The New England Journal of Statistics in Data Science Volume 0, 1–18 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS110

Cost-Effective Biomarker Studies for Cox Regression with Serial Biospecimens: Bias Mitigation and Improved Efficiency

Jianchu CHEN* and Richard J. COOK

Abstract

Disease registries often involve cohorts of patients seen repeatedly at tertiary care centers with biospecimens collected at periodic clinic visits. With such registries it is possible to study the association between time-dependent biomarkers and a failure time of interest through Cox regression. However it can be prohibitively expensive or labor-intensive to assay all biospecimens for all members of the registry and in practice it is common to select a single biosample to assay for each individual. We derive the asymptotic bias of estimators obtained from the Cox model based on common strategies. We investigate the nature of this bias from a joint multistate model for the marker and failure processes. Alternative selection methods are then developed for consistent and efficient sub-sampling under budgetary constraints. The asymptotic relative efficiency of regression coefficients obtained from the Fisher information is then explored and an optimal design is identified within a class of designs. The proposed selection method is illustrated in a registry of patients with psoriatic arthritis where we investigate the association between a time-dependent biomarker (ESR_CRP) and the risk of developing arthritis mutilans.

KEYWORDS AND PHRASES: Misspecified Cox models, Time-dependent covariates, Intermittent observation, Multistate models, Cost-effective sampling, Survival analysis.

1. INTRODUCTION

In studies of chronic diseases, interest often lies in studying the association between time-dependent biomarkers and an event such as disease progression or death. We consider the setting of a disease registry in which recruited patients provide biospecimens (e.g. urine or serum samples) at periodic clinic visits to measure established markers of disease activity – residual material from each biospecimen is then stored in a biobank. Disease registries are often maintained by research groups, and novel biomarkers are routinely proposed for study. In such settings, stored biospecimens can be assayed to provide values of a novel biomarker at the times the samples were taken. The Cox regression models [8] are then used for modeling the association between the marker values and the failure time of interest.

While disease registries like this offer an unprecedented opportunity to study chronic disease processes, it is often too labour intensive or costly to assay all available biospecimens when a new biomarker is of interest. It is common, instead, to select a single biospecimen to assay per individual. *Ad hoc* strategies include:

a) assaying the biospecimen collected most recently before failure or censoring;
b) assaying the biospecimen collected most recently before failure or censoring and using this sampling time as a left-truncation time;
c) assaying the first biospecimen collected,
d) assaying the first biospecimen collected and treating this sampling time as a left-truncation time.

The (misguided) rationale for a) is that this marker value is likely the closest to the "true" biomarker value at the failure time of the respective individuals. This rationale does not address the fact that the censored individuals have their marker measured at a random time unrelated to their (censored) failure time. Strategy b) is put forward to avoid carrying "backward" measurements made later in the course of follow-up to the time of recruitment to the registry. The rationale for c) is that it may be viewed as approximating a "baseline" marker value – it offers a consistent strategy for those failing and those who are censored, but involves carrying this marker value backward to the time of recruitment and forward to the time of failure or censoring. Strategy d) simply eliminates the carrying backward of the first marker assessment to the time of clinic entry. All strategies yield biased estimators, but the precise nature of these biases have not been systematically investigated to date. Using large sample theory of Cox regression coefficients from these selection and carry-forward approaches (without or with left-truncation) [17, 12] we describe a framework to study the limiting biases of regression coefficients. We focus on a particular data generation procedure for our detailed investigation but lay out the general theory for this investigation so that other models can be investigated.

*Corresponding author.

A valid analytical approach is to jointly model for the marker and failure processes. We formulate such a joint model for the case of a dynamic binary marker and a fixed baseline covariate acting on a failure time through a Cox regression model. We then propose and investigate simple cost-preserving selection strategies for efficient choice of individuals and biospecimens to assay for consistent parameter estimation through this joint model. We compare different strategies in terms of the efficiency of the resulting estimators of the regression coefficients in the Cox model.

The remainder of this paper is organized as follows. In Section 2 we define notation and formulate a three-state multistate model for joint modeling of the marker and the failure processes. The joint marker-failure multistate model is then extended to incorporate random censoring and marker-dependent visit processes. In Section 3 we present the Cox model and review the large sample theory associated with misspecified Cox models. We also consider the four aforementioned *ad hoc* cost-effective strategies and calculate limiting values of the corresponding estimators from the fitted Cox models. Determinants of the asymptotic biases of estimators based on these strategies are investigated through simulation studies. In Section 4, we describe an alternative simple cost-effective selection strategy where the final dataset for inference consists of a subsample of individuals for whom biospecimens collected at every visit are assayed, and a subsample of individuals for whom only the first collected biospecimen is assayed. A joint model is fitted to these selected individuals and the effect of the fraction of the number of serial assays of biospecimens (to the total number of biospecimen assays) on the efficiency of estimating the covariate effects is explored via numerical studies. In Section 5 we illustrate the proposed selection strategy through analysis of data from the University of Toronto Psoriatic Arthritis Cohort Study where we consider the inflammatory biomarker c-reactive protein and investigate its association with the development of arthritis mutilans. Concluding remarks are provided in Section 6.

2. A FRAMEWORK FOR JOINT MODELING

2.1 The Marker-Failure Process

Consider a time-dependent binary marker process and suppose the primary aim is to study the association between the marker and a failure time of interest. Binary markers are common and may represent an indicator that a particular symptom or condition is present, but they also arise when clinicians dichotomize continuous markers to define two levels representing normal or elevated marker values. More categories can of course be considered, as can settings with multiple discrete markers, but we retain this simple model for convenience when describing this framework. Failure may be defined to represent any adverse clinical event or death - we use the term failure for generality but it is best viewed as any terminal event of interest. To define and model the association we consider a 3-state process for the marker and the failure time depicted in Figure 1 where states 0 and 1 represent possible marker values held by individuals who are alive, and state 2 is an absorbing state entered upon failure at time T. We let $Z(s)$ denote the state occupied at time s, $s \geq 0$, and $\{Z(s), 0 \leq s\}$ denote this joint 3-state marker-failure process. Let $H(t) = \{Z(s), 0 \leq s < t, X_2\}$ denote the history of the joint marker-failure process up to any time t^- along with a $p \times 1$ vector of fixed covariates X_2.

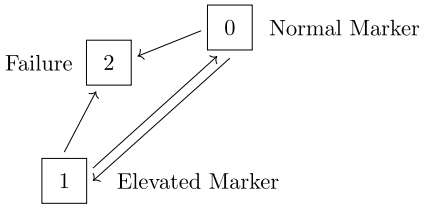


Figure 1: A joint marker-failure process for a binary marker (0 =normal, 1 =elevated) and a single absorbing state (2) entered upon failure.

The intensity $\lambda_k(s \mid H(s))$ governing $1-k \to k$ transitions between the marker states is defined as

$$\lim_{\Delta s \downarrow 0} \frac{P(Z(s+\Delta s^-) = k \mid Z(s^-) = 1-k, H(s))}{\Delta s} \tag{2.1a}$$

for $k = 0, 1$. Likewise, the intensity for failure from state k, $\lambda_{k2}(s \mid H(s))$, is

$$\lim_{\Delta s \downarrow 0} \frac{P(Z(s+\Delta s^-) = 2 \mid Z(s^-) = k, H(s))}{\Delta s} \tag{2.1b}$$

for $k = 0, 1$. Cox regression models are often formed by constraining $\lambda_{12}(s \mid H(s))$ and $\lambda_{02}(s \mid H(s))$ to be proportional. If we let $X_1(s) = \mathbb{1}(Z(s^-) = 1)$ represent the internal covariate indicating an elevated marker and X_2 the fixed covariates, then $X(s) = (X_1(s), X_2')'$ is the full set of covariates. We may then specify the failure intensity as

$$\lambda_2(s \mid H(s)) = \lambda_{02}(s) \exp(X'(s)\beta) \tag{2.2}$$

where $X'(s)\beta = X_1(s)\beta_1 + X_2'\beta_2$, with $\beta_2 = (\beta_{21}, \ldots, \beta_{2p})'$. Here $\exp(\beta_1)$ is the relative risk of failure associated with an elevated marker while controlling for the baseline covariates X_2; likewise $\exp(\beta_{2j})$ is the relative risk associated with a one unit increase in X_{2j} while all other covariates are fixed, $j = 1, \ldots, p$.

2.2 The Observation and Coarsening Processes

Let τ denote an administrative censoring time, W a random censoring (withdrawal) time, and $C = \min(\tau, W)$; we assume an individual is followed continuously for failure over $(0, V]$ where $V = \min(T, C)$. Let $Y(s) = \mathbb{1}(s \leq C)$ indicate that an individual is uncensored at time s and $Y^\dagger(s) = \mathbb{1}(s \leq T)$. Then $\bar{Y}(s) = Y(s)Y^\dagger(s)$ indicates that

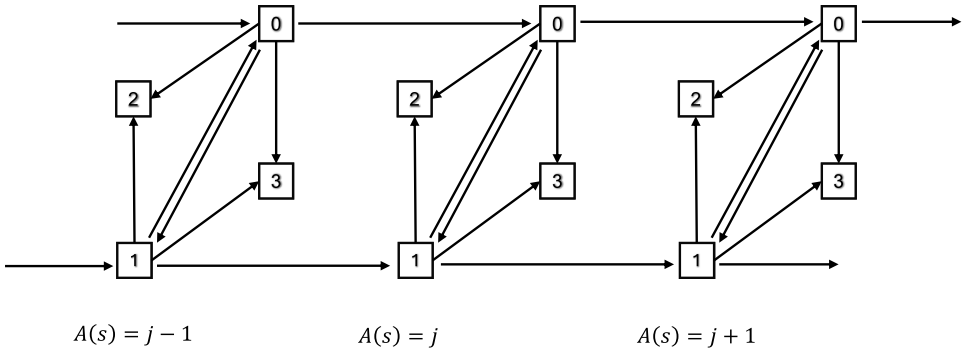


Figure 2: An expanded joint process $\mathcal{Z}(s) = (A(s), Z(s))$ incorporating the visit process and the marker-failure-censoring process.

an individual is under follow-up and at risk of failure at time $s > 0$.

The marker process is not under continuous observation but marker values can be determined by laboratory tests. We consider the case where biospecimens are taken at clinic visits and a marker value is ascertained if a biospecimen is assayed. In what follows we initially assume that all biospecimens are assayed at all visits but in Sections 3 and 4 we consider the case where biospecimens are available but only some are selected to be assayed. Let A_j denote the random time of the jth clinic visit and biospecimen collection and a_j denote its realized value, $j = 1, \ldots$. We define the visit counting process by letting $dA(s) = \sum_{j=1}^{\infty} \mathbb{1}(s = A_j)$, with the cumulative visit count $A(s) = \int_0^s \bar{Y}(u) dA(u)$; $\{A(s), 0 \leq s\}$ denotes the counting process for visits. We denote the full history of the joint marker-failure process and the censoring and visit processes at t^- as $\mathcal{H}(t) = \{\bar{Y}(s), dA(s), Z(s), 0 \leq s < t, X_2\}$, which includes the baseline covariates. We refer to this as the "full history" since it includes information which will not be available, such as the number and times of transitions for the marker process. We define it to help define and carefully discuss independence conditions.

To fully characterize the observation process, we model the random censoring time. We let $dC(s) = \mathbb{1}(W = s)$, $C(s) = \int_0^s \bar{Y}(u) dC(u)$ and $\{C(s), 0 \leq s\}$ denote the counting process for the random censoring. Let $\Delta C(s) = C(s + \Delta s^-) - C(s^-)$, the intensity for the random censoring time is then

$$\lim_{\Delta s \downarrow 0} \frac{P(\Delta C(s) = 1 \mid \mathcal{H}(s))}{\Delta s} = \bar{Y}(s) \lambda_3(s \mid \mathcal{H}(s)) \,. \qquad (2.3)$$

The intensity for the visit process governing the biospecimen collection is most generally defined as

$$\lim_{\Delta s \downarrow 0} \frac{P(\Delta A(s) = 1 \mid \mathcal{H}(s))}{\Delta s} = \bar{Y}(s) \lambda_4(s \mid \mathcal{H}(s)) \qquad (2.4)$$

where $\Delta A(s) = A(s + \Delta s^-) - A(s^-)$.

