 Incomplete covariate data arising from a two-phase design

Two-phase designs offer an appealing framework to facilitate efficient selection of a subsample from recruited individuals to further ascertain X_1 while satisfying budgetary constraints. We let $\Delta_2 = I(X_1 \text{ is observed})$ indicate whether the individual is selected for the phase II subsample to have their covariate X_1 measured. As in all two-phase designs, the sampling probabilities can be controlled through specification of the phase II selection model

$$\pi_2(\boldsymbol{\rho}_2) = P(\Delta_2 = 1 | \bar{\mathcal{Z}}(\infty), A_0, \mathbf{X}_2, \Delta_1 = 1);$$

see Cook and Lawless (2018, Section 7.2). Note that the covariate X_1 is missing at random (Little and Rubin, 2002) and this selection model is expressed in a compatible way in the sense that it covers various phase II sub-sampling schemes. For example, $\pi_2(\boldsymbol{\rho}_2) = P(\Delta_2 = 1 | \Delta_1 = 1) = \tau$ corresponds to phase II simple random sampling, where τ $(0 \leqslant \tau \leqslant 1)$ is fixed.

Let $\mathbf{X}^\circ = \mathbf{X} = (X_1, \mathbf{X}_2')'$ if $\Delta_2 = 1$ and $\mathbf{X}^\circ = \mathbf{X}_2$, otherwise; so data from a recruited individual is denoted by $O = \{\bar{\mathcal{Z}}(\infty), A_0, \mathbf{X}^\circ, \Delta_1 = 1, \Delta_2\}$. When considering the data from all N individuals in the phase I sample, we let i index individuals and write the observed data following phase II sampling as $\{O_i, i = 1, \ldots, N\}$.

5.3.1.3 Maximum Likelihood

Given the observed data $\{O_i, i = 1, \ldots, N\}$, efficient estimation for $\boldsymbol{\vartheta}$ can be approached based on the partial likelihood $L(\boldsymbol{\vartheta}) \propto \prod_{i=1}^N L_i(\boldsymbol{\vartheta})$ with

$$L_i(\boldsymbol{\vartheta}) = \mathcal{L}_i(\boldsymbol{\vartheta})^{\Delta_{i2}} \left[\int \mathcal{L}_i(\boldsymbol{\vartheta}) dX_{i1} \right]^{1-\Delta_{i2}}, \tag{5.3.1}$$

where

$$\mathcal{L}_i(\boldsymbol{\vartheta}) = \frac{P(\bar{\mathcal{Z}}_i(\infty), Z(A_{i0}) \in \{0,1\} | A_{i0}, \mathbf{X}_i; \boldsymbol{\theta}) \cdot P(X_{i1} | \mathbf{X}_{i2}; \boldsymbol{\alpha})}{\int P(Z_i(A_{i0}) \in \{0,1\} | A_{i0}, \mathbf{X}_i; \boldsymbol{\theta}) P(X_{i1} | \mathbf{X}_{i2}; \boldsymbol{\alpha}) dX_{i1}}.$$

The score function based on (5.3.1) is $\mathrm{S}(\boldsymbol{\vartheta}) = \sum_{i=1}^N \mathrm{S}_i(\boldsymbol{\vartheta})$ where $\mathrm{S}_i(\boldsymbol{\vartheta}) = \partial \log L_i(\boldsymbol{\vartheta}) / \partial \boldsymbol{\vartheta}$ and the observed information matrix is

$$I(\boldsymbol{\vartheta}) = -N^{-1} \partial \mathrm{S}(\boldsymbol{\vartheta}) / \partial \boldsymbol{\vartheta}'.$$

The ML estimate $\hat{\boldsymbol{\vartheta}}$ may be found by solving the score equation $\mathrm{S}(\boldsymbol{\vartheta}) = \mathbf{0}$ directly or via an EM algorithm (Dempster et al., 1977) adopting a Turnbull-type likelihood (Turnbull, 1976). The limiting distribution of $\hat{\boldsymbol{\vartheta}}$ is asymptotically

$$\sqrt{N}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}) \xrightarrow{d} \mathrm{MVN}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\Omega}))$$

(Boos and Stefanski, 2013), where $\boldsymbol{\Omega} = (\boldsymbol{\vartheta}', \boldsymbol{\rho}_2', \boldsymbol{\xi}')'$ denotes the full parameter set for $P(\bar{\mathcal{Z}}(\infty), A_0, \mathbf{X}, \Delta_2)$ with $P(A_0, \mathbf{X}_2)$ indexed by $\boldsymbol{\xi}$, and $\mathcal{I}(\boldsymbol{\Omega}) = E[I(\boldsymbol{\vartheta}); \boldsymbol{\Omega}]$ denotes the expected information. Note that the functional dependence of $\mathcal{I}(\boldsymbol{\Omega})$ on $\boldsymbol{\rho}_2$ provides the basis for two-phase designs. In practice, we usually use an empirical estimate to approximate the expected information $\mathcal{I}(\boldsymbol{\Omega})$ by evaluating the information matrix $I(\boldsymbol{\vartheta})$ at the ML estimator $\hat{\boldsymbol{\vartheta}}$.

5.3.1.4 IPW estimating equations with estimated weights

It can be undesirable to specify the nuisance covariate model when X_1 is continuous or high dimensional (Lawless et al., 1999), so a more robust alternative is achieved by restricting attention to individuals with complete data. When $\pi_2(\boldsymbol{\rho}_2)$ is bounded away from zero, a consistent estimator of $\boldsymbol{\theta}$ can be achieved through the use of the following inverse probability weighted (IPW) estimating function

$$U_1(\boldsymbol{\theta}; \boldsymbol{\rho}_2) = \sum_{i=1}^{N} \frac{\Delta_{i2}}{\pi_{i2}(\boldsymbol{\rho}_2)} \mathcal{S}_1(\boldsymbol{\theta}),$$

where

$$\mathcal{S}_1(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \Big[\log P(\bar{\mathcal{Z}}_i(\infty), Z(A_{i0}) \in \{0,1\} | A_{i0}, \mathbf{X}_i; \boldsymbol{\theta}) - \log P(Z(A_{i0}) \in \{0,1\} | A_{i0}, \mathbf{X}_i; \boldsymbol{\theta}) \Big].$$

Note that it is not necessary to model the nuisance covariate distribution for $X_1 | \mathbf{X}_2$ to construct $U_1(\boldsymbol{\theta}; \boldsymbol{\rho}_2)$, so the analysis via IPW estimating equations is potentially more robust however less efficient than analysis via parametric ML.

If $\boldsymbol{\rho}_2$ is treated as known, we can set $U_1(\boldsymbol{\theta}; \boldsymbol{\rho}_2) = 0$ and obtain an IPW estimate of $\boldsymbol{\theta}$. However, the estimation efficiency can be improved if we estimate the weights even when they are known *a priori* (Robins et al., 1994; Lawless et al., 1999). We therefore extend the inverse probability weighted estimating functions with another score function for $\boldsymbol{\rho}_2$, denoted by $U_2(\boldsymbol{\rho}_2)$, giving

$$U(\boldsymbol{\varphi}) = (U_1'(\boldsymbol{\theta}; \boldsymbol{\rho}_2), U_2'(\boldsymbol{\rho}_2))'$$

(Robins et al., 1994), where $\boldsymbol{\varphi} = (\boldsymbol{\theta}', \boldsymbol{\rho}_2')'$. If $\tilde{\boldsymbol{\varphi}}$ is the IPW estimator of $\boldsymbol{\varphi}$, the limiting distribution of $\tilde{\boldsymbol{\varphi}}$ is asymptotically

$$\sqrt{N}(\tilde{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \xrightarrow{d} \text{MVN}(\mathbf{0}, \mathcal{A}^{-1}(\boldsymbol{\Omega}) \mathcal{B}(\boldsymbol{\Omega}) \mathcal{A}^{-1}(\boldsymbol{\Omega})),$$

where $\mathcal{A}(\boldsymbol{\Omega}) = -E[\partial U(\boldsymbol{\varphi}) / \partial \boldsymbol{\varphi}'; \boldsymbol{\Omega}]$ and $\mathcal{B}(\boldsymbol{\Omega}) = E[U(\boldsymbol{\varphi}) U'(\boldsymbol{\varphi}); \boldsymbol{\Omega}]$. The empirical estimates of these expectations are

$$\tilde{\mathcal{A}}(\tilde{\boldsymbol{\varphi}}) = -N^{-1} \frac{\partial U(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}'} \Big|_{\boldsymbol{\varphi} = \tilde{\boldsymbol{\varphi}}} \text{ and } \tilde{\mathcal{B}}(\tilde{\boldsymbol{\varphi}}) = N^{-1} U(\boldsymbol{\varphi}) U'(\boldsymbol{\varphi}) \Big|_{\boldsymbol{\varphi} = \tilde{\boldsymbol{\varphi}}}.$$

5.3.2 Augmentation with incomplete cross-sectional data

5.3.2.1 Identifiability issue with pooled prevalent data

Here we consider the problem of Chapter 4 and the challenge of estimating the full six-state model depicted in Figure 4.1 under the sampling schemes discussed there. Specifically with the data pooled from *Registry 1* and *Registry 2*, there is an identifiability problem that arises if interest lies in estimating the $0 \rightarrow 1$ intensity or the $0 \rightarrow D_0$ intensity in the absence of observations on disease-free individuals. Current status data on the disease state are available from the National Psoriasis Foundation (NPF) survey data (Gelfand et al., 2005) and can be utilized for this purpose, provided data on the phase I covariates $\mathbf{V}$ (e.g. gender, ethnic, etc.) are available. In addition, population mortality data can be used as well if it is available for strata defined by $\mathbf{V}$. Here we discuss such an extension to augmented likelihood incorporating an auxiliary cross-sectional sample.

5.3.2.2 Auxiliary cross-sectional data with incomplete biomarker

Suppose a sample of N_0 individuals are recruited under cross-sectional sampling from the base population; let $\mathcal{R}_\circ$ denote the sample of recruited individuals. Upon recruitment, current status and inexpensive covariate information are collected from these individuals. The data collected from a recruited individual include $\{H(A_0), \mathbf{V}, Z(A_0) \in \{0,1,2\}\}$. Let $i \in \mathcal{R}_\circ$ index a recruited individual, the likelihood contribution from this cross-sectional sample is $L_0(\boldsymbol{\xi}) = \prod_{i\in\mathcal{R}_\circ} L_{i0}(\boldsymbol{\xi})$ with elements

$$L_{i0}(\boldsymbol{\xi}) \propto \frac{P(H_i(A_{i0})|A_{i0}, \mathbf{V}_i)}{\sum_{j=0}^{2} P(Z_i(A_{i0}) = j|A_{i0}, \mathbf{V}_i)}.$$

5.3.2.3 The augmented likelihood

Let $\mathcal{R}_j$ denote the set of individuals selected into registry j with $N_j = ||\mathcal{R}_j||$ the size of the sample ($j = 1, 2$), the conditional likelihood based on the pooled registry data $\{H_i(A_i^\dagger), Z_i(A_{i0}) = j, X_i^\circ, \Delta_i, i \in \mathcal{R}_j, j = 1, 2\}$ is

$$L(\boldsymbol{\xi}) = \prod_{j=1}^{2} \prod_{i\in\mathcal{R}_j} L_{ij}(\boldsymbol{\xi}).$$

The augmented likelihood becomes

$$AL(\boldsymbol{\xi}) \propto \prod_{i\in\mathcal{R}_\circ} L_{i0}(\boldsymbol{\xi}) \prod_{i\in\mathcal{R}_1} L_{i1}(\boldsymbol{\xi}) \prod_{i\in\mathcal{R}_2} L_{i2}(\boldsymbol{\xi}).$$

The mortality rates are best dealt with by using population mortality data if it is available for different strata; otherwise model assumptions in this regard are necessary.

References

C. Anderson-Bergman. icenReg: regression models for interval censored data in R. *Journal of Statistical Software*, 81(12), 2017.

J. Barthelemy and T. Suesse. mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *Journal of Statistical Software*, 86, 2018.

K A Bauer, B I Eriksson, M R Lassen, and A GG Turpie. Fondaparinux compared with enoxaparin for the prevention of venous thromboembolism after elective major knee surgery. *New England Journal of Medicine*, 345(18):1305–1310, 2001.

P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

D.D. Boos. On generalized score tests. *The American Statistician*, 46(4):327–333, 1992.

D.D. Boos and L. Stefanski. *Essential Statistical Inference: Theory and Methods*. Springer, Berlin, 2013.

Ø. Borgan and S. O. Samuelsen. *Handbook of Survival Analysis*, chapter Nested Case-Control and Case-Cohort Studies. CRC Press, Boca Raton London New York, 2014.

Ø Borgan, B Langholz, S O Samuelsen, L Goldstein, and J Pogoda. Exposure stratified case-cohort designs. *Lifetime data analysis*, 6(1):39–58, 2000.

N. E. Breslow and J. A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34:86–102, 2007.

N. E. Breslow, T. Lumley, C.M. Ballantyne, L. E. Chambless, and M. Kulich. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences*, 1:32–49, 2009.

M. Bukhari, M. Lunt, B. J. Harrison, D. G. I. Scott, D. P. M. Symmons, and A. J. Silman. Erosions in inflammatory polyarthritis are symmetrical regardless of rheumatoid factor status: results from a primary care-based inception cohort of patients. *Rheumatology*, 41:246–252, 2002.

K. C. Cain and N. T. Lange. Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, 40(2):493–499, 1984.

V. Chandran. The genetics of psoriasis and psoriatic arthritis. *Clinical reviews in allergy & immunology*, 44(2):149–156, 2013.

