Published in final edited form as:
Biostatistics & Epidemiology (2019), **3**(1): 178–200
DOI: 10.1080/24709360.2019.1699341

Cohort study design for illness-death processes with disease status under intermittent observation

NATHALIE C. MOON

Department of Statistical Sciences,
University of Toronto, Toronto, ON, M5G 1X6, Canada
E-mail: nmoon@uwaterloo.ca

LEILEI ZENG

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Cohort studies are routinely conducted to learn about the incidence or progression rates of chronic diseases. The illness-death model offers a natural framework for joint consideration of non-fatal events in the semi-competing risks setting. We consider the design of prospective cohort studies where the goal is to estimate the effect of a marker on the risk of a non-fatal event which is subject to interval-censoring due to an intermittent observation scheme. The sample size is shown to depend on the effect of interest, the number of assessments, and the duration of follow-up. Minimum-cost designs are also developed to account for the different costs of recruitment and follow-up examination. We also consider the setting where the event status of individuals is observed subject to misclassification; the consequent need to increase the sample size to account for this error is illustrated through asymptotic calculations.

Keywords: censored data, illness-death model, intermittent assessment, sample size, study design

1 Introduction

Longitudinal cohort studies are commonly used to study the progression of individuals through stages of a chronic disease. Multistate models offer an appealing framework for modeling such disease processes (Andersen and Keiding, 2002) but the infeasibility of continuous monitoring of individuals means that disease status is typically only determined at intermittent assessment times; the resulting data are referred to as panel data. Kalbfleisch and Lawless (1985) developed a Fisher-scoring method for fitting multistate Markov models with time homogeneous or piecewise-constant transition intensities to panel data. In some settings transition times into some states may be observed subject only to right censoring; Jackson (2011) shows how to fit multistate models when processes are under a hybrid panel and continuous observation scheme and these methods are implemented in the R packages `msm`. Missing scheduled assessments and time trends in transition intensities often arise in longitudinal cohort studies of chronic disease processes, and while these pose challenges for analysis, methods have been proposed to accommodate missingness (both ignorable and non-ignorable) for non-homogeneous Markov processes through use of a time transformation (Kalbfleisch and Lawless, 1985; Chen and Zhou, 2011).

The cost of conducting a longitudinal cohort study is often appreciable, in great part due to the cost of repeatedly assessing individuals, often via in-person examination by a physician and/or expensive clinical tests; as such, there is great interest in the design of longitudinal studies which allocate resources most efficiently (Moskowitz et al., 2017; Timmons and Preacher, 2015; Collins and Graham, 2002; Singer and Willett, 1991). Design considerations depend on the stochastic process generating the response of interest and the precise objectives of the study. Much work has been carried out on the design of studies where interest lies in detecting differences in the mean of a continuous outcome between two groups, where the outcome is measured intermittently over time (Galbraith et al., 2002; Diggle, 2002; Kirby et al., 1994). This work has focused on the optimal frequency and timing of assessments in relation to the expected trajectory of the response. In other contexts, the focus of longitudinal studies has been on time-to-event outcomes, and oftentimes these are interval-censored; in this case, the lag between assessments is of critical importance at the design stage and has been investigated by a number of authors (Cook, 2000; Lawless and Rad, 2015; Kim et al., 2016; Jóźwiak and Moerbeek, 2012). More generally, and as described above, longitudinal studies may also be used to monitor individuals as they progress through various stages of disease, and some work has been done to consider the impact of assessment frequency on the precision of resulting estimators in a few special cases; for example for a simple progressive disease process (Hwang and Brookmeyer, 2003) and for a two-state reversible process (Mehtälä et al., 2015).

In this paper, we develop design criteria for a longitudinal study for an illness-death process in which individuals are under intermittent observation according to a protocol. Although missingness is a common issue arising in longitudinal studies, we restrict attention to the case of complete assessments in order to evaluate the impact of the schedule of planned visits on study cost and the precision of estimates. We consider the case in which disease progression status is observed intermittently, but transitions into the death state are observed subject to right censoring. Such an observation scheme is routinely used when monitoring a cohort of individuals at high risk for the onset of a disease, or of diseased individuals where interest lies in monitoring them for the development of a complication from the disease.

The remainder of this paper is organized as follows. In Section 2, we consider the design of such cohort studies and consider the impact of the frequency of assessments, in relation to the disease incidence rate and the frequency of a covariate of interest. We consider both statistical power and the minimal cost design. In Section 3, we derive the form of the Fisher information matrix for longitudinal studies in which there is misclassification in the states recorded at inspection times and use this to evaluate the impact of misclassification on study design subject to cost constraints. Concluding remarks are made in Section 4.

2 Prospective Cohort Studies

2.1 Multistate Markov Models and Maximum Likelihood Estimation

The multistate model is a flexible and powerful framework for modeling a chronic disease process with multiple stages (Cook and Lawless, 2018). Let $\{Z(t), t > 0\}$ be a continuous time stochastic process with state space $\mathcal{S} = \{0, 1, 2\}$ as described in Figure 1, where state 0 is the initial state of the process, state 1 denotes the presence of some condition of interest (e.g. onset or progression of a disease), and state 2 is an absorbing state, typically representing death. Without loss of generality, the states 0, 1 and 2 are often referred to as "disease-free", "disease" and "death" states accordingly.

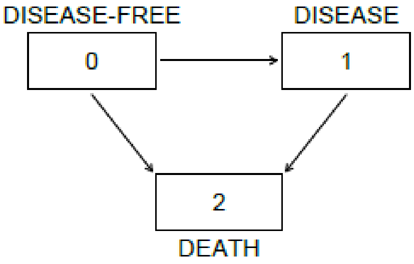


Figure 1: Multistate diagram for the three-state illness-death process

Let $\mathcal{H}(t) = \{Z(s), X; 0 \leq s \leq t\}$ be the history of the multistate process, and the intensity for $k \to \ell$ transitions is defined as

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

where $k < \ell \in \mathcal{S}$. Markov models are among the most commonly used types of multistate models due to the broad range of conditions it can represent, their simplicity and tractibility. Under such models, all dependence of transition intensities on the history of the process are encompassed in the current state, so we may write $\lambda_{k\ell}(t|\mathcal{H}(t^-)) = \lambda_{k\ell}(t)$. The transition probabilities $p_{k\ell}(s,t) = P(Z(t) = \ell | Z(s) = k)$ relate to intensities via the Kolmogorov forward differential equation

$$\frac{\partial}{\partial t}\mathbb{P}(s,t) = \mathbb{P}(s,t)\mathbb{A}(t) \qquad s < t \ , \tag{1}$$

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

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

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

$$\mathbb{P}(s,t) = \prod_{r=r_s}^{r_t} \mathbb{P}\big(\max\{s, b_{r-1}\}, \min\{t, b_r\}\big) = \prod_{r=r_s}^{r_t} \exp\big\{\big(\min\{t, b_r\} - \max\{s, b_{r-1}\}\big)\mathbb{A}_r\big\} \tag{2}$$

where the matrix exponential is used to obtain transition probabilities within each piece intersecting the interval of interest $[s,t]$. The effect of an explanatory variable X on the transition intensities is modeled using proportional intensities such that $\lambda_{k\ell}(t|X) = \lambda_{k\ell}(t)\exp(\beta_{k\ell}X)$, $k < \ell \in \mathcal{S}$.