We now define counting process notation and intensities for the marker and failure processes *under this observation scheme*. For the marker process let $dN_k(s) = \mathbb{1}(Z(s^-) = 1 - k, Z(s) = k)$, $d\bar{N}_k(s) = \bar{Y}(s) dN_k(s)$, $\bar{N}_k(s) = \int_0^s d\bar{N}_k(u)$ and $\{\bar{N}_k(s), 0 \leq s\}$ denote the counting process for observable $k - 1 \to k$ transitions, $k = 0, 1$. The intensities for transitions between marker states are

$$\lim_{\Delta s \downarrow 0} \frac{P(\Delta \bar{N}_k(s) = 1 \mid Z(s^-) = 1 - k, \mathcal{H}(s))}{\Delta s} = \bar{Y}(s) \lambda_k(s \mid \mathcal{H}(s))$$

where $\Delta \bar{N}_k(s) = \bar{N}_k(s + \Delta s^-) - \bar{N}_k(s^-)$, for $k = 0, 1$. If the observation process does not affect the transition intensities between the marker values then $\lambda_k(s \mid \mathcal{H}(s)) = \lambda_k(s \mid H(s))$ where the latter is defined in (2.1a). We let $dN_2(s) = \mathbb{1}(T = s)$, $d\bar{N}_2(s) = \bar{Y}(s) dN_2(s)$ indicate a failure occurred and observed at s, $\bar{N}_2(s) = \int_0^s d\bar{N}_2(u)$, and $\{\bar{N}_2(s), 0 \leq s\}$ denote the counting process for observed failures. The failure intensity is then

$$\lim_{\Delta s \downarrow 0} \frac{P(\Delta \bar{N}_2(s) = 1 \mid \mathcal{H}(s))}{\Delta s} = \bar{Y}(s) \lambda_2(s \mid \mathcal{H}(s))$$

where $\Delta \bar{N}_2(s) = \bar{N}_2(s + \Delta s^-) - \bar{N}_2(s^-)$. If the failure process is conditionally independent of the observation processes [6],

$$\lambda_2(s \mid \mathcal{H}(s)) = \lambda_2(s \mid H(s))$$

with $\lambda_2(s \mid H(s))$ defined in (2.2).

Figure 2 presents a state space diagram for the 3-state process of interest, augmented to reflect the visit process and random censoring. Specifically we introduce state 3, which is entered upon withdrawal from follow-up. Transitions from the left collection of four states to the middle set indicate occurrence of the jth visit. We represent this joint visit, censoring, and marker-failure process by $\{\mathcal{Z}(s), 0 < s\}$ where $\mathcal{Z}(s) = (A(s), Z(s))$. Under the observation scheme described above, the observed history of the marker-failure process and the observation processes up to time t^- is

$$\bar{\mathcal{H}}(t) = \Big\{ \bar{Y}(s), dA(s), d\bar{N}_2(s), 0 \leq s < t,$$

$$\left\{(a_j, Z(a_j)), j = 1, \ldots, A(t^-)\right\}, X_2\Big\}. \quad (2.5)$$

A conventional analysis of data from processes under such an intermittent observation scheme for the marker process involves carrying forward recorded marker values from one assessment to the next and to update the covariate values at that time. This means using the most recently assessed marker value $X_1^\circ(s) := X_1(a_{A(s)})$ in lieu of $X_1(s)$ and fitting a standard Cox model for the failure time with covariates $X_1^\circ(s)$ and X_2. Cook et al. [7] investigate the asymptotic bias of covariate effects estimated from this conventional method. This presumes that all biospecimens are assayed which of course is not feasible in many settings due to the expense and time constraints. We next turn to commonly proposed *ad hoc* strategies when cost constraints limit the number of assays possible.

3. BIASES FROM NAIVE STRATEGIES

3.1 Ad Hoc Strategies for Cox Regression

In a large database biospecimens may be collected periodically over follow-up and stored in a biobank, but it may be prohibitively expensive to assess all of them. In settings with electronic medical records it may likewise be necessary to delve into medical charts to retrieve more detailed clinical information than is routinely recorded. This can be laborious and time-consuming. In such cases, as discussed in Section 1, researchers often select only one visit or biospecimen per individual for the measurement of expensive biomarkers. Common approaches include using the biospecimen collected at the first visit to measure the marker value, or the biospecimen collected at the last visit before failure or censoring at time V. We let X_1° denote the marker value assessed from the biospecimen collected at the selected visit. Specifically, $X_1^\circ = X_1^\circ(V) = X_1(a_{A(V)})$ when the biospecimen assayed is collected at the last visit before censoring or failure, and $X_1^\circ = X_1^\circ(a_1) = X_1(a_1)$ when the biospecimen collected at the first visit is assayed.

The observed history of the marker-failure and observation processes up to time t^- under such a scheme is now

$$\bar{\mathcal{H}}^\circ(t) = \left\{\bar{Y}(s), dA(s), d\bar{N}_2(s), 0 \le s < t, X_1^\circ, X_2\right\}.$$

A conventional analysis involves using X_1° in place of $X_1(s)$, treating the value as fixed from time 0 to V which involves both "carrying backward" and "carrying forward" the marker value. Alternatively one can treat the time the assayed biospecimen was collected as a left truncation time for the failure time analysis and only carry forward to V. Fitting a working Cox model for the failure time is then carried based on

$$\Gamma(s \mid X^\circ) = \Gamma_0(s)\exp(\psi' X^\circ) \quad (3.1)$$

where $X^\circ = (X_1^\circ, X_2')'$, $\psi = (\psi_1, \psi_2')'$, and $\Gamma_0(s)$ is the baseline intensity for failure under this specification. We consider four *ad hoc* strategies described in Section 1. The deviation of X_1° from the true marker process $X_1(s)$ under these strategies is illustrated in Figure 3 through an example. The latent marker process $X_1(s)$ is shown by the blue dotted line, whereas X_1° is shown by the black solid line. The top panels illustrate strategies that assay the marker value using the first biospecimen sample, while the bottom panels illustrate strategies that use the last biospecimen sample. Within each row, the left panel presents strategies that carry the assayed marker value both forward and backward, whereas the right panel presents strategies that treat the marker value as left-truncated. The shaded regions reflect the time intervals when the covariate is misclassified. Such a misclassification generally results in asymptotic bias in the estimation of parameters of interest. In particular, we will explore the asymptotic bias of this misspecification on estimating parameters β_1 and β_2 when the true model for the failure process is (2.2).

Suppose we have already selected a visit (e.g., the last visit) at which the collected biospecimen sample will be assayed to measure the marker value, the observed data from following a sample of n independent processes to the administrative censoring time τ is $\{\bar{\mathcal{H}}_i^\circ(\tau), i = 1, 2, \ldots, n\}$. Treating the marker value fixed with X_1° from time 0 to the time until failure or censoring at V, the Cox partial likelihood score functions based on (3.1) can be formulated as

$$U_1(s) = \sum_{i=1}^n \mathbb{1}(A_i(\tau) \ge 1)\bar{Y}_i(s)\left\{dN_{i2}(s) - d\Gamma(s \mid X_i^\circ)\right\} \quad (3.2a)$$

for $s > 0$, and

$$U_2 = \sum_{i=1}^n \mathbb{1}(A_i(\tau) \ge 1)\int_0^\tau \bar{Y}_i(s)\left\{dN_{i2}(s) - d\Gamma(s \mid X_i^\circ)\right\} X_i^\circ \quad (3.2b)$$

since each individual must have at least one visit to provide a biospecimen for measuring the marker value X_{i1}°. Solving $U_1(s) = 0$ for $d\Gamma_0(s)$ at a fixed value of ψ gives the Breslow profile estimate [13] modified to reflect the requirement of at least one visit

$$d\tilde{\Gamma}_0(s;\psi) = \frac{\sum_{i=1}^n \mathbb{1}(A_i(\tau) \ge 1)\bar{Y}_i(s)dN_{i2}(s)}{\sum_{i=1}^n \mathbb{1}(A_i(\tau) \ge 1)\bar{Y}_i(s)\exp(\psi' X_i^\circ)} \quad (3.3)$$

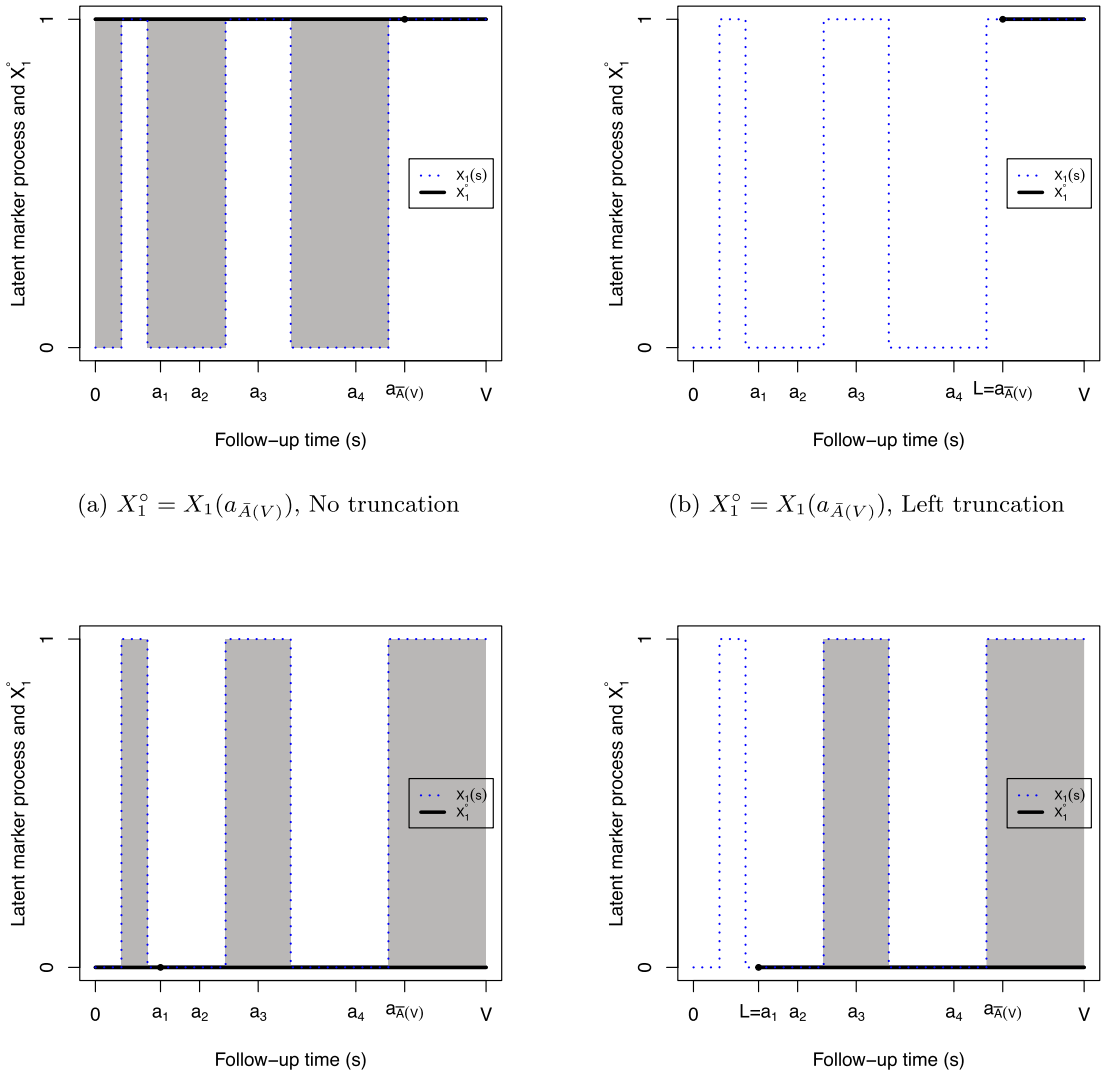
for $s > 0$. Upon substitution into (3.2b) we obtain a partial profile score function $U(\psi)$ for ψ as

$$\sum_{i=1}^n \int_0^\tau \mathbb{1}(A_i(\tau) \ge 1)\bar{Y}_i(s)\left\{X_i^\circ - \frac{\sum_{i=1}^n R_i^{(1)}(s;\psi)}{\sum_{i=1}^n R_i^{(0)}(s;\psi)}\right\} dN_{i2}(s) \quad (3.4)$$

where

$$R_i^{(k)}(s;\psi) = \mathbb{1}(A_i(\tau) \ge 1)\bar{Y}_i(s)(X_i^\circ)^{(k)}\exp(\psi' X_i^\circ) \quad (3.5)$$

for $k = 0, 1$, with $a^{(k)} = 1$ for $k = 0$ and $a^{(k)} = a$ for $k = 1$. Setting $U(\psi) = 0$ and solving for ψ gives $\tilde{\psi}$. Upon substituting $\tilde{\psi}$ into (3.3), we obtain $d\tilde{\Gamma}_0(s;\tilde{\psi})$ and thus $\tilde{\Gamma}_0(s) = \int_0^s d\tilde{\Gamma}_0(u)$.



(a) $X_1^\circ = X_1(a_{\bar{A}(V)})$, No truncation

(b) $X_1^\circ = X_1(a_{\bar{A}(V)})$, Left truncation

(c) $X_1^\circ = X_1(a_1)$, No truncation

(d) $X_1^\circ = X_1(a_1)$, Left truncation

Figure 3: An illustration of the discordance between X_1° and the true value $X_1(s)$ under four *ad hoc* strategies when minimizing the number of biospecimen assays to one per individual.

