Published in final edited form as:
Statistics in Medicine (2018), **37**(15): 2354–2366
DOI: 10.1002/sim.7646

Tracing studies in cohorts with attrition: Selection models for efficient sampling

NATHALIE C. MOON
Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

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

RICHARD J. COOK
Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Cohort studies of chronic diseases involve recruitment and longitudinal followup of affected individuals with a view to studying the effect of risk factors on disease progression and death. When the time to withdrawal from the cohort is conditionally independent of the disease process the primary consequence is a loss of information on the parameters of interest. This loss can sometimes be mitigated through the conduct of tracing studies in which a subsample of those lost to follow up are contacted and some information is obtained on their disease and survival status. We describe the use of selection models to sample individuals for tracing who will yield more efficient estimators than those obtained by simple random sampling. Efficient sampling schemes featuring cost constraints are also developed and shown to perform well. An application to data from the University of Toronto Psoriatic Arthritis Cohort illustrates how to apply the method in a real setting.

Keywords: attrition, inception cohort, Markov model, selection model, sequentially missing at random, tracing

1 INTRODUCTION

Considerable investments are being made in health research to support the conduct of large cohort studies geared towards understanding the relationships between diverse features (e.g. exposure to toxins, genetic and serological biomarkers) and disease incidence, progression, and mortality. Birth cohorts are typically directed at measuring the impact of maternal exposures on neonatal and early life outcomes (Kobayashi et al., 2016), whereas studies in infants and young children may be directed at the impact of early diet and care on cognitive and physical development (Lakshman et al., 2015). The Canadian Longitudinal Study on Aging (Raina et al., 2009) focuses on disease processes in later life; 50,000 individuals aged 45 to 85 were recruited and are to be followed for 20 years to gain insight into the complex relationships between behaviour, biomarkers and disease incidence. The EPIC Norfolk study (Riboli, 1992) and many others have broadly similar objectives. In other settings, attention may be directed at diseased individuals and interest lies in studying the incidence of complications or comorbidities in affected individuals; studies in diabetics are particularly ubiquitous (Early Treatment Diabetic Retinopathy Study Research Group and others, 1991). While interest may lie primarily in biomarkers and their effect on the development of complications from disease and the onset of comorbidities, mortality rates may be appreciable and joint models incorporating survival times are required for valid inferences. Multistate models offer a convenient and powerful framework for the joint consideration of disease incidence, progression, and mortality.

We consider the setting in which individuals are recruited to an inception cohort and are followed prospectively to learn about the disease process and identify risk factors for the occurrence of disease complications and the development of comorbidities. Clinically important events are often self-evident (e.g. strokes, heart attacks, and death) but their observation times are subject to right-censoring. Some complications, however, are asymptomatic and so will only be detected at the time of clinical examination or radiographic assessment. For example, asymptomatic fractures among individuals with osteoporosis are only detected upon x-ray (Kreiger et al., 1999), progression in retinopathy in diabetics is only detected upon examination by an ophthalmologist (Control et al., 1993), and progression in fibrosis of the liver among individuals with hepatitis C infection is only assessable by biopsy (Sweeting et al., 2006). In settings where interest lies in the development of conditions or complications which are not self-evident, data become available at periodic clinic visits; such data can be naturally analyzed using multistate models.

At the planning stage, it is natural to investigate the efficiency implications of different sample sizes, planned durations of follow-up, and assessment schedules, which will depend on the form of the underlying model and the parameter setting of interest. Albert and Brown (1991) consider different sampling schemes and schedules for the assessments in a two-state process. Cook (2000) assessed the impact of the assessment schedule on the precision of estimates of transition intensities and occupancy probabilities; Lawless and Nazeri Rad (2015) studied this and more general three-state processes. Mehtälä et al. (2011) consider sample size and the optimal scheduling of assessments for time-homogeneous two-state Markov processes; Hwang and Brookmeyer (2003) consider similar issues for progressive $K-$state processes. Unlike this work, here our focus is on a secondary design issue which arises in longitudinal studies with attrition. In some such settings, it is possible to trace individuals who are lost to follow-up to obtain information on their disease status. We develop a framework for efficiently selecting individuals lost to follow-up for such a tracing study in order to yield more efficient estimators of the key parameters of interest.

This work is motivated by a research program at the Centre for Prognosis Studies in the Rheumatic Diseases at the University Health Network in Toronto, Canada. Individuals with psoriatic arthritis have been recruited into the University of Toronto Psoriatic Arthritis Cohort since 1976. Upon completing a detailed examination upon clinic entry individuals are scheduled for annual clinical assessments and biannual radiographic assessment. A primary scientific objective is to estimate the

incidence of arthritis mutilans, a severe form of arthritis, along with the effect of human leukocyte antigen (HLA) markers.

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

The remainder of the paper is organized as follows. In Section 2, we introduce notation for multi-state models and the likelihood for panel data with attrition under a SMAR mechanism, define the tracing selection model, and describe the idea of optimal selection for tracing. Asymptotic calculations demonstrating the efficiency gains from optimal tracing compared with simple random sampling are also given. When the cost of securing information on disease progression status is different from the cost of simply obtaining survival status, the cost implications of optimal tracing are also provided. In Section 3, a more general optimization process is described with cost constraints, which leads to different optimal selection models; the efficiency gains are also illustrated in this setting based on asymptotic results. An application of the proposed methodology to data collected from a cohort study conducted at the University of Toronto Psoriatic Arthritis Clinic is presented in Section 4 and general remarks are given in Section 5.

2 Model Formulation and the Design of Tracing Studies

2.1 A Multi-state Markov Model for Disease Progression

The K-state process depicted in Figure 1 offers a powerful framework for joint consideration of disease complications, comorbidities, and death in progressive conditions. Let $Z(t)$ represent which state in the state space $\mathcal{S} = \{1, 2, \ldots, K\}$ is occupied at time t since disease onset. We let $\{Z(s), 0 \leq s\}$ denote the corresponding stochastic process and write the complete process history as $\mathcal{H}(t) = \{Z(s), 0 \leq s < t\}$ which contains information on the times and types of transitions over $[0, t)$. The intensity of a transition from state j to state k at time t is

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

for all $j < k \in \mathcal{S}$. For Markov processes the transition intensities depend only on $Z(t^-)$ and we write $\lambda_{jk}(t|\mathcal{H}(t)) = \lambda_{jk}(t)$.