V. Chandran, D. C. Tolusso, R. J. Cook, and D. D. Gladman. Risk factors for axial inflammatory arthritis in patients with psoriatic arthritis. *The Journal of rheumatology*, 37(4):809–815, 2010.

V. Chandran, L. Stecher, V. Farewell, and D. D. Gladman. Patterns of peripheral joint involvement in psoriatic arthritis - symmetric, ray and/or row? *Seminars in Arthritis and Rheumatism*, 48(3):430–435, 2018.

N. Chatterjee and J. Shih. A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics*, 57:779–786, 2001.

N. Chatterjee, Y. Chen, and N. Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168, 2003.

C. Chen, T. Lu, M. Chen, and C. Hsu. Semiparametric transformation models for current status data with informative censoring. *Biometrical journal*, 54(5):641–656, 2012.

K. Chen and S. Lo. Case-cohort and case-control analysis with Cox's model. *Biometrika*, 86(4):755–764, 1999.

T. Chen and T. Lumley. Optimal multiwave sampling for regression modeling in two-phase designs. *Statistics in Medicine*, 00(0):1–10, 2020.

D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–145, 1978.

Canadian Longitudinal Study on Aging CLSA. The canadian longitudinal study on aging scientific executive summary. https://clsa-elcv.ca/doc/1090, 2015.

D. Commenges. Multi-state models in epidemiology. *Lifetime data analysis*, 5(4):315–327, 1999.

D. Commenges. Inference for multi-state models from interval-censored data. *Statistical methods in medical research*, 11(2):167–182, 2002.

R. Cook and D. Tolusso. Second-order estimating equations for the analysis of clustered current status data. *Biostatistics*, 10(4):756–772, 2009.

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

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

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

R.J. Cook, B. J. G. White, G. Y. Yi, K. Lee, and T. E. Warkentin. Analysis of a nonsusceptible fraction with current status data. *Statistics in Medicine*, 27:2715–2730, 2008.

A. J. Copas and V. T. Farewell. Incorporating retrospective data into an analysis of time to illness. *Biostatistics*, 2(1):1–12, 2001.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

L. Cresswell and V. Farewell. Assessment of joint symmetry in arthritis. *Statistics in Medcine*, 30:973–983, 2011.

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)*, 39(1):1–22, 1977.

A. Derkach. Statistical methodologies for genetic association studies with rare variants. *Unpublished Ph.D. Dissertation*, 2014.

l. D. Diamond, J. W. McDonald, and I. H. Shah. Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography*, pages 607–620, 1986.

J. Ding, T. Lu, J. Cai, and H. Zhou. Recent progresses in outcome-dependent sampling with failure time data. *Lifetime Data Analysis*, 23:57–82, 2017.

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, 2011.

L. Eder, V. Chandran, F. Pellet, S. Shanmugarajah, C. F. Rosen, S. B. Bull, and D. D. Gladman. Human leucocyte antigen risk alleles for psoriatic arthritis among patients with psoriasis. *Annals of the rheumatic diseases*, 71(1):50–55, 2012.

B I Eriksson, K A Bauer, M R Lassen, and A GG Turpie. Fondaparinux compared with enoxaparin for the prevention of venous thromboembolism after hip-fracture surgery. *New England Journal of Medicine*, 345(18):1298–1304, 2001.

O. Espin-Garcia, R. V. Craiu, and S. B. Bull. Two-phase designs for joint quantitative-trait-dependent and genotype-dependent sampling in post-GWAS regional sequencing. *International Genetic Epidemiology Society*, 42(1):104–116, 2017.

V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046, 1982.

V. T. Farewell and L. Su. A multistate model for events defined by prolonged observation. *Biostatistics*, 12(1):102–111, 2011.

M. J. Frank. On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae*, 19:194–226, 1979.

M. Friedman. Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1):101–112, 1982.

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):573–577, 2005.

D. Gladman and V. Chandran. Observational cohort studies: lessons learnt from the university of toronto psoriatic arthritis program. *Rheumatology*, 50(1):25–31, 2011.

D. D. Gladman and V. Chandran. Observational cohort studies: lessons learnt from the University of Toronto Psoriatic Arthritis Program. *Rheumatology*, 50(1):25–31, 2010.

D. D. Gladman and V. T. Farewell. The role of HLA antigens as indicators of disease progression in psoriatic arthritis. *Arthritis and Rheumatism*, 38:845–850, 1995.

D. D. Gladman, K. A. Anhorn, R. K. Schachter, and H. Mervart. HLA antigens in psoriatic arthritis. *The Journal of rheumatology*, 13(3):586–592, 1986.

D. D. Gladman, R. Shuckett, M. L. Russell, J. C. Thorne, and R. K. Schachter. Psoriatic Arthritis (PsA) - an analysis of 220 patients. *Quarterly Journal of Medicine*, 62(238): 127–141, 1987.

D. D. Gladman, C. Antoni, P. Mease, D. O. Clegg, and P. Nash. Psoriatic arthritis: epidemiology, clinical features, course, and outcome. *Annals of the Rheumatic Diseases*, 24(suppl 2):ii14–ii17, 2005.

M. Goodman, Y. Li, and R. Tiwari. Detecting multiple change points in piecewise constant hazard functions. *Journal of Applied Statistics*, 38(11):2523–2532, 2011.

B. T. Grenfell and RM Anderson. The estimation of age-related rates of infection from case notifications and serological data. *Epidemiology & Infection*, 95(2):419–436, 1985.

P. Groeneboom and J. A. Wellner. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Seminar, Band 19, Birkhauser, New York, 1992.

L. Hanin and L. Huang. Identifiability of cure models revisited. *Journal of Multivariate Analysis*, 130:261–274, 2014.

H. Hebert, F. Ali, J. Bowes, C. Griffiths, A. Barton, and R. Warren. Genetic susceptibility to psoriasis and psoriatic arthritis: implications for therapy. *British Journal of Dermatology*, 166(3):474–482, 2012.

P. S. Helliwell, J. Hetthen, K. Sokoll, M. Green, A. Marchesoni, E. Lubrano, D. Veale, and P. Emery. Joint symmetry in early and late rheumatoid and psoriatic arthritis: comparison with a mathematical model. *Arthritis Rheum*, 43:865–871, 2000.

J. W. Hogan, J. Roy, and C. Korkontzelou. Tutorial in Biostatistics handling drop-out in longitudinal studies. *Statistics in Medicine*, 23:1455–1497, 2004.

P. Hougaard. A class of multivariate failure time distributions. *Biometrika*, 73(3):671–678, 1986.

S. Jiang and R. J. Cook. A mixture model for bivariate interval-censored failure times with dependent susceptibility. *Statistics in Biosciences*, 12:37–62, 2020.

H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall, London, 1997.

P Karmacharya, R Chakradhar, and A Ogdie. The epidemiology of psoriatic arthritis: a literature review. *Best Practice & Research Clinical Rheumatology*, page 101692, 2021.

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. Age-period-cohort analysis in the 1870s: diagrams, stereograms, and the basic differential equation. *Can J Stat.*, 39(3):405–420, 2011.

J P Klein and M L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer, 2003.

J P Klein, H C Van Houwelingen, J G Ibrahim, and T H Scheike. *Handbook of survival analysis*. CRC Press Boca Raton, FL, 2014.

M R Lassen, K A Bauer, B I Eriksson, and A GG for the European Pentasaccharide Hip Elective Surgery Study (EPHESIS) Steering Committee Turpie. Postoperative fondaparinux versus preoperative enoxaparin for prevention of venous thromboembolism in elective hip-replacement surgery: a randomised double-blind comparison. *The Lancet*, 359(9319):1715–1720, 2002.

J. Lawless, J. Kalbfleisch, and C. Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438, 1999.

J. F. Lawless. Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis*, 24:28–44, 2018.

C. Li, J. M. G. Taylor, and J. P. Sy. Identifiability of cure models. *Statistics and Probability Letters*, 54:389–395, 2001.

Z. Li and B. Nan. Relative risk regression for current status data in case-cohort studies. *The Canadian Journal of Statistics*, 39(4):557–577, 2011.

Z. Li, P. Gilbert, and B. Nan. Weighted likelihood method for grouped survival data in case-cohort studies with application to hiv vaccine trails. *Biometrics*, 64:1247–1255, 2008.

Bruce G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data (2nd ed.)*. John Wiley Sons, New York, 2002.

T. Lumley, P. A. Shaw, and J. Y. Dai. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79, 2: 200–220, 2011.

M. McIsaac. *Statistical Methods for Incomplete Covariates and Two-Phase Designs*. PhD thesis, University of Waterloo, 2012.

M. A. McIsaac and R. J. Cook. Response-dependent two-phase sampling designs for biomarker studies. *The Canadian Journal of Statistics*, 42(2):268–284, 2014.

M. A. McIsaac and R. J. Cook. Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Statistics in Medicine*, 34:2899–2912, 2015.

G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer Science+Business Media, Inc., New York, 2006.

R. B. Nelsen. *An Introduction to Copulas*. Springer Science+Business Media, Inc., New York, 2006.

J. Neyman. Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association*, 33:101–116, 1938.

Y. Peng and J. M.G. Taylor. Mixture cure model with random effects for the analysis of a multi-center Tonsil cancer study. *Statistics in Medicine*, 30:211–223, 2011.

R. L. Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11, 1986.

R. L. Prentice and L. P. Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47(3):825–839, 1991.

B. F. Qaqish and K. Liang. Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. *Biometrics*, 48(3):939–950, 1992.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

P. Rahman and J. T. Elder. Genetic epidemiology of psoriasis and psoriatic arthritis. *Annals of the rheumatic diseases*, 64(suppl 2):ii37–ii39, 2005.

P. Rahman, D. D. Gladman, R. J. Cook, Y. Zhou, G. Young, and D. Salonen. Radiological assessment in psoriatic arthritis. *British Journal of Rheumatology*, 37:760–765, 1998.

P. S. Raina, C. Wolfson, S. A. Kirkland, and L. E. Griffith. The Canadian Longitudinal Study on Aging (CLSA). *Canadian Journal on Aging/ La Revue canadienne du vieillissement*, 28(3):221–229, 2009.

M. Reilly. Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology*, 143(1):92–100, 1996.

M. Reilly and M. S. Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314, 1995.

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

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

A. Rossini and A. Tsiatis. A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91(434): 713–721, 1996.

E. M. Ruderman and S. Tambar. Psoriatic arthritis: prevalence, diagnosis, and review of therapy for the dermatologist. *Dermatol Clin*, 22:477–486, 2004.

O. Saarela, S. Kulathinal, and J. Karvanen. Joint analysis of prevalence and incidence data using conditional likelihood. *Biostatistics*, 10(3):575–587, 2009.

P. Sasieni. Maximum weighted partial likelihood estimators for the Cox model. *Journal of the American Statistical Association*, 88(421):144–152, 1993.

J. H. Shih and T. A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4):1384–1399, 1995.

A. Sklar. Fonctions de repartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

P. X. Song. Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27:305–320, 2000.

C. Su and F. Lin. Analysis of clustered failure time data with cure fraction using copula. *Statistics in Medicine*, 38:3961–3973, 2019.

J. Sun. *The Statistical Analysis of Interval-censored Failure Time Data*. Springer Science+Business Media, Inc., New York, 2006.

R Tao, D Zeng, and D Lin. Optimal designs of two-phase studies. *Journal of the American Statistical Association*, 115(532):1946–1959, 2020.

T. Therneau. *A package for survival analysis in R*, 2020. URL https://CRAN.R-project.org/package=survival. R package version 3.2-7.

T Therneau and P Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.

B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976.

A GG Turpie, K A Bauer, B I Eriksson, M R Lassen, PENTATHLON 2000 Study Steering Committee, et al. Postoperative fondaparinux versus postoperative enoxaparin for prevention of venous thromboembolism after elective hip-replacement surgery: a randomised double-blind trial. *The Lancet*, 359(9319):1721–1726, 2002.

University Health Network. Psoriatic arthritis clinic at Toronto Western hospital brochure. https://www.uhn.ca/PatientsFamilies/Health_Information/Health_Topics/Documents/Psoriatic_Arthritis_Clinic_at_Toronto_Western_Hospital.pdf, 2019.

B Van Es, C AJ Klaassen, and K Oudshoorn. Survival analysis under cross-sectional sampling: length bias and multiplicative censoring. *Journal of Statistical Planning and Inference*, 91(2):295–312, 2000.

C. Varin. On composite marginal likelihoods. *Advances in Statistical Analysis*, 92:1–28, 2008.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

M. Wang. Gap time bias in incident and prevalent cohorts. *Statistica Sinica*, pages 999–1010, 1999.

C. Wen and Y. Chen. Assessing age-at-onset risk factors with incomplete covariate current status data under proportional odds models. *Statistics in Medicine*, 32(12):2001–2012, 2013.

C. Wen and C. Lin. Analysis of current status data with missing covariates. *Biometrics*, 67(3):760–769, 2011.

R H White, P S Romano, H Zhou, J Rodrigo, and W Bargar. Incidence and time course of thromboembolic outcomes following total hip or knee arthroplasty. *Archives of internal medicine*, 158(14):1525–1531, 1998.

D. B. Wolfson, A. F. Best, V. Addona, J. Wolfson, and S. M. Gadalla. Benefits of combining prevalent and incident cohorts: An application to myotonic dystrophy. *Statistical methods in medical research*, 28(10-11):3333–3345, 2019.

T. Wright. Exact optimal sample allocation: More efficient than Neyman. *Statistics & Probability Letters*, 129:50–57, 2017.

Y. Wu and R. J. Cook. Variable selection and prediction in biased samples with censored outcomes. *Lifetime data analysis*, 24(1):72–93, 2018.