Struthers and Kalbfleisch [17] show that the limiting value of $\tilde{\psi}$, denoted by ψ^*, is the solution to

$$\int_0^\tau \left\{ r^{(1)}(s) - \frac{r^{(1)}(s;\psi)}{r^{(0)}(s;\psi)} r^{(0)}(s) \right\} = 0\,, \tag{3.6}$$

where

$$r^{(k)}(s) = E\{R_i^{(k)}(s)\} \tag{3.7}$$

with

$$R_i^{(k)}(s) = \mathbb{1}(A_i(\tau) \geq 1)\bar{Y}_i(s)(X_i^\circ)^{(k)} dN_{i2}(s) \tag{3.8}$$

and

$$r^{(k)}(s;\psi) = E\{R_i^{(k)}(s;\psi)\} \tag{3.9}$$

with $R_i^{(k)}(s;\psi)$ defined in (3.5) for $k = 0, 1$. The expectations are taken with respect to the true data generating model which we described in general in Section 2 for the current setting. As $n \to \infty$,

$$\sqrt{n}(\tilde{\psi} - \psi^*) \sim \text{MVN}\left(0, \mathcal{A}^{-1}(\psi^*)\mathcal{B}(\psi^*)[\mathcal{A}^{-1}(\psi^*)]'\right) \tag{3.10}$$

where $\mathcal{A}(\psi) = E\{-\partial U(\psi)/\partial\psi\}$ and $\mathcal{B}(\psi) = E\{U(\psi)U'(\psi)\}$ [1, 17, 14].

Alternatively, if we treat the failure time as left-truncated at the selected visit time which we denote generally by L_i (i.e., the time of the measurement) and only carry forward X_{i1}° from L_i to V_i, then the Cox partial likelihood score functions, the estimators $\tilde{\psi}$ and $\tilde{\Gamma}_0(s)$, as well as the limiting value ψ^* and its asymptotic property, follow the results given

in (3.2a) to (3.10), but with $\bar{Y}_i(s) = Y_i(s)Y_i^{\dagger}(s)Y_i^L(s)$ where $Y_i^L(s) = \mathbb{1}(s \geq L_i)$. If the last visit is selected and the biospecimen collected at that visit is used to measure the marker value, then $L_i = a_{iA_i(V_i)}$; if the first visit is selected, then $L_i = a_{i1}$.

3.2 Investigations for a Joint Markov Process

For what follows we focus on the scenario where the random censoring is independent and the visit process does not affect the nature of the transition intensities between the joint marker-failure process, so that

$$\lambda_2(s \mid \mathcal{H}(s)) = \lambda_2(s \mid H(s))$$

and

$$\lambda_k(s \mid \mathcal{H}(s)) = \lambda_k(s \mid H(s))$$

for $k = 0, 1$.

Under a Markov assumption [5] given the fixed covariates X_2, $\lambda_k(s \mid H(s))$ in (2.1a) simplifies to $\lambda_k(s \mid X_2)$ and $\lambda_2(s \mid H(s))$ in (2.1b) simplifies to $\lambda_2(s \mid X(s))$. Multiplicative intensity-based regression models enable investigation of covariate effects within the joint marker-failure process. For the marker process we set

$$\lambda_k(s \mid X_2; \theta_k) = \lambda_k(s; \alpha_k) \exp(X_2' \gamma_k) \tag{3.11a}$$

where $\lambda_k(s; \alpha_k)$ denotes the baseline intensity for a $1-k \to k$ transition indexed by α_k, γ_k is a $p \times 1$ vector of regression coefficients, and $\theta_k = (\alpha_k', \gamma_k')'$, $k = 0, 1$. For the failure process we set

$$\lambda_2(s \mid X(s); \theta_2) = \lambda_2(s; \alpha_2) \exp(X'(s)\beta)\,; \tag{3.11b}$$

here $\lambda_2(s; \alpha_2)$ is a baseline failure intensity corresponding to a $0 \to 2$ transition indexed by α_2 and $\beta = (\beta_1, \beta_2')'$ is a $(1+p) \times 1$ vector of regression coefficients with β_1 characterizing the association between the marker and the failure time; $\theta_2 = (\alpha_2', \beta')'$.

Censoring can be marker-dependent via (2.3) but for simplicity we consider a completely independent random censoring process with

$$\lambda_3(s \mid \mathcal{H}(s); \theta_3) = \lambda_3(s; \alpha_3)\,, \tag{3.12}$$

where $\theta_3 = \alpha_3$ with α_3 indexing the censoring intensity. For the visit process we consider a modulated Poisson process [4] given the covariates $X(s)$ with a proportional intensity function so that $\lambda_4(s \mid \mathcal{H}(s))$ in (2.4) simplifies to

$$\lambda_4(s \mid X(s); \theta_4) = \lambda_4(s; \alpha_4) \exp(X'(s)\eta)\,, \tag{3.13}$$

where $\lambda_4(s; \alpha_4)$ denotes the baseline visit intensity indexed by α_4 and $\eta = (\eta_1, \eta_2')'$ is a $(1+p) \times 1$ vector of regression coefficients. Under (3.13)

$$E\{dA(s) \mid X(s)\} = \lambda_4(s; \alpha_4) ds \, \exp(X'(s)\eta)\,,$$

and we note that (3.13) accommodates a marker-dependent visit process.

We discuss the implications of specifying (3.1) instead of the true model (3.11b) under the observation scheme described in Section 2.2. We consider the aforementioned four *ad hoc* strategies and investigate the percent asymptotic bias of $\hat{\psi}$ on the true parameter β defined here as $100(\psi^* - \beta)/\beta$.

In order to calculate the expectations required to evaluate the limiting value ψ^* through (3.6), we describe a further expansion to the joint process of Figure 2, where states are labeled as $\mathcal{Z}^\circ(s) := (\mathcal{Z}(s), X_1^\circ(s)) = (A(s), Z(s), X_1^\circ(s))$; the last entry is the marker value assessed from the biospecimen collected at the most recent visit before time s. The corresponding state-space diagram is provided in Appendix A. We next define classes of states which are convenient for future calculations. We let $\mathcal{C}_j^a = \{(a, z, x_1^\circ) : a = j\}$ denote a set of states occupied by individuals with the jth follow-up visit, $\mathcal{C}_k^z = \{(a, z, x_1^\circ) : z = k\}$ denote a set of states occupied by individuals who are in state k, $k = 0, 1, 2, 3$, for the marker-failure-censoring process, and $\mathcal{C}_l^{x_1^\circ} = \{(a, z, x_1^\circ) : x_1^\circ = l\}$ denote a set of states occupied by individuals whose most recently assessed marker value is l.

Based on this general multistate model $\mathcal{Z}^\circ(s)$, we calculate $r^{(k)}(s)$ and $r^{(k)}(s; \psi)$, $k = 0, 1$, under each of the four *ad hoc* strategies. Note $r^{(0)}(s)$ can be written as

$$r^{(0)}(s) = E_{Z_i(0), X_{i2}} \left\{ E \left\{ R_i^{(0)}(s) \mid Z_i(0), X_{i2} \right\} \right\},$$

and $r^{(1)}(s)$ can be partitioned into $r^{(1)}(s) = (r_1^{(1)}(s), [r_2^{(1)}(s)]')'$ where

$$r_1^{(1)}(s) = E_{Z_i(0), X_{i2}} \left\{ E \left\{ R_{i1}^{(1)}(s) \mid Z_i(0), X_{i2} \right\} \right\}$$

and

$$\begin{aligned} r_2^{(1)}(s) &= E_{Z_i(0), X_{i2}} \left\{ E \left\{ R_{i2}^{(1)}(s) \mid Z_i(0), X_{i2} \right\} \right\} \\ &= E_{Z_i(0), X_{i2}} \left\{ X_{i2}\, E \left\{ R_i^{(0)}(s) \mid Z_i(0), X_{i2} \right\} \right\}, \end{aligned}$$

with $R_i^{(0)}(s)$ defined in (3.8),

$$R_{i1}^{(1)}(s) = \mathbb{1}(A_i(\tau) \geq 1) \bar{Y}_i(s) X_{i1}^\circ dN_{i2}(s)\,,$$

and

$$R_{i2}^{(1)}(s) = \mathbb{1}(A_i(\tau) \geq 1) \bar{Y}_i(s) X_{i2}\, dN_{i2}(s)\,.$$

Similarly we can write $r^{(0)}(s; \psi)$ as

$$r^{(0)}(s; \psi) = E_{Z_i(0), X_{i2}} \left\{ E \left\{ R_i^{(0)}(s; \psi) \mid Z_i(0), X_{i2} \right\} \right\}$$

and partition $r^{(1)}(s; \psi) = (r_1^{(1)}(s; \psi), [r_2^{(1)}(s; \psi)]')'$ with

$$r_1^{(1)}(s; \psi) = E_{Z_i(0), X_{i2}} \left\{ E \left\{ R_{i1}^{(1)}(s; \psi) \mid Z_i(0), X_{i2} \right\} \right\}$$

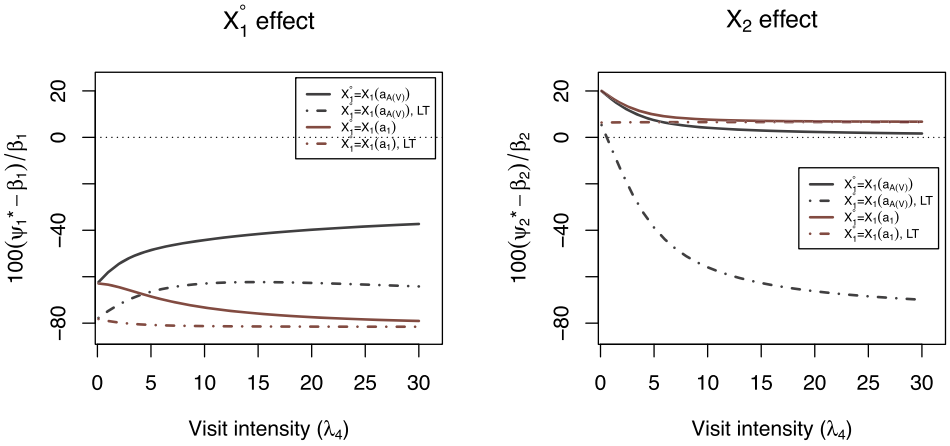


Figure 4: Percent asymptotic biases of $\hat{\psi}_1$ (left panel) and $\hat{\psi}_2$ as a function of the baseline visit intensity λ_4 with an independent visit process under four *ad hoc* strategies.

and

$$r_2^{(1)}(s;\psi) = E_{Z_i(0),X_{i2}}\left\{E\left\{R_{i2}^{(1)}(s;\psi) \mid Z_i(0), X_{i2}\right\}\right\}$$
$$= E_{Z_i(0),X_{i2}}\left\{X_{i2}\, E\left\{R_i^{(0)}(s;\psi) \mid Z_i(0), X_{i2}\right\}\right\},$$

where $R_i^{(0)}(s;\psi)$ is defined in (3.5),

$$R_{i1}^{(1)}(s;\psi) = \mathbb{1}(A_i(\tau) \geq 1)\bar{Y}_i(s)X_{i1}^{\circ}\exp(\psi' X_i^{\circ}),$$

and

$$R_{i2}^{(1)}(s;\psi) = \mathbb{1}(A_i(\tau) \geq 1)\bar{Y}_i(s)X_{i2}\exp(\psi' X_i^{\circ}).$$

Note the terms $E\{R_i^{(0)}(s) \mid Z_i(0), X_{i2}\}$, $E\{R_{i1}^{(1)}(s) \mid Z_i(0), X_{i2}\}$, $E\{R_i^{(0)}(s;\psi) \mid Z_i(0), X_{i2}\}$ and $E\{R_{i1}^{(1)}(s;\psi) \mid Z_i(0), X_{i2}\}$ may take different forms when different *ad hoc* strategies are used. Detailed calculations of these terms can be found in Appendix B.