The transition intensity matrix $\mathbb{A}(t)$ has entries $\lambda_{jk}(t)$ for $j < k \in \mathcal{S}$, $-\sum_{k \neq j} \lambda_{jk}$ along the diagonal, and zero elsewhere. The transition probability matrix $\mathbb{P}(s, t)$ has (j, k) entry $[\mathbb{P}(s, t)]_{jk} = P(Z(t) = k | Z(s) = j)$ and is obtained by solving the Kolmogorov forward differential equations $\partial \mathbb{P}(s, t)/\partial t = \mathbb{P}(s, t)\mathbb{A}(t)$. For time-homogeneous processes, $\lambda_{jk}(t) = \lambda_{jk}$ for all $j < k \in \mathcal{S}$ and so $\mathbb{A}(t) = \mathbb{A}$; we may then obtain $\mathbb{P}(t) = \exp(\mathbb{A}t)$ and $\exp(\cdot)$ here denotes the matrix exponential (Cox and Miller, 1965).

Multiplicative intensity-based models are used to characterize the effect of prognostic variables

on the dynamics of the disease process. Modulated Markov models are obtained by specifying

$$\lambda_{jk}(t|X) = \lambda_{jk0}(t)\exp(X'\beta_{jk}), \;\; j < k \in \mathcal{S},$$

where $\lambda_{jk0}(t)$ is the baseline transition intensity, $X = (X_1, \ldots, X_p)'$ represents a $p \times 1$ covariate vector, and $\beta_{jk} = (\beta_{jk1}, \ldots, \beta_{jkp})'$ is a vector of regression coefficients specific to $j \to k$ transitions. We let θ be the vector of parameters indexing all baseline intensity functions and regression coefficients; for a time-homogeneous illness-death process $K = 3$ and $\theta = (\lambda', \beta')'$ where $\lambda = (\lambda_{120}, \lambda_{130}, \lambda_{230})'$ and $\beta = (\beta'_{12}, \beta'_{13}, \beta'_{23})'$ is the vector of all regression coefficients.

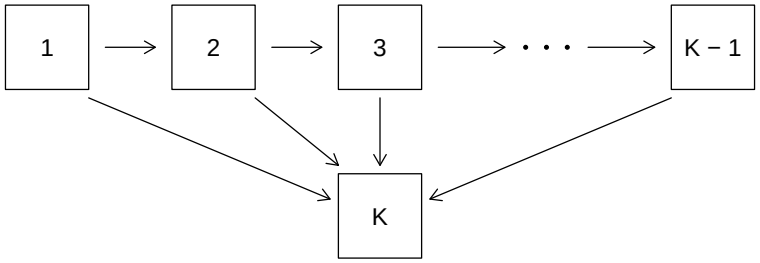


Figure 1: A K-state progressive model with $K - 1$ transient states and one absorbing state

2.2 Intermittent Assessment with Dropout

It is often not possible to monitor disease status continuously in cohort studies but rather only examine individuals at periodic assessment times. Consider an inception cohort of individuals recruited and examined at the time of disease onset ($t = 0$, say), and let V_j, $j = 1, \ldots, J$ represent common planned assessment times measured from the time of disease onset; in this case $V_J = \tau$ is a common administrative censoring time. To simplify the notation, we consider the contributions from a generic individual and let $Z_j = Z(V_j)$ denote the state occupied at the jth assessment, where $Z_0 = Z(V_0) = Z(0) = 1$. Let $\bar{Z}_j = \{Z_0, Z_1, \ldots, Z_j\}$ denote the history of the process up to the jth assessment, $j = 0, 1, \ldots, J$; $\bar{Z}_J$ then represents the complete response vector we aim to observe.

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

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

where $P(\bar{Y}_J \mid \bar{Z}_J; X; \gamma)$ is the conditional probability of the missing data indicators $\bar{Y}_J$ given the full response vector $\bar{Z}_J$ and covariate vector X, parameterized in terms of γ. Note that if $Z_C = K$ then the process has been observed to completion and the sum in (1) is degenerate. With a monotone SMAR mechanism (Hogan et al., 2004), the probability of becoming lost to follow-up at a given assessment

only depends on an individual's disease status and covariates observed at the previous assessments, that is

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

So under a SMAR mechanism, we assume that whether an individual who is present at assessment j (i.e. $Y_j = 1$) is available at assessment $j + 1$ depends only on the data available at assessment j. In the event that death occurs over (V_j, V_{j+1}), the individual will by definition *not* be present at visit $j + 1$, but his/her vital status could still be ascertained, either from family members or other sources (e.g. death registries/newspapers). We assume here that the probability that vital status is ascertained is governed by a SMAR mechanism, in which case (1) can be factored as the product of two terms of the form

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

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

$$L_1(\theta) = \prod_{j=1}^{C} P(Z_j|\bar{Z}_{j-1}, X; \theta) \ , \tag{2}$$

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

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

$$P(R = 1|\mathcal{D}, \Delta = 0) = P(R = 1|\bar{Z}_C, X, C, \Delta = 0; \rho) \ , \tag{3}$$

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

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

We assume

$$P(Z_J|R, \bar{Z}_C, \bar{Y}_J, X, C, \Delta) = P(Z_J|\bar{Z}_C, X; \theta) \ , \tag{4}$$

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

$$L_2(\theta) = P(Z_J|\bar{Z}_C, X; \theta)^{R(1-\Delta)} \ . \tag{5}$$

We can then augment the likelihood $L_1(\theta)$ in (2) by incorporating data from the tracing study and use

$$L(\theta) = L_1(\theta)L_2(\theta) \ . \tag{6}$$

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

2.3 Optimal Designs for Tracing

Now consider a sample of size n where we use the subscript i to label individuals, $i = 1, \ldots, n$. Let $\mathcal{D}_i = \{\bar{Z}_{iC}, \bar{Y}_{iJ}, X_i, C_i, \Delta_i\}$ denote the observed data from individual i from the regular follow-up process in phase I and $\mathcal{D} = \{\mathcal{D}_i, i = 1, \ldots, n\}$ denote the phase I data. Then, we write $L_1(\theta) = \prod_{i=1}^n L_{i1}(\theta)$ where $L_{i1}(\theta) = \prod_{j=1}^{C_i} P(Z_{ij}|\bar{Z}_{i,j-1}, X_i; \theta)$ as in (2) and let $\tilde{\theta}$ be the MLE of θ obtained by maximizing the likelihood $L_1(\theta)$ from the phase I data. The observed information matrix from phase I is

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

If $L_{i2}(\theta) = P(Z_{iJ}|\bar{Z}_{iC}, X_i; \theta)^{R_i(1-\Delta_i)}$ is the contribution from individual i from (5), then conditioning on their phase I data, their contribution to the expected information matrix from tracing is