L. Xiang, X. Ma, and K. K.W. Yau. Mixture cure model with random effects for clustered interval-censored survival data. *Statistics in Medicine*, 30:995–1006, 2011.

C. Yang, L. Diao, and R. J. Cook. Adaptive two-phase designs: some results on robustness and efficiency. unpublished, 2021.

K. K.W. Yau and A. S.K. Ng. Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma. *Statistics in Medicine*, 20:1591–1607, 2001.

D. Zeng and D. Lin. Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, 109(505):371–383, 2014.

L. Zeng, R.J. Cook, and Y. Zhong. Multistate analysis from cross-sectional and auxiliary samples. *Statistics in Medicine*, 39:387–408, 2020.

M. Zhan. Analysis of incomplete event history data. *Unpublished Ph.D. Dissertation*, 1999.

Y. Zhong and R. J. Cook. Augmented composite likelihood for copula modeling in family studies under biased sampling. *Biostatistics*, 17(3):437–452, 2016.

Q. Zhou, H. Zhou, and J. Cai. Case-cohort studies with interval-censored failure time data. *Biometrika*, 104(1):17–29, 2017.

Q. Zhou, J. Cai, and H. Zhou. Outcome-dependent sampling with interval-censored failure time data. *Biometrics*, 74:58–67, 2018.

Q. Zhou, J. Cai, and H. Zhou. Semiparametric inference for a two-stage outcome-dependent sampling design with interval-censored failure time data. *Lifetime Data Analysis*, 26: 85–108, 2020.

APPENDICES

Appendix A

Appendix to Chapter 2

A.1 Derivation of $\mathscr{L}_{jkj'k'2}$

By definition of Z_{jk}, if $a_{1jk} < \infty$ were observed, $P(T_{jk} \in \mathscr{B}_{jk}|\mathbf{X}_{jk}) = \pi_{jk}(\mathcal{F}_j(a_{0jk}) - \mathcal{F}_j(a_{1jk}))$. Let $\xi_{jk} = I(a_{1jk} < \infty)$, we then have

$$\begin{aligned}
\mathscr{L}_{jkj'k'2} =& P(T_{jk} \in \mathscr{B}_{jk}, T_{j'k'} \in \mathscr{B}_{j'k'}|V_{jkj'k'}) \\
=& \left[\omega_{jkj'k'}\mathscr{P}^{11}_{jkj'k'}\right]^{\xi_{jk}\xi_{j'k'}} \left[\omega_{jkj'k'}\mathscr{P}^{11}_{jkj'k'} + (\pi_{jk} - \omega_{jkj'k'})\mathscr{P}^{10}_{jkj'k'}\right]^{\xi_{jk}(1-\xi_{j'k'})} \\
& \times \left[\omega_{jkj'k'}\mathscr{P}^{11}_{jkj'k'} + (\pi_{j'k'} - \omega_{jkj'k'})\mathscr{P}^{01}_{jkj'k'}\right]^{(1-\xi_{jk})\xi_{j'k'}} \\
& \times \Big[\omega_{jkj'k'}\mathscr{P}^{11}_{jkj'k'} + (\pi_{j'k'} - \omega_{jkj'k'})\mathscr{P}^{01}_{jkj'k'} + \\
& (\pi_{jk} - \omega_{jkj'k'})\mathscr{P}^{10}_{jkj'k'} + (1 - \pi_{jk} - \pi_{j'k'} + \omega_{jkj'k'})\Big]^{(1-\xi_{jk})(1-\xi_{j'k'})},
\end{aligned}$$

where

$$\begin{aligned}
\mathscr{P}^{11}_{jkj'k'} =& \Phi_2(\Phi^{-1}(\mathcal{F}_j(a_{0jk}); \boldsymbol{\theta}_j), \Phi^{-1}(\mathcal{F}_{j'}(a_{0j'k'}); \boldsymbol{\theta}'_j); \rho_{jkj'k'}) \\
& - \Phi_2(\Phi^{-1}(\mathcal{F}_j(a_{0jk}); \boldsymbol{\theta}_j), \Phi^{-1}(\mathcal{F}_{j'}(a_{1j'k'}); \boldsymbol{\theta}'_j); \rho_{jkj'k'}) \\
& - \Phi_2(\Phi^{-1}(\mathcal{F}_j(a_{1jk}); \boldsymbol{\theta}_j), \Phi^{-1}(\mathcal{F}_{j'}(a_{0j'k'}); \boldsymbol{\theta}'_j); \rho_{jkj'k'}) \\
& + \Phi_2(\Phi^{-1}(\mathcal{F}_j(a_{1jk}); \boldsymbol{\theta}_j), \Phi^{-1}(\mathcal{F}_{j'}(a_{1j'k'}); \boldsymbol{\theta}'_j); \rho_{jkj'k'}),
\end{aligned}$$

$$\mathscr{P}^{10}_{jkj'k'} = \mathcal{F}_j(a_{0jk}; \boldsymbol{\theta}_j) - \mathcal{F}_j(a_{1jk}; \boldsymbol{\theta}_j), \;\; \mathscr{P}^{01}_{jkj'k'} = \mathcal{F}_{j'}(a_{0j'k'}; \boldsymbol{\theta}_{j'}) - \mathcal{F}_{j'}(a_{1j'k'}; \boldsymbol{\theta}_{j'}),$$

and $\mathscr{P}^{00}_{jkj'k'} = 1$ since T_{jk} and $T_{j'k'}$ are taken to be infinite.

A.2 Intermittent assessments and conditionally independent visit process conditions

Here we review the ideas of Cook and Lawless (2021) with a view to the current setting and to formulate pairwise and working-independence composite likelihoods. We first introduce some notation in counting process and then specify the independence conditions analogous to but stronger than a conditionally independent visiting process (CIVP) assumption, introduced by Cook and Lawless (2018).

A.2.1 Intermittent assessments and counting process notation

Instead of continuous inspection, status of all joints is assessed intermittently over a period of fixed length C_A, where C_A is the administrative censoring time from the time of disease onset. Let C_R represent the random censoring time which leads to a potential premature loss to follow up if $C_R < C_A$. The observed censoring time is then $C = \min\{C_A, C_R\}$. In many settings the visit times are either assumed to be pre-specified (i.e. fixed) at the study entry, or governed by a random visit process. In the former case, a missing data problem arises if subjects miss some visits; a common assumption for the missing pattern is the sequential missing at random (SMAR) assumption introduced by Hogan et al. (2004). For the latter, a conditionally independent visiting process (CIVP) assumption, introduced by Cook and Lawless (2018), is required as a continuous-time analogue to the SMAR assumption to use the simplified partial likelihood of interest.

We restrict our attention to the later setting and discuss the conditions needed to simplify the construction of likelihood from intermittent observation of a multivariate failure time process. For clarity of exposition, counting process notation is employed here. Let $N_{jk}(t) = I(T_{jk} \leqslant t)$ indicate the occurrence of damage in joint (j,k) over $(0,t]$, where if $Z_{jk} = 0$ then $N_{jk}(t) = 0$ for any finite $t > 0$. Let $\Delta N_{jk}(t) = N_{jk}(t + \Delta t^-) - N_{jk}(t^-)$ and $dN_{jk}(t) = \lim_{\Delta t \to 0} \Delta N_{jk}(t)$ indicate whether the damage occurred at time t in joint (j,k); $dN_{jk}(t) = 1$ if so, and $dN_{jk}(t) = 0$ otherwise, and $N_{jk}(t) = \int_0^t dN_{jk}(s)$. We let $\mathbf{N}_j(t) = (N_{j1}(t), \ldots, N_{jK_j}(t))'$ and $\mathbf{N}(t) = (\mathbf{N}_1'(t), \ldots, \mathbf{N}_J'(t))'$. We can also write the full vector $\mathbf{N}(t) = (N_{jk}(t); (j,k) \in \mathcal{S})'$. Similarly, we define $d\mathbf{N}_j(t) = (dN_{j1}(t), \ldots, dN_{jK_j}(t))'$ and $d\mathbf{N}(t) = (d\mathbf{N}_1'(t), \ldots, d\mathbf{N}_J'(t))'$.

Let $\dot{A}(t)$ count the number of follow-up visits over $(0,t]$ and $Y(t) = I(t \leqslant C)$ indicate whether the individual is still under observation at time t. The observed visit process hence has increments $dA(t) = Y(t)d\dot{A}(t)$ for $t > 0$ and we let $A(t) = \int_0^t dA(t)$ count the number of observed follow-up visits over $(0,t]$ and $\{A(t), t \geqslant 0\}$ denote the observed

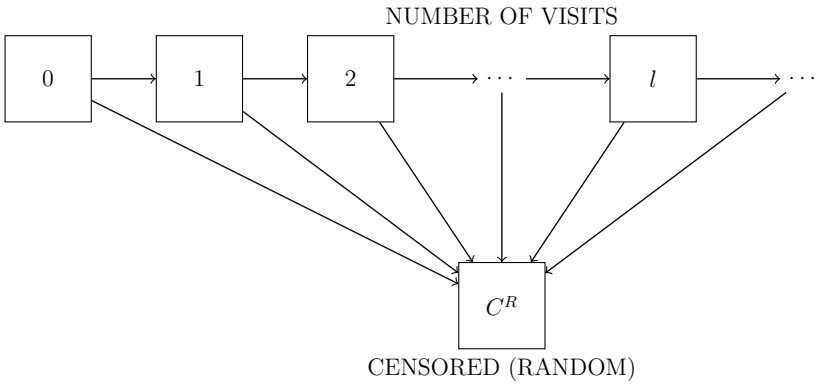


Figure A.1: A state space diagram for joint consideration of the visiting and random censoring processes with C^R representing an absorbing state of being censored.

visit process. Thus, to observe $\mathbf{N}(t)$, the individual must not have withdrawn from the study before t and must have a visit at time t. We let $dC^R(t) = Y(t)I(C_R = t)$ be the indicator of whether random censoring occurred at time t, $C^R(t) = \int_0^t dC^R(s)$, and denote the corresponding counting process as $\{C^R(t), t \geqslant 0\}$. A state space diagram for joint consideration of $\{C^R(t), A(t), t \geqslant 0\}$ is portrayed in Web Figure A.1.

Let $\mathscr{H}_{jk}(t^-) = \{dC^R(s), dA(s), dN_{jk}(s), 0 \leqslant s < t; X_{jk}\}$ contain the complete history joint (j, k) and the covariate X_{jk}, and let $\mathcal{N}_{jk}(t^-) = \{N_{jk}(s), 0 \leqslant s < t\}$ be the history of the counting process for joint (j, k). Let the observed histories be denoted by $H_{jk}(t^-) = \{dC^R(s), dA(s), 0 \leqslant s < t; N_{jk}(a_l), l = 0, 1, \ldots, A(t^-), X_{jk}\}$ and $\mathcal{N}^\circ_{jk}(t^-) = \{N_{jk}(a_l), a_l, l = 0, 1, \ldots, A(t^-)\}$.

A.2.2 A conditionally independent observation scheme

Following the construction of the CIVP conditions in Cook and Lawless (2018), we present the conditionally independent visit process conditions here as having two components. Let

$$\mathscr{H}(t^-) = \{dC^R(t), dA(s), d\mathbf{N}(s), 0 \leqslant s < t, \mathcal{X}\}$$

contain the complete history of the observation and failure time processes and the full covariate information. Note that the history $\mathscr{H}(t^-)$ contains information on the failure time process in continuous time which is not available. We let $H(t^-) = \{dC^R(s), dA(s), 0 \leqslant s < t, \mathbf{N}(a_l), l = 0, 1, \ldots, A(t^-), \mathbf{X}\}$ denote the observed history and $\mathcal{N}^\circ(t^-) = \{(\mathbf{N}(a_l), a_l), l = 0, 1, \ldots, A(t^-)\}$ denote the observed history of the multivariate counting process alone.

We first assume that

$$\{dC^R(s), dA(s), s > a_{l-1}\} \perp \{d\mathbf{N}(s), s > a_{l-1}\} | H(a_{l-1}), \tag{A.2.1}$$

which implies that the random censoring and visit processes from a_{l-1} onward does not depend on the failure process beyond a_{l-1} given the observed history at a_{l-1}.

We consider the likelihood contribution from the time of the visit at a_{l-1} for the case in which a visit occurred at a_l. The conditional likelihood over $(a_{l-1}, a_l]$ based on joint consideration of the failure time, censoring and visit processes is given by

$$\begin{aligned}&\Big\{P(\mathbf{N}(a_l)|C^R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1, H(a_{l-1}))\\&\times P(C_R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1|H(a_{l-1}))\Big\}^{(1-C^R(a_l))dA(a_l)}\\&\times P(C_R(a_l)=1, dA(a_l)=0, A(a_l^-)=l-1|H(a_{l-1}))^{C^R(a_l)(1-dA(a_l))}.\end{aligned} \tag{A.2.2}$$

Note that under (A.2.1) and the assumption that the observation process is non-informative, we can focus on the partial likelihood contribution of the first line in (A.2.2),

$$P(\mathbf{N}(a_l)|C^R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1, H(a_{l-1})). \tag{A.2.3}$$

To write (A.2.3) in terms of the multivariate failure time process of interest, we further require the assumption

$$P(\mathbf{N}(a_l)|C^R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1, H(a_{l-1})) = P(\mathbf{N}(a_l)|\mathcal{N}^\circ(a_{l-1}), a_l, \mathbf{X}). \tag{A.2.4}$$

This condition ensures that a particular observation scheme does not alter the evolution of the underlying multivariate failure time process and enables expression of the partial likelihood in terms of the model of interest. The partial likelihood is then proportional to

$$P(\mathbf{T} \in \mathscr{B}|\mathbf{X}),$$

where $\mathcal{B} = \prod_{(j,k)\in\mathcal{S}} \mathscr{B}_{jk}$ denotes the censoring region of all joints with $\mathscr{B}_{jk} = (a_{0jk}, a_{1jk}]$ denote the interval within which T_{jk} occurs ($0 \leqslant a_{0jk} < a_{1jk} \leqslant \infty$). In particular, $a_{0jk} = a_{l-1}$ and $a_{1jk} = a_l$, where

$$l = \begin{cases}\min\{l : \mathbf{N}(a_l) = 1; l = 2, \ldots, r\} & \text{if } \sum_{l=1}^{r} \mathbf{N}(a_l) > 0;\\ r+1 & \text{otherwise.}\end{cases}$$

A.2.3 A pairwise conditionally independent visit process

In the following we present the pairwise conditionally independent visit conditions to write down a pairwise composite likelihood. We first assume that

$$\{dC^R(s), dA(s), s > a_{l-1}\} \perp \{dN_{jk}(s), dN_{j'k'}(s), s > a_{l-1}\}|H_{jk}(a_{l-1}), H_{j'k'}(a_{l-1}). \tag{A.2.5}$$

Note that the history presented in the condition of (A.2.5) contains information only on joints (j,k) and (j',k'). The fact that this condition needs to hold for every pair of joints (j,k) and (j',k') means it is a very strong condition approaching that of a completely independent visit process.