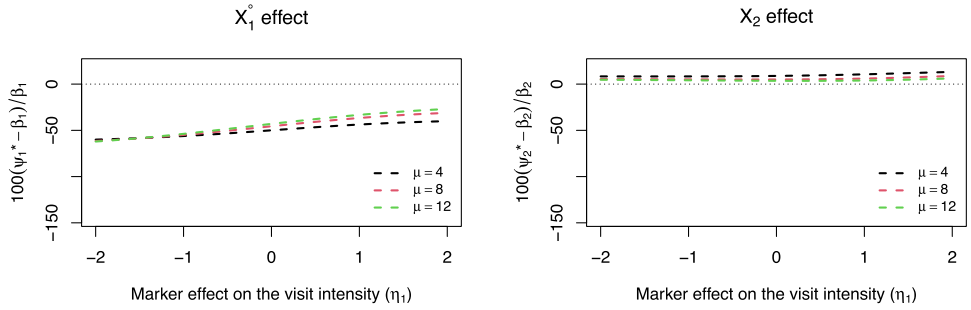
3.3 Limiting Values of ψ^*

Next, we investigate the limiting value ψ^* as a function of the transition intensity parameters in equations (3.11a) to (3.13) under the above *ad hoc* strategies. We set the initial state occupancy of the low marker value to be $P(Z(0) = 0) = \pi_0 = 0.5$ and set the administrative censoring time to be $\tau = 1$ without loss of generality. We consider a single binary fixed covariate X_2 with $P(X_2 = 1) = 0.5$. For the conditional Markov marker-failure process with proportional hazard transition intensities specified in (3.11a) and (3.11b), we set the baseline transition intensities to be time homogeneous so that $\lambda_2(s;\alpha_2) = \lambda_2$ and $\lambda_k(s;\alpha_k) = \lambda_k$ for $k = 0, 1$. For the failure intensities we set $\beta_1 = \log 1.5$ so that an elevated marker confers a 50% increase in the intensity to failure, and we set $\beta_2 = \log 1.25$ so that X_2 is also a risk factor. For the marker process we set $\gamma_0 = \log 0.9$ and $\gamma_1 = \log 1.1$ so $X_2 = 1$ elevates the $0 \to 1$ transition intensity but reduces the $1 \to 0$ intensity. We let $\pi_1 = \lambda_1/(\lambda_1+\lambda_2)$ denote the odds of a $0 \to 1$ transition when the process enters state 0 with $X_2 = 0$, and set $\pi_1 = 0.75$. We then set $r_{01} = (\lambda_0 + \lambda_2\exp(\beta_1))/(\lambda_1 + \lambda_2) = 2$ so that the mean sojourn time in state 0 is twice the mean sojourn time in state 1 when $X_2 = 0$. With all other parameters specified as described above, we control the marginal failure rate at the administrative censoring time for the marker-failure process by letting $\pi_2 = P(Z(\tau) = 2) = E_{Z(0),X_2}\{P(Z(\tau) = 2 \mid Z(0), X_2\}$ and solving for λ_2 to satisfy $\pi_2 = 0.8$.

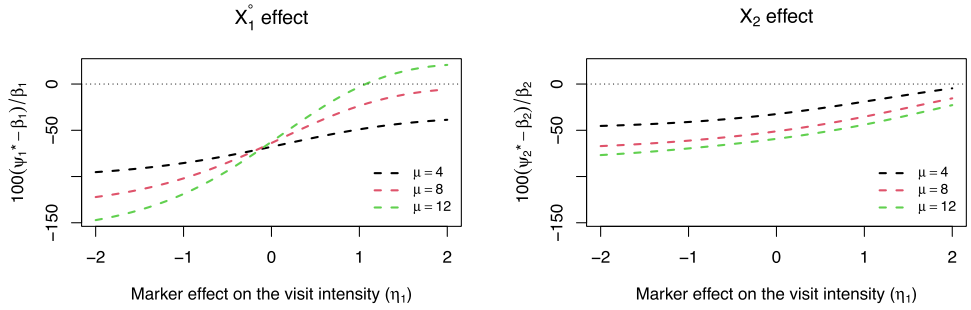
Having specified the parameters for the three-state marker-failure process in Figure 1, we now consider the observation processes. We consider an independent random censoring with an intensity specified in (3.12) with a time-homogeneous model so that $\lambda_3(s;\alpha_3) = \lambda_3$. Let $\pi_3 = P(Z(\tau) = 3) = E_{Z(0),X_2}\{P(Z(\tau) = 3 \mid Z(0), X_2\}$ denote the marginal random censoring rate by the administrative censoring time, and we solve $\pi_3 = 0.2$ for λ_3 to ensure a 20% random censoring rate. Lastly we consider a conditional Poisson visit process with a proportional intensity for visits specified in (3.13). We let the visit process be time-homogeneous so that $\lambda_4(s;\alpha_4) = \lambda_4$. We consider both a completely independent visit process ($\eta_1 = \eta_2 = 0$ in (3.13)) and a marker-dependent visit process ($\eta_1 \neq 0, \eta_2 = 0$). For the former case we investigate the limiting value ψ^* as a function of λ_4. For the latter case, we investigate ψ^* as a function of the marker effect on the visits η_1; we let

$$\mu = E\{A(\tau)\} = E_{Z(0),X_2}\left\{\sum_{j=0}^{\infty} j \cdot P(\mathcal{Z}^{\circ}(\tau) \in \mathcal{C}_j^a \mid Z(0), X_2\right\}$$

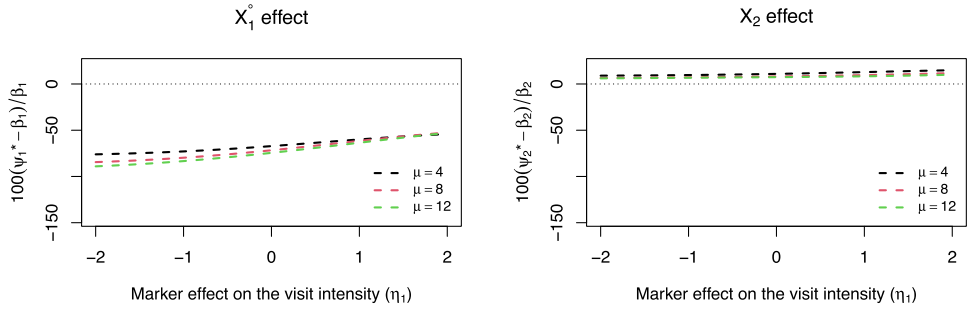
and for each specified value of η_1, we solve for λ_4 such that $\mu = 4, 8, 12$. The percent asymptotic bias of $\hat{\psi}$ obtained from the four *ad hoc* strategies with either an independent or marker-dependent visit process is plotted in Figures 4 and 5 respectively.



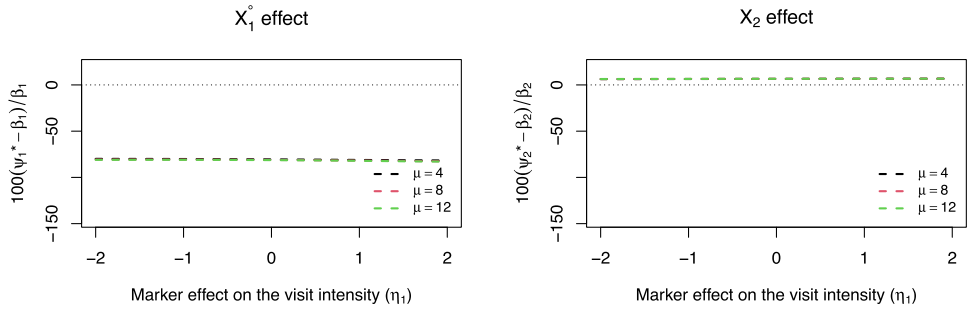
(a) $X_1^\circ = X_1(a_{A(V)})$, No truncation



(b) $X_1^\circ = X_1(a_{A(V)})$, Left truncation



(c) $X_1^\circ = X_1(a_1)$, No truncation



(d) $X_1^\circ = X_1(a_1)$, Left truncation

Figure 5: Percent asymptotic biases of $\hat{\psi}_1$ (left panel) and $\hat{\psi}_2$ as a function of the marker effect on the visit intensity (η_1) for $\mu = E\{A(\tau)\} = 4, 8, 12$ with a marker-dependent visit process under four *ad hoc* strategies.

Figure 4 contains plots of the percent asymptotic bias of estimators of the effects of X_1° (left column) and X_2 (right column) on the failure time process, defined as $100(\psi_1^* - \beta_1)/\beta_1$ and $100(\psi_2^* - \beta_2)/\beta_2$ respectively. Black lines correspond to strategies that assay the last biospecimen, while red lines correspond to those that assay the first biospecimen. Dot-dashed lines represent strategies that incorporate left truncation. The results demonstrate that even with an independent visit process, selecting only a single visit to measure the marker value and carrying it forward or backward as well renders the estimator $\hat{\psi}$ inconsistent for β. All four strategies severely underestimate the marker effect. For the effect of fixed covariate X_2, all strategies overestimate the effect except when X_1° is measured from the biospecimen collected at the first visit and treated as left-truncated. Specifically, when X_1° is measured from the biospecimen collected at the last visit and carried both backward and forward, there is a trend of decreasing bias of both $\hat{\psi}_1$ and $\hat{\psi}_2$ as the visit intensity λ_4 increases. When X_1° is measured from the last biospecimen but carried only forward, we still observe a decreasing bias of $\hat{\psi}_1$ as λ_4 increases, but the bias of $\hat{\psi}_2$ goes to the other direction and significantly increases when λ_4 increases. For the strategy where X_1° is measured from the biospecimen collected at the first visit and carried both backward and forward, the asymptotic bias of $\hat{\psi}_1$ tends to increase while the bias of $\hat{\psi}_2$ tends to decrease with an increasing λ_4. There is no trend of change of the biases of $\hat{\psi}_1$ and $\hat{\psi}_2$ when X_1° is carried only forward from the first visit.

Figure 5 presents plots of the percent asymptotic bias of $\hat{\psi}_1$ and $\hat{\psi}_2$ when the visit process is marker-dependent, with each row corresponding to one *ad hoc* strategy. The figure shows analogous results that in general all strategies provide a conservative estimate of the marker effect and a progressive estimate for the fixed covariate effect, except when X_1° is carried only forward from the last visit. A negative value of η_1 indicates a negative association between the marker and the visit intensity. As η_1 increases from -2 to 2, the association between the marker and the visit process goes from negative to positive, and the strength of the association first decreases (to 0) and then increases. When X_1° is carried only forward from the first visit, no trend is observed as η_1 increases. In contrast, a trend of increasing ψ_1^* and ψ_2^* is apparent for all other strategies, with the rate of increase being higher with greater visit intensities.

4. DESIGN FOR COST-EFFECTIVE SELECTION

Having established that any of the *ad hoc* approaches that we considered yield inconsistent parameter estimators with appreciable asymptotic bias, we now turn to the joint modeling framework described in Section 2 to mitigate this bias. For the remainder of this paper, we assume the marker-failure process follows the conditional Markov model and is conditionally independent of the observation process, and that the observation process is non-informative. The bias induced by *ad hoc* strategies can, in principle, be mitigated by fitting the full joint process $\mathcal{Z}(s)$ represented by the state space diagram in Figure 2. Using intermittently observed data $\bar{\mathcal{H}}(V_i^+)$ defined in (2.5), together with the transition intensities governing this process, the observed data likelihood can be formulated assuming biospecimens collected at all visits are assayed. The corresponding observed data likelihood contribution for the marker-failure process is then given by

$$\prod_{i=1}^{n}\left\{\prod_{j=2}^{A_i(V_i)} P(Z_i(a_{ij}) \mid Z_i(a_{i,j-1}), X_{i2}) \times \sum_{k=0}^{1}\left\{P\left(Z_i(V_i^-)=k \mid Z_i(a_{iA_i(V_i)}), X_{i2}\right) \times \lambda_2(V_i \mid X_{i1}(V_i^-)=k, X_{i2})^{\delta_i}\right\}\right\}.$$

However, such complete assessment may be prohibitively expensive in many studies. We therefore seek cost-effective sampling strategies that retain the key advantages of this joint modeling framework while remaining feasible under realistic assay constraints. Specifically, we propose a subsampling scheme in which some individuals are selected for complete longitudinal assessment of their biospecimens over visits, and some are selected for a single assessment of the baseline (first-visit) biospecimen value. The rationale for this strategy is that individuals selected for complete longitudinal assessment of their samples enable estimation of the marker transition intensities, which in turn facilitates suitable use of data from those for whom only a single biospecimen is assayed. There are, however, individuals who experience events that are not sub-sampled in this scheme so there is the potential for some loss of information.

Suppose we have n individual processes, each with $m_i \geq 1$ clinical visits by the administrative censoring time for individual i, $i = 1, \ldots, n$. For all individuals, biospecimens are sampled and stored at each clinic visit, yielding a total of $m_\cdot = \sum_{i=1}^n m_i$ biospecimen samples across all n individuals. Suppose that instead of assaying all $m_\cdot$ biospecimens, budgetary constraints permit only an average of μ_b assays ($\mu_b < m_\cdot$) across these individuals. In Section 3, we investigated strategies allocating only one assay per individual. We let n_L represent the number of individuals for whom all biospecimens are assayed and let n_S be the number of individuals from the remainder of the cohort for whom the first available biospecimen will be assayed. Let R_{iL} indicate individual i is selected to be followed longitudinally for complete assays of biospecimens over time, R_{iS} indicate the individual is selected for a single assay of the biospecimen collected at the first visit, $i = 1, \ldots, n$, then $n_L = \sum_{i=1}^n R_{iL}$ and $n_S = \sum_{i=1}^n R_{iS}$. Note under this selection strategy, an

individual is either selected to be in the longitudinal subsample ($R_{iL} = 1, R_{iS} = 0$), selected to be in the single-measurement subsample ($R_{iL} = 0, R_{iS} = 1$), or not selected at all ($R_{iL} = 0, R_{iS} = 0$). Let

$$m_L = \sum_{i=1}^{n} R_{iL} m_i$$

denote the total number of assays conducted among those in the longitudinal subsample, we have the total number of assays $m_b = m_L + n_S$ with $E(m_b) = \mu_b$ to satisfy the budget constraint. A question then naturally arises: how should we allocate m_b assays between the longitudinal and single-measurement subsamples? In other words, what fraction of m_b should be allocated to m_L?