Prospective cohort studies are commonly employed to collect data on life history processes. This involves acquiring a sample of size n from a population of individuals and tracking the occurrence of the event of interest (generally referred to as disease event) longitudinally over a certain follow-up period. It is generally infeasible to monitor individuals' disease status continuously, thus assessments are made intermittently at J specified time points $0 = a_0 < a_1 < \cdots < a_J = \tau$ over the study period $(0, \tau]$ although the exact event times are not available. On the other hand, the vital status is often tracked in continuous time and the exact death time is typically known or subject to right censoring if the participants become lost to follow-up at time C before the end of study. As such, the multistate data arising from longitudinal cohort studies may be mixed in its nature: disease status data may be available under a panel observation scheme, along with exact or right-censored death data. Let T_1 be the time to disease progression, T_2 be the time to death, $T^\dagger = \min(T_2, C, \tau)$ denote the minimum of the time to death and censoring and $\delta = I(T^\dagger = T_2)$ indicate that death is observed (see Figure 2). Assume $Z(a_0) = 0$ and let $\bar{Z}_j = (Z(a_1), \dots, Z(a_j))$ denote the history of the observed disease status up to and including assessment j, $j = 1, \dots, J$ and $M = \max\{j\,;\, a_j < T^\dagger, j = 0, \dots, J\}$ be the random number of assessments for an individual prior to right censoring or death. Under a Markov model indexed by the parameter vector θ in general, the likelihood contribution from a single subject is written as

$$\begin{aligned} L(\theta) &= P(\bar{Z}_m, t^\dagger, \delta \mid X) \\ &= \prod_{j=0}^{m-1} P\big(Z(a_{j+1}) \mid Z(a_j), X\big) \sum_{\ell=0}^{1} P\big(Z(t^\dagger) = \ell \mid Z(a_m), X\big)\lambda_{\ell 2}^{\delta}(t^\dagger \mid X) \end{aligned} \tag{3}$$

where the summation accounts for the fact that the disease status right before death or censoring may not be known due to the intermittent observation scheme. The estimates of θ can be obtained by maximizing the product of terms having the form of (3) over a sample of independent subjects. Instead of using a Newton-Raphson algorithm, a simple Fisher scoring method was proposed by Kalbfleisch and Lawless (1985) for obtaining the MLEs in which only first derivatives are required; this can be adapted to deal with observed times of death as shown by Zeng et al. (2018).

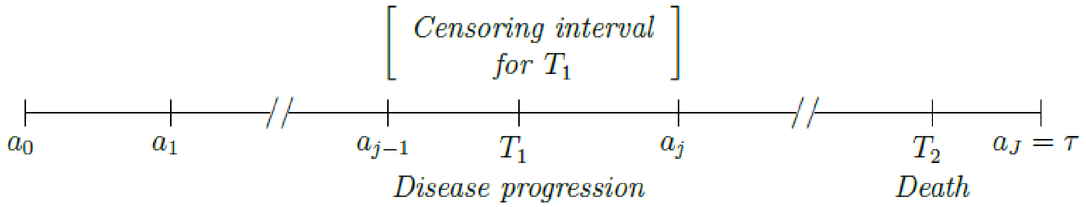


Figure 2: Schematic for mixed observation scheme, where the time of disease progression (T_1) is subject to interval-censoring and the time of death (T_2) is subject to right censoring.

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

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

The Fisher information matrix thus takes the form

$$\mathcal{I} = \sum_{x=0}^{1}\sum_{j=1}^{J+1}\int_{a_{j-1}}^{a_j}\sum_{q=1}^{j}\sum_{k=0}^{1} E\left[Y_k^\dagger(a_{q-1)}\frac{\partial \ell_{kj}}{\partial \theta}\frac{\partial \ell_{kj}}{\partial \theta'} \mid C = c, X = x\right] dG(c)P(X = x), \qquad (4)$$

where we suppose X is a binary explanatory variable, and let $G(\cdot;\rho)$ be the distribution function for censoring time C indexed by parameter ρ and $a_{J+1} = \infty$. The calculation details of the conditional expectation in the inner summation of (4) can be found in Zeng et al. (2018). Note that the construction of the Fisher information relies on transition probabilities and their first derivatives. Under the piecewise-constant model, the transition probability matrix can be obtained using (2), and its first derivatives can be taken in a straightforward manner. For the illness-death process, for example, suppose the derivatives are taken with respect to the vector of constant transition intensities associated with the rth piece $\mathcal{B}_r$, $\lambda^{(r)} = (\lambda_{01}^{(r)}, \lambda_{02}^{(r)}, \lambda_{12}^{(r)})'$, then we will simply have

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

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

2.2 Design Choices: Sample Size and Number of Assessments

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

Suppose the primary interest of a cohort study lies in the estimation of the effect of a covariate on the $0 \to 1$ transition (e.g. disease progression). We assume a binary covariate X has a multiplicative effect on the $0 \to 1$ transition, with intensity $\lambda_{01}(t \mid X) = \lambda_{01}(t)\exp(\beta X)$ under a Markov model. The estimator obtained from fitting the Markov models described in Section 2.1 has the following asymptotic distribution

$$\sqrt{n}(\widehat{\beta} - \beta) \sim N\big(0, \mathcal{I}^{\beta\beta}(\theta, \rho, J, \tau)\big)\ , \qquad (5)$$

where n is the sample size, and $\mathcal{I}^{\beta\beta}(\theta, \rho, J, \tau)$ is the (β, β) element of the inverse of the Fisher information $\mathcal{I}(\theta, \rho, J, \tau)$ given in (4). The asymptotic variance depends on θ and ρ from the disease and censoring processes respectively, as well as the number of assessments (J), the asessment times (a_j,

$j = 1, \ldots, J$), and the administrative censoring time (τ). The dependence on the actual assessment times is suppressed for convenience since we assume here that the assessment times are fixed and evenly scheduled over the interval $(0, \tau]$. It is straightforward to extend this work to irregular assessment times as long as the visit process is independent of the disease process. Response-dependent assessment processes and their impact on estimation and study design are further discussed in Section 5. Following the argument of Demidenko (2007), the power for a two-sided Wald-test of H_0: $\beta = \beta_0$ vs H_1: $\beta \neq \beta_0$ at a significance level of α_1 for detecting an effect of size $\beta = \beta_A$ is

$$\text{power}(\beta) = \Phi\left(-z_{\alpha_1/2} - \frac{\beta_0 - \beta_A}{\sqrt{\mathcal{I}_A^{\beta\beta}(\theta, \rho, J, \tau)/n}}\right) + \Phi\left(-z_{\alpha_1/2} + \frac{\beta_0 - \beta_A}{\sqrt{\mathcal{I}_A^{\beta\beta}(\theta, \rho, J, \tau)/n}}\right), \qquad (6)$$

where $\mathcal{I}_A^{\beta\beta}(\cdot)$ is evaluated at $\beta = \beta_A$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The power is a function of (i) the sample size (n), (ii) the number of evenly scheduled assessments (J), and (iii) the maximum duration of follow-up (τ). When τ is fixed and assessments are evenly scheduled, the study properties can be determined in terms of (i) sample size n and (ii) the frequency of assessments J for different desired levels of power; different pairs of design factors (n, J) may achieve the same power. Furthermore, if either n or J is fixed, the other can be solved by using (6). For example, if the number of assessments J is fixed, the required sample size for a Wald-test at significance level α_1 and power $1 - \alpha_2$ to detect an effect $\beta = \beta_A$ is

$$n = \left(\frac{z_{\alpha_1/2} + z_{\alpha_2}}{\beta_A}\right)^2 \mathcal{I}_A^{\beta\beta}(\theta, \rho, J, \tau). \qquad (7)$$