We consider the likelihood contribution from the time of the visit at a_{l-1} for the case in which a visit occurred at a_l. For a generic pair of joints (j,k) and (j',k'), the conditional likelihood over $(a_{l-1}, a_l]$ based on joint consideration of the failure time, censoring and visit processes is given by

$$\begin{aligned}&\Big\{P(N_{jk}(a_l), N_{j'k'}(a_l)|C^R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1, H_{jk}(a_{l-1}), H_{j'k'}(a_{l-1}))\\&\times P(C_R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1|H_{jk}(a_{l-1}), H_{j'k'}(a_{l-1}))\Big\}^{(1-C^R(a_l))dA(a_l)}\\&\times P(C_R(a_l)=1, dA(a_l)=0, A(a_l^-)=l-1|H_{jk}(a_{l-1}), H_{j'k'}(a_{l-1}))^{C^R(a_l)(1-dA(a_l))}.\end{aligned} \tag{A.2.6}$$

Note that under (A.2.5) and the assumption that the observation process is non-informative, we can focus on the partial likelihood contribution of the first line in (A.2.6),

$$P(N_{jk}(a_l), N_{j'k'}(a_l)|C^R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1, H_{jk}(a_{l-1}), H_{j'k'}(a_{l-1})). \tag{A.2.7}$$

To write (A.2.7) in terms of the bivariate failure time process of interest, we further require the assumption

$$\begin{aligned}&P(N_{jk}(a_l), N_{j'k'}(a_l)|C^R(a_l)=0, dA(a_l)=1, A(a_l^-)=l-1, H_{jk}(a_{l-1}), H_{j'k'}(a_{l-1}))\\=&P(N_{jk}(a_l), N_{j'k'}(a_l)|\mathcal{N}^{\circ}_{jk}(a_{l-1}), \mathcal{N}^{\circ}_{j'k'}(a_{l-1}), a_l, \mathbf{V}_{jkj'k'}).\end{aligned} \tag{A.2.8}$$

This condition ensures that a particular observation scheme does not alter the evolution of any underlying bivariate failure time process and enables expression of the pairwise composite likelihood in terms of the model of interest. In summary, the observation process beyond a_{l-1} cannot depend on

i) the multivariate counting process $\mathbf{N}^{(-j,-k,-j',-k')}(s), s > a_{l-1}$ where $\{\mathbf{N}^{(-j,-k,-j',-k')}(s), s > 0\}$ contains the counting process for all joints except (j,k) and (j',k'), or

ii) the observed history $H^{(-j,-k,-j',-k')}(a_{l-1})$ which is given by

$$\{dC^R(s), dA(s), 0 \leqslant s \leqslant a_{l-1}; \mathbf{N}^{(-j,-k,-j',-k')}(a_h), h=0,1,\ldots,l-1, \mathbf{X}^{(-j,-k,-j',-k')}\}.$$

Under the conditions (A.2.5) and (A.2.8), the partial likelihood contribution from the information collected at the lth observed visit for joints labeled (j,k) and (j',k') is

$$P(N_{jk}(a_l), N_{j'k'}(a_l)|\mathcal{N}^{\circ}_{jk}(a_{l-1}), \mathcal{N}^{\circ}_{j'k'}(a_{l-1}), a_l, \mathbf{V}_{jkj'k'}).$$

Letting $\boldsymbol{\psi} = (\boldsymbol{\vartheta}', \boldsymbol{\varphi}')'$ denote the parameter indexing the joint distribution of the failure time processes of interest given covariates, the composite likelihood for $\boldsymbol{\psi}$ concerning the joint process of $N_{jk}(t)$ and $N_{j'k'}(t)$ over $[0, C^A]$ is

$$\prod_{l=1}^{r} P(N_{jk}(a_l), N_{j'k'}(a_l)|\mathcal{N}^{\circ}_{jk}(a_{l-1}), \mathcal{N}^{\circ}_{j'k'}(a_{l-1}), a_l, \mathbf{V}_{jkj'k'}). \tag{A.2.9}$$

The pairwise composite likelihood contribution expressed in counting process notation in (A.2.9) is equivalent to the probability of $(T_{jk}, T_{j'k'})$ falling in the censoring region $\mathscr{B}_{jk} \times \mathscr{B}_{j'k'}$ given covariates $\mathbf{V}_{jkj'k'}$ associated with this pair of joints, that is

$$P(T_{jk} \in \mathscr{B}_{jk}, T_{j'k'} \in \mathscr{B}_{j'k'}|\mathbf{V}_{jkj'k'}).$$

A.2.4 A working-independent conditionally independent visit process

In this section we present a working-independent version of CIVP conditions to enable the utilization of a working independent composite likelihood in stage I of the two-stage estimation procedure. Similarly to the formulation of (A.2.5) and (A.2.8), we assume

$$\{dC^R(s), dA(s), s > a_{l-1}\} \perp \{dN_{jk}(s), s > a_{l-1}\}|H_{jk}(a_{l-1}), \tag{A.2.10}$$

and

$$P(N_{jk}(a_l)|C^R(a_l) = 0, dA(a_l) = 1, H_{jk}(a_{l-1})) = P(N_{jk}(a_l)|\mathcal{N}^{\circ}_{jk}(a_{l-1}), a_l, X_{jk}), \tag{A.2.11}$$

and that the observation process is non-informative about the parameter to be estimated in stage I. These are even stronger assumptions then the pairwise conditionally independence conditions (A.2.5) and (A.2.8) given above, but enable us to derive the working-independence composite likelihood $\mathscr{L}_1$ arising from intermittent assessments, given by

$$\prod_{(j,k)\in\mathcal{S}} P(T_{jk} \in \mathscr{B}_{jk}|X_{jk}).$$

A.3 Simulation results with specified higher-order dependencies for the susceptibility indicator

Web Table A.1 contains additional simulation results for the first set of simulation studies where we consider three types ($J = 3$) comprised of $(K_1, K_2, K_3) = (2, 2, 2)$ or $(K_1, K_2, K_3) = (2, 6, 6)$ joints for each type. The details on data generation and model formulation are described in the main paper.

A.4 Application to hand joint data from the UTPAC

Here we explain the details in obtaining the nonparametric estimates of cumulative probabilities for damage times of, for example, symmetric hand joints; and estimations for other kind of joint pairs can be similarly obtained.

(a) Prepare the $n \times 4$ matrix `dt_cen` containing 4 columns collecting left and right censoring points of $\mathscr{B}_{jk}$ the censoring interval for T_{jk} and left and right censoring points of $\mathscr{B}_{j'k'}$ the censoring interval for $T_{j'k'}$, where $n = 14 \times N$ is the total number of symmetric pairs;

Note: Due to the constraints for computing time and storage space memory, the entries in `dt_cen` *corresponding to ray, row and other joint pairs are rounded to year.*

(b) Obtain NPMLE of the joint distribution function of T_{jk} and $T_{j'k'}$ along with the regions or rectangles of possible support for the joint distribution function through calling `ICNPMLE(dt_cen)` function;

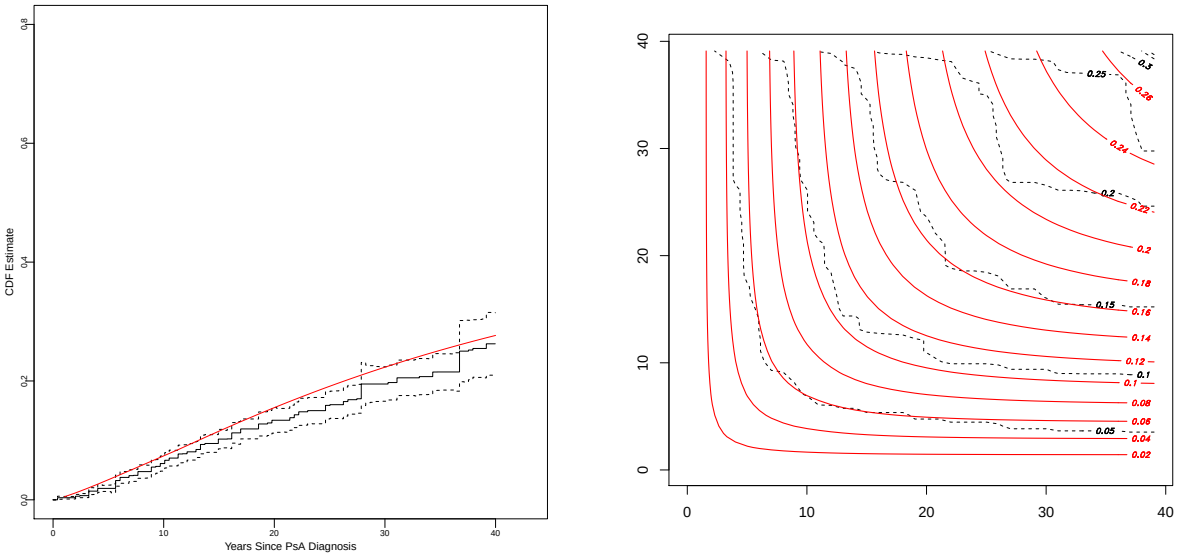
(c) Obtain empirical estimate of the bivariate cumulative probabilities $P(T_{jk} \leqslant t_1, T_{j'k'} \leqslant t_2)$ by summing over the probability masses (NPMLE) on the rectangles that are completely covered by $[0, t_1] \times [0, t_2]$.

Web Figures A.2, A.3, and A.4, referenced in Section 5 of the main manuscript, present distributional characteristics for ray/row/other pairs of hand joints.

Table A.1: Empirical properties of two-stage composite likelihood estimates based on one thousand simulated samples of size $N = 6000$ with $J = 3$ and zero d-order dependencies ($d \geqslant 3$); two-piece piecewise constant proportional hazards models for $T_{jk}|Z_{jk} = 1$ without covariates and the logistic models for $Z_{jk}|\mathbf{X}_{jk}$ in stage I.