We consider independent random sampling, where the probability of being selected into the longitudinal subsample is $P(R_{iL} = 1, R_{iS} = 0) = \pi_L$, the probability of being selected into the single-measurement sample is $P(R_{iL} = 0, R_{iS} = 1) = \pi_S$, and the probability of not being selected into either sample is $P(R_{iL} = 0, R_{iS} = 0) = 1 - \pi_L - \pi_S$, with $\pi_L + \pi_S < 1$. Selection into both subsamples is not allowed, so $P(R_{iL} = 1, R_{iS} = 1) = 0$. Moreover, we have $P(R_{iL} = 1) = \pi_L$ and $P(R_{iS} = 1) = \pi_S$. Under this framework and sampling scheme, the observed data likelihood contribution from one individual for the marker-failure process can be written as

$$L_i \propto \{\pi_L L_{iL}\}^{R_{iL}} \{\pi_S L_{iS}\}^{R_{iS}}$$

where L_{iL}, L_{iS} denote the contributions from the longitudinal and single-measurement subsamples respectively, with explicit forms

$$\begin{aligned} L_{iL} \propto & \prod_{j=2}^{m_i} P\left(Z_i(a_{ij}) \mid Z_i(a_{i,j-1}), X_{i2}\right) \\ & \times \sum_{k=0}^{1} \Big\{ P\left(Z_i(V_i^-) = k \mid Z_i(a_{i,m_i}), X_{i2}\right) \\ & \qquad \times \lambda_2(V_i \mid X_{i1}(V_i^-) = k, X_{i2})^{\delta_i} \Big\} \end{aligned}$$

and

$$\begin{aligned} L_{iS} \propto & \sum_{k=0}^{1} \Big\{ P\left(Z_i(V_i^-) = k \mid Z_i(a_{i1}), X_{i2}\right) \\ & \qquad \times \lambda_2(V_i \mid X_{i1}(V_i^-) = k, X_{i2})^{\delta_i} \Big\} \end{aligned}$$

where $V_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{1}(V_i = T_i)$. With a sample of n independent individual processes, the overall observed data likelihood is then $L(\theta) = \prod_{i=1}^{n} L_i(\theta)$ where $\theta = (\theta_0', \theta_1', \theta_2')'$ denotes parameters associated with the marker-failure process defined in (3.11a) and (3.11b). A maximum likelihood estimator $\hat{\theta}$ can then be obtained by maximizing the observed data likelihood. The Fisher information $\mathcal{I}(\theta) = E\{-\partial \log L_i(\theta)/\partial\theta\partial\theta'\}$ can be written as a weighted sum of the Fisher information contributed from the longitudinal and single-measurement subsamples

$$\mathcal{I}(\theta) = \pi_L \mathcal{I}_L(\theta) + \pi_S \mathcal{I}_S(\theta)\,,$$

where

$$\mathcal{I}_L(\theta) = E\{-\partial \log L_{iL}(\theta)/\partial\theta\partial\theta'\}$$

and

$$\mathcal{I}_S(\theta) = E\{-\partial \log L_{iS}(\theta)/\partial\theta\partial\theta'\}\,.$$

The asymptotic covariance matrix of $\hat{\theta}$ is given by $\mathcal{I}^{-1}(\theta)$. To identify an optimal sampling design under the assay budget constraint, we study the asymptotic variance of the regression coefficient estimators. Since $\mathcal{I}(\theta)$ depends on the sampling probabilities π_L and π_S, the optimal design is determined by the specification of π_L and π_S under the budget constraint.

To characterize this constraint, we next relate the sampling probabilities π_L and π_S to the assay budget. Taking the expectation of $m_b = m_L + n_S$ on both sides gives

$$\mu_b = \mu_{bL} + \mu_{bS} \tag{4.1}$$

where

$$\mu_{bL} = E(m_L) = n\pi_L\bar{\mu}$$

and

$$\mu_{bS} = E(n_S) = n\pi_S\,,$$

with $\bar{\mu} = E(m_i) = E\{A_i(\tau) \mid A_i(\tau) \geq 1\}$ denoting the expected number of visits (and thus biospecimens) per individual by the administrative censoring time conditioning on having at least one visit. We can thus re-write (4.1) as

$$\mu_b/n = \pi_L\bar{\mu} + \pi_S\,. \tag{4.2}$$

With pre-specified values of μ_b/n and $\bar{\mu}$, the budget constraint in (4.2) implies that π_S is uniquely determined by π_L. Consequently, the Fisher information $\mathcal{I}(\theta)$ and corresponding asymptotic variance of the regression coefficient estimators can both be expressed as functions of π_L alone. This allows us to identify the optimal value of π_L that minimizes the asymptotic variance under the specified assay budget.

For practical interpretation, we re-parameterize this allocation strategy in terms of

$$\phi_L := E(m_L)/\mu_b = n\pi_L\bar{\mu}/\mu_b\,, \tag{4.3}$$

which represents the fraction of total assays allocated to the longitudinal subsample. We next investigate the asymptotic variance of $\hat{\beta}$ as a function of ϕ_L. We follow the model specifications and parameter settings of Section 3.3 and let the failure rate be $\pi_2 = 0.1, 0.25, 0.5$, or 0.8. We set $\mu_b/n = 0.5$

or 1 and $\bar{\mu} = 4$ or 8. For each value of ϕ_L, we compute the asymptotic relative efficiency of $\hat{\beta}$, defined as

$$\text{ARE}(\hat{\beta}) = \frac{\text{asvar}(\hat{\beta})}{\min_{\phi_L} \text{asvar}(\hat{\beta})},$$

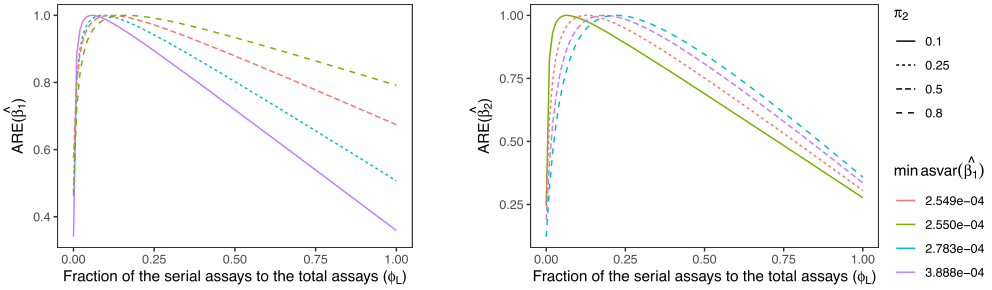
and plot $\text{ARE}(\hat{\beta})$ against ϕ_L in Figure 6. The optimal values of ϕ_L vary between 5% to 25% and the corresponding asymptotic variances of $\hat{\beta}$ are provided in the legends of the plots. Across all parameter settings, there is a common pattern: the asymptotic relative efficiency of β increases rapidly and then gradually decreases when the fraction of m_L increases from 0 to 1. This is because when ϕ_L is very small, we have more individuals with fewer repeated measurements of the marker value so that we lose information about the marker transitions; when ϕ_L is larger, on the other hand, we have fewer individuals with more repeated marker measurements, so we lose information on the failure time. Comparing various levels of failure rate π_2, the optimal fraction ϕ_L is larger with higher π_2; additionally, the rate of change in the relative efficiency is also larger. The asymptotic variances are overall smaller when we have more budget on the number of assays ($\mu_b/n = 0.5$ vs $\mu_b/n = 1$). Although these numerical results are based on specific simulation settings, the proposed design framework is general: for any given study setting and assay budget, investigators can vary ϕ_L over its feasible range, evaluate the corresponding asymptotic variance, and identify the value that minimizes asymptotic variance to determine the corresponding sub-sampling design.

5. ARTHRITIS MUTILANS IN PSA

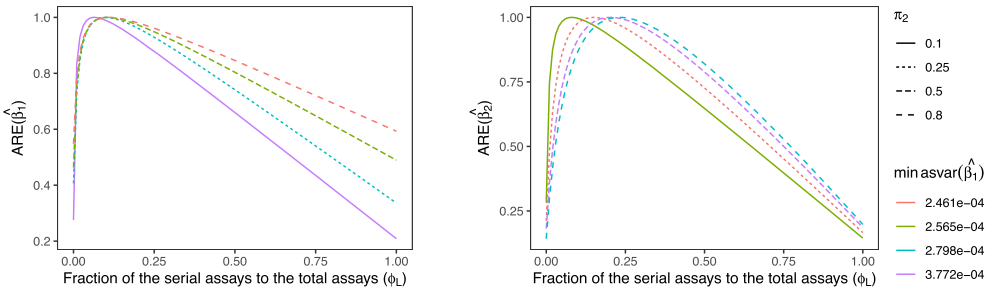
An illustrative example of the proposed selection strategy is given here by applying the method to data from the University of Toronto Psoriatic Arthritis Cohort Study. Psoriatic arthritis (PsA) is a complex rheumatic disease defined as a chronic inflammatory arthritis associated with psoriasis. The University of Toronto PsA Clinic at the Centre for Prognosis Studies in The Rheumatic Diseases at Toronto Western Hospital maintains a comprehensive database of patients diagnosed with PsA [10]. Individuals with a prior diagnosis of PsA are recruited to participate in the cohort and are followed prospectively over time to monitor the severity of disease progression. C-reactive protein (CRP) is a substance produced by the liver in response to inflammation, and a CRP blood test measures its level to detect inflammation from acute conditions or to monitor the severity of disease in chronic conditions. Erythrocyte sedimentation rate (ESR) is another blood test used to detect inflammation in the body. While similar to CRP in its purpose, ESR measures different aspects and can sometimes be used together with CRP for a more comprehensive view of inflammation. Both CRP and ESR measurements can only be taken during periodic clinical visits through blood tests. The primary interest here lies in studying the effect of the time-dependent covariate denoted ESR_CRP (which indicates if either ESR or CRP is elevated) on the time to the development of five or more damaged joints in patients with PsA. This represents a sufficient degree of damage that it satisfies the condition for a severe form of arthritis called arthritis mutilans [11].

We focus on a total of $n = 1123$ patients with a prior diagnosis of PsA who had made at least one follow-up visit after recruitment into the study. For each individual, the reported age at PsA diagnosis is taken as the time origin. At each subsequent clinical visit, data including the ESR and CRP results from the blood tests, the date of the blood test, and the number of cumulative damaged joints were recorded. The age of onset of PsA varies from 4 to 49 years (with an average of 39.07 and S.D. of 13.57) across patients in this cohort, and the number of visits made per individual during their follow-up period ranges from 1 to 62 (with an average of $\bar{\mu} = 9.98$ and S.D. of 12.42). Additionally, 274 (24.4 %) patients were observed to develop five or more damaged joints.

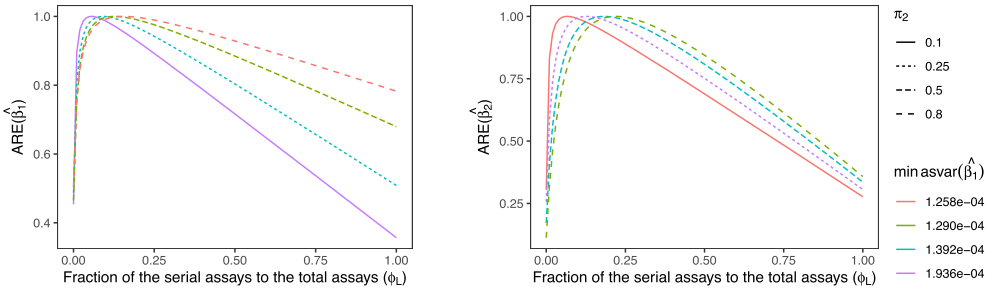
Instead of using all $m_\cdot = 11204$ measurements of the ESR_CRP marker value, we imagine that the measurements were not available and consider a budgetary constraint. We subsample the visit times at which samples are available by our proposed cost-effective sampling strategy for a joint analysis of the ESR_CRP marker and the development of arthritis mutilans. We consider a fixed covariate of HLA-B27 and consider budgets μ_b to be $0.5n$, n or $2n$. With a pre-specified value of ϕ_L (i.e., the fraction of the marker measurements in the longitudinal subsample to the total budget), π_L and π_S can be determined via (4.2) and (4.3). We then select a subsample of marker measurements accordingly for the joint analysis. Our goal is to empirically investigate the variance of the coefficients of both ESR_CRP and HLA-B27 covariates under different specifications of ϕ_L. When $\mu_b = 0.5n$ or n, we choose $\phi_L = 0.2, 0.4, 0.6$, or 0.8; when $\mu_b = 2n$, we choose $\phi_L = 0.6, 0.7, 0.8$, or 0.9, since we have to assign at least half of the budget to the longitudinal subsample (i.e., $\phi_L > 0.5$) to ensure $\pi_S < 1$. Time-homogeneous transition intensities are fitted for the joint multistate model, and the results of the estimates of the regression coefficients under various specifications of μ_b and ϕ_L are summarized in Table 1. For $\mu_b = 0.5n$ or n, a minimum variance of the effect of the time-dependent marker ESR_CRP on failure is achieved when we assign 40% of the measurement budget to the longitudinal subsample, while a minimum variance of the effect of the fixed covariate HLA-B27 is achieved at 60%. For $\mu_b = 2n$, the minimum variance of both the marker effect and the fixed covariate effect is achieved when $\phi_L = 60\%$. Additionally, the estimated variance of both effects decreases as we have more measurement budget (data) for the time-dependent ESR_CRP value. The point estimates of the ESR_CRP effect are comparable when $\phi_L < 0.6$ but not when $\phi_L > 0.6$. This may be due to the small censoring rate of the original data, meaning that selecting fewer individuals (and thus having more