We provide empirical examples of the sample size calculation and relationship between power, sample size and the number of scheduled assessments for prospective cohort studies targeting the estimation of the effect of a binary covariate X on disease incidence. We assume all subjects are in state 0 (i.e. progression-free) at the time origin (i.e. $Z(0) = 0$), disease status is determined at J equally spaced assessments, and survival status is monitored continuously over the study period $(0, \tau]$ subject to random right censoring. Without loss of generality, we let $\tau = 1$. For simplicity, we consider a time-homogeneous disease process with transition intensities $\lambda_{k\ell}(t|X) = \lambda_{k\ell}\exp(X\beta_{k\ell})$ where λ_{01}, λ_{02}, and λ_{12} are baseline transition intensities and there is a covariate effect on disease progression denoted by $\beta_{01} = \beta$ but no covariate effects are assumed on death (i.e. $\beta_{02} = \beta_{12} = 0$). Let $\beta = \log 0.75$ indicate a covariate (biomarker) which is protective for disease progression, and $P(X = 1) = \{0.05, 0.25\}$. The values for parameters $(\lambda_{01}, \lambda_{02}, \lambda_{12})$ are set to satisfy the following constraints: (i) $P_1 = P(T_1 < \tau \mid X = 0) = \{0.10, 0.25, 0.50\}$, (ii) $P_2 = P(T_2 < \tau \mid X = 0) = \{0.10, 0.25, 0.50\}$, and (iii)$\lambda_{12}/\lambda_{02} = \{1.10\}$. We assume individuals may become lost to follow-up at a random time C which follows an exponential distribution with a rate ρ with the value of ρ set to satisfy $P(T_2 < \min(C, \tau)|X = 0) = \{0.05, 0.20\}$, where

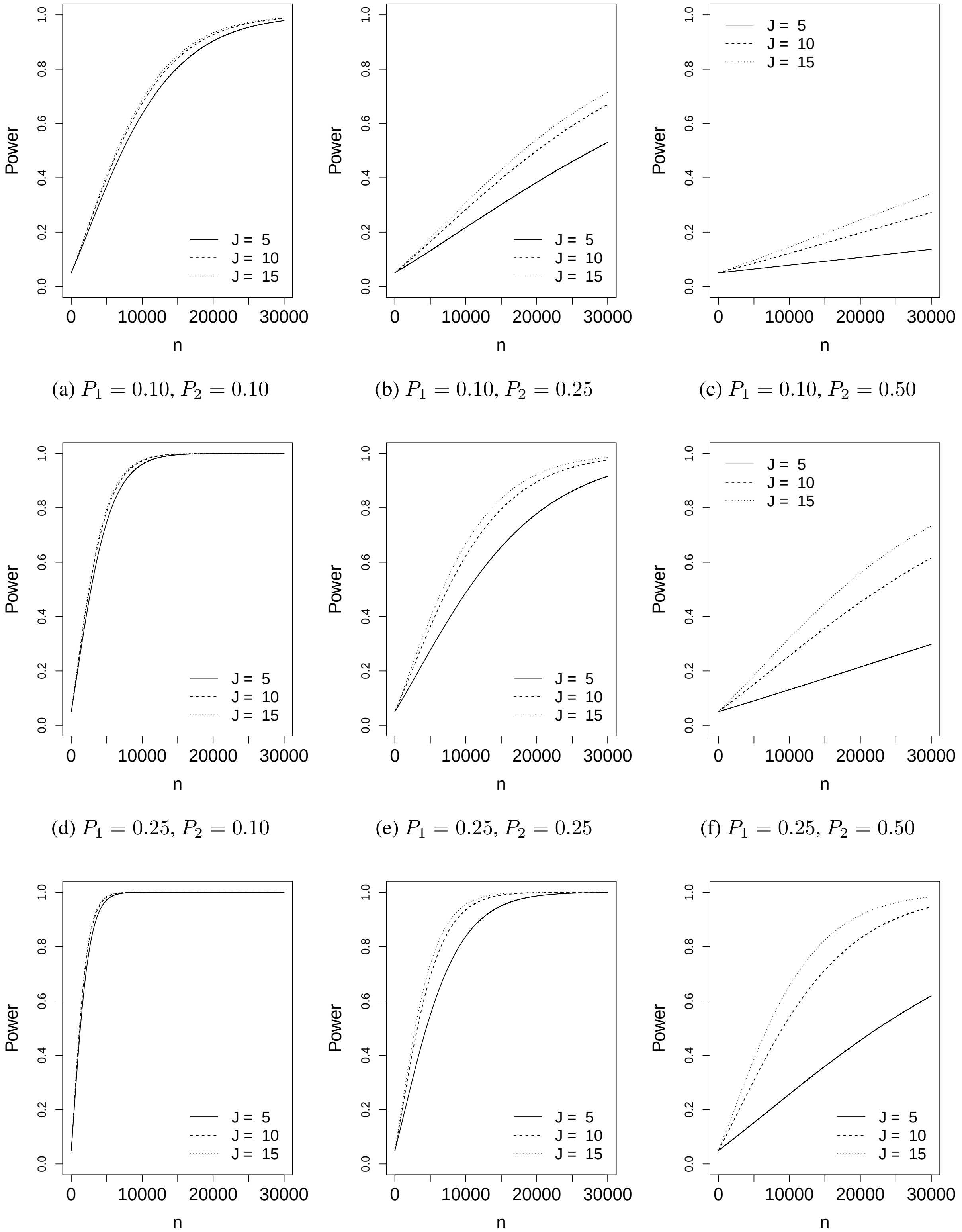
$$P(T_2 < \min(C, \tau)|X = 0) = P(T_2 < \tau|X = 0)(1 - G(\tau)) + \int_0^\tau P(T_2 < c|X = 0)g(c)dc\ .$$

In Table 1, we report the sample size n for testing H_0: $\beta = 0$ vs H_A: $\beta \neq 0$ calculated using formula (7), when the frequency of the assessments is fixed at $J = \{5, 10\}$, power at $\{80\%, 90\%\}$ and significance level $\alpha_1 = 0.05$. To validate these sample size calculations, for each scenario we simulate $2,000$ datasets as described in Cook and Lawless (2018), get point estimates $\hat{\beta}_{01}$ and their variance estimates using the `msm` package in R (Jackson, 2011), and report the empirical power. Sample sizes were calculated under the assumption that follow-up visits occur precisely at times $a_j = j\tau/J$ for all individuals at risk. Code for the computation of the Fisher information and calculation of the required sample size/number of assessments is available from the first author upon request, both for the settings with $\beta_{02} = \beta_{12} = 0$ and without these constraints. We simulated data under this assumption and calculated the empirical power ($EP^1\%$ in Table 1). Since this assumption is generally not plausible in practice, we conducted sensitivity analyses by simulating 2,000 datasets with modest variation in the visit times by adding a stochastic variation in the form of independent error terms with $\epsilon_j \sim N(0, \sigma_e)$ for visit j with $\sigma_e = \tau/(20J)$; we calculated the resulting empirical power and denote it by $EP^2\%$ in Table 1. The empirical power is close to the nominal level across all scenarios, although it can be slightly elevated when the covariate X is rare (i.e. when $P(X = 1) = 0.05$).

Table 1: Empirical power (EP%) for detecting an effect of covariate X on disease progression at the significance level $\alpha = 0.05$, when $\beta_{01} = \log 0.75$, $\lambda_{12}/\lambda_{02} = 1.1$, $P_2 = 0.25$, and $P(T_2 < \min(C, \tau)|X = 0) = 0.2$, based on $2,000$ simulated datasets of size n and sample sizes n are calculted as in (7); for EP^1%, the follow-up assessments occur exactly at times $a_j = j\tau/J$, while for EP^2% the follow-up assessments vary slightly from the scheduled times, with $a_j \sim \mathcal{N}(j\tau/J, \tau/(20J))$

			$P(X = 1) = 0.25$					$P(X = 1) = 0.05$				
Power	P_1	J	n	AVSE	ASE	EP^1%	EP^2%	n	AVSE	ASE	EP^1%	EP^2%
80%	0.10	5	8,442	0.1031	0.1027	81.3	81.9	35,025	0.1029	0.1027	82.0	84.1
		10	8,102	0.1030	0.1027	82.0	82.1	33,615	0.1030	0.1027	81.9	83.6
	0.25	5	3,290	0.1031	0.1027	82.0	82.9	13,601	0.1031	0.1027	81.0	82.2
		10	3,157	0.1031	0.1027	82.1	81.2	13,052	0.1032	0.1027	82.5	83.3
90%	0.10	5	11,301	0.0890	0.0887	90.8	91.7	46,888	0.0889	0.0887	91.3	92.7
		10	10,846	0.0890	0.0887	91.4	91.6	45,001	0.0889	0.0887	91.2	92.3
	0.25	5	4,405	0.0890	0.0887	91.7	91.4	18,208	0.0890	0.0887	92.1	92.2
		10	4,227	0.0890	0.0887	92.1	91.6	17,472	0.0890	0.0887	92.2	92.9