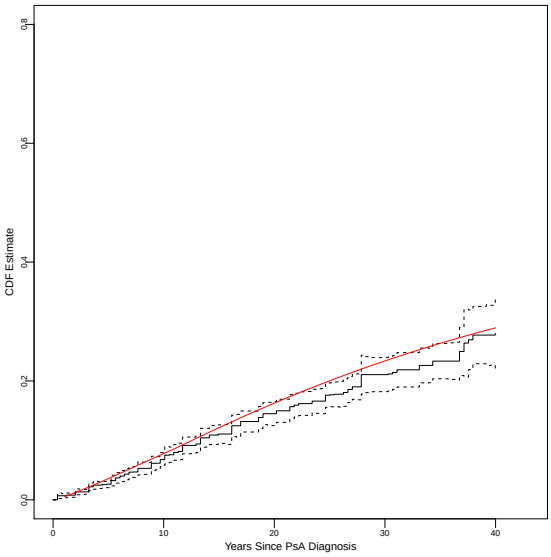
	STAGE I										STAGE II									
	$(K_1, K_2, K_3) = (2, 2, 2)$					$(K_1, K_2, K_3) = (2, 6, 6)$						$(K_1, K_2, K_3) = (2, 2, 2)$					$(K_1, K_2, K_3) = (2, 6, 6)$				
PARA	TRUE	BIAS	ESE	ASE	ECP	TRUE	BIAS	ESE	ASE	ECP	PARA	TRUE	BIAS	ESE	ASE	ECP	TRUE	BIAS	ESE	ASE	ECP
(a) $p_2 = 0.10, \eta_2 = 0.00$																					
$\log \alpha_{11}$	0.834	−0.008	0.083	0.088	0.930	×	−0.007	0.084	0.083	0.935	$\rho^{\dagger}_{11}$	0.386	0.012	0.228	0.233	0.952	×	0.005	0.227	0.238	0.941
$\log \alpha_{12}$	0.834	−0.007	0.272	0.287	0.909	×	−0.002	0.276	0.269	0.941	$\rho^{\dagger}_{12}$	0.124	−0.003	0.132	0.127	0.955	×	−0.007	0.133	0.133	0.957
$\log \alpha_{21}$	0.640	−0.015	0.129	0.126	0.939	×	−0.015	0.128	0.124	0.934	$\rho^{\dagger}_{13}$	0.124	−0.010	0.131	0.129	0.947	×	−0.008	0.132	0.130	0.958
$\log \alpha_{22}$	0.640	−0.005	0.346	0.333	0.929	×	−0.012	0.340	0.336	0.932	$\rho^{\dagger}_{22}$	0.386	0.005	0.342	0.368	0.933	×	0.010	0.262	0.262	0.934
$\log \alpha_{31}$	0.640	−0.018	0.129	0.126	0.943	×	−0.016	0.129	0.132	0.928	$\rho^{\dagger}_{23}$	0.124	−0.006	0.159	0.160	0.955	×	−0.009	0.161	0.154	0.961
$\log \alpha_{32}$	0.640	−0.024	0.340	0.332	0.928	×	−0.008	0.342	0.346	0.922	$\rho^{\dagger}_{33}$	0.386	0.007	0.342	0.347	0.947	×	0.010	0.258	0.265	0.945
η_{10}	−1.299	0.009	0.066	0.069	0.940	×	0.007	0.067	0.066	0.947	γ_{11}	0.182	−0.013	0.102	0.100	0.960	×	−0.009	0.102	0.102	0.955
η_{20}	−1.648	0.013	0.103	0.102	0.940	×	0.012	0.091	0.090	0.938	γ_{12}	0.049	−0.009	0.061	0.061	0.948	×	−0.011	0.057	0.056	0.949
η_{30}	−1.648	0.016	0.103	0.102	0.940	×	0.011	0.092	0.093	0.939	γ_{13}	0.049	−0.008	0.061	0.060	0.942	×	−0.008	0.056	0.057	0.945
η_1	−0.200	0.000	0.035	0.035	0.947	×	0.000	0.034	0.034	0.954	γ_{22}	0.182	−0.014	0.143	0.142	0.952	×	−0.015	0.089	0.088	0.952
η_2	0.000	−0.002	0.056	0.055	0.959	×	−0.002	0.054	0.052	0.958	γ_{23}	0.049	−0.008	0.072	0.073	0.939	×	−0.011	0.061	0.060	0.960
											γ_{33}	0.182	−0.010	0.143	0.147	0.945	×	−0.018	0.090	0.090	0.956
(b) $p_2 = 0.10, \eta_2 = 0.10$																					
$\log \alpha_{11}$	0.834	−0.005	0.084	0.084	0.928	×	−0.009	0.084	0.083	0.933	$\rho^{\dagger}_{11}$	0.386	0.003	0.227	0.236	0.945	×	0.004	0.226	0.228	0.948
$\log \alpha_{12}$	0.834	0.002	0.274	0.280	0.911	×	−0.009	0.273	0.274	0.926	$\rho^{\dagger}_{12}$	0.124	−0.001	0.132	0.126	0.963	×	−0.007	0.133	0.131	0.948
$\log \alpha_{21}$	0.640	−0.018	0.128	0.132	0.930	×	−0.022	0.130	0.130	0.939	$\rho^{\dagger}_{13}$	0.124	−0.009	0.131	0.131	0.947	×	−0.005	0.133	0.134	0.946
$\log \alpha_{22}$	0.640	−0.011	0.339	0.348	0.921	×	−0.024	0.342	0.345	0.930	$\rho^{\dagger}_{22}$	0.386	0.011	0.343	0.365	0.936	×	0.008	0.262	0.264	0.946
$\log \alpha_{31}$	0.640	−0.018	0.128	0.126	0.942	×	−0.019	0.130	0.133	0.926	$\rho^{\dagger}_{23}$	0.124	−0.004	0.158	0.159	0.951	×	−0.006	0.162	0.162	0.950
$\log \alpha_{32}$	0.640	−0.025	0.339	0.331	0.940	×	−0.020	0.343	0.345	0.937	$\rho^{\dagger}_{33}$	0.386	0.009	0.343	0.340	0.955	×	0.007	0.263	0.262	0.951
η_{10}	−1.310	0.007	0.067	0.067	0.928	×	0.008	0.067	0.066	0.935	γ_{11}	0.182	−0.011	0.102	0.102	0.944	×	−0.012	0.102	0.099	0.951
η_{20}	−1.658	0.016	0.102	0.107	0.934	-2.859	0.017	0.093	0.096	0.933	γ_{12}	0.049	−0.009	0.061	0.060	0.954	×	−0.010	0.057	0.056	0.951
η_{30}	−1.658	0.016	0.102	0.103	0.947	-2.869	0.014	0.093	0.093	0.937	γ_{13}	0.049	−0.009	0.061	0.060	0.953	×	−0.010	0.057	0.055	0.944
η_1	−0.200	−0.001	0.035	0.035	0.953	×	0.000	0.034	0.034	0.954	γ_{22}	0.182	−0.017	0.144	0.147	0.946	×	−0.015	0.090	0.088	0.950
η_2	0.100	−0.001	0.052	0.050	0.963	×	−0.001	0.049	0.047	0.961	γ_{23}	0.049	−0.009	0.072	0.073	0.943	×	−0.011	0.062	0.061	0.943
											γ_{33}	0.182	−0.012	0.144	0.145	0.952	×	−0.018	0.090	0.087	0.964

Note. PARA, parameter; TRUE, true value of the parameter; ESE, sample standard deviation; ASE, model-based standard error; × means the value is the same as that specified in the "TRUE" column when $(K_1, K_2, K_3) = (2, 2, 2)$; $\rho^{\dagger}_{jj'} = \tan(.5\pi\rho_{jj'})$, $\gamma_{jj'} = \log \zeta_{jj'}$, $j \leqslant j'; j, j' \in \{1, 2\}$.

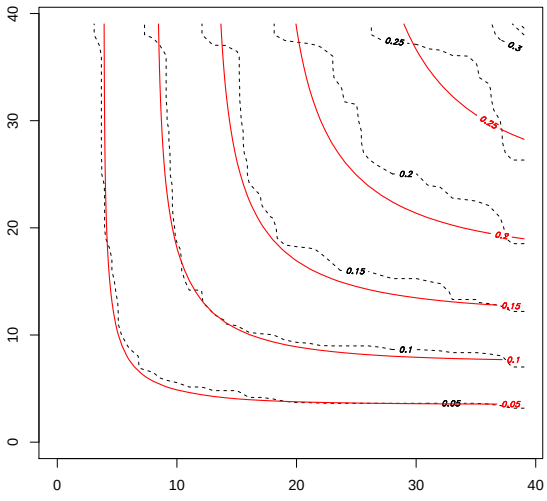


(a) Marginal CDF of $\max(T_{jk}, T_{j'k'})$ **(b)** Bivariate CDF of $(T_{jk}, T_{j'k'})$

Figure A.2: Plots of parametric and non-parametric estimates for **ray** pairs: (a) estimation of the marginal concordance function of damage based on the fitted second order models for susceptibility and failure times of susceptibles (solid smooth line) and nonparametric estimates (solid nonsmooth line) with a 95% CI band (dashed lines); and (b) bivariate cumulative probability of damage based on the fitted second order models for susceptibility and failure times of susceptibles (solid line) and nonparametric estimates (dashed lines).

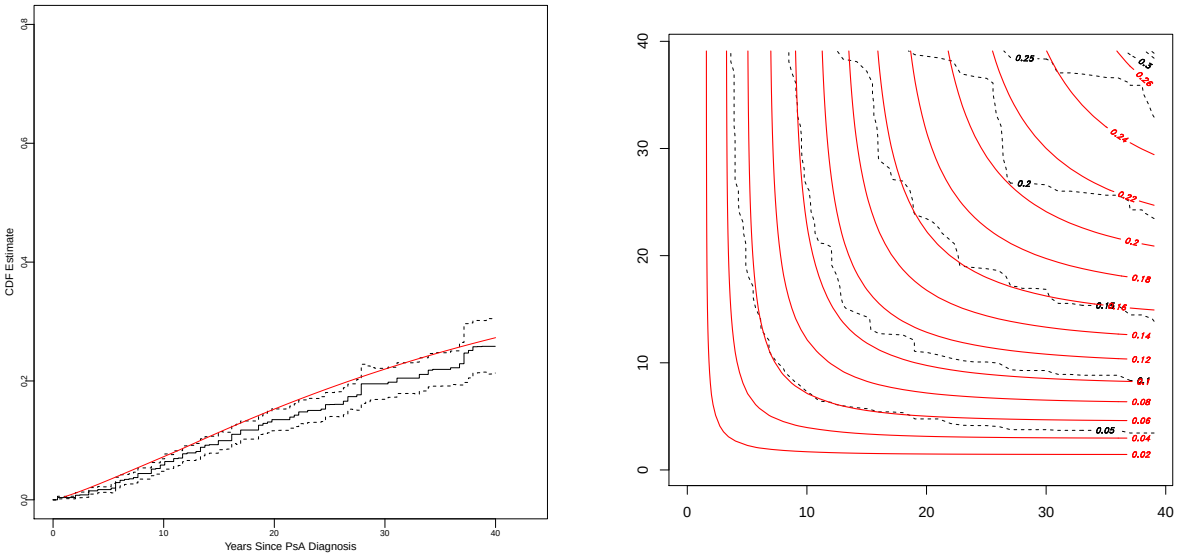


(a) Marginal CDF of $\max(T_{jk}, T_{j'k'})$



(b) Bivariate CDF of $(T_{jk}, T_{j'k'})$

Figure A.3: Plots of parametric and non-parametric estimates for **row** pairs: (a) estimation of the marginal concordance function of damage based on the fitted second order models for susceptibility and failure times of susceptibles (solid smooth line) and nonparametric estimates (solid nonsmooth line) with a 95% CI band (dashed lines); and (b) bivariate cumulative probability of damage based on the fitted second order models for susceptibility and failure times of susceptibles (solid line) and nonparametric estimates (dashed lines).



(a) Marginal CDF of $\max(T_{jk}, T_{j'k'})$ **(b)** Bivariate CDF of $(T_{jk}, T_{j'k'})$

Figure A.4: Plots of parametric and non-parametric estimates for **other** pairs: (a) estimation of the marginal concordance function of damage based on the fitted second order models for susceptibility and failure times of susceptibles (solid smooth line) and nonparametric estimates (solid nonsmooth line) with a 95% CI band (dashed lines); and (b) bivariate cumulative probability of damage based on the fitted second order models for susceptibility and failure times of susceptibles (solid line) and nonparametric estimates (dashed lines).

Appendix B

Appendix to Chapter 3

B.1 Derivation of the score-type residual M_μ

Following Tao et al. (2020), we let $M_\mu = \partial \log P(Y|A,\mathbf{X})/\partial\mu$ which can be expressed as

$$M_\mu = \left(-Y\frac{\mathcal{F}(A|\mathbf{X})}{1-\mathcal{F}(A|\mathbf{X})} + 1 - Y\right)\frac{\partial \log \mathcal{F}(A|\mathbf{X})}{\partial\mu} = Q(\mathbf{Z}, X_1)\log \mathcal{F}(A|\mathbf{X}), \tag{B.1.1}$$

where $Q(A|\mathbf{X}) = -F^{-1}(A|\mathbf{X})[Y - F(A|\mathbf{X})]$. Since $E[Q(\mathbf{Z}, X_1)|A,\mathbf{X}] = 0$ it follows that $E[M_\mu] = 0$. Next we note that

$$M_{\mu\mu} = \frac{\partial M_\mu}{\partial\mu} = \frac{\partial}{\partial\mu}\left\{\left(1 - \frac{Y}{1-\mathcal{F}}\right)\log\mathcal{F}\right\} = M_\mu + \frac{Y\mathcal{F}}{(1-\mathcal{F})^2}\log\mathcal{F}^2, \tag{B.1.2}$$

giving

$$\begin{aligned}\frac{\partial^2 \log L(\vartheta)}{\partial\beta_1^2} =& RM_{\mu\mu}X_1^2 + (1-R)\left\{\frac{\sum_{X_1} M_{\mu\mu}X_1^2(1-\mathcal{F})^Y\mathcal{F}^{1-Y}P(X_1|\mathbf{X}_2)}{\sum_{X_1}(1-\mathcal{F})^Y\mathcal{F}^{1-Y}P(X_1|\mathbf{X}_2)}\right.\\ &\left. + \frac{\sum_{X_1} M_\mu^2X_1^2(1-\mathcal{F})^Y\mathcal{F}^{1-Y}P(X_1|\mathbf{X}_2)}{\sum_{X_1}(1-\mathcal{F})^Y\mathcal{F}^{1-Y}P(X_1|\mathbf{X}_2)} - \left(\frac{\sum_{X_1} M_\mu X_1(1-\mathcal{F})^Y\mathcal{F}^{1-Y}P(X_1|\mathbf{X}_2)}{\sum_{X_1}(1-\mathcal{F})^Y\mathcal{F}^{1-Y}P(X_1|\mathbf{X}_2)}\right)^2\right\}\end{aligned}$$

which under $\beta_1 = o(1)$ is

$$RM_{\mu\mu}X_1^2 + (1-R)\Big[M_{\mu\mu}E[X_1^2|\mathbf{X}_2] + M_\mu^2 Var(X_1|\mathbf{X}_2)\Big].$$

This gives

$$\mathcal{I}_{\beta_1\beta_1} = E_{\mathrm{Z}}\{M_{\mu\mu}E[X_1^2|X_2] + M_\mu^2\mathrm{Var}(X_1|X_2)\} + E_{\mathrm{RZ}}\{RM_\mu^2\mathrm{Var}(X_1|X_2)\}.$$

Letting $\theta_{\circ j}$ denote the jth element of $\boldsymbol{\theta}_\circ$, we have $M_{\theta_{\circ j}} = \partial \log P(Y|A, \mathbf{X}; \boldsymbol{\theta})/\partial\theta_{\circ j}$, $M_{\theta_{\circ j}\theta_{\circ k}} = \partial M_{\theta_{\circ j}}/\partial\theta_{\circ k}$, and $M_{\mu\theta_{\circ j}} = \partial M_\mu/\partial\theta_{\circ j}$. Moreover

$$\frac{\partial \log L(\boldsymbol{\vartheta})}{\partial\beta_1\partial\theta_{\circ j}}\Big|_{\theta=(0,\boldsymbol{\theta}'_\circ)'} = \left\{RM_{\mu\boldsymbol{\theta}_{\circ j}}X_1 + (1-R)M_{\mu\boldsymbol{\theta}_{\circ j}}E[X_1|\mathbf{X}_2]\right\},$$

and

$$\frac{\partial \log L(\boldsymbol{\vartheta})}{\partial\theta_{\circ j}\partial\theta_{\circ k}}\Big|_{\boldsymbol{\theta}=(0,\boldsymbol{\theta}'_\circ)'} = M_{\theta_{\circ j}\theta_{\circ k}}.$$

Hence under $\beta_1 = o(1)$, neither $\mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}$ nor $\mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}$ depend on the phase II subsampling schemes.