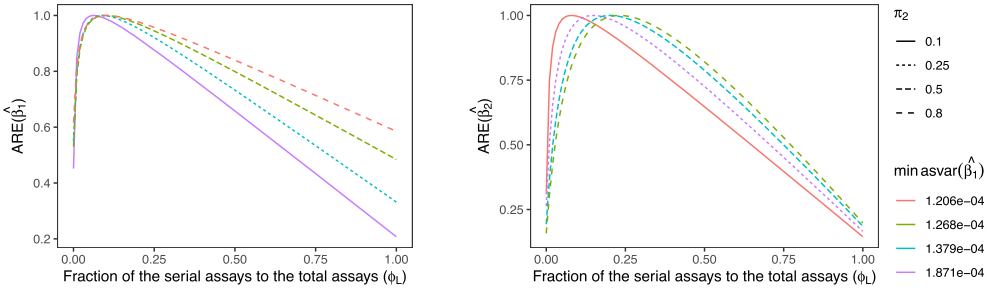
(a) Expected number of assays to the sample size $\mu_b/n = 0.5$; Expected number of biospecimens per individual $\bar{\mu} = 4$



(b) Expected number of assays to the sample size $\mu_b/n = 0.5$; Expected number of biospecimens per individual $\bar{\mu} = 8$



(c) Expected number of assays to the sample size $\mu_b/n = 1$; Expected number of biospecimens per individual $\bar{\mu} = 4$



(d) Expected number of assays to the sample size $\mu_b/n = 1$; Expected number of biospecimens per individual $\bar{\mu} = 8$

Figure 6: Asymptotic relative efficiency of $\hat{\beta}_1$ (left panel) and $\hat{\beta}_2$ as a function of ϕ_L for $\pi_2 = 0.1, 0.25, 0.5, 0.8$, $\mu_b/n = 0.5, 1$ and $\bar{\mu} = 4, 8$; the asymptotic variances are obtained from the expected information matrix $\mathcal{I}(\theta)$ with the expectation carried out by Monte-Carlo integration with 500,000 replications.

Table 1. Joint analyses of the ESR_CRP and failure processes with a fixed covariate of HLA-B27 using subsamples selected under different specifications of μ_b and ϕ_L; time-homogeneous baseline transition intensities are fitted; variance estimates are obtained through observed information matrix averaged over 500 repetitions.

		Covariate			
		ESR_CRP (β_1)		HLA-B27 (β_2)	
μ_b/n	ϕ_L	EST	VAR	EST	VAR
0.5	0.2	1.639	2.848	0.175	0.590
	0.4	1.464	2.449	0.293	0.588
	0.6	1.361	2.608	0.225	0.532
	0.8	0.941	2.634	0.291	0.574
1	0.2	1.660	1.343	0.231	0.417
	0.4	1.522	1.288	0.270	0.339
	0.6	1.355	1.305	0.285	0.312
	0.8	0.917	1.324	0.258	0.351
2	0.6	1.229	0.705	0.283	0.200
	0.7	1.134	0.732	0.270	0.211
	0.8	0.974	0.974	0.268	0.231
	0.9	0.798	0.791	0.263	0.273

longitudinal marker measurements) results in less data on observed failures.

6. DISCUSSION

The bias in Cox regression coefficients arising from the carry-forward approach has recently been studied formally [7, 12]. To address this issue, joint modeling of continuous marker processes and failure times is needed, with most models for the marker trajectory having an additive form with subject-specific random effects accommodating heterogeneity in the marker paths and a serial dependence. The association between the marker process and the failure process is accommodated by incorporating shared or correlated random effects to the failure time model [15, 20]. Conditioning on the random effects, the failure time process and the marker processes are typically assumed to be independent, but a joint model can be obtained by integrating over the random effects. Such joint models often require a strong assumption about the distributions of the marker processes [3, 21] and are less appealing when the failure event is death, as the marker process does not terminate upon death [19]. Other strategies for dealing with intermittent observation of marker processes do not involve joint modeling. de Bruijne et al. [9] proposed a way of attenuating the bias using a weighted Cox regression method where weights decay as a function of time from the most recent marker measurement; this reduces the bias by down-weighting contributions to the partial score function when it is a long time since the marker measurement. Other methods involving using smoothing techniques to impute the missing "current" marker values at failure times have been explored [16, 18, 2].

In medical research, it is common to discretize continuous biomarkers by specifying cut-points agreed upon by researchers; in some settings these may be thresholds delineating meaningfully different disease states. Such discretization can be appealing since it enables a joint multistate model of the marker and failure processes by specification of a discrete state spacing incorporating different levels of marker states and an absorbing state for the terminal effect of death [5]. Moreover, in settings where marker values are recorded at clinic visits or when an individual seeks medical care, the visit times are often associated with the values of the marker process. It is then crucial to account for such association between the clinical visit process and the disease process for valid inference of the marker effect as the observation process conveys information about the disease process [5, 6].

We studied the association between a time-dependent biomarker and a failure time of interest in settings where biomarker measurements are only available at intermittent clinical visits. We introduced a multistate model for the joint marker-failure process by discretizing the continuous biomarker into two levels. This model formulation allows for natural extensions to accommodate dependent observation processes [7], and can be generalized to accommodate more marker states and higher-dimensional discrete markers. Markov transition intensities with proportional hazard models were adopted to govern the disease dynamics. With marker values available at each clinical visit, a joint analysis can be conducted for the marker-failure process. We also considered budgetary constraints on assessing marker values at clinical visits and investigated the asymptotic bias of four commonly used but naive *ad hoc* strategies for statistical inference arising from misspecified Cox models. Our findings indicated that there is generally an attenuation of the estimated regression coefficients, particularly for the marker effect, even when the visit process is completely independent of the disease process. To address this, an alternative cost-effective strategy has been proposed to mitigate bias under budgetary constraints which involves efficient selections of individuals. The asymptotic relative efficiency of estimators obtained from this proposed strategy under different designs has been explored and an optimal one can thus be determined.

As we only have focused on simple random sampling for individual selection, a natural alternative to the selection strategy is to consider outcome-dependent sampling for individual selection to the two subsamples. Specifically, we might intentionally select all individuals with an observed failure and assess their marker values using all biospecimen samples. Besides selecting individuals with either a single assay or complete assays of biospecimens, we could also consider selecting several specific visits at which the collected biospecimen is assayed for each individual, and the visits selected could vary across individuals. In general, if we let $R_{ij} = 1, j = 1, \ldots, m_i$, indicate the biospecimen collected at the jth visit for individual i is assayed for a marker value,

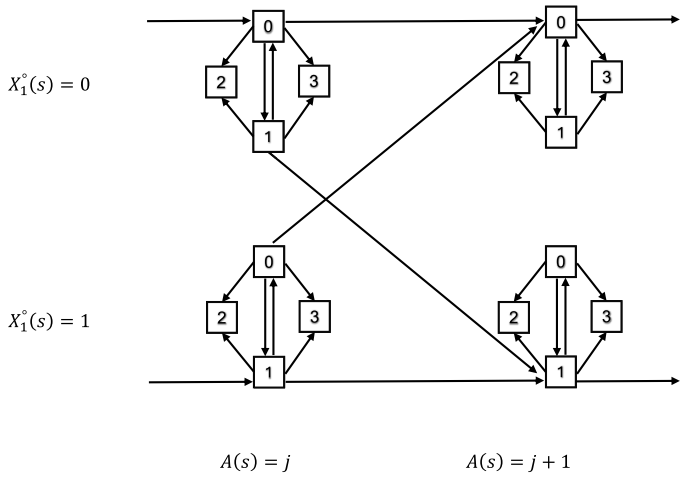


Figure A.1: A state space diagram for a multistate model of $\mathcal{Z}^\circ(s) = (\mathcal{Z}(s), X_1^\circ(s)) = (A(s), Z(s), X_1^\circ(s))$ depicting a more general joint model for the marker-failure process and the observation processes.

the budgetary constraint becomes $m_b = \sum_{i=1}^{n}\sum_{j=1}^{m_i} R_{ij}$. Various models can be specified for the joint distribution of the selection indicators R_{ij}, allowing for the formulation of a joint likelihood of the selection indicators and the marker-failure process. Fisher information can then be derived for determining an optimal design of the parameters related to the selection distributions.

The robustness of the multistate model for the joint marker-failure process can be improved by adopting piecewise constant baseline transition intensities. Random effects could be incorporated into the transition intensities to account for unexplained heterogeneity at the individual level. More complex multistate disease process can be considered when there is more than one clinical event of interest. Additionally, the multistate model can be extended to accommodate a higher-dimensional marker process via the use of more cut-points for the marker value or through the consideration of multiple time-dependent markers. Mixture multistate models might also be of interest when there is a portion of the population not susceptible to the event of interest. These models can separate the susceptible and non-susceptible subpopulations, providing more accurate estimates of the effects of covariates on the failure risk. This approach could be particularly useful in settings where heterogeneity in the population's response to covariates or failure susceptibility is expected.

APPENDIX A. STATE SPACE DIAGRAM FOR $\mathcal{Z}^\circ(s) = (A(s), Z(s), X_1^\circ(s))$

The states in Figure A.1 are arranged so that individuals with $X_1^\circ(s) = X_1(a_{A(s)}) = 0$ occupy the upper set of states, whereas those with $X_1^\circ(s) = 1$ occupy the lower set. Transitions from the left to the right panels indicate that a visit has occurred. Transitions between the upper and lower sets of states occur only when a new visit takes place and the marker value assessed at that visit differs from the value assessed at the previous visit. This figure thus contains information on the latent marker process, the observable failure and observation processes, and the marker values assessed at each visit.

APPENDIX B. CONDITIONAL EXPECTATIONS

Here we present detailed calculations of the expectations required for the four naive *ad hoc* strategies described in Section 3.2.

B.1 Assaying the Last Sample

When the marker value is assessed using the biospecimen sample collected at the last visit and treated without left truncation, we have $X_{i1}^\circ = X_{i1}^\circ(V_i) = X_{i1}(a_{iA_i(V_i)})$ and $\bar{Y}_i(s) = Y_i(s)Y_i^\dagger(s)$. $E\{R_i^{(0)}(s) \mid Z_i(0), X_{i2}\}$ can be calculated as

$$
\begin{aligned}
&E\left\{\mathbb{1}(A_i(\tau) \geq 1)\bar{Y}_i(s)dN_{i2}(s) \mid Z_i(0), X_{i2}\right\} \\
&= \sum_{k=0}^{1} P\left(A_i(\tau) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1 \mid Z_i(0), X_{i2}\right) \\
&= \sum_{k=0}^{1} P\left(A_i(s^-) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1 \mid Z_i(0), X_{i2}\right) \\
&= \sum_{k=0}^{1} P\left(A_i(s^-) \geq 1, Z_i(s^-) = k \mid Z_i(0), X_{i2}\right) \times \\
&\qquad P\left(dN_{i2}(s) = 1 \mid Z_i(s^-) = k, X_{i2}\right) \\
&= \lambda_2(s)ds \sum_{k=0}^{1}\sum_{j=1}^{\infty} P\left(\mathcal{Z}_i^\circ(s^-) \in \mathcal{C}_{jk}^{az} \mid Z_i(0), X_{i2}\right) \\
&\qquad \exp(k\beta_1 + X_{i2}'\beta_2)
\end{aligned}
\tag{B.1}
$$

where $\mathcal{C}_{jk}^{az} = \mathcal{C}_j^a \cap \mathcal{C}_k^z$. Recall $\mathcal{C}_j^a = \{(a, z, x_1^\circ) : a = j\}$ denotes a set of states occupied by individuals with the jth follow-up visit, $\mathcal{C}_k^z = \{(a, z, x_1^\circ) : z = k\}$ denotes a set of states occupied by individuals who are in state k for the four-state process depicted in Figure 1, and $\mathcal{C}_l^{x_1^\circ} = \{(a, z, x_1^\circ) : x_1^\circ = l\}$ denotes a set of states occupied by individuals whose most recently recorded marker value is l. $E\{R_{i1}^{(1)}(s) \mid Z_i(0), X_{i2}\}$ can be calculated as