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

which over all n individuals gives expected information matrix

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

under the assumption in (4). Consider a hybrid information matrix defined as the sum of the observed information matrix from the phase I data, and the expected information matrix arising from a phase II tracing study, given by

$$I_H(\tilde{\theta}, \rho) = I_1(\tilde{\theta}) + \mathcal{I}_2^{\dagger}(\tilde{\theta}, \rho) \ . \tag{7}$$

We propose to use (7) with θ replaced by the estimate $\tilde{\theta}$ from phase I to set the value of ρ for the selection model. If interest lies in making inference for a particular parameter θ_k, for example, the so-called "optimal" tracing selection parameters ρ^{opt} may be obtained by minimizing $[I_H^{-1}(\tilde{\theta}, \rho)]_{kk}$, the (k, k) element of the inverse of (7), subject to a constraint on $\pi = P(R = 1|\Delta = 0)$, the overall proportion of individuals lost to follow-up who are traced. This can be implemented by minimizing

$$\big[I_H^{-1}(\tilde{\theta}, \rho)\big]_{kk} + \zeta \Big[\sum_{i:\Delta_i=0} P(R_i = 1|\mathcal{D}_i, \Delta_i = 0; \rho)/(n - \dot{\Delta}) - \pi \Big] \tag{8}$$

with respect to ρ to get ρ^{opt}, where ζ is a Lagrange multiplier, the first term in square brackets is the empirical expectation of the selection probabilities averaging over the observed data with $\dot{\Delta} = \sum_{i=1}^n \Delta_i$, and the entire term in square brackets is a constraint which ensures the expected proportion of individuals lost to follow-up to be traced is satisfied. The delta method may be used to consider situations when estimation of a function $g(\theta)$ is the focus. The optimal criteria in (8) can be

generalized to involve any linear function $h(\cdot)$ of the elements of $I_H(\tilde{\theta}, \rho)$. In particular, analogs of A-optimality and C-optimality (Emery and Nenarokomov, 1998) can be achieved, but we do not pursue this here as we focus on the case the tracing study is conducted with a specific scientific question in mind.

Let $\widehat{\theta}$ denote the final estimates obtained based on the augmented likelihood (6) once the tracing study is completed. The asymptotic variance of $\widehat{\theta}$ is thus asvar$(\sqrt{n}(\widehat{\theta} - \theta)) = \mathcal{I}^{-1}(\theta, \gamma, \rho)$ where

$$\mathcal{I}(\theta, \gamma, \rho) = E\Big[- \frac{\partial^2 \log L_i(\theta)}{\partial\theta\partial\theta'}\Big] = E\big[I_{i1}(\theta)\big] + E\big[(1 - \Delta_i)\mathcal{I}^{\dagger}_{i2}(\theta, \rho)\big] \ . \tag{9}$$

The above expectation is taken with respect to the phase II tracing information by first conditioning on the phase I (incomplete) data and then taking the unconditional expectation. Note that to determine ρ^{opt} in applications, as in the analysis of Section 4, we use (8); but for the calculation of the asymptotic relative efficiency that follow, we use (9) in lieu of $I_H(\tilde{\theta}, \rho)$ in (8) for computational efficiency; the results agree extremely well with the more computationally demanding results based on (8).

2.4 Assessing the Efficiency Gains from "Optimal" Tracing

We now study the properties of estimators obtained following the proposed tracing procedure, highlighting the efficiency gains over selection models involving simple random sampling (SRS). We consider a time-homogeneous illness-death model and assume a binary covariate X with $P(X = 1) = 0.25$ modulating the $1 \to 2$ transition intensity, which gives a parameter vector $\theta = (\lambda_{120}, \lambda_{130}, \lambda_{230}, \beta_{12})'$. For an inception cohort, without loss of generality, we consider the period of observation $[0, \tau]$ with $\tau = 1$. We let $N_{12}(\tau)$ indicate that a $1 \to 2$ transition occurred over $[0, \tau]$. We set $\beta_{12} = \log 1.5$ and the values of the baseline intensities to satisfy the following constraints:

(i) $P_2 = P(N_{12}(\tau) = 1|X = 0) = \lambda_{120}/(\lambda_{120} + \lambda_{130})\{1 - \exp[-(\lambda_{120} + \lambda_{130})\tau]\} = \{0.25, 0.75\}$;

(ii) $\lambda_{230}/\lambda_{130} = 1.5$; and

(iii) $P_3 = P(Z(\tau) = 3|X = 0) = \{0.1, 0.5\}$.

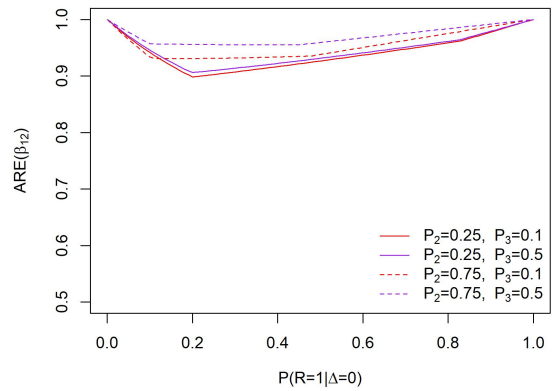
We assume the progression status is assessed intermittently at $J = 5$ equally spaced scheduled assessments over $[0, \tau]$. For the dropout process, we set the dropout indicator $Y_0 = 1$ at baseline (e.g. time V_0) for all individuals and generate Y_j given (Y_{j-1}, Z_{j-1}) sequentially for $j = 1, 2, \ldots, J$. As described in Section 2.2, $P(Y_j = 1|Y_{j-1} = 0) = 0$ and $P(Y_j = 1|Y_{j-1} = 1, Z_{j-1} = K) = 1$. For the SMAR mechanism, we set logit $P(Y_j|Y_{j-1} = 1, \bar{Z}_{j-1}, X; \gamma) = \gamma_0 + \gamma_1 I(Z_{j-1} = 2)$, that is the odds of drop-out at a given assessment depends on the disease status at the previous assessment. The value of the parameters (γ_0, γ_1) are set to achieve an overall percentage of dropout of $P(\Delta = 0) = \{0.4, 0.8\}$ and an odds ratio of dropout for individuals with previous disease status $Z_{j-1} = 2$ vs $Z_{j-1} = 1$ to be $\exp(\gamma_1) = 2$.