Following the projection method in Tao et al. (2020), we have

$$\mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}_{\eta\eta}^{-1}\mathcal{I}_{\boldsymbol{\eta}\beta_1} = E\left\{\frac{(E[RM_\mu|A,\mathbf{X}_2])^2}{E[R|A,\mathbf{X}_2]}\mathrm{Var}(X_1|\mathbf{X}_2)\right\}. \tag{B.1.3}$$

so under $\beta_1 = o(1)$, V_{β_1} can be expressed as

$$\begin{aligned}&\Sigma_1 + E\left\{\left[E[RM_\mu^2|A,\mathbf{X}_2] - \frac{(E[RM_\mu|A,\mathbf{X}_2])^2}{E[R|A,\mathbf{X}_2]}\right]\mathrm{Var}(X_1|\mathbf{X}_2)\right\}\\ =&\Sigma_1 + E\left\{E[R|A,\mathbf{X}_2]^{-1}\mathrm{Var}[M_\mu|R=1,A,\mathbf{X}_2]\mathrm{Var}(X_1|\mathbf{X}_2)\right\},\end{aligned} \tag{B.1.4}$$

where

$$\Sigma_1 = E_{\mathbf{Z}}\{M_{\mu\mu}E[X_1^2|\mathbf{X}_2] + M_\mu^2\mathrm{Var}(X_1|\mathbf{X}_2)\} - \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}^{-1}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1}$$

does not depend on the phase II sub-sampling rules. Tao et al. (2020) derived this for right censored data and replaced M_μ by martingale-type residuals in (B.1.4). Here we make use of this same general result to gain insights into efficient phase II sub-sampling with current status data in Section 3.1.

B.2 Neyman and adaptive approximate Neyman allocation

Steps of implementing NEY:

1. compute the approximate influence functions $\{\tilde{\Delta}_i(\beta_1), i = 1, \ldots, N\}$ using (4.3) and true θ;

2. compute $\tilde{\sigma}_j$, the standard deviation of $\{\tilde{\Delta}_i(\beta_1), i \in \text{ stratum } j\}$ obtained in step 1;

3. select $n_j \propto n N_j \tilde{\sigma}_j$ subjects from stratum j using an integer-valued algorithm (Wright, 2017), with $\sum_{j=1}^{N} n_j = n$.

Steps of implementing NEYA:

1. select n_a^j from stratum j using balanced sampling with $n_a = \sum_j n_j^a < n$ in phase IIa and obtain $\tilde{\theta}^a$ an IPW estimator of $\boldsymbol{\theta}$;

2. compute the approximate influence functions $\{\tilde{\Delta}_i(\beta_1), i \in \text{ phase IIa sample}\}$ using (4.3) and $\tilde{\boldsymbol{\theta}}^a$;

3. compute $\tilde{\sigma}_j$, the standard deviation of $\{\tilde{\Delta}_i(\beta_1), i \in \text{ stratum } j \cap \text{ phase IIa sample}\}$ obtained in step 2;

4. select $n_j^b \propto (n - n_a)(N_j - n_j^a)\tilde{\sigma}_j$ subjects from the rest of stratum j (i.e. those not selected in step 1) using an integer-valued algorithm (Wright, 2017), with $\sum_{j=1}^{N} n_j^b = n - n_a$.

Therefore $n_j = n_j^a + n_j^b$ subjects are selected from stratum j and $n = \sum_j n_j$.

B.3 TAO-OPTA to EXT-M_μ and EXT-(A, Y)

Table B.1 presents different choices for the proportions of the phase IIa samples n_a/n for TAO-OPTA and their comparison to EXT-M_μ and EXT-(A, Y). The relative efficiency of each other design compared to TAO-OPT is calculated as the ratio of the mean asymptotic variance of $\hat{\beta}_1$ under TAO-OPT to that under other designs. TAO-OPTA performs better when n_a is smaller, but a sufficient phase IIa sample is necessary to make reliable estimates for $\text{Var}(X_1|X_2)$. The details on data generation and model formulation are described in Section 3.3.2.

Table B.1: Relative efficiency (%) to TAO-OPT for the estimated log hazard ratio in X_1 under maximum likelihood based on 1000 samples of size 2000; $\beta_2 = -0.2$.

			$X_1 \perp X_2$ (i.e. constant Var$(X_1\|X_2)$)					$X_1 \not\perp X_2$ (i.e. non-constant Var$(X_1\|X_2)$)				
			Non-adaptive Design		TAO-OPTA (n_a/n)			Non-adaptive Design		TAO-OPTA (n_a/n)		
ψ	β_1	n	EXT-M_μ	EXT-(A,Y)	0.2	0.4	0.6	EXT-M_μ	EXT-(A,Y)	0.2	0.4	0.6
(a) Nonrare event: $q_1 = 0.6, q_2 = 0.3$												
0.1	-0.2	300	99.6	98.9	88.1	79.6	69.1	95.0	96.3	88.3	80.0	69.2
		600	100.0	99.6	92.5	85.8	77.0	97.2	97.5	92.6	85.9	77.0
	0	300	99.9	99.0	88.3	80.0	69.1	95.9	96.8	88.2	79.8	69.1
		600	99.9	99.5	92.5	86.0	77.1	97.8	98.0	92.6	86.0	77.1
	0.2	300	100.1	99.0	88.2	80.0	69.1	96.9	97.5	88.5	80.3	69.3
		600	100.1	99.6	93.0	86.6	77.8	98.3	98.4	93.0	86.5	77.8
0.3	-0.2	300	99.7	99.2	86.6	76.9	65.0	95.7	96.6	86.1	77.0	65.5
		600	99.9	99.6	92.5	85.2	75.0	98.6	98.3	92.3	85.3	75.1
	0	300	99.8	99.3	86.2	76.6	64.4	96.5	97.4	86.0	76.8	65.2
		600	100.0	99.6	92.8	85.3	75.1	98.7	98.5	92.5	85.4	75.2
	0.2	300	100.0	99.1	86.3	76.2	64.1	97.3	97.7	86.2	76.7	65.0
		600	100.0	99.6	92.9	85.7	75.3	99.1	98.7	92.6	85.8	75.3
(b) Rare event: $q_1 = 0.1, q_2 = 0.05$												
0.1	-0.2	300	99.9	99.5	96.5	88.9	70.9	99.3	99.5	96.6	89.7	72.8
		600	100.0	99.8	99.2	98.2	96.8	99.7	99.8	99.1	98.2	96.8
	0	300	100.0	99.4	96.4	89.0	71.5	99.5	99.6	96.5	89.9	73.7
		600	100.0	99.8	99.1	98.0	96.4	99.8	99.8	99.1	98.1	96.5
	0.2	300	100.0	99.4	96.2	89.3	72.0	99.6	99.6	96.3	90.0	74.2
		600	100.0	99.7	99.0	97.8	96.1	99.9	99.8	99.0	97.9	96.1
0.3	-0.2	300	99.9	99.4	96.2	88.9	70.5	99.4	99.5	96.6	89.7	73.3
		600	100.0	99.8	98.8	97.3	95.4	99.8	99.8	98.8	97.4	95.5
	0	300	99.9	99.3	96.1	88.9	71.3	99.4	99.5	96.5	89.8	73.8
		600	100.0	99.8	98.7	97.0	94.9	99.8	99.8	98.7	97.1	94.9
	0.2	300	99.9	99.3	95.9	88.9	71.7	99.5	99.5	96.3	89.7	74.1
		600	99.9	99.8	98.6	96.7	94.3	99.8	99.8	98.6	96.8	94.4

Note. The relative efficiency to TAO-OPT is computed by ratio of mean asymptotic variance of $\hat{\beta}_1$ under TAO-OPT to that under other designs; TAO-OPTA selects a simple random sample of size n_a in phase IIa; $\psi = \text{Var}(A^\dagger)$; When $X_1 \not\perp X_2$, the odds ratio between X_1 and X_2 is set by $\varphi = 2$.

Appendix C

Appendix to Chapter 4

C.1 Derivation of the general likelihood

Following Phase II sub-sampling a single individual in *Registry 1* provide would provide data $\{Z_0 = 1, H(A^\dagger), X^\circ, \Delta\}$, or equivalently, $\{Z_0 = 1, B, A_1, S_1, S_2, \delta_1 = 1, \delta_2, \delta_D, \mathbf{X}^\circ, \Delta\}$ where $\mathbf{X}^\circ = (X, \mathbf{V}')'$. The likelihood contribution relevant for the life history process from such an individual is then proportional to

$$\begin{aligned} &P(S_1, S_2, \delta_2, \delta_D, X | Z(A_0) = 1, A_0, A_1, \delta_1 = 1, \mathbf{V})^{\Delta} \\ &\times P(S_1, S_2, \delta_2, \delta_D | Z(A_0) = 1, A_0, A_1, \delta_1 = 1, \mathbf{V})^{1-\Delta}, \end{aligned} \tag{C.1.1}$$

where

$$\begin{aligned} &P(S_1, S_2, \delta_2, \delta_D, X | Z(A_0) = 1, A_0, A_1, \delta_1 = 1, \mathbf{V}) \\ =&\frac{P(S_1, S_2, \delta_2, \delta_D, | A_0, A_1, \delta_1 = 1, \mathbf{V}, X)}{P(Z(A_0) = 1 | A_0, A_1, \delta_1 = 1, \mathbf{V}, X)} \times P(X | Z(A_0) = 1, B, A_0, A_1, \delta_1 = 1, \mathbf{V}) \end{aligned} \tag{C.1.2}$$

The sample distribution X in this registry is $P(X|Z(A_0) = 1, B, A_0, A_1, \delta_1 = 1, \mathbf{V})$, given by

$$\begin{aligned} &\frac{P(A_1, \delta_1 = 1, Z(A_0) = 1 | \mathbf{V}, X)}{P(A_1, \delta_1 = 1, Z(A_0) = 1 | \mathbf{V})} P(X | \mathbf{V}) \\ =&\left[\frac{\lambda_0(A_1 | \mathbf{V}, X) P_{00}(0, A_1 | \mathbf{V}, X) P_{11}(A_1, A_0 | \mathbf{V}, X, A_1)}{\int \lambda_0(A_1 | \mathbf{V}, x) P_{00}(0, A_1 | \mathbf{V}, x) P_{11}(A_1, A_0 | \mathbf{V}, x, A_1) P(x | \mathbf{V}) dx} \right] P(X | \mathbf{V}), \end{aligned} \tag{C.1.3}$$

where $P_{jj}(a_1, a_2 | \cdot) = P(Z(a_2) = j | Z(a_1) = j, \cdot)$ denotes the probability of no transition determined by $\lambda_j(\cdot)$ and $\gamma_j(\cdot)$. The bracketed term in (C.1.3) is rather complex and its

evaluation relies on the specification of $\lambda_0(\cdot)$, $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$. Identifiability issues arise when trying to estimate $\gamma_0(\cdot)$ without additional information or imposing an assumption on the mortality rates among disease-free individuals (Cook and Lawless, 2018, Chapter 7). While under Assumption 1 and Assumption 2, (C.1.3) can be simplified to be

$$\left[\frac{\mathcal{F}_1(A_0 - A_1|\mathbf{V}, X, A_1)}{\int \mathcal{F}_1(A_0 - A_1|\mathbf{V}, X, A_1)P(x|\mathbf{V})dx}\right] P(X|\mathbf{V}), \tag{C.1.4}$$

where $\mathcal{F}_1(A_0 - A_1|A_1, \mathbf{V}, X) = \exp(-\int_0^{A_0-A_1} \lambda_1(t|A_1, \mathbf{V}, X))$, and we informally write $P(X|\mathbf{V})$ as the conditional probability model for $X|\mathbf{V}$.

The sample distribution $P(S_1, S_2, \delta_2, \delta_D|A_0, A_1, \delta_1 = 1, \mathbf{V}, X)$ in *Registry 1* is given by

$$\begin{aligned}
&\mathcal{F}_1(S_1|A_1, \mathbf{V}, X)\lambda_1(S_1|A_1, \mathbf{V}, X)^{\delta_2} \\
\times&\mathcal{G}_1(S_1|A_1, \mathbf{V}, X)\gamma_1(A_1 + S_1|H(A_1 + S_1), \mathbf{V}, X)^{(1-\delta_2)\delta_D} \\
&\times \left[P_{22}(A_2, A_2 + S_2|A_1, S_1, \mathbf{V}, X)\gamma_2(A_2 + S_2|H(A_2 + S_2), \mathbf{V}, X)^{\delta_D}\right]^{\delta_2},
\end{aligned} \tag{C.1.5}$$

where $\mathcal{G}_1(S_1|A_1, \mathbf{V}, X) = \exp(-\int_{A_1}^{A_1+S_1} \gamma_1(a|H(a^-), \mathbf{V}, X)da)$.