The figures in the remainder of this section are plots based on the asymptotic variance. Figure 3 displays power curves to illustrate the impact of features of the process (P_1 and P_2) and of the study design (n and J) on power; with $P_1 = \{0.10, 0.25, 0.50\}$ across rows, and $P_2 = \{0.10, 0.25, 0.50\}$ across columns. For all panels, we have $\beta = \log 0.75$. As before, the focus is on testing H_0: $\beta = 0$ vs H_A: $\beta \neq 0$ at a significance level of $\alpha = 0.05$. As expected, the power increases monotonically with the frequency of scheduled assessments over $[0, \tau]$. More generally, the power for detecting a covariate effect on progression is also higher when more precise information about disease progression is available, which, by comparing across the plots, can be seen to be driven by factors such as the proportion of disease progression events (P_1) and deaths (P_2) over $[0, \tau]$. As the proportion of deaths over $[0, \tau]$ increases, the number of realized clinical visits (M) decreases, and with it the extent of information on disease progression is reduced which leads to a large reduction in power; this can be seen by comparing across panels from left to right. For example, when $P_1 = P_2 = 0.10$, a sample of size $n = 15,000$ with $J = 5$ planned assessments over $(0, \tau]$ yields approximately 80% power for rejecting H_0: $\beta = 0$ vs H_A: $\beta \neq 0$, but if P_2 increases to 0.25 and 0.50, the power decreases substantially to 30% and 10% respectively. When interest lies in estimating the effect of a covariate on disease progression (β), we intuitively expect that an increase in the probability of progression should lead to an increase in power and these figures confirm this. When P_2 is fixed at 0.10, prospectively following a sample of $n = 5,000$ individuals for $J = 5$ planned assessments over $(0, \tau]$ leads to approximately 40% power when $P_1 = 0.10$, and this increases to 80% and 95% for $P_1 = 0.25$ and 0.50 respectively (comparing across rows in Figure 3).



(a) $P_1 = 0.10$, $P_2 = 0.10$ (b) $P_1 = 0.10$, $P_2 = 0.25$ (c) $P_1 = 0.10$, $P_2 = 0.50$

(d) $P_1 = 0.25$, $P_2 = 0.10$ (e) $P_1 = 0.25$, $P_2 = 0.25$ (f) $P_1 = 0.25$, $P_2 = 0.50$

(g) $P_1 = 0.50$, $P_2 = 0.10$ (h) $P_1 = 0.50$, $P_2 = 0.25$ (i) $P_1 = 0.50$, $P_2 = 0.50$

Figure 3: Plots of power curves for testing H_0: $\beta = 0$ vs H_A: $\beta \neq 0$, where $\beta = \log 0.75$ and type-I error rate is $\alpha_1 = 0.05$; across all panels, we have $\lambda_{12}/\lambda_{02} = 1.1$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$.

2.3 Cost-effective Design of Cohort Studies

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

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

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

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

$$\begin{aligned} P(M=j|X) \;\; = \;\; & \sum_{k=0}^{1} p_{0k}(0, a_j|X) \sum_{l=0}^{1} \Bigg\{ \int_{a_j}^{a_{j+1}} p_{kl}(a_j, t|X)\lambda_{l2}(t|X)dt \; (1 - G(a_{j+1})) \\ & + \int_{a_j}^{a_{j+1}} \Bigg[p_{kl}(a_j, c|X) \; + \; \int_{a_j}^{c} p_{kl}(a_j, t|X)\lambda_{l2}(t|X)dt \Bigg] dG(c) \Bigg\} \end{aligned}$$

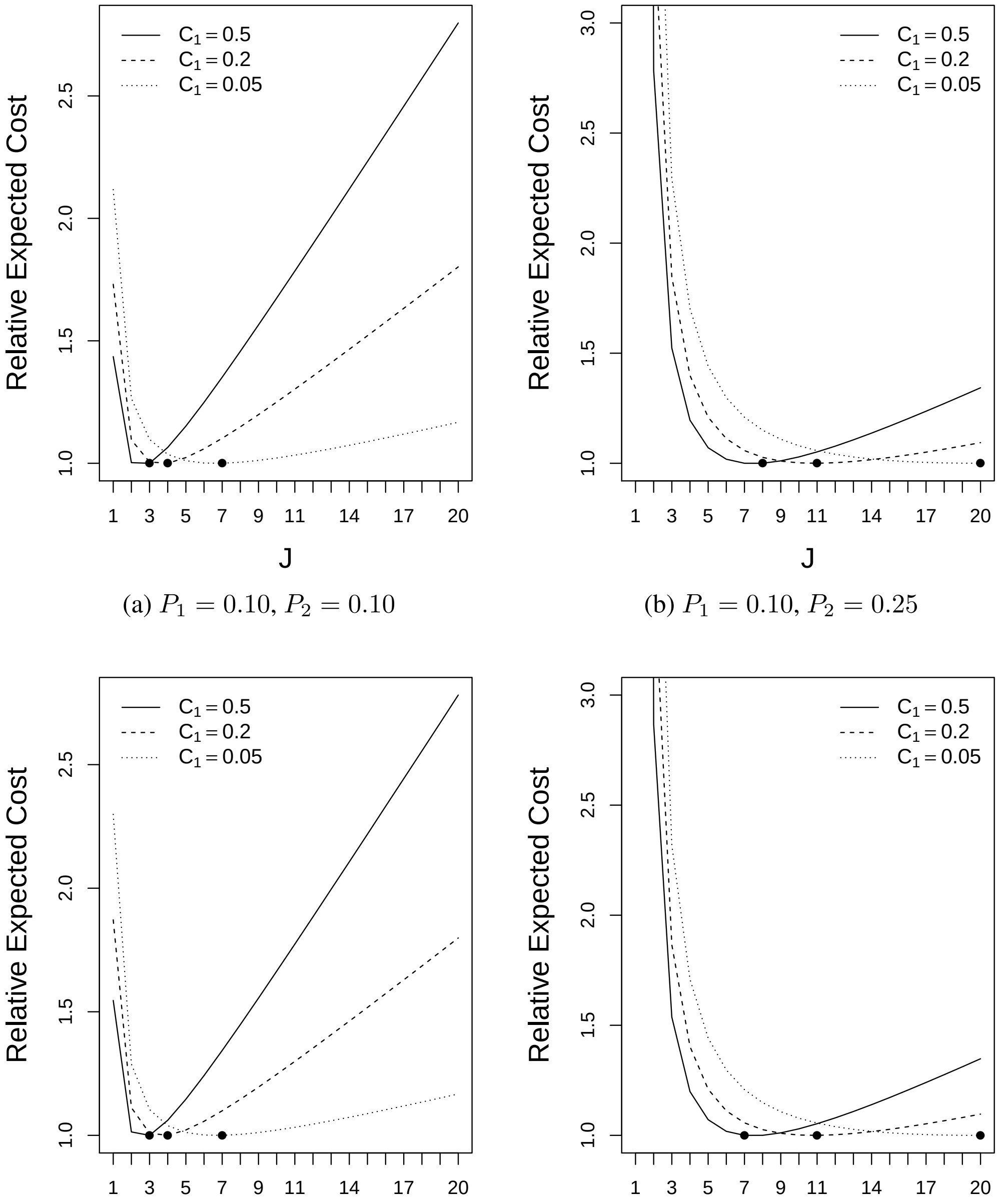
for $j < J$, and $P(M = J|X) = \sum_{l=0}^{1} p_{0l}(0, \tau|X)(1 - G(\tau))$.