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

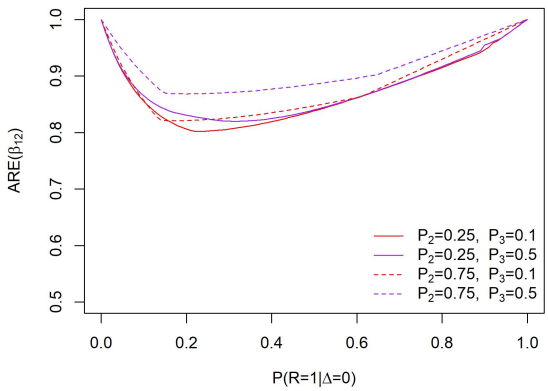
$$\text{logit}\, P(R = 1|Z_C, X, \Delta = 0; \rho) = \rho_0 + \rho_1 I(Z_C = 2) + \rho_2 X + \rho_3 I(Z_C = 2)X \tag{M1}$$

where X is the same binary covariate related to the $1 \to 2$ transition. To illustrate the magnitude of potential efficiency gains from tracing as well as influential factors, we compare the asymptotic variance of estimates of response parameters based on an optimal design versus a simple random sampling (SRS) design (which is equivalent to setting the tracing model parameters to be $\rho^{srs} = (\rho_0, 0, 0, 0)$). The optimal tracing parameter ρ^{opt} results in the minimal asymptotic relative efficiency

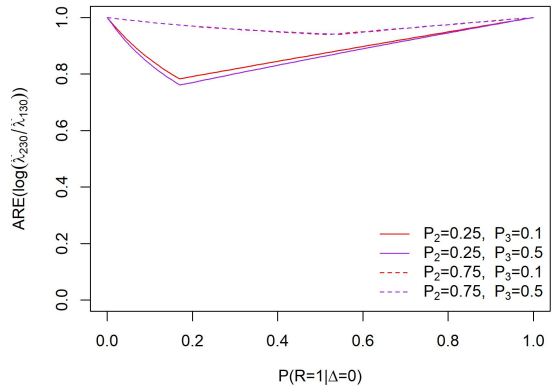
$$ARE(\widehat{\theta}_k) = \frac{[\mathcal{I}^{-1}(\theta, \gamma, \rho^{opt})]_{kk}}{[\mathcal{I}^{-1}(\theta, \gamma, \rho^{srs})]_{kk}} \ , \tag{10}$$



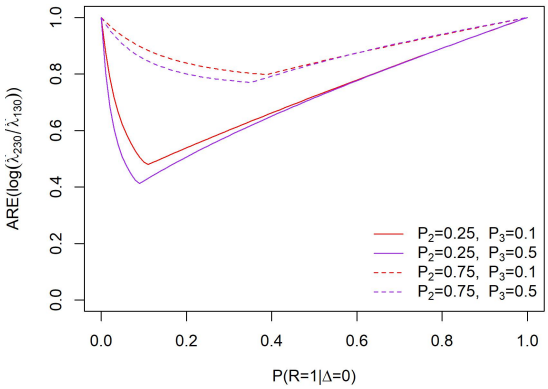
(a) $P(\Delta = 0) = 0.4$



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



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



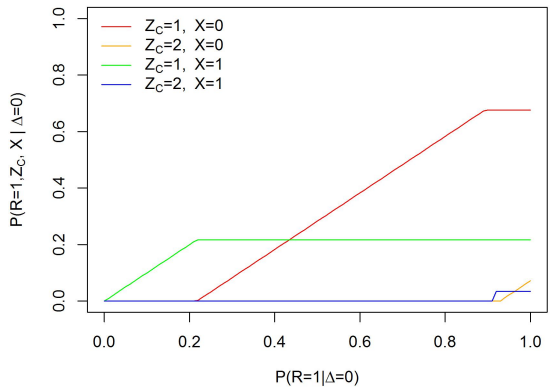
(d) $P(\Delta = 0) = 0.8$

Figure 2: Asymptotic relative efficiency (10) of the estimator $\widehat{\beta}_{12}$ (top panels) and $\log(\widehat{\lambda}_{230}/\widehat{\lambda}_{130})$ (bottom panels) with a tracing study under an optimal design versus a SRS design of the same expected size; $P_2 = P(N_{12}(\tau) = 1 \mid X = 0)$, $P_3 = P(Z(\tau) = 3 \mid X = 0)$, $\lambda_{230}/\lambda_{130} = 1.5$, $\beta_{12} = \log 1.5$

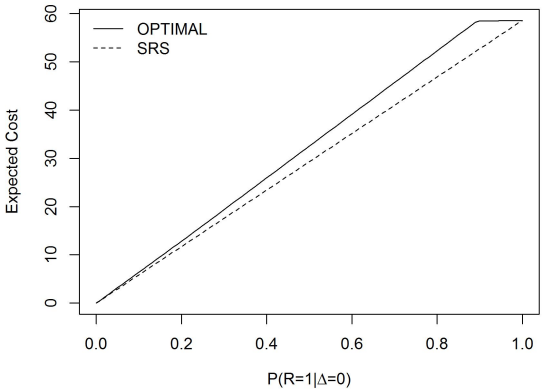
subject to a pre-specified proportion of tracing $\pi = P(R = 1|\Delta = 0)$.

As expected, the optimal tracing designs lead to more precise estimates than the SRS designs across all scenarios. This is depicted in Figure 2 for the estimation of covariate effect β_{12} (top two panels) and $\log(\lambda_{230}/\lambda_{130})$ (bottom two panels). Across all parameter configurations considered, the gain in efficiency increases with the probability of dropout $P(\Delta = 0)$. The magnitude of the gain in efficiency also varies as a function of the parameters of the disease process (as represented by the multiple curves in each panel) and the marginal tracing probability π. While these relationships are complex and dependent on properties of the disease process, we describe some general trends apparent in the present examples. When interest lies in estimating the covariate effect modulating the $1 \to 2$ transition (β_{12}), the smaller the percentage of progression by the administrative censoring time τ (e.g. P_2), the greater the gain in efficiency achieved by the optimal tracing scheme relative to the SRS approach. This is due to the fact that the optimal design for estimation of β_{12} prioritizes tracing progression-free individuals ($Z_C = 1$) over those who have already progressed ($Z_C = 2$) as the former may potentially provide new information on the $1 \to 2$ transition directly; this can be seen in panels (a) and (b) of Figure 2 when contrasting the solid ($P_2 = 0.25$) and dashed ($P_2 = 0.75$) lines of