If only data from *Registry 1* are available for analysis, we do not need Assumption 2 and Assumption 3 for proceeding via partial likelihood. However, the sample distribution $P(S_1, S_2, \delta_2, \delta_D, X|A_0, A_1, \delta_1 = 1, \mathbf{V})$ in the PsA cohort (*Registry 2*) can only be simplified to

$$\lambda_1(S_1|A_1, \mathbf{V}, X)\gamma_2(A^\dagger|H(A^\dagger-), \mathbf{V}, X)^{\delta_D} \frac{P_{11}(A_1, A_2|\mathbf{V}, X, A_1)P_{22}(A_2, A^\dagger|A_1, S_1, \mathbf{V}, X)}{\int P_{12}(A_1, A_0|\mathbf{V}, x, A_1)P(x|\mathbf{V})dx} P(X|\mathbf{V}) \tag{C.1.6}$$

under Assumption 1 and Assumption 2. With Assumption 2 and Assumption 3, it becomes

$$\frac{\lambda_1(S_1|A_1, \mathbf{V}, X)\mathcal{F}_1(S_1|A_1, \mathbf{V}, X)}{1 - \int \mathcal{F}_1(A_0 - A_1|A_1, \mathbf{V}, x)P(x|\mathbf{V})dx} P(X|\mathbf{V}) \times C^*, \tag{C.1.7}$$

where C^* only involves the intensity models for mortality which, assumed to be non-informative about $\boldsymbol{\vartheta}$.

C.2 Impact of violations of Assumption 3

Here we present simulation results of the full data analysis in the sensitivity study in sub-Section 4.5.3. The purpose is to investigate the impact of violations to the nondifferential mortality assumption (Assumption 3 in Section 4.3) on estimation based on the partial likelihood approach in the absence of the missing covariate problem. Table C.1 summarizes

the finite sample characteristics of the maximum partial likelihood estimators of $\boldsymbol{\vartheta}$ under a full data analysis, where $\gamma_2(a|V)/\gamma_1(a|V) = \exp(\nu) \neq 1$ such that Assumption 3 is violated. For the settings considered there is modest bias in the estimated coefficients but larger biases for the parameter indexing the baseline hazard when $\nu = \log 1.2$ or $\log 1.5$.

C.3 Information for $\hat{\beta}_1$ when $\beta_1 = o(1)$

Here we first present the information bound for $\hat{\beta}_1$ under the partial likelihood assuming Assumptions 1-3 and then note the efficiency for our proposed weighted residual-dependent designs of Section 4.4 under the setting where $\beta_1 = o(1)$.

Let $\boldsymbol{\theta}$ index the $1 \to 2$ transition intensity and $\boldsymbol{\eta}$ index the covariate distribution $P(X|\mathbf{V})$; we write $\lambda_1(t|A_1, \mathbf{V}, X; \boldsymbol{\theta})$ and $P(X|\mathbf{V}; \boldsymbol{\eta})$. To estimate $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\eta}')$, we consider a partial likelihood

$$L_p(\boldsymbol{\vartheta}) \propto \left[\frac{\{\mathscr{L}_{\boldsymbol{\theta}} \mathscr{L}_{\boldsymbol{\eta}}\}^{\Delta} (E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}])^{1-\Delta}}{E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]} \right]^{I(Z_0=1)} \left[\frac{\{\mathscr{L}_{\boldsymbol{\theta}} \mathscr{L}_{\boldsymbol{\eta}}\}^{\Delta} (E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}])^{1-\Delta}}{1 - E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]} \right]^{I(Z_0=2)},$$

where $\mathscr{L}_{\boldsymbol{\theta}} = \mathcal{F}_1(S_1|A_1, \mathbf{V}, X; \boldsymbol{\theta}) \lambda_1^{\delta_2}(S_1|A_1, \mathbf{V}, X; \boldsymbol{\theta})$ and $\mathscr{L}_{\boldsymbol{\eta}} = P(X|\mathbf{V}; \boldsymbol{\eta})$.

Let $S_p(\boldsymbol{\vartheta}) = \partial \log L_p(\boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}$ denote the partial score vector. We consider a partition of $S_p(\boldsymbol{\vartheta}) = (S'_{\boldsymbol{\theta}}, S'_{\boldsymbol{\eta}})'$ with $S_{\boldsymbol{\theta}} = \partial \log L_p(\boldsymbol{\vartheta})/\partial \boldsymbol{\theta}$ given by

$$I(Z_0 = 1) \left[\Delta \mathscr{S}_{\boldsymbol{\theta}} + (1 - \Delta) \frac{E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\theta}} \mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]}{E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]} - \frac{E_{X|\mathbf{V}}[\partial \mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta})/\partial \boldsymbol{\theta}; \boldsymbol{\eta}]}{E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]} \right]$$
$$+ I(Z_0 = 2) \left[\Delta \mathscr{S}_{\boldsymbol{\theta}} + (1 - \Delta) \frac{E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\theta}} \mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]}{E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]} + \frac{E_{X|\mathbf{V}}[\partial \mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta})/\partial \boldsymbol{\theta}; \boldsymbol{\eta}]}{1 - E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]} \right]$$

and $S_{\boldsymbol{\eta}} = \partial \log L_p(\boldsymbol{\vartheta})/\partial \boldsymbol{\eta}$ equaling

$$I(Z_0 = 1) \left[\Delta \mathscr{S}_{\boldsymbol{\eta}} + (1 - \Delta) \frac{E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}} \mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]}{E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]} - \frac{E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}} \mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]}{E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]} \right]$$
$$+ I(Z_0 = 2) \left[\Delta \mathscr{S}_{\boldsymbol{\eta}} + (1 - \Delta) \frac{E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}} \mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]}{E_{X|\mathbf{V}}[\mathscr{L}_{\boldsymbol{\theta}}; \boldsymbol{\eta}]} + \frac{E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}} \mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]}{1 - E_{X|\mathbf{V}}[\mathcal{F}_1(T_0|A_1, \mathbf{V}, X; \boldsymbol{\theta}); \boldsymbol{\eta}]} \right]$$

with $\mathscr{S}_{\boldsymbol{\theta}} = \partial \log \mathscr{L}_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$ and $\mathscr{S}_{\boldsymbol{\eta}} = \partial \log \mathscr{L}_{\boldsymbol{\eta}}/\partial \boldsymbol{\eta}$. The expected information matrix is $\mathcal{I} = -E[\partial S_p(\boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}'] = E[S_p(\boldsymbol{\vartheta}) S'_p(\boldsymbol{\vartheta})]$, giving

$$\begin{pmatrix} \mathcal{I}_{\beta_1\beta_1} & \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ} & \mathcal{I}_{\beta_1\boldsymbol{\eta}} \\ \mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} & \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ} & \mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\eta}} \\ \mathcal{I}_{\boldsymbol{\eta}\beta_1} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\theta}_\circ} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} \end{pmatrix}$$

where $\boldsymbol{\theta}_\circ = (\beta_2', \beta_3, \alpha')'$. The asymptotic variance of the maximum (partial) likelihood estimator $\hat{\beta}_1$ is given by $V_{\beta_1}^{-1}$ with

$$V_{\beta_1} = \mathcal{I}_{\beta_1\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}^{-1}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathcal{I}_{\boldsymbol{\eta}\beta_1}. \tag{C.3.1}$$

Next we show the detailed expressions of each component when $\beta_1 = o(1)$.

Let $\mu = \beta_1 X + \beta_2'\mathbf{V} + \beta_3 \log A_1$ denote the linear predictor of interest and $M_\mu = \delta_2 - \Lambda(S_1;\alpha)\exp(\mu)$, where $\Lambda(t;\alpha) = \int_0^t \lambda(u;\alpha)$. For ease of presentation, let $Z_0 = Z(A_0)$, $\mathbf{W} = (B, A_1, A^\dagger, \delta_1, \delta_D, \mathbf{V}')'$, $\Lambda_0 = \Lambda(T_0;\alpha)$ and $\mathcal{F}_1 = \mathcal{F}_1(T_0|A_1, \mathbf{V}, X)$ in what follows.

C.3.1 Derivation of $\mathcal{I}_{\beta_1\beta_1}$

When $\beta_1 = o(1)$, we have $\mathscr{S}_{\beta_1} = \partial \log \mathscr{L}_{\boldsymbol{\theta}}/\partial\beta_1 = M_\mu X$ and $S_{\beta_1}(\boldsymbol{\vartheta}) = \partial \log L_p(\boldsymbol{\vartheta})/\partial\beta_1$ is given by

$$\begin{aligned}
S_{\beta_1}(\boldsymbol{\vartheta}) =& I(Z_0 = 1)\left(\Delta\mathscr{S}_{\beta_1} + (1-\Delta)E_{X|\mathbf{V}}[\mathscr{S}_{\beta_1}] - E_{X|\mathbf{V}}\left[\frac{\partial \log \mathcal{F}_1(T_0|A_1,\mathbf{V},X)}{\partial\beta_1}\right]\right) \\
&+ I(Z_0 = 2)\left(\Delta\mathscr{S}_{\beta_1} + (1-\Delta)E_{X|\mathbf{V}}[\mathscr{S}_{\beta_1}] + \frac{\mathcal{F}_1(T_0|A_1,\mathbf{V},X)}{1-\mathcal{F}_1(T_0|A_1,\mathbf{V},X)}E_{X|\mathbf{V}}\left[\frac{\partial \log \mathcal{F}_1(T_0|A_1,\mathbf{V},X)}{\partial\beta_1}\right]\right) \\
=& I(Z_0 \in \{1,2\})[\Delta M_\mu(X - E[X|\mathbf{V}])] + I(Z_0 = 1)\left[M_\mu + \Lambda_0\exp(\mu)\right]E[X|\mathbf{V}] \\
&+ I(Z_0 = 2)\left[M_\mu - \frac{\mathcal{F}_1}{1-\mathcal{F}_1}\Lambda_0\exp(\mu)\right]E[X|\mathbf{V}]
\end{aligned}$$

and

$$\begin{aligned}
S_{\beta_1}^2(\boldsymbol{\vartheta}) =& I(Z_0 \in \{1,2\})\Delta M_\mu^2(X - E[X|\mathbf{V}])^2 \\
&+ \left\{I(Z_0 = 1)\left[M_\mu + \Lambda_0\exp(\mu)\right] + I(Z_0 = 2)\left[M_\mu - \frac{\mathcal{F}_1}{1-\mathcal{F}_1}\Lambda_0\exp(\mu)\right]\right\}^2 E[X|\mathbf{V}]^2 \\
&+ 2 \times \Delta M_\mu(X - E[X|\mathbf{V}]) \times E[X|\mathbf{V}]\Big\{I(Z_0 = 1)\left[M_\mu + \Lambda_0\exp(\mu)\right] + I(Z_0 = 2)\left[M_\mu - \frac{\mathcal{F}_1}{1-\mathcal{F}_1}\Lambda_0\exp(\mu)\right]\Big\} \\
=& K_1 + K_2 + K_3
\end{aligned}$$

with K_j corresponding to the term in the jth line.

Since under $\beta_1 = o(1)$, $P(X|\Delta, \mathbf{W}, S_1, \delta_2, Z_0) = P(X|\mathbf{V})$ we have

$$\mathcal{I}_{\beta_1\beta_1} = E[S_{\beta_1}^2(\boldsymbol{\vartheta})] = \mathcal{I}_{\beta_1\beta_1}^{\Delta} + \mathcal{I}_{\beta_1\beta_1}^{0},$$

where $\mathcal{I}_{\beta_1\beta_1}^{\Delta} = E[K_1]$ is given by

$$E\Big[\Delta I(Z_0 \in \{1,2\})M_\mu^2\mathbf{V}\mathrm{ar}(X|\mathbf{V})\Big] = E_{\mathbf{W}}\Big[E[\Delta|\mathbf{W}]E[I(Z_0 \in \{1,2\})M_\mu^2|\Delta = 1, \mathbf{W}]\mathbf{V}\mathrm{ar}(X|\mathbf{V})\Big], \tag{C.3.2}$$

with $E[\Delta|\mathbf{W}, X] = E[\Delta|\mathbf{W}]$ and $E[I(Z_0 \in \{1,2\})M_\mu|\Delta = 1, \mathbf{W}, X] = E[I(Z_0 \in \{1,2\})M_\mu|\Delta = 1, \mathbf{W}]$. We also have $\mathcal{I}^0_{\beta_1\beta_1} = E[K_2]$ given by

$$E\left[\left\{I(Z_0 = 1)\left[M_\mu + \Lambda_0 \exp(\mu)\right] + I(Z_0 = 2)\left[M_\mu - \frac{\mathcal{F}_1}{1-\mathcal{F}_1}\Lambda_0 \exp(\mu)\right]\right\}^2 E[X|\mathbf{V}]^2\right]$$

and $E[K_3] = 0$ since $E_{X|\mathbf{V}}[K_3] = 0$.