A minimum-cost design is a design which minimizes expected total cost among all the designs (n, J) that achieve the same desired power to detect an effect of size $\beta = \log 0.75$. Figure 4 shows the relative expected cost of a design (n, J) versus the optimal one (n^{opt}, J^{opt}) represented by the dot on each line, when the power is fixed at 80%. The lines correspond to different values of cost ratio $C_1/C_0 = \{0.50, 0.20, 0.05\}$ and we set $C_0 = 1$ without loss of generality. As expected, the optimal frequency of assessments J^{opt} increases as the cost of conducting a follow-up assessment (C_1) decreases. As the probability of death over $[0, \tau]$ increases (comparing across columns in Figure 4), minimum-cost designs are achieved by scheduling more visits (e.g. increasing J^{opt}); this is sensible given that death terminates the observation process, and hence limits expected costs even when assessments are frequent. This observation is consistent with the power profile plots in Figure 3. On the other hand, the probability of progression (P_1) has little effect on the determination of the frequency of assessment J^{opt} in minimum-cost designs, as can be seen by comparing across rows in Figure 4. While Figure 3 demonstrated the large effect of P_1 on power for testing H_0: $\beta = 0$ vs H_A: $\beta \neq 0$, J^{opt} is far less sensitive to it. However, this does imply an increase in n^{opt} as P_1 decreases, which would in turn lead to an increase in expected study cost. Finally, note that the above discussion extends to any desired (fixed) level of power, as we can easily show that

$$\frac{n_{80}(J_1)}{n_{80}(J_2)} = \frac{n_{90}(J_1)}{n_{90}(J_2)},$$

where $n_p(J)$ is the sample size obtained from (7) to achieve $p\%$ power with J regular assessments over $(0, \tau]$. This implies that given the cost of follow-up assessments C_1, the value J^{opt} minimizing the expected total study cost does not change as a function of power.

Rather than identifying the most cost-effective design achieving a desired level of power, this approach may also be used to identify the design (n, J) which is most powerful among all designs



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

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

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

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

Figure 4: Ratio of expected study cost (to achieve 80% power for testing H_0: $\beta = 0$ vs H_A: $\beta \neq 0$ at a significance level of $\alpha = 0.05$ when $\beta = \log 0.75$) given J and the expected cost of the minimum-cost design; minimum-cost designs identified by dots for $C_1/C_0 = \{0.5, 0.2, 0.05\}$, $\lambda_{12}/\lambda_{02} = 1.1$, and $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$

satisfying a budgetary constraint. This is illustrated in Figure 5, where the budget is set at 3,500 times the cost of recruiting one individual, and the cost of each follow-up assessment is 20% of this (i.e. $C_0 = 1$ and $C_1 = 0.2$); the scenario considered here is the same as in the bottom-left panel of Figure 4. As the cost of each follow-up assessment increases, we see that the number of follow-up assessments maximizing power subject to the budgetary constraint decreases, which mirrors the trend observed in the minimum-cost designs, as shown in Figure 4.

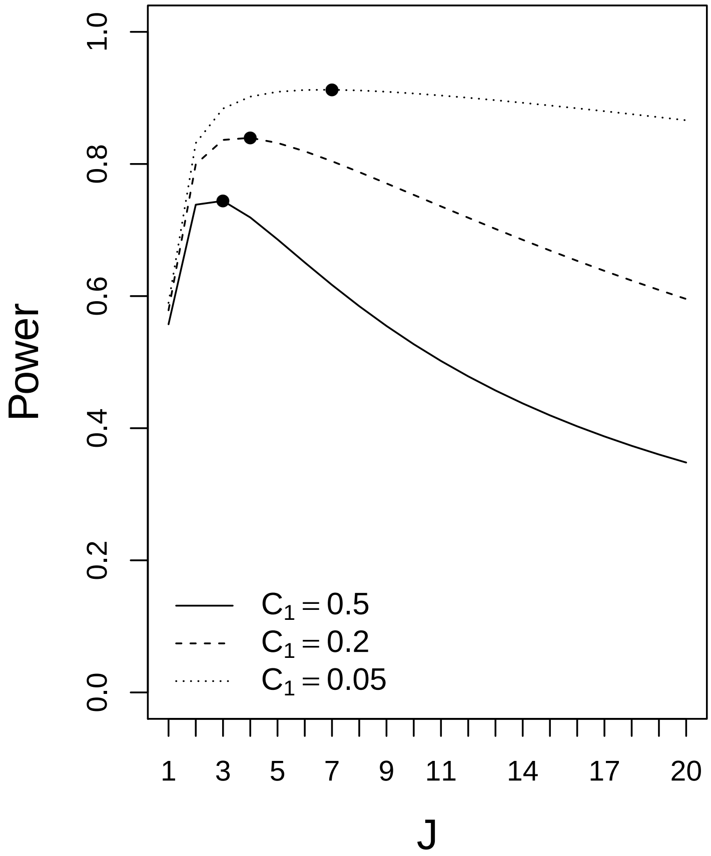


Figure 5: Power of designs subject to a budgetary constraint on the expected study cost, with the most powerful design identified by a dot for each of $C_1/C_0 = \{0.5, 0.2, 0.05\}$; $P_1 = 0.50$, $P_2 = 0.10$, $\lambda_{12}/\lambda_{02} = 1.1$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$

3 Accommodating Misclassification of Disease Status

3.1 Likelihood and EM Algorithm

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

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

$$L_o \propto \sum_{\bar{Z}_m} P(\bar{Z}_m, t^\dagger, \delta \mid X) P(\bar{W}_m \mid \bar{Z}_m, t^\dagger, \delta, X). \tag{8}$$

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

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

where $\mathcal{A}_j = [a_{j-1}, a_j)$ is the jth intermittent observation interval as before but an extra interval is defined as $\mathcal{A}_{m+1} = [a_m, \infty)$. For the misclassification process, we assume $W(a_j)$ depends only on the current true state $Z(a_j)$ but not on the classification history or the true disease status in the past and future, thus the second term in (8) becomes

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

where $\pi = (\pi_0, \pi_1)$ with $\pi_1 = P(W(a_j) = 1|Z(a_j) = 1)$, $\pi_0 = P(W(a_j) = 0|Z(a_j) = 0)$; the misclassification rates $FP = 1 - \pi_0$ and $FN = 1 - \pi_1$ are often assumed to be known (Ma et al., 2016). Given the above, the observed likelihood (8) can be expressed as

$$L_o(\theta, \pi) = \sum_{j=1}^{m+1} P(t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\dagger, \delta | X; \theta) P(\bar{W}_m | t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\delta, X; \pi). \tag{9}$$

For the estimation of disease process parameters θ, direct maximization of the observed likelihood (9) is difficult in general. An EM algorithm can alternatively be employed by casting the problem into a missing-data framework where the time of entry into state 1, T_1, is viewed as missing. We define the complete log-likelihood as

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

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

$$E\left[\ell_c(\theta) \mid \mathcal{D}; \widehat{\theta}^{(r)}\right] = \sum_{i=1}^{n} \int_0^\infty \log \left[f(t_{i1}, t_i^\dagger, \delta_i \mid X_i; \theta) \right] f(t_{i1} | \bar{W}_{M_i}, t_i^\dagger, \delta_i, X_i; \widehat{\theta}^{(r)}, \pi)\, dt_{i1} \ , \tag{10}$$

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

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

The M-step then requires maximizing the conditional expectation in (10) to get updated estimates of θ, and the iteration between the E- and M-steps continues until convergence. Variance estimation for the estimates $\widehat{\theta}$ from the EM algorithm is done by calculating the observed information via Louis (1982). The details of an EM algorithm procedure for the estimation of a time-homogeneous three-state model with observed disease status subject to misclassification are provided in Appendix A.