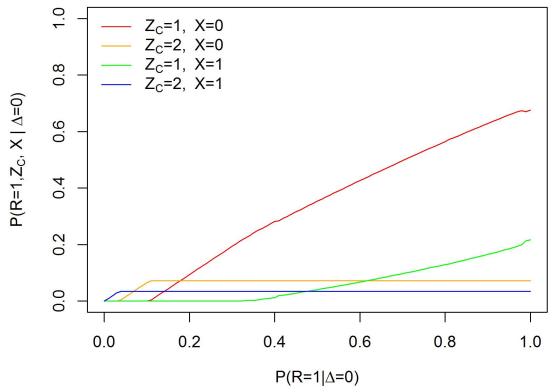
the same colours with the fixed P_3. This trend is much clearer for estimating the relative risk of death $\log(\lambda_{230}/\lambda_{130})$ as shown in Figure 2 (c), (d) . The percentage of death observed by the administrative censoring time (i.e. P_3) also has some impact on the estimation of a covariate effect on progression, β_{12}. The lower P_3 is (e.g. 0.1 versus 0.5) the bigger the gain in efficiency by adopting the optimal design for tracing, although such a difference is only appreciable when the percentage of progression is high ($P_2 = 0.75$) as shown in Figure 2 (a), (b). Interestingly, P_3 seems to have a different impact on efficiency gain for the estimation of the relative risk of death $\log(\lambda_{230}/\lambda_{130})$. When the drop-out rate is high ($P(\Delta = 0) = 0.8$), Figure 2 (d) shows slightly greater benefit of the optimal tracing scheme over SRS as the percentage of death P_3 increases, but such a pattern is only noticeable if percentage of tracing is low to moderate (e.g. $\pi < 0.4$).



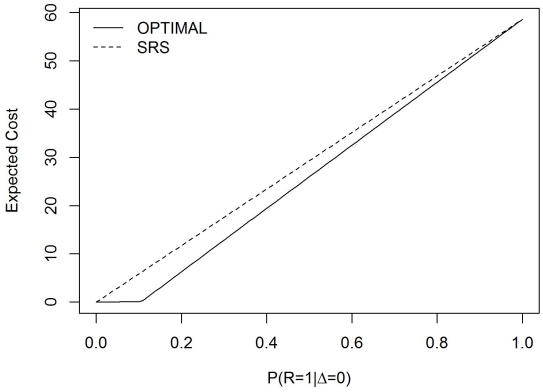
(a) Optimal Design for Estimation of β_{121}



(b) Expected Cost under Designs for Estimation of β_{121}



(c) Optimal Design for Estimation of $\log(\lambda_{230}/\lambda_{130})$



(d) Expected Cost under Designs for Estimation of $\log(\lambda_{230}/\lambda_{130})$

Figure 3: Optimal tracing design (left-hand panels) and expected cost (right-hand panels) under an optimal vs a SRS design for the estimation of β_{12} (top panels) and $\log(\lambda_{230}/\lambda_{130})$ (bottom panels), with $P_2 = 0.25$ and $P_3 = 0.1$, $\lambda_{230}/\lambda_{130} = 1.5$, $\beta_{12} = \log 1.5$, $P(\Delta = 0) = 0.8$, and cost ratio $\xi = C_d/C_s = 100$

In summary, efficiency gains for the estimation of both covariate effect on progression β_{12} and relative risk of death $\log(\lambda_{230}/\lambda_{130})$ are primarily driven by observing instances of disease progres-

sion, e.g P_2. The percentage of death P_3 has some additional impact depending on which quantity is of interest for estimation. Slightly larger gains in efficiency for the estimation of β_{12} can be obtained when P_3 is low, because as P_3 increases the likelihood of gaining information about progression at the time of tracing decreases. However, when interest lies in estimating $\log(\lambda_{230}/\lambda_{130})$, observation of death events could be more valuable and so larger gains in efficiency are achieved by the proposed approach when the probability of death is higher.

Figure 3 focuses on the setting with $P_2 = 0.25$, $P_3 = 0.1$, and $P(\Delta = 0) = 0.8$ (e.g. the solid red line in the right-hand panels of Figure 2), again considering estimation of β_{12} and $\log(\lambda_{230}/\lambda_{130})$ in the top and bottom panels respectively. The left-hand panels contain plots of the joint probability $P(R = 1, Z_C, X|\Delta = 0)$ against the marginal probability of tracing π under an optimal design. As will be discussed in Section 3, it is generally reasonable to assume that the cost of tracing individuals for vital status (C_s) is substantially lower than that of assessing disease status (C_d), so $\xi = C_d/C_s \geq 1$. In the right-hand panels of Figure 3, we fix $\xi = 100$ and observe that the expected cost of the proposed optimal tracing scheme (solid line) is greater than that of a SRS tracing scheme (dashed line) for the estimation of β_{12}, whereas it is lesser for the estimation of $\log(\lambda_{230}/\lambda_{130})$. This follows directly from the patterns exhibited in the left-hand panels: the optimal tracing scheme for β_{12} preferentially selects individuals with $Z_C = 1$ (more expensive) over those with $Z_C = 2$ (less expensive), while the optimal scheme for $\log(\lambda_{230}/\lambda_{130})$ prioritizes individuals with $Z_C = 2$ over those with $Z_C = 1$. It is interesting to note that for the latter case, the optimal design not only leads to substantial gains in efficiency, but is also more economical than the SRS design of the same size. In addition, the optimal scheme for β_{12} sequentially draws upon individuals with $(Z_C = 1, X = 1)$, $(Z_C = 1, X = 0)$, $(Z_C = 2, X = 1)$, and $(Z_C = 2, X = 0)$; preferring the former subgroups to the exclusion of the latter, as the marginal probability of selection for tracing (π) increases. However, this is not true to the same extent in the optimal scheme for $\log(\lambda_{230}/\lambda_{130})$; in this case, the proposed tracing scheme allows for the optimal equilibrium to be identified, which would not otherwise be possible. The results of extensive simulation studies (not shown) demonstrate excellent agreement between the asymptotic and empirical efficiency gains.

2.5 A Tracing Selection Model Incorporating the Time of Study Withdrawal

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

$$\text{logit}\, P(R = 1|Z_C, X, \Delta = 0; \rho) = \rho_0 + \rho_1 I(Z_C = 2) + \rho_2 X + \rho_3 I(Z_C = 2)X + \rho_4 D \qquad \text{(M2)}$$

to allow tracing selection probabilities to further depend on $D = \tau - V_C$, the time from loss-to-follow-up to tracing. Since the tracing selection model in (M1) is nested in (M2), greater efficiency gains may be realized under the latter model. The benefit of including time since loss-to-follow-up in the tracing selection model is most appreciable for the estimation of relative risk of death with, versus without, progression given by $\log(\lambda_{23}/\lambda_{13})$. A summary of results comparing asymptotic efficiency gains under these two tracing selection models is presented in the left-hand columns of Table 1 under the heading "Size Constraint"; we defer the discussion of the right-hand side under the heading "Cost Constraint" to Section 3.2. Here, we find the efficiency gains can be appreciable for both β_{12} and $\log(\lambda_{230}/\lambda_{130})$. We also see a non-monotonic trend in relative efficiency of "optimal" versus simple random sampling when viewed as a function of the marginal selection probability for tracing, which are similar to the trends of the red solid curves in Figure 2 (b), (d).