$$
\begin{aligned}
&E\left\{\mathbb{1}(A_i(\tau) \geq 1)\bar{Y}_i(s)X_{i1}^\circ dN_{i2}(s) \mid Z_i(0), X_{i2}\right\} \\
&= \sum_{k=0}^{1} P\left(A_i(\tau) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1,\right. \\
&\qquad \left. X_{i1}^\circ(V_i) = 1 \mid Z_i(0), X_{i2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=0}^{1} P\left(A_i(s^-) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1,\right. \\
&\qquad \left. X^{\circ}_{i1}(s^-) = 1 \mid Z_i(0), X_{i2}\right) \\
&= \sum_{k=0}^{1} \left\{ P\left(A_i(s^-) \geq 1, Z_i(s^-) = k,\right.\right. \\
&\qquad \left. X^{\circ}_{i1}(s^-) = 1 \mid Z_i(0), X_{i2}\right) \times \\
&\qquad \left. P\left(dN_{i2}(s) = 1 \mid Z_i(s^-) = k, X_{i2}\right) \right\} \\
&= \lambda_2(s) ds \sum_{k=0}^{1} \sum_{j=1}^{\infty} P\left(\mathcal{Z}^{\circ}_i(s^-) \in \mathcal{C}^{azx^{\circ}_1}_{jk1} \mid Z_i(0), X_{i2}\right) \\
&\qquad \exp(k\beta_1 + X'_{i2}\beta_2) \qquad \text{(B.2)}
\end{aligned}
$$

with $\mathcal{C}^{azx^{\circ}_1}_{jkl} = \mathcal{C}^{az}_{jk} \cap \mathcal{C}^{x^{\circ}_1}_{l}$. Similarly, we can calculate $E\{R^{(0)}_i(s;\psi) \mid Z_i(0), X_{i2}\}$ as

$$
\begin{aligned}
&E\left\{\mathbb{1}(A_i(\tau) \geq 1)\bar{Y}_i(s)\exp(\psi' X^{\circ}_i) \mid Z_i(0), X_{i2}\right\} \\
&= \sum_{l=0}^{1}\sum_{k=0}^{1} P\left(A_i(\tau) \geq 1, Z_i(s^-) = k, X^{\circ}_{i1}(\tau) = l \mid Z_i(0), X_{i2}\right) \\
&\qquad \exp(\psi_1 l + \psi'_2 X_{i2}) \\
&= \sum_{l=0}^{1}\sum_{k=0}^{1} \left\{ P\left(Z_i(s^-) = k \mid Z_i(0), X_{i2}\right) \times \right. \\
&\qquad P\left(A_i(\tau) \geq 1, X^{\circ}_{i1}(\tau) = l \mid Z_i(s^-) = k, X_{i2}\right) \\
&\qquad \left. \exp(\psi_1 l + \psi'_2 X_{i2}) \right\} \\
&= \sum_{l=0}^{1}\sum_{k=0}^{1} P\left(\mathcal{Z}^{\circ}_i(s^-) \in \mathcal{C}^{z}_{k} \mid Z_i(0), X_{i2}\right) \times \\
&\quad \sum_{j=1}^{\infty} P\left(\mathcal{Z}^{\circ}_i(\tau) \in \mathcal{C}^{ax^{\circ}_1}_{jl} \mid \mathcal{Z}^{\circ}_i(s^-) \in \mathcal{C}^{z}_{k}, X_{i2}\right) \\
&\qquad \exp(\psi_1 l + \psi'_2 X_{i2}),
\end{aligned}
$$

and $E\{R^{(1)}_{i1}(s;\psi) \mid Z_i(0), X_{i2}\}$ as

$$
\begin{aligned}
&E\left\{\mathbb{1}(A_i(\tau) \geq 1)\bar{Y}_i(s)X^{\circ}_{i1}\exp(\psi' X^{\circ}_i) \mid Z_i(0), X_{i2}\right\} \\
&= \sum_{k=0}^{1} P\left(A_i(\tau) \geq 1, Z_i(s^-) = k, X^{\circ}_{i1}(\tau) = 1 \mid Z_i(0), X_{i2}\right) \\
&\qquad \exp(\psi_1 + \psi'_2 X_{i2}) \\
&= \sum_{k=0}^{1} P\left(\mathcal{Z}^{\circ}_i(s^-) \in \mathcal{C}^{z}_{k} \mid Z_i(0), X_{i2}\right) \times \\
&\quad \sum_{j=1}^{\infty} P\left(\mathcal{Z}^{\circ}_i(\tau) \in \mathcal{C}^{ax^{\circ}_1}_{j1} \mid \mathcal{Z}^{\circ}_i(s^-) \in \mathcal{C}^{z}_{k}, X_{i2}\right) \\
&\qquad \exp(\psi_1 + \psi'_2 X_{i2}).
\end{aligned}
$$

B.2 Assaying the Last Sample; Left-Truncation

When the marker value is assessed using the biospecimen collected at the last visit and the failure time is considered as left-truncated, we have $X^{\circ}_{i1} = X^{\circ}_{i1}(V_i) = X_{i1}(a_{iA_i(V_i)})$ and $\bar{Y}_i(s) = Y_i(s)Y^{\dagger}_i(s)Y^{L}_i(s)$ where $L_i = a_{iA_i(V_i)}$. $E\{R^{(0)}_i(s) \mid Z_i(0), X_{i2}\}$ is now

$$
\begin{aligned}
&\sum_{k=0}^{1} P\big(A_i(\tau) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1, \\
&\qquad s \geq a_{iA_i(V_i)} \mid Z_i(0), X_{i2}\big) \\
&= \sum_{k=0}^{1} P\big(A_i(\tau) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1, \\
&\qquad A_i(s) = A_i(\tau) \mid Z_i(0), X_{i2}\big) \\
&= \sum_{k=0}^{1}\sum_{j=1}^{\infty} P\big(A_i(\tau) = j, Z_i(s^-) = k, dN_{i2}(s) = 1, \\
&\qquad A_i(s) = A_i(\tau) \mid Z_i(0), X_{i2}\big) \\
&= \sum_{k=0}^{1}\sum_{j=1}^{\infty} P\big(A_i(\tau) = j, Z_i(s^-) = k, dN_{i2}(s) = 1, \\
&\qquad A_i(s) = j \mid Z_i(0), X_{i2}\big) \\
&= \sum_{k=0}^{1}\sum_{j=1}^{\infty} P\big(A_i(s^-) = j, Z_i(s^-) = k, \\
&\qquad dN_{i2}(s) = 1 \mid Z_i(0), X_{i2}\big) \\
&= \sum_{k=0}^{1} P\left(A_i(s^-) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1 \mid Z_i(0), X_{i2}\right)
\end{aligned}
$$

and $E\{R^{(1)}_{i1}(s) \mid Z_i(0), X_{i2}\}$ is equal to

$$
\begin{aligned}
&\sum_{k=0}^{1} P\big(A_i(\tau) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1, X^{\circ}_{i1}(s^-) = 1, \\
&\qquad A_i(s) = A_i(\tau) \mid Z_i(0), X_{i2}\big) \\
&= \sum_{k=0}^{1}\sum_{j=1}^{\infty} P\big(A_i(\tau) = j, Z_i(s^-) = k, \\
&\qquad dN_{i2}(s) = 1, X^{\circ}_{i1}(s^-) = 1, A_i(s) = j \mid Z_i(0), X_{i2}\big) \\
&= \sum_{k=0}^{1} P\big(A_i(s^-) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1, \\
&\qquad X^{\circ}_{i1}(s^-) = 1 \mid Z_i(0), X_{i2}\big).
\end{aligned}
$$

Note these two conditional expectations have the same forms as in (B.1) and (B.2) where the analysis is conducted without left-truncation. This makes sense since $dN_{i2}(s) = 1$ implies no more visits after failure, which means the process has made its last visit by time s, i.e., $s \geq a_{iA_i(V_i)}$. The other two conditional expectations $E\{R^{(0)}_i(s;\psi) \mid Z_i(0), X_{i2}\}$ and $E\left\{R^{(1)}_{i1}(s;\psi) \mid Z_i(0), X_{i2}\right\}$ can be calculated as

$$
\sum_{l=0}^{1}\sum_{k=0}^{1}\sum_{j=1}^{\infty} \Big\{ P\big(A_i(\tau) = j, Z_i(s^-) = k, X^{\circ}_{i1}(\tau) = l,
$$

$$
\begin{aligned}
&A_i(s^-) = j \mid Z_i(0), X_{i2}\big) \\
&\exp(\psi_1 l + \psi_2' X_{i2})\Big\} \\
= \sum_{l=0}^{1}\sum_{k=0}^{1}\sum_{j=1}^{\infty} &\Big\{P\left(A_i(s^-) = j, Z_i(s^-) = k \mid Z_i(0), X_{i2}\right) \\
&\exp(\psi_1 l + \psi_2' X_{i2}) \\
&P\big(A_i(\tau) = j, X_{i1}^{\circ}(\tau) = l \mid A_i(s^-) = j, \\
&Z_i(s^-) = k, X_{i2}\big)\Big\} \\
= \sum_{l=0}^{1}\sum_{k=0}^{1}\sum_{j=1}^{\infty} &\Big\{P\left(\mathcal{Z}_i^{\circ}(s^-) \in \mathcal{C}_{jk}^{az} \mid Z_i(0), X_{i2}\right) \\
&\exp(\psi_1 l + \psi_2' X_{i2}) \\
&P\left(\mathcal{Z}_i^{\circ}(\tau) \in \mathcal{C}_{jl}^{ax_1^{\circ}} \mid \mathcal{Z}_i^{\circ}(s^-) \in \mathcal{C}_{jk}^{az}, X_{i2}\right)\Big\}
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{k=0}^{1}\sum_{j=1}^{\infty}\Big\{P\big(A_i(\tau) = j, Z_i(s^-) = k, X_{i1}^{\circ}(\tau) = 1, \\
&A_i(s^-) = j \mid Z_i(0), X_{i2}\big) \\
&\exp(\psi_1 + \psi_2' X_{i2})\Big\} \\
= \sum_{k=0}^{1}\sum_{j=1}^{\infty} &\Big\{P\left(\mathcal{Z}_i^{\circ}(s^-) \in \mathcal{C}_{jk}^{az} \mid Z_i(0), X_{i2}\right)\exp(\psi_1 + \psi_2' X_{i2}) \\
&P\left(\mathcal{Z}_i^{\circ}(\tau) \in \mathcal{C}_{j1}^{ax_1^{\circ}} \mid \mathcal{Z}_i^{\circ}(s^-) \in \mathcal{C}_{jk}^{az}, X_{i2}\right)\Big\}
\end{aligned}
$$

respectively.

B.3 Assaying the First Sample

Now we use the marker value measured using the first biospecimen, i.e., $X_{i1}^{\circ} = X_{i1}^{\circ}(a_{i1}) = X_{i1}(a_{i1})$ and the Cox regression analysis is conducted without left-truncation, i.e., $\bar{Y}_i(s) = Y_i(s)Y_i^{\dagger}(s)$. $E\{R_i^{(0)}(s) \mid Z_i(0), X_{i2}\}$ can be written as

$$
\begin{aligned}
&\sum_{k=0}^{1} P\big(A_i(\tau) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1, \\
&\quad X_{i1}^{\circ}(a_{i1}) = 1 \mid Z_i(0), X_{i2}\big) \\
= &\sum_{k=0}^{1} P\big(A_i(s^-) \geq 1, Z_i(s^-) = k, dN_{i2}(s) = 1, \\
&\quad X_{i1}^{\circ}(a_{i1}) = 1 \mid Z_i(0), X_{i2}\big) \\
= &\sum_{k=0}^{1}\int_0^s P\big(Z_i(s^-) = k, dN_{i2}(s) = 1, A_i(a_1^-) = 0, \\
&\quad Z_i(a_1^-) = 1, dA_i(a_1) = 1 \mid Z_i(0), X_{i2}\big)da_1 \\
= &\sum_{k=0}^{1}\int_0^s \Big\{P\big(Z_i(s^-) = k, A_i(a_1^-) = 0, Z_i(a_1^-) = 1, \\
&\quad dA_i(a_1) = 1 \mid Z_i(0), X_{i2}\big) \\
&\quad P\big(dN_{i2}(s) = 1 \mid Z_i(s^-) = k\big)\Big\}da_1 \\
= \lambda_2(s)ds &\sum_{k=0}^{1}\Big\{\exp(k\beta_1 + X_{i2}'\beta_2) \\
&\int_0^s P\big(Z_i(s^-) = k, A_i(a_1^-) = 0, Z_i(a_1^-) = 1, \\
&dA_i(a_1) = 1 \mid Z_i(0), X_{i2}\big)da_1\Big\} \qquad \text{(B.3)}
\end{aligned}
$$

where $P(Z_i(s^-) = k, A_i(a_1^-) = 0, Z_i(a_1^-) = l, dA_i(a_1) = 1 \mid Z_i(0), X_{i2})$ is equal to