3.2 Fisher Information and Design

Obtaining the Fisher information matrix with misclassified disease status requires taking derivatives of the logarithm of the observed likelihood $L_o(\bar{W}_M, T^\dagger, \delta, X)$ given in (8). Let $\ell_o = \log L_o(\bar{W}_M, T^\dagger, \delta, X)$ and $S_o = \partial \ell_o / \partial \theta$ be the vector of first-order derivatives. In general, the form of the observed-data score S_o may be complicated due to taking the logarithm of a sum. It is helpful to write it as an expectation of the complete-data score given the observed data,

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

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

$$\begin{aligned} \mathcal{I}(\theta) &= \sum_{x=0}^{1} \sum_{j=0}^{J} \int_{a_j}^{a_{j+1}} E\left[S_o S_o' \Big| C = c \in \mathcal{A}_j, X = x \right] dG(c) P(X = x) \qquad (11) \\ &= \sum_{x=0}^{1} \sum_{j=0}^{J} \int_{a_j}^{a_{j+1}} \left[H(j, \min(c, \tau), 0, x) + \sum_{q=1}^{j} \int_{a_{q-1}}^{\min(a_q, c)} H(q, t_2, 1, x) dt_2 \right] dG(c) P(X = x), \end{aligned}$$

where

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

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

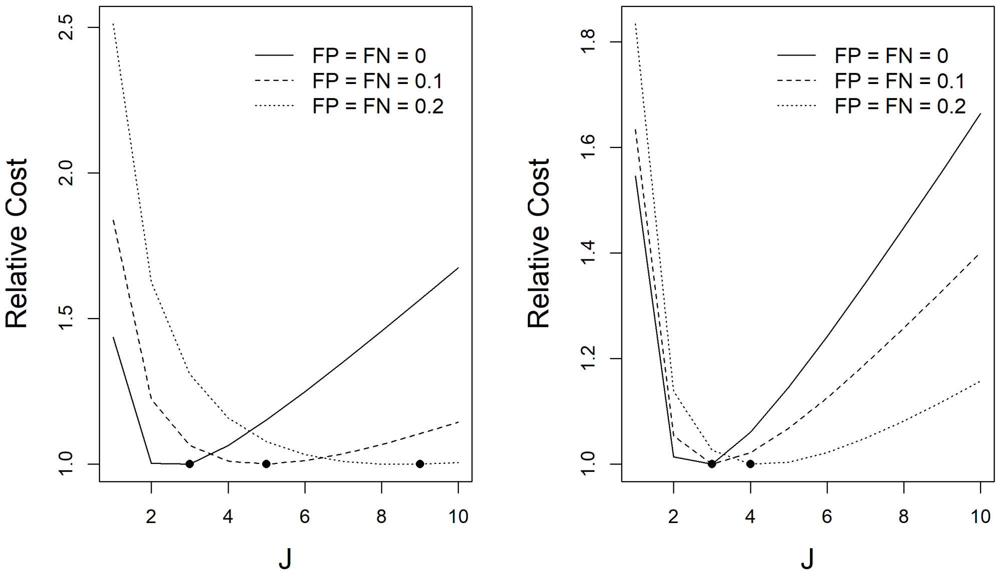
We validate the asymptotic variance obtained from the Fisher information (11) by comparing it to the empirical variance and average estimated variance of MLEs obtained via the EM algorithm for each of $2,000$ simulated datasets. These results are reported in Table 2; the parameter settings mirror those in Section 2.2. Note the excellent agreement between the empirical standard error and the asymptotic standard error, as well as the coverage achieving the nominal level of 95%, even in presence of slight and moderate misclassification.

Intuitively, it is clear that the scheduling of more frequent assessments mitigates, to some degree, the loss of information due to potential state misclassification. However, when considering both the cost of increasing the sample size (C_0) and the cost of follow-up assessments (C_1), it is not obvious whether it would be more cost-effective to increase n or J to achieve a desired level of power. In Figure 6, we see that as the degree of misclassification increases, the minimum-cost design is achieved by increasing the frequency of assessments over $[0, \tau]$. This is particularly apparent when disease progression is rare in the cohort (i.e. when P_1 is low), in which case even a modest rate of false positives/negatives has a significant impact on J^{opt}.

Finally, we consider the differential impact of false positive and false negative errors on features of the minimum-cost design (see Figure 7). For example, when P_1 is low (that is when progression events are rare in the cohort) and interest lies in detecting a covariate effect on disease progression, an increase in the rate of false positives (FP = $1 - \pi_0$) has a much larger impact on J^{opt} than does an increase in the rate of false negatives (FN = $1 - \pi_1$).

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

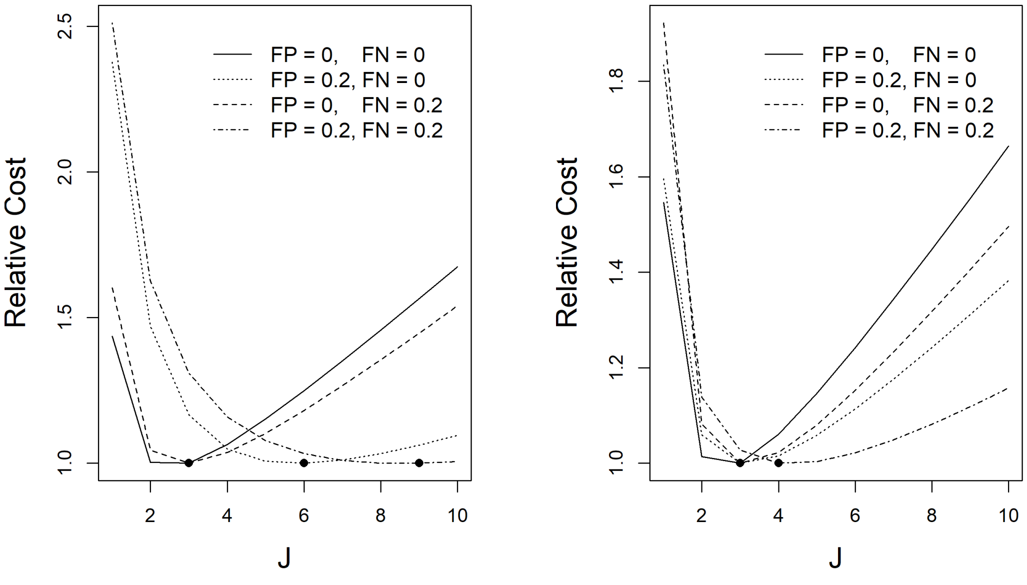
		$\log \lambda_{01}$ (-1.0928)				$\log \lambda_{02}$ (-1.2607)				$\log \lambda_{12}$ (-1.1654)				β (-0.28768)			
J	MIS	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
5	0	-0.004	0.059	0.060	94.9	3×10^{-4}	0.056	0.055	95.4	-0.005	0.156	0.154	95.4	-0.004	0.130	0.132	95.6
	0.10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



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

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

Figure 6: Ratio of expected study cost (to achieve 80% power for testing H_0: $\beta = 0$ vs H_A: $\beta \neq 0$ at a significance level of $\alpha = 0.05$ when $\beta = \log 0.75$) given J and the expected cost of the minimum-cost design; minimum-cost designs identified by dots for different degrees of misclassification where we assume equal false positive (FP) and false negative (FN) rates; $C_1/C_0 = 0.5$, $\lambda_{12}/\lambda_{02} = 1.1$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$



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

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

Figure 7: Ratio of expected study cost (to achieve 80% power for testing H_0: $\beta = 0$ vs H_A: $\beta \neq 0$ at a significance level of $\alpha = 0.05$ when $\beta = \log 0.75$ given J and the expected cost of the minimum-cost design; minimum-cost designs identified by dots for various combinations of false positive rate (FP= $1 - \pi_0$) and false negative rate (FN= $1 - \pi_1$); $C_1/C_0 = 0.5$, $\lambda_{12}/\lambda_{02} = 1.1$, $P(T_2 < \min(C, \tau)|X = 0) = 0.05$, and $P(X = 1) = 0.25$