Table 1: Asymptotic relative efficiency (10) of estimators (optimal versus SRS tracing design) under tracing selection models in (M1) and (M2); with $P_2 = 0.25$ and $P_3 = 0.1$, $\lambda_{230}/\lambda_{130} = 1.5$, $\beta_{12} = \log 1.5$, $P(\Delta = 0) = 0.8$

		Size Constraint (π^a)			Cost Constraint (ξ^b)		
Estimand	Tracing Model	0.05	0.25	0.50	5	20	100
β_{12}	M1	0.908	0.803	0.838	0.836	0.852	0.951
	M2	0.881	0.787	0.811	0.795	0.800	0.918
$\log(\lambda_{23}/\lambda_{13})$	M1	0.620	0.571	0.721	0.664	0.487	0.418
	M2	0.543	0.570	0.719	0.664	0.487	0.418

[a] $\pi = P(R = 1 \mid \Delta = 0)$ is the marginal probability of selection for tracing

[b] $\xi = C_d/C_s$ is the relative cost of determining disease status compared to survival status

3 Design with a Budgetary Constraint

3.1 Formulation of the Optimization Problem

In general, the cost associated with tracing individuals known to be diseased before loss-to-follow-up (i.e. those with $Z_C = 2$) is lower than that for individuals without the disease (i.e. $Z_C = 1$); in this section, we exploit this fact to design optimal tracing schemes subject to more realistic budget constraints. For the former group, the only information that we can learn is the survival status at the time of tracing, but for the latter group, disease status may also be ascertained for individuals who are still alive at tracing. Let C_s and C_d denote the cost for tracing survival status and disease status, respectively, and let $\xi = C_d/C_s$ be the cost ratio; we assume $\xi \geq 1$ in general.

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

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

Note that the right side of this equation depends on the parameter ρ in the tracing selection model, whereas the expected number of individuals eligible for tracing, $nP(\Delta = 0)$, and the distribution of observed data among the eligible individuals, $P(\mathcal{D}|\Delta = 0)$, are known after collection of phase I data. In addition, under the assumption (4) the probability $P(Z_J|R, \mathcal{D}, \Delta = 0; \theta)$ can be estimated by $P(Z_J \mid \bar{Z}_C, X; \tilde{\theta})$ where $\tilde{\theta}$ is the MLE obtained from phase I. This implies that if one is interested in precise estimation of θ_k, for a given fixed total budget B, cost C_s and ratio ξ, we can optimize the selection model by minimizing

$$\big[I_H^{-1}(\tilde{\theta}, \rho)\big]_{kk} + \zeta\big[\dot{C}(\rho; C_s, \xi) - B\big] \tag{11}$$

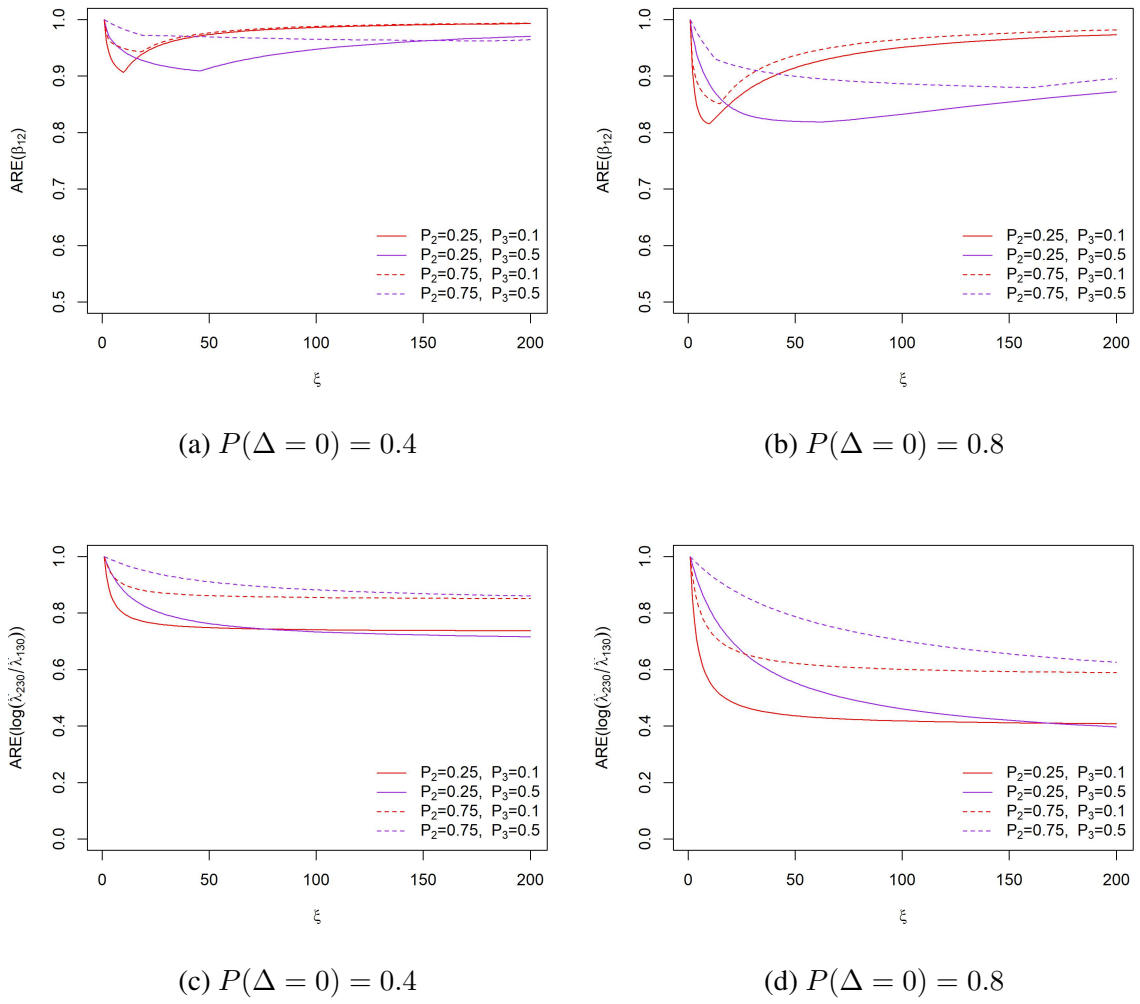
which is like (8) but with a cost constraint in place of a constraint simply on the expected sample size.