$$
\begin{aligned}
&P(\mathcal{Z}_i^{\circ}(a_1^-) \in \mathcal{C}_{0l}^{az} \mid Z_i(0), X_{i2}) \times \\
&P\left(\mathcal{Z}_i^{\circ}(s^-) \in \mathcal{C}_k^{z} \mid \mathcal{Z}_i^{\circ}(a_1) \in \mathcal{C}_{1ll}^{azx_1^{\circ}}, X_{i2}\right) \qquad \text{(B.4a)} \\
&\quad \times \lambda_4(a_1)\exp(l\alpha_1 + X_{i2}'\alpha_2)
\end{aligned}
$$

for $a_1 \leq s$, and

$$
\begin{aligned}
&P(\mathcal{Z}_i^{\circ}(s^-) \in \mathcal{C}_{0k}^{az} \mid Z_i(0), X_{i2}) \times \\
&P\left(\mathcal{Z}_i^{\circ}(a_1^-) \in \mathcal{C}_{0l}^{az} \mid \mathcal{Z}_i^{\circ}(s^-) \in \mathcal{C}_{0k}^{az}, X_{i2}\right) \qquad \text{(B.4b)} \\
&\quad \times \lambda_4(a_1)\exp(l\alpha_1 + X_{i2}'\alpha_2)
\end{aligned}
$$

for $a_1 > s$. Moreover, we can write $E\{R_i^{(0)}(s;\psi) \mid Z_i(0), X_{i2}\}$ as

$$
\begin{aligned}
&\sum_{l=0}^{1}\sum_{k=0}^{1} P\big(A_i(\tau) \geq 1, Z_i(s^-) = k, X_{i1}^{\circ}(a_{i1}) = l \mid Z_i(0), X_{i2}\big) \\
&\quad \exp(\psi_1 l + \psi_2' X_{i2}) \\
= &\sum_{l=0}^{1}\sum_{k=0}^{1}\exp(\psi_1 l + \psi_2' X_{i2}) \times \\
&\int_0^{\tau} P\big(A_i(\tau) \geq 1, Z_i(s^-) = k, A_i(a_1^-) = 0, Z_i(a_1^-) = l, \\
&\quad dA_i(a_1) = 1 \mid Z_i(0), X_{i2}\big)da_1 \\
= &\sum_{l=0}^{1}\sum_{k=0}^{1}\exp(\psi_1 l + \psi_2' X_{i2}) \times \\
&\int_0^{\tau} P\big(Z_i(s^-) = k, A_i(a_1^-) = 0, Z_i(a_1^-) = l, \\
&\quad dA_i(a_1) = 1 \mid Z_i(0), X_{i2}\big)da_1
\end{aligned}
$$

and $E\left\{R_{i1}^{(1)}(s;\psi) \mid Z_i(0), X_{i2}\right\}$ as

$$
\begin{aligned}
&\sum_{k=0}^{1}\exp(\psi_1 + \psi_2' X_{i2}) \times \\
&\int_0^{\tau} P\big(Z_i(s^-) = k, A_i(a_1^-) = 0, Z_i(a_1^-) = 1, \\
&\quad dA_i(a_1) = 1 \mid Z_i(0), X_{i2}\big)da_1\,,
\end{aligned}
$$

where $P(Z_i(s^-)=k, A_i(a_1^-)=0, Z_i(a_1^-)=l, dA_i(a_1)=1 \mid Z_i(0), X_{i2})$ can be obtained from (B.4a) and (B.4b).

B.4 Assaying the First Sample; Left-Truncation

We still have $X_{i1}^\circ = X_{i1}^\circ(a_{i1}) = X_{i1}(a_{i1})$. The left truncation time L_i is now the first visit time, i.e., $L_i = a_{i1}$, and we have $\bar{Y}_i(s) = Y_i(s)Y_i^\dagger(s)Y_i^L(s)$ under this strategy. Note $E\{R_i^{(0)}(s) \mid Z_i(0), X_{i2}\}$ equals to

$$\begin{aligned}
&\sum_{k=0}^{1} P\big(A_i(\tau)\geq 1, Z_i(s^-)=k, dN_{i2}(s)=1,\\
&\qquad s\geq a_{i1} \mid Z_i(0), X_{i2}\big)\\
=&\sum_{k=0}^{1} P\big(A_i(\tau)\geq 1, Z_i(s^-)=k, dN_{i2}(s)=1,\\
&\qquad A_i(s)\geq 1 \mid Z_i(0), X_{i2}\big)\\
=&\sum_{k=0}^{1} P\left(A_i(s^-)\geq 1, Z_i(s^-)=k, dN_{i2}(s)=1 \mid Z_i(0), X_{i2}\right)
\end{aligned}$$

which is calculated in (B.1), and $E\{R_{i1}^{(1)}(s) \mid Z_i(0), X_{i2}\}$ can be written as

$$\begin{aligned}
&\sum_{k=0}^{1} P\big(A_i(\tau)\geq 1, Z_i(s^-)=k, dN_{i2}(s)=1, X_{i1}^\circ(a_{i1})=1,\\
&\qquad A_i(s)\geq 1 \mid Z_i(0), X_{i2}\big)\\
=&\sum_{k=0}^{1} P\big(A_i(s^-)\geq 1, Z_i(s^-)=k, dN_{i2}(s)=1,\\
&\qquad X_{i1}^\circ(a_{i1})=1 \mid Z_i(0), X_{i2}\big),
\end{aligned}$$

which can be obtained by (B.3). $E\{R_i^{(0)}(s;\psi) \mid Z_i(0), X_{i2}\}$ can be written as

$$\begin{aligned}
&\sum_{l=0}^{1}\sum_{k=0}^{1} P\big(A_i(\tau)\geq 1, Z_i(s^-)=k, X_{i1}^\circ(a_{i1})=l,\\
&\qquad A_i(s)\geq 1 \mid Z_i(0), X_{i2}\big)\exp(\psi_1 l+\psi_2' X_{i2})\\
=&\sum_{l=0}^{1}\sum_{k=0}^{1} P\big(A_i(s^-)\geq 1, Z_i(s^-)=k,\\
&\qquad X_{i1}^\circ(a_{i1})=l \mid Z_i(0), X_{i2}\big)\exp(\psi_1 l+\psi_2' X_{i2})\\
=&\sum_{l=0}^{1}\sum_{k=0}^{1}\exp(\psi_1 l+\psi_2' X_{i2})\times\\
&\int_0^s P\big(A_i(s^-)\geq 1, Z_i(s^-)=k, A_i(a_1^-)=0, Z_i(a_1^-)=l,\\
&\qquad dA_i(a_1)=1 \mid Z_i(0), X_{i2}\big)da_1\\
=&\sum_{l=0}^{1}\sum_{k=0}^{1}\exp(\psi_1 l+\psi_2' X_{i2})\times\\
&\int_0^s P\big(Z_i(s^-)=k, A_i(a_1^-)=0, Z_i(a_1^-)=l,\\
&\qquad dA_i(a_1)=1 \mid Z_i(0), X_{i2}\big)da_1
\end{aligned}$$

and $E\{R_{i1}^{(1)}(s;\psi) \mid Z_i(0), X_{i2}\}$ is calculated as

$$\begin{aligned}
&\sum_{k=0}^{1} P\big(A_i(\tau)\geq 1, Z_i(s^-)=k, X_{i1}^\circ(a_{i1})=1,\\
&\qquad A_i(s)\geq 1 \mid Z_i(0), X_{i2}\big)\exp(\psi_1+\psi_2' X_{i2})\\
=&\sum_{k=0}^{1}\exp(\psi_1+\psi_2' X_{i2})\times\\
&\int_0^s P\big(Z_i(s^-)=k, A_i(a_1^-)=0, Z_i(a_1^-)=1,\\
&\qquad dA_i(a_1)=1 \mid Z_i(0), X_{i2}\big)da_1\,.
\end{aligned}$$

Similarly we can calculate $P(Z_i(s^-)=k, A_i(a_1^-)=0, Z_i(a_1^-)=l, dA_i(a_1)=1 \mid Z_i(0), X_{i2})$ via (B.4a) and (B.4b).

ACKNOWLEDGEMENTS

Richard J. Cook thanks the University of Waterloo for support through the Faculty of Mathematics Research Chair program.

FUNDING

This work was funded by grants to Richard J. Cook from the Canadian Institutes for Health Research (FRN 13887) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-04207).

Accepted 25 July 2026

REFERENCES

[1] Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* **10**(4) 1100–1120. MR0673646

[2] Andersen, P. K. and Liestøl, K. (2003). Attenuation caused by infrequently updated covariates in survival analysis. *Biostatistics* **4**(4) 633–649. https://doi.org/10.1093/biostatistics/4.4.633.

[3] Binquet, C., Chêne, G., Jacqmin-Gadda, H., Journot, V., Savès, M., Lacoste, D., Dabis, F. and the Groupe d'Epidémiologie Clinique du SIDA en Aquitaine (2001). Modeling Changes in CD4-positive T-Lymphocyte Counts after the Start of Highly Active Antiretroviral Therapy and the Relation with Risk of Opportunistic Infections The Aquitaine Cohort, 1996–1997. *American Journal of Epidemiology* **153**(4) 386–393. https://doi.org/10.1093/aje/153.4.386.

[4] Cook, R. J. and Lawless, J. F. (2007) *The Statistical Analysis of Recurrent Events*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-69810-6. MR3822124

[5] Cook, R. J. and Lawless, J. F. (2018) *Multistate Models for the Analysis of Life History Data*. Chapman and Hall/CRC, New York, NY. https://doi.org/10.1201/9781315119731.

[6] Cook, R. J. and Lawless, J. F. (2021). Independence conditions and the analysis of life history studies with intermittent observation. *Biostatistics* **22**(3) 455–481. https://doi.org/10.1093/biostatistics/kxz047. MR4287163

[7] Cook, R. J., Lawless, J. F. and Xie, B. (2022). Marker-dependent observation and carry-forward of internal covariates in Cox regression. *Lifetime Data Analysis* **28**(4) 560–584. https://doi.org/10.1007/s10985-022-09561-9. MR4484929

[8] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2) 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. MR0341758

[9] de Bruijne, M. H. J., le Cessie, S., Kluin-Nelemans, H. C. and van Houwelingen, H. C. (2001). On the use of Cox regression in the presence of an irregularly observed time-dependent covariate. *Statistics in Medicine* **20**(24) 3817–3829. https://doi.org/10.1002/sim.1083.

[10] Gladman, D. D. and Chandran, V. (2011). Observational cohort studies: lessons learnt from the University of Toronto Psoriatic Arthritis Program. *Rheumatology* **50**(1) 25–31. https://doi.org/10.1093/rheumatology/keq262.

[11] Gladman, D. D., Shuckett, R., Russell, M. L., Thorne, J. C. and Schachter, R. K. (1987). Psoriatic Arthritis (PSA) - An Analysis of 220 Patients. *QJM: An International Journal of Medicine* **62**(2) 127–141. https://doi.org/10.1093/oxfordjournals.qjmed.a068085.

[12] Jiang, S., Cook, R. J. and Zeng, L. (2020). Mitigating bias from intermittent measurement of time-dependent covariates in failure time analysis. *Statistics in Medicine* **39**(13) 1833–1845. https://doi.org/10.1002/sim.8517. MR4098524

[13] Lawless, J. F. (2003) *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, Hoboken, NJ. https://doi.org/10.1002/9781118033005. MR1940115

[14] Lin, D. Y. and Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association* **84**(408) 1074–1078. https://doi.org/10.1080/01621459.1989.10478874. MR1134495

[15] Papageorgiou, G., Mauff, K., Tomer, A. and Rizopoulos, D. (2019). An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes. *Annual Review of Statistics and Its Application* **6**(1) 223–240. https://doi.org/10.1146/annurev-statistics-030718-105048. MR3939519

[16] Raboud, J., Reid, N., Coates, R. A. and Farewell, V. T. (1993). Estimating Risks of Progressing to Aids when Covariates are Measured with Error. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **156**(3) 393–406. https://doi.org/10.2307/2983065.

[17] Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**(2) 363–369. https://doi.org/10.1093/biomet/73.2.363. MR0855896

[18] Tsiatis, A. A., Degruttola, V. and Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association* **90**(429) 27–37. https://doi.org/10.1080/01621459.1995.10476485

[19] Tsiatis, A. A. and Davidian, M. (2004). Joint Modeling of Longitudinal and Time-to-Event Data: an OVERVIEW. *Statistica Sinica* **14**(3) 809–834. MR2087974

[20] Wulfsohn, M. S. and Tsiatis, A. A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics* **53**(1) 330–339. https://doi.org/10.2307/2533118. MR1450186

[21] Xu, J. and Zeger, S. L. (2002). Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Events. *Journal of the Royal Statistical Society Series C: Applied Statistics* **50**(3) 375–387. https://doi.org/10.1111/1467-9876.00241. MR1856332

Jianchu Chen. Department of Statistics and Actuarial Science, University of Waterloo, Canada. E-mail address: j773chen@uwaterloo.ca

Richard J. Cook. Department of Statistics and Actuarial Science, University of Waterloo, Canada. E-mail address: rjcook@uwaterloo.ca