4 Illustration Motivated by the Canadian Longitudinal Study on Aging

Here we consider an illustrative calculation motivated by the kinds of questions investigators involved in large cohort studies find of critical scientific interest. The Canadian Longitudinal Study on Aging (CLSA) (Raina et al., 2018) is a large longitudinal cohort study with the objective of better understanding various facets of aging in Canada. Approximately 50,000 Canadians aged 45-85 have been recruited into the CLSA, with planned follow-up assessments every 3 years over a period of 20 years. Data from the CLSA will allow investigators to study the effect of several genetic and environmental risk factors on the incidence and progression of several diseases such as diabetes and dementia. To reflect the design of the CLSA, we consider a study with $J = 7$ (corresponding to visits every 3 years over a period of 21 years) and calculate the sample size that would be required to achieve 80% and 90% power to detect a covariate effect β. Given the target age range of individuals at recruitment, we set $P_2 = P(T_2 < 21|X = 0) = 0.35$ and assume $P(T_2 < \min(C, 21)|X = 0) = 0.2$. We set $P_1 = P(T_1 < 21|X = 0) = \{0.14, 0.02\}$ to reflect the prevalence of diabetes and dementia respectively in the population (Ma et al., 2016), and $\lambda_{12}/\lambda_{02} = 1.1$ as both diabetes and dementia are associated with slight increases in the mortality rate. Finally, we consider $P(X = 1) = \{0.05, 0.25\}$ to reflect the prevalence of potential risk factors (e.g. genetic and environmental). In Table 3, we see that sample sizes of less than 30,000 are required to achieve 90% power to detect a covariate effect β when the disease and the covariate of interest are both relatively common (e.g. $P_1 = 0.14$ and $P(X = 1) = 0.25$), even when the ascertainment of disease status is subject to 20% misclassification. However, we see that when the disease is less common (e.g. $P_1 = 0.02$) or the covariate of X less frequent (e.g. $P(X = 1) = 0.05$), the sample sizes required to achieve 80% and 90% power become much larger, even when disease status may be determined without error. These findings are consistent with those of Ma et al. (2016).

Table 3: Sample sizes required to achieve 80% and 90% power to detect a covariate effect β with $J = 7$, where MIS=$1 - \pi$ is the misclassification proportion, $P_2 = 0.35$, $\lambda_{12}/\lambda_{02} = 1.1$, and $P(T_2 < \min(C, 21)|X = 0) = 0.2$

			Power = 0.80			Power = 0.90		
$P_1^{\dagger}$	$P(X = 1)$	β	MIS=0	MIS=0.1	MIS = 0.2	MIS=0	MIS=0.1	MIS = 0.2
0.14	0.25	$\log 0.75$	8,734	13,850	22,251	11,692	18,542	29,788
		$\log 0.6$	3,295	5,385	8,881	4,411	7,208	11,889
		$\log 0.5$	2,075	3,482	5,883	2,778	4,661	7,876
	0.05	$\log 0.75$	36,215	57,821	93,401	48,481	77,406	125,036
		$\log 0.6$	14,111	23,293	38,734	18,890	31,183	51,854
		$\log 0.5$	9,090	15,437	26,329	12,168	20,665	35,248
0.02	0.25	$\log 0.75$	62,971	140,990	317,798	84,300	188,746	425,441
		$\log 0.6$	23,938	55,661	131,126	32,046	74,514	175,541
		$\log 0.5$	15,155	36,428	89,291	20,288	48,767	119,536
	0.05	$\log 0.75$	261,729	590,992	1,244,226	350,381	791,171	1,799,541
		$\log 0.6$	102,839	242,123	577,870	137,671	324,134	773,606
		$\log 0.5$	66,624	162,464	404,134	89,190	217,493	541,021

[†] $P_1 = 0.14$ is specified to correspond roughly to the onset of diabetes; and $P_1 = 0.02$ is specified for the onset of dementia in such a cohort study.

5 Discussion

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

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

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

More generally, longitudinal cohort studies are designed under idealized assumptions which may not always be realized in practice; participants may not adhere to the visit schedule specifed in the protocol, some may not take treatments as directed, and responses may be misreported or missing. While some departures from design assumptions are unavoidable, it is important to capture as many realistic features of the study as possible at the design stage. In longitudinal studies where responses are collected on an individual repeatedly over time, it is reasonable to expect that the schedule of assessments and/or the treatment protocol may change over time, and this is an area that warrants future work. For example, while we assumed the assessments were scheduled at regular intervals over $(0, \tau]$, this need not be the case. The asymptotic variance can be calculated with (11) with unequally spaced assessment times, as long as their adaptive scheduled is described in the protocol. This allows for the assessment schedule to be optimized at the design stage; this may be of particular interest if transition intensities are vary with time, for example via piecewise constant intensities, and scheduling more frequent assessments for participants in high-risk periods may lead to more cost-effective designs.

Further generalizations of this design work could be considered to accommodate departures from Markov models. Models involving shared or correlated random effects acting multiplicatively on state transition intensities have been developed by Satten (1999), Cook et al. (2004), Chen and Zhou (2013), and O'Keeffe et al. (2013). Such models yield information matrices of the sort considered here, conditional on random effects, so the design would need to be modified to accommodate this additional source of variation; Louis' method (Louis, 1982) could again be useful in this calculation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by an Alexandar Graham Bell Canada Graduate Scholarship and an Ontario 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). This work was conducted while R.J. Cook held a Canada Research Chair in Statistical Methods for Health Research.

Notes on contributors

Nathalie C Moon is an Assistant Professor, Teaching Stream in the department of Statistical Sciences at the University of Toronto

Leilei Zeng is an Associate Professor in the department of Statistics and Actuarial Science at the University of Waterloo.

Richard J Cook is a Professor in the department of Statistics and Actuarial Science at the University of Waterloo and a Faculty Research Chair.

References

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

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

Chen, B. and Zhou, X.-H. (2011). Non-homogeneous markov process models with informative observations with an application to alzheimer's disease. *Biometrical Journal*, 53(3):444–463.

Chen, B. and Zhou, X.-H. (2013). A correlated random effects model for non-homogeneous markov processes with nonignorable missingness. *Journal of Multivariate Analysis*, 117:1–13.

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

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

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

Cook, R. J., Yi, G. Y., Lee, K.-A., and Gladman, D. D. (2004). A conditional markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*, 60(2):436–443.

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

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

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

Diggle, P. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

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

Galbraith, S., Stat, M., and Marschner, I. C. (2002). Guidelines for the design of clinical trials with longitudinal outcomes. *Controlled Clinical Trials*, 23(3):257–273.

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.

Jóźwiak, K. and Moerbeek, M. (2012). Cost-effective designs for trials with discrete-time survival endpoints. *Computational Statistics & Data Analysis*, 56(6):2086–2096.

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

Kim, H.-Y., Williamson, J. M., and Lin, H.-M. (2016). Power and sample size calculations for interval-censored survival analysis. *Statistics in Medicine*, 35(8):1390–1400.

Kirby, A. J., Galai, N., and Muñoz, A. (1994). Sample size estimation using repeated measurements on biomarkers as outcomes. *Controlled Clinical Trials*, 15(3):165–172.

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

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. (2015). Optimal designs for epidemiologic longitudinal studies with binary outcomes. *Statistical Methods in Medical Research*, 24(6):803–818.

Moskowitz, A. L., Krull, J. L., Trickey, K. A., and Chorpita, B. F. (2017). Quality vs. quantity: assessing behavior change over time. *Journal of Psychopathology and Behavioral Assessment*, 39(3):1–20.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In Collins, L. M. and Horn, J. L., editors, *Best methods for the analysis of change: recent advances, unanswered questions, future directions*, chapter 6, pages 92–105. American Psychological Association.

O'Keeffe, A. G., Tom, B. D., and Farewell, V. T. (2013). Mixture distributions in multi-state modelling: some considerations in a study of psoriatic arthritis. *Statistics in Medicine*, 32(4):600–619.