3.2 Efficiency Gains from Optimal Tracing with Cost Constraints

The study setting here parallels that of Section 2.4, with the exception that the constraint is imposed on the budget rather than the size of the sample selected for tracing. We set the maximum budget $B = \dot{C}(\rho; C_s, \xi = 1)$ to equal the expected cost of tracing all eligible individuals when $\xi = 1$. The budget constraint in (11) then becomes

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

which only depends on the cost ratio ξ and the selection parameter ρ. We consider values of ξ from 1 to 200 and the same values of $(\theta', \gamma')'$ as in Section 2.4, where $\theta = (\lambda', \beta')'$.



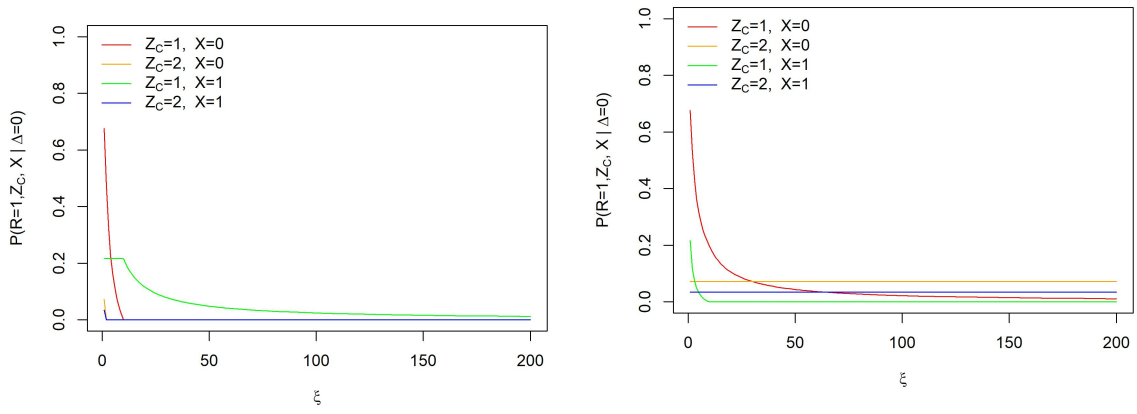
(a) $P(\Delta = 0) = 0.4$ (b) $P(\Delta = 0) = 0.8$

(c) $P(\Delta = 0) = 0.4$ (d) $P(\Delta = 0) = 0.8$

Figure 4: Asymptotic relative efficiency (10) of estimators for biomarker effect $\widehat{\beta}_{12}$ (top panels) and $\log(\widehat{\lambda}_{230}/\widehat{\lambda}_{130})$ (bottom panels) with a tracing study under an optimal design versus a SRS design of the same expected cost; $P_2 = P(N_{12}(\tau) = 1 \mid X = 0)$, $P_3 = P(Z(\tau) = 3 \mid X = 0)$, $\lambda_{230}/\lambda_{130} = 1.5$, $\beta_{12} = \log 1.5$; cost ratio $\xi = C_d/C_s$ is the relative cost of determining disease status compared to survival status

Figure 4 displays the patterns of relative efficiency exhibited by the optimal tracing selection

probabilities under a cost constraint under the selection model (M1), which are similar to those observed under the size constraint in the previous section (see Figure 2). In fact, in some sense this cost constraint amounts to a transformation of the size constraint. That is, due to the choice of budget constraint B, setting $\xi = 1$ implies that all eligible individuals may be traced (e.g. $\pi = P(R = 1|\Delta = 0) = 1$); thus, the left-most points in each panel of Figure 4 correspond to the right-most points in the analogous panels of Figure 2. On the other hand, as $\xi \to \infty$, the cost of tracing individuals with $Z_C = 1$ becomes prohibitively expensive, and $\lim_{\xi\to\infty} P(R = 1|\Delta = 0, Z_C = 1) = 0$. Thus, if individuals with $Z_C = 1$ furnish more informations upon tracing, as is the case for β_{12} (see Figure 5 (a)), then $\lim_{\xi\to\infty} P(R = 1|\Delta = 0) = 0$. On the other hand, if the optimal scheme prioritizes tracing individuals with $Z_C = 2$, as is the case for $\log(\lambda_{230}/\lambda_{130})$ (see Figure 5 (b)) , $\lim_{\xi\to\infty} P(R = 1|\Delta = 0) = P(Z_C = 2|\Delta = 0)$. Although the plots in Figures 4 and 5 only extend to $\xi = 200$, the limits in the relative efficiency gain and the percentage of eligible individuals that can be traced under a fixed budget are apparent.



(a) Optimal Design for Estimation of β_{12}

(b) Optimal Design for Estimation of $\log(\lambda_{230}/\lambda_{130})$

Figure 5: Optimal tracing design under a fixed budget constrain for the estimation of β_{12} (left panel) and $\log(\lambda_{230}/\lambda_{130})$ (right panel), with $P_2 = 0.25$ and $P_3 = 0.1$, $\lambda_{230}/\lambda_{130} = 1.5$, $\beta_{12} = \log 1.5$, $P(\Delta = 0) = 0.8$.

To compare the two selection models (M2) and (M1), the right-hand columns of Table 1 contain the asymptotic relative efficiencies under both models with the cost constraints. We see that using either of these two models results in appreciable efficiency gains, where the gain decreases as the cost ratio ξ increases from 5 to 100 for β_{12} but it increases as ξ increases for $\log(\lambda_{23}/\lambda_{13})$. These are consistent with the red solid curves showed in Figure 4 (b), (d). We also see the efficiency gains under (M2) are greater than those under (M1) for the estimation of β_{12} in most cases and they are very similar for $\log(\lambda_{230}/\lambda_{130})$. This is because for the latter, the optimal tracing scheme prioritizes tracing individuals with $Z_C = 2$, and the optimal selection probability of those is 1 under the settings considered here. As the cost for tracing disease status becomes more expensive (e.g. ξ increases), the optimal selection probability for individuals with $Z_C = 1$ quickly approaches 0. As a consequence the time from loss-to-follow-up to tracing has very little room to influence the selection probabilities under the optimal design, leading to comparable results under the two selection models.