C.3.2 Derivation of $\mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}$ and $\mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}$

When $\beta_1 = o(1)$, we have $\mathscr{S}_{\boldsymbol{\theta}_\circ} = \partial \log \mathscr{L}_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}_\circ$ and $S_{\boldsymbol{\theta}_\circ}(\boldsymbol{\vartheta}) = \partial \log L_p(\boldsymbol{\vartheta})/\partial \boldsymbol{\theta}_\circ$ given by

$$I(Z_0 = 1)\left[\mathscr{S}_{\boldsymbol{\theta}_\circ} + \frac{\partial \log \mathcal{F}_1}{\partial \boldsymbol{\theta}_\circ}\right] + I(Z_0 = 2)\left[\mathscr{S}_{\boldsymbol{\theta}_\circ} - \frac{\mathcal{F}_1}{1-\mathcal{F}_1}\frac{\partial \log \mathcal{F}_1}{\partial \boldsymbol{\theta}_\circ}\right].$$

Hence, $\mathcal{I}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ} = E[S_{\boldsymbol{\theta}_\circ}S'_{\boldsymbol{\theta}_\circ}]$ does not rely on the design rules of Δ. In addition, we have $\mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ} = E[S_{\beta_1}(\boldsymbol{\vartheta})S'_{\boldsymbol{\theta}_\circ}(\boldsymbol{\vartheta})]$ given by

$$E\Big\{\Big[I(Z_0 = 1)\left[M_\mu + \Lambda_0 \exp(\mu)\right] + I(Z_0 = 2)\left[M_\mu - \frac{\mathcal{F}_1}{1-\mathcal{F}_1}\Lambda_0 \exp(\mu)\right]\Big]E[X|\mathbf{V}]S_{\boldsymbol{\theta}_\circ}(\boldsymbol{\vartheta})\Big\}$$
$$+ E\Big\{I(Z_0 \in \{1,2\})\Delta M_\mu S_{\boldsymbol{\theta}_\circ}(\boldsymbol{\vartheta})(X - E[X|\mathbf{V}])\Big\}$$

and the second component equals zero, so $\mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}$ does not rely on Δ either.

C.3.3 Derivation of $\mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\beta_1}$

We use a projection approach to find $\mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\beta_1}$. First, when $\beta_1 = o(1)$,

$$\begin{aligned}
S_{\boldsymbol{\eta}}(\boldsymbol{\vartheta}) =& I(Z_0 = 1)\Big[\Delta\mathscr{S}_{\boldsymbol{\eta}} + (1-\Delta)E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}}] - E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}}]\Big] \\
&+ I(Z_0 = 2)\Big[\Delta\mathscr{S}_{\boldsymbol{\eta}} + (1-\Delta)E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}}] + E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}}]\frac{\mathcal{F}_1}{1-\mathcal{F}_1}\Big] \\
=& \Delta I(Z_0 \in \{1,2\})\mathscr{S}_{\boldsymbol{\eta}}.
\end{aligned}$$

as $E_{X|\mathbf{V}}[\mathscr{S}_{\boldsymbol{\eta}}] = 0$. We aim to project S_{β_1} onto the linear span of all possible $S_{\boldsymbol{\eta}}$ (Tao et al., 2020). Note that given $\mathbf{W}$, $I(Z_0 \in \{1,2\})$ is known. Let

$$S^*_{\boldsymbol{\eta}} = \Delta I(Z_0 \in \{1,2\})J(X, \mathbf{W})$$

with $S_{\beta_1} - S^*_{\boldsymbol{\eta}}$ orthogonal to this linear span, $J(X, \mathbf{W})$ is solved by

$$E[S_{\boldsymbol{\eta}}S_{\beta_1}] = E[S_{\boldsymbol{\eta}}S^*_{\boldsymbol{\eta}}].$$

Specifically,

$$\begin{aligned} S_{\boldsymbol{\eta}} S_{\beta_1} =& \Delta \mathscr{S}_{\boldsymbol{\eta}} \times \Big\{ I(Z_0 \in \{1,2\}) \Delta M_\mu (X - E[X|\mathbf{V}]) \\ &+ \left(I(Z_0 = 1) \left[M_\mu + \Lambda_0 \exp(\mu) \right] + I(Z_0 = 2) \left[M_\mu - \frac{\mathcal{F}_1}{1-\mathcal{F}_1} \Lambda_0 \exp(\mu) \right] \right) E[X|\mathbf{V}] \Big\}. \end{aligned}$$

We also have

$$\begin{aligned} E[S_{\boldsymbol{\eta}} S_{\beta_1}] =& E_{\mathbf{W},X} \Big\{ I(Z_0 \in \{1,2\}) E[M_\mu | \Delta = 1, \mathbf{W}] (X - E[X|\mathbf{V}]) E[\Delta|\mathbf{W}] \mathscr{S}_{\boldsymbol{\eta}} \Big\} \\ &+ E_{\mathbf{W}} \Big\{ \left[I(Z_0 \in \{1,2\}) E[M_\mu | \Delta = 1, \mathbf{W}] + E\left[\left(I(Z_0 = 1) - I(Z_0 = 2) \frac{\mathcal{F}_1}{1-\mathcal{F}_1} \right) \Lambda_0 \exp(\mu) \Big| \mathbf{W} \right] \right] \\ &\times E[X|\mathbf{V}] E[\Delta|\mathbf{W}] E[\mathscr{S}_{\boldsymbol{\eta}}|\mathbf{W}] \Big\} \\ =& E_{\mathbf{W},X} \Big\{ I(Z_0 \in \{1,2\}) E[\Delta|\mathbf{W}] E[M_\mu | \Delta = 1, \mathbf{W}] \mathscr{S}_{\boldsymbol{\eta}} (X - E[X|\mathbf{V}]) \Big\}. \end{aligned} \tag{C.3.3}$$

And

$$S_{\boldsymbol{\eta}} S_{\boldsymbol{\eta}}^* = I(Z_0 \in \{1,2\}) \Delta \mathscr{S}_{\boldsymbol{\eta}} J(X, \mathbf{W}).$$

We have

$$\begin{aligned} E[S_{\boldsymbol{\eta}} S_{\boldsymbol{\eta}}^*] =& E\Big\{ I(Z_0 \in \{1,2\}) \Delta \mathscr{S}_{\boldsymbol{\eta}} J(X, \mathbf{W}) \Big\} \\ =& E_{\mathbf{W},X} \Big\{ E[\Delta|\mathbf{W}] \mathscr{S}_{\boldsymbol{\eta}} I(Z_0 \in \{1,2\}) J(X, \mathbf{W}) \Big\}. \end{aligned} \tag{C.3.4}$$

By comparing formulas (C.3.3) and (C.3.4), we obtain

$$J(X, \mathbf{W}) = E[M_\mu | \Delta = 1, \mathbf{W}](X - E[X|\mathbf{V}]).$$

Therefore,

$$\begin{aligned} S_{\beta_1} S_{\boldsymbol{\eta}}^* =& \Delta J(X, \mathbf{W}) S_{\beta_1} \\ =& \Delta J(X, \mathbf{W}) \Big\{ I(Z_0 \in \{1,2\}) M_\mu X + \left(I(Z_0 = 1) - I(Z_0 = 2) \frac{\mathcal{F}_1}{1-\mathcal{F}_1} \right) \Lambda_0 \exp(\mu) E[X|\mathbf{V}] \Big\} \end{aligned}$$

Then $\mathcal{I}_{\beta_1 \boldsymbol{\eta}} \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1} \mathcal{I}_{\boldsymbol{\eta}\beta_1} = E[S_{\beta_1} S_{\boldsymbol{\eta}}^*]$ is given by

$$\begin{aligned} & E_{\mathbf{W},X} \Big[I(Z_0 \in \{1,2\}) E[\Delta|\mathbf{W}] J(\mathbf{W}, X)^2 \Big] \\ =& E_{\mathbf{W}} \left\{ I(Z_0 \in \{1,2\}) E[\Delta|\mathbf{W}] E[M_\mu | \Delta = 1, \mathbf{W}]^2 \mathrm{Var}(X|\mathbf{V}) \right\} \end{aligned} \tag{C.3.5}$$

C.3.4 Design efficiency when $\beta_1 = o(1)$

We rewrite V_{β_1} defined in (C.3.1) as $\mathcal{I}^{\Delta}_{\beta_1\beta_1} + \mathcal{I}^{0}_{\beta_1\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}^{-1}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\beta_1}$. The components $\mathcal{I}^{0}_{\beta_1\beta_1}$ and $\mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}^{-1}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1}$ does not depend on Δ. Given formulas (C.3.2) and (C.3.5), $\mathcal{I}^{\Delta}_{\beta_1\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\beta_1}$, the component depending on Δ, is given by

$$E_{\mathbf{W}}\Big\{I(Z_0 \in \{1,2\})E[\Delta|\mathbf{W}]\mathrm{Var}[M_\mu|\mathbf{W},\Delta=1]\mathrm{Var}(X|\mathbf{V}).\Big\}. \tag{C.3.6}$$

An optimal sampling rule of Δ would maximize (C.3.6) so to minimize the variance of $\hat{\beta}_1$,

$$\left[\mathcal{I}^{\Delta}_{\beta_1\beta_1} + \mathcal{I}^{0}_{\beta_1\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\theta}_\circ}\mathcal{I}^{-1}_{\boldsymbol{\theta}_\circ\boldsymbol{\theta}_\circ}\mathcal{I}_{\boldsymbol{\theta}_\circ\beta_1} - \mathcal{I}_{\beta_1\boldsymbol{\eta}}\mathcal{I}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\beta_1}\right]^{-1}.$$

With the combined prevalent cohort data, formula (C.3.6) suggests we select subjects with the largest and the smallest values of $M_\mu\mathrm{Var}(\mathrm{X}|\mathbf{V})^{1/2}$.

Table C.1: Full data analysis using partial likelihood when Assumption 3 is violated: estimation results, including empirical bias (BIAS, $\times$ 100), empirical standard deviation (ESE), average sandwich standard error (ASE) and covarage probability (ECP), are summarized over 1000 replicated samples with $N_1 = N_2 = 1000$, $\beta_2 = 0.4$, and $\mathrm{OR}(X, V) = 2$.

			$100P(\delta_2 = 1 \mid Z_0 = 1) = 10$				$100P(\delta_2 = 1 \mid Z_0 = 1) = 30$			
$\exp(\nu)$		TRUE	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
					$(\beta_1, \beta_3) = (0, 0)$					
1.2	α_0	-3.55	-9.08	0.17	0.18	0.95	-5.92	0.14	0.13	0.94
	β_1	0	0.38	0.10	0.10	0.94	0.17	0.09	0.09	0.93
	β_2	0.4	1.11	0.13	0.13	0.96	-0.08	0.09	0.09	0.97
	β_3	0	1.00	0.05	0.05	0.96	1.01	0.04	0.04	0.96
					$(\beta_1, \beta_3) = (0, 1)$					
	α_0	-6.51	-15.38	0.34	0.33	0.94	-4.13	0.21	0.21	0.95
	β_1	0	1.16	0.10	0.10	0.93	0.61	0.09	0.09	0.94
	β_2	0.4	0.68	0.11	0.11	0.95	-0.11	0.09	0.09	0.92
	β_3	1	3.05	0.09	0.09	0.94	0.50	0.06	0.06	0.96
					$(\beta_1, \beta_3) = (0.7, 0)$					
	α_0	-3.61	-10.14	0.18	0.17	0.93	-6.6	0.13	0.13	0.92
	β_1	0.7	2.21	0.10	0.09	0.94	1.42	0.09	0.08	0.94
	β_2	0.4	1.55	0.13	0.13	0.95	0.36	0.09	0.09	0.94
	β_3	0	1.15	0.05	0.05	0.94	0.98	0.04	0.04	0.95
					$(\beta_1, \beta_3) = (0.7, 1)$					
	α_0	-6.58	-14.02	0.32	0.32	0.94	-4.34	0.20	0.20	0.95
	β_1	0.70	2.09	0.10	0.09	0.95	1.08	0.09	0.09	0.96
	β_2	0.4	0.93	0.12	0.11	0.94	-0.05	0.09	0.09	0.92
	β_3	1	2.67	0.08	0.08	0.95	0.60	0.06	0.06	0.95
					$(\beta_1, \beta_3) = (0, 0)$					
1.5	α_0	-3.35	-17.5	0.17	0.18	0.88	-11.81	0.12	0.13	0.90
	β_1	0	0.18	0.10	0.10	0.95	-0.09	0.09	0.09	0.94
	β_2	0.4	1.82	0.12	0.13	0.95	-0.29	0.09	0.09	0.97
	β_3	0	1.49	0.05	0.05	0.96	1.64	0.04	0.04	0.95
					$(\beta_1, \beta_3) = (0, 1)$					
	α_0	-6.29	-34.09	0.32	0.32	0.84	-11.25	0.20	0.20	0.92
	β_1	0	0.11	0.09	0.10	0.95	0.14	0.09	0.09	0.95
	β_2	0.4	2.72	0.11	0.11	0.94	1.00	0.09	0.08	0.95
	β_3	1	6.81	0.08	0.09	0.9	1.51	0.05	0.06	0.95
					$(\beta_1, \beta_3) = (0.7, 0)$					
	α_0	-3.42	-17.43	0.18	0.17	0.82	-11.7	0.14	0.13	0.86
	β_1	0.7	4.57	0.09	0.09	0.93	2.39	0.08	0.08	0.94
	β_2	0.4	1.86	0.13	0.12	0.95	-0.67	0.10	0.09	0.94
	β_3	0	1.41	0.05	0.05	0.94	1.60	0.04	0.04	0.94
					$(\beta_1, \beta_3) = (0.7, 1)$					
	α_0	-6.35	-28.55	0.3	0.3	0.87	-9.74	0.20	0.20	0.92
	β_1	0.7	3.38	0.10	0.09	0.93	1.50	0.09	0.09	0.95
	β_2	0.4	1.73	0.11	0.10	0.96	-0.01	0.09	0.08	0.94
	β_3	1	5.42	0.08	0.08	0.91	1.28	0.06	0.05	0.93