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

Satten, G. A. (1999). Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics*, 55(4):1228–1231.

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

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

Yi, G. Y. (2016). *Statistical Analysis with Measurement Error or Misclassification*. Springer-Verlag New York.

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

Appendix A: An EM algorithm when disease status is misclassified

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

$$\begin{aligned}
&p_{00}(t \mid X) = \exp\{-(\lambda_{01}(X) + \lambda_{02}(X))(t)\}\ ,\\
&p_{11}(t \mid X) = \exp\{-\lambda_{12}(X)(t)\}\ ,\\
&p_{01}(t \mid X) = \frac{\lambda_{01}(X)}{\lambda_{01}(X) + \lambda_{02}(X) - \lambda_{12}(X)}\Big[\exp\{-t\lambda_{12}(X)\} - \exp\{-t(\lambda_{01}(X) + \lambda_{02}(X))\}\Big]\ .
\end{aligned}$$

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

$$\begin{aligned}
L_o(\theta, \pi) = {} & \frac{\lambda_{01}(X)\lambda_{12}(X)^{\delta}e^{-\lambda_{12}(X)t^\dagger}}{\lambda_{01}(X) + \lambda_{02}(X) - \lambda_{12}(X)}\sum_{j=1}^{m+1}\left[e^{-\lambda^*(X)a_{j-1}} - e^{-\lambda^*(X)\min\{a_j, t^\dagger\}}\right]\\
& \times P(\bar{W}_m \mid t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\dagger, \delta, X)\\
& + \lambda_{02}(X)^{\delta}e^{-(\lambda_{01}(X)+\lambda_{02}(X))t^\dagger}P(\bar{W}_m \mid t_1 \in \mathcal{A}_{m+1}, \bar{Z}_m, t^\dagger, \delta, X) \qquad \text{(A.1)}
\end{aligned}$$

and

$$P(\bar{W}_m \mid t_1 \in \mathcal{A}_j, \bar{Z}_m, t^\dagger, \delta, X) = \prod_{\ell=1}^{j-1}\pi_0^{1-W(a_\ell)}(1 - \pi_0)^{W(a_\ell)}\prod_{\ell=j}^{m}\pi_1^{W(a_\ell)}(1 - \pi_1)^{1-W(a_\ell)}\ , \qquad \text{(A.2)}$$

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

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

The E-step:

The joint distribution of the complete data $\{t_1, t^\dagger, \delta\}$ is

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

The complete log-likelihood then can be written as

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

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

$$\begin{aligned}Q(\theta; \widehat{\theta}^{(r)}) =& w_1^{(r)}\Big\{ \log \lambda_{01}(X) + \delta \log \lambda_{12}(X) - \lambda_{12}(X)t^\dagger\Big\} - w_2^{(r)}\Big\{\lambda_{01}(X) + \lambda_{02}(X) - \lambda_{12}(X)\Big\} \\ &+ \big[1 - w_1^{(r)}\big]\Big\{\delta \log \lambda_{02}(X) - \big[\lambda_{01}(X) + \lambda_{02}(X)\big]t^\dagger\Big\} \qquad \text{(A.3)}\end{aligned}$$

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

$$\begin{aligned}w_1(\theta, \pi) &= \int_0^{t^\dagger} f(t_1 \mid \mathcal{D}; \theta, \pi)dt_1 \\ &= 1 - \frac{\lambda_{02}(X)^{\delta} e^{-(\lambda_{01}(X)+\lambda_{02}(X))t^\dagger} P(\bar{W}_m|t_1 \in \mathcal{A}_{m+1}, t^\dagger, \delta, X)}{L_o(\theta, \pi)} \\ w_2(\theta, \pi) &= \int_0^{t^\dagger} t_1 f(t_1 \mid \mathcal{D}; \theta, \pi)dt_1 \\ &= \frac{\sum_{k=1}^{M+1} \int_{a_{k-1}}^{\min\{a_k, t^\dagger\}} t_1 f(t_1, t^\dagger, \delta|X; \theta)dt_1 P(\bar{W}_m|t_1 \in \mathcal{A}_k, t^\dagger, \delta, X)}{L_o(\theta, \pi)}\end{aligned}$$

where $P(\bar{W}_m|t_1 \in \mathcal{A}_k, t^\dagger, \delta, X)$ is given in $(A.2)$, $L_o(\theta, \pi)$ is given in $(A.1)$, and the integration $\int_a^b t_1 f(t_1, t^\dagger, \delta|X; \theta)dt_1$ takes following form

$$\frac{\lambda_{01}(X)\lambda_{12}(X)^{\delta} e^{-\lambda_{12}(X)t^\dagger}}{\lambda^*(X)^2}\Big\{\big[a\lambda^*(X) + 1\big]e^{-a\lambda^*(X)} - \big[b\lambda^*(X) + 1\big]e^{-b\lambda^*(X)}\Big\} ,$$

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

The M-step:

The updated estimates of θ are obtained by maximizing the conditional expectation $Q(\theta; \widehat{\theta}^{(r)})$. Note that under time-homogeneous model, the $Q(\theta; \widehat{\theta}^{(r)})$ function given in $(A.3)$ can be re-organized as

$$\begin{aligned} Q(\theta; \widehat{\theta}^{(r)}) = {} & w_1^{(r)} \left[\log \lambda_{01}(X) - v^{(r)} \lambda_{01}(X)\right] \;+\; \left(1 - w_1^{(r)}\right) \left[\delta \log \lambda_{02}(X) - t^\dagger \lambda_{02}(X)\right] \\ & +\; w_1^{(r)} \left[\delta \log \lambda_{12}(X) - \left(t^\dagger - v^{(r)}\right) \lambda_{12}(X)\right] \;+\; \left(1 - w_1^{(r)}\right) \left[-t^\dagger \lambda_{01}(X)\right] \\ & +\; w_1^{(r)} \left[-v_1^{(r)} \lambda_{02}(X)\right] \end{aligned} \tag{A.4}$$

Note that $Q(\theta; \widehat{\theta}^{(r)})$ function resembles the sum of weighted log-likelihood function of Poisson observations with offsets. This implies that the maximization can be done by generating a pseudo dataset and fitting log-linear Poisson models. More specifically for each subject i, we create five pseudo responses y_{ij} $(j = 1, \ldots, 5)$, and we assume $y_{ij} \sim$ Poisson$(\lambda_{ij} u_{ij})$ with a pseudo offset u_{ij} and a log-linear model for the rate $\log \lambda_{ij} = z'_{ij}\theta$ where z_{ij} is a vector of pseudo covariates associated with y_{ij}. The conditional expectation $(A.4)$ is then equivalent to the weighted log-likelihood of a pseudo dataset

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

where we let $\theta = (\log \lambda_{01}, \log \lambda_{02}, \log \lambda_{12}, \beta_{01}, \beta_{02}, \beta_{12})'$, and the values of weights (i.e. $\widehat{w}_{ij}$), responses (i.e. y_{ij}), covariates (i.e. z_{ij}) and offsets (i.e. u_{ij}) of this pseudo dataset are given in Table A.1

Table A.1: Pseudo-data for loglinear model

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

In other words, we can use the `glm` function in `R` to do the maximization by fitting a Poisson log-linear model on a pseudo dataset where each individual in the original sample will give rise to a number of 'pseudo-individuals' with their weights, responses, and associated covariates and offsets generated as described in Table A.1.

Observed Information:

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

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

Evaluating the above amounts to calculating $w_1(\theta) = P(T_1 < t^\dagger|\mathcal{D}; \hat{\theta}, \pi)$, $w_2(\theta) = E\big[T_1|t_1 < t^\dagger, \mathcal{D}; \hat{\theta}, \pi\big]$, and $w_3(\theta) = E\big[T_1^2|t_1 < t^\dagger, \mathcal{D}; \hat{\theta}, \pi\big]$ for each individual, where $w_1(\theta)$ and $w_2(\theta)$ are as in the E-step of the EM algorithm and

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

where

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