4 APPLICATION TO UNIVERSITY OF TORONTO PSORIATIC ARTHRITIS COHORT STUDY

Scientists at the University of Toronto Psoriatic Arthritis Clinic have created and maintained a registry of individuals with psoriatic arthritis which continues to be an invaluable resource in deepening understanding of the progression of psoriatic arthritis and related comorbidities. A scientific question of primary interest is in estimating the incidence of arthritis mutilans in individuals with psoriatic arthritis, and estimating the effect of the marker HLA-B27 on the disease progression taking into account the full disease process including death.

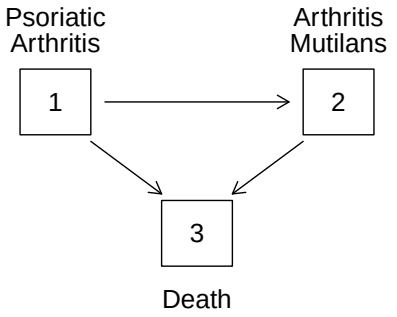


Figure 6: Multistate diagram for the onset of arthritis mutilans and death in individuals with psoriatic arthritis

The cohort we focused on consists of 870 individuals with psoriatic arthritis and they are scheduled to come to the clinic for assessments on an annual basis. We take December 2016 as the end of phase I follow-up, and use the patients records until then as phase I data. While variability arises in practice, this protocol informs the decision to view individuals who have not been seen for 2+ years as being lost to follow-up; this leads to 72% of the cohort being eligible for tracing. In total, 152 (17.5%) are observed to develop arthritis mutilans, 147 deaths are recorded (16.9%), including 36 among individuals whose disease progressed. Further, 56 individuals (6.4%) are positive for the HLA-B27 marker. Phase-I maximum likelihood estimates were obtained using the `R` package `msm` (Jackson, 2011). We assume the visit times are uninformative and that data are missing sequentially at random. The proposed approach is applied to demonstrate possible optimal designs for a tracing study conducted in January 2017.

Table 2 reports the optimal tracing probabilities $P(R = 1|Z_C, X, \Delta = 0)$ arising from selection model (M1) under the constraint of a fixed sample size or cost, respectively. It is apparent that if interest lies in estimating β_{12} one should first select all individuals who were not observed to progress before they withdrew from the study in phase I (i.e. $Z_C = 1$) and then individuals who have progressed (i.e. $Z_C = 2$), as long as the fixed sample size permits; this trend also holds when the budget is constrained. On the other hand, when interest lies in estimating $\log(\lambda_{230}/\lambda_{130})$, it always prioritizes individuals known to have progressed (e.g. with $Z_C = 2$) under both the sample size and budget constraints, since only survival status, which is less expensive, needs to be determined. We also considered using tracing selection model (M2), which leads to very similar gains in efficiency as when using model (M1). We did not report the optimal tracing probability here as they vary continuously with respect to time since loss-to-follow-up, $D = \tau - V_C$. For the psoriatic arthritis cohort, the proposed optimal tracing study design can lead to gains in efficiency of 10-30% relative to using a SRS design.

Table 2: Optimal selection probabilities by strata based on model (M1) for tracing psoriatic arthritis/mutilans cohort

		Size Constraint						Cost Constraint				
		Strata (Z_C, X)						Strata (Z_C, X)				
	π^a	$(1,0)$	$(1,1)$	$(2,0)$	$(2,1)$	RE^b	ξ^c	$(1,0)$	$(1,1)$	$(2,0)$	$(2,1)$	RE(%)
β_{12}	5%	0	0.95	0	0	71.1	3	0.54	1	0	0	85.4
	25%	0.25	1	0	0	76.2	5	0.34	1	0	0	80.1
	50%	0.57	1	0	0	84.0						
$\log(\lambda_{23}/\lambda_{13})$	5%	0	0	0.37	0	89.7	3	0.55	0	1	1	87.3
	25%	0.12	0	1	1	80.5	5	0.37	0	1	1	82.4
	50%	0.43	0	1	1	87.1						
Stratum size		495	33	84	14			495	33	84	14	

[a] $\pi = P(R = 1|\Delta = 0)$ is the marginal probability of selection for tracing

[b] RE is the relative efficiency of adopting the proposed tracing design as opposed to SRS (%)

[c] $\xi = C_d/C_s$ is the relative cost of determining disease status compared to survival status

5 Discussion

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

We have assumed a time-homogeneous Markov model with regularly scheduled assessment times. Tracing studies can of course be designed for non-homogeneous Markov models using piecewise-constant baseline intensities. The regularity of scheduled assessments makes it reasonably straightforward to determine which individuals are lost to follow-up and therefore eligible for tracing. In settings where assessment times are less regular and left to the discretion of patients, it is more challenging to define the set of individuals who are lost to follow-up and eligible for tracing. One can discretize time in such settings and declare individuals not seen in several potential periods as lost to follow-up. We also assumed a progressive disease process in which all states can only be entered

once. It is well-known that transition intensities involving recurring states are more poorly estimated under panel observation schemes (Lange and Minin, 2013; Ma et al., 2016). Moreover, when assessments are far apart in time (relative to dynamic features of the process of interest) estimates of transition intensities are less efficient compared to when the assessments are closer in time; the effect of widely spaced assessment times is smaller on other features such as state occupancy probabilities. These issues pertain to the conduct of tracing studies for non-progressive processes so one should expect a smaller gain in efficiency from "optimally" tracing individuals for in reversible processes.

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

ACKNOWLEDGEMENTS

This research was supported by an Alexander Graham Bell Canada Graduate Scholarship to N. Moon, Discovery Grants from the Natural Science and Engineering Research Council of Canada to L. Zeng (RGPIN 115928) and R. J. Cook (RGPIN 155849) and from the Canadian Institutes for Health Research to R. J. Cook (FRN 13887). R. J. Cook is a Canada Research Chair in Statistical Methods for Health Research. The authors would like to thank Drs. Dafna Gladman, Vinod Chandran and Lihi Eder of the Centre for Prognosis Studies in Rheumatic Diseases for helpful discussions.

REFERENCES

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

Control, D., Group, C. T. R., et al. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med*, 1993(329):977–986.

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

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. CRC Press, Boca Raton, FL.

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

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

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

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

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

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

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

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

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

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

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

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

Mehtälä, J., Auranen, K., and Kulathinal, S. (2011). Optimal designs for epidemiologic longitudinal studies with binary outcomes. *Statistical Methods in Medical Research*, 24(6):803–818.

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

Riboli, E. (1992). Nutrition and cancer: background and rationale of the European prospective investigation into cancer and nutrition (epic). *Annals of Oncology*, 3(10):783–791.

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

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