



Evaluation of the natural history of disease by combining incident and prevalent cohorts: application to the Nun Study

Daewoo Pak¹ · Jing Ning² · Richard J. Kryscio³ · Yu Shen² 

Received: 5 October 2022 / Accepted: 22 April 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

The Nun study is a well-known longitudinal epidemiology study of aging and dementia that recruited elderly nuns who were not yet diagnosed with dementia (i.e., incident cohort) and who had dementia prior to entry (i.e., prevalent cohort). In such a natural history of disease study, multistate modeling of the combined data from both incident and prevalent cohorts is desirable to improve the efficiency of inference. While important, the multistate modeling approaches for the combined data have been scarcely used in practice because prevalent samples do not provide the exact date of disease onset and do not represent the target population due to left-truncation. In this paper, we demonstrate how to adequately combine both incident and prevalent cohorts to examine risk factors for every possible transition in studying the natural history of dementia. We adapt a four-state nonhomogeneous Markov model to characterize all transitions between different clinical stages, including plausible reversible transitions. The estimating procedure using the combined data leads to efficiency gains for every transition compared to those from the incident cohort data only.

Keywords Combined cohort data · Incident cohort · Interval censoring · Left truncation · Multistate model · Prevalent cohort

✉ Yu Shen
yshen@mdanderson.org

¹ Division of Data Science, Yonsei University, Wonju, Korea

² Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

³ Department of Statistics, University of Kentucky, Lexington, KY, USA

1 Introduction

In the field of epidemiology, understanding the natural history of a chronic disease, which refers to the clinical course of a disease progress, is important for formulating prevention and control strategies for the disease. In dementia studies, the cognitive function of a senior is assessed periodically and summarized roughly as intact cognition, mild cognitive impairment (MCI), or dementia. The transitions among these states describe the natural history of the disease. An example is the Nun Study of aging and Alzheimer's disease (Nun Study), which monitored nuns aged 75 years or older at entry for future disease progression with approximately annual examination up to 10 years. In this study cohort, some individuals were already found to be in the dementia state in their screening tests, yielding two sub-cohorts: the incident cohort of subjects without dementia at entry and followed for the potential diagnosis of dementia and death, and the prevalent cohort of subjects who were diagnosed with dementia at entry and followed up to death.

Of particular interest is the study of risk factors for cognitive decline and the duration from dementia onset to death. In the analyses using incident cohort data from the Nun Study, Tyas et al. (2007) concluded that the presence of apolipoprotein E4 allele (ApoE4) and low education increased the risk of transition to MCI but not the risk of the transition from MCI to dementia or death. Meanwhile, Wei et al. (2014) showed that both risk factors were positively associated with the transitions to MCI and dementia. Although the incident cohort is the collection of random samples from the target population, it is possibly limited in its ability to observe sufficient events of death. This factor can be mitigated by including the prevalent cohort with more failure events. Analyzing the combined data from the incident and prevalent cohorts leads to more efficiency in estimating baseline intensity functions and evaluating risk factors. In this paper, we revisit the association of risk factors with the transitions to MCI, dementia, and death by using both incident and prevalent cohorts from the Nun Study.

Analyzing the combined incident and prevalent cohort data is complex, owing to some unique and challenging data features. First, cognitive functioning is often assessed by intermittent visits, resulting in interval censored times for transitions. Second, since reverting from MCI back to intact cognition is fairly common in the natural history of dementia, the exact disease trajectory for a subject is difficult to ascertain. Third, the patients sampled in the midst of dementia usually cannot provide the exact starting date of the dementia. Only the remaining period of the progression from dementia to death, the time from study enrollment to death or loss to follow-up, is observed from the prevalent cohort. Lastly, the subjects from the prevalent cohort are likely to have longer durations from dementia to death compared to the dementia patients in the incidence cohort, which causes selection bias, i.e., the left-truncation problem. Failure to address these theoretical challenges when combining incident and prevalent cohorts may result in biased inference on the natural history of the disease and its association with risk factors.

With a single survival endpoint of interest, a few works from the literature have shown the benefits of combining both incident and prevalent cohorts. For

example, Lee et al. (2019) proposed an efficient estimation procedure that uses both the incident and prevalent cohorts under a proportional mean residual life model with right-censored data. In other areas, Wolfson et al. (2019) suggested approaches for estimating the non-parametric estimators of the survival function with the combined data in myotonic dystrophy by assuming that the disease onset of a prevalent case was known to be within an interval. McVittie et al. (2020) constructed the parametric likelihood of the combined data to estimate hospital stay durations, subject to right-censoring and the assumption of stationarity. Although important, their methods are not applicable to describe the natural history of disease that is represented with transitions between multiple states and often observed with data subject to selection bias. Joint modeling of the combined data with multiple transitions has been proposed in the literature. For example, Kulathinal et al. (2020) showed a gain in efficiency by combining both the incident and prevalent cohorts under the multistate model framework. They postulated that the incidence of the event of interest could be retrospectively observed in the prevalent cohort. On the other hand, Saarela et al. (2009) and Gorfine et al. (2021) proposed estimation procedures to jointly model both prevalent and incident cases subject to right censoring when important information from the prevalent cases is uncertain. To the best of our knowledge, there is no prior work that addresses the topic of modeling interval-censored natural history data obtained from a combination of incident and prevalent cohorts, particularly when the onset of prevalent disease cannot be observed retrospectively.

We propose a multistate approach for analyzing interval-censored event history data that are observed from both incident and prevalent cohorts to examine the association of risk factors with the transitions that are involved in the natural history of the disease. The modelling of cognitive changes and death in dementia is illustrated with a nonhomogeneous multistate Markov model, allowing for the reversible transition between intact cognition and MCI and the transition from any cognitive state to death. Within a multistate modeling framework, we derive the distribution of dementia onset using event-history observations, and we use this to model the observable residual periods of prevalent samples in a general truncation structure. We present a likelihood-based estimation procedure for jointly modelling the combined data from incident and prevalent cohorts in Sect. 2. In Sect. 3, intensive simulation studies are conducted to investigate the finite sample performance of the estimation procedure and to show the efficiency gains of the estimators compared to those that are obtained using the incident cohort only. We use the proposed method to analyze the Nun Study in Sect. 4. Section 5 contains some concluding remarks.

2 Statistical methodologies

2.1 Notations and model

We denote the participant's health status at a given time t ($t \geq 0$) by a multistate process $X(t)$ that takes a finite number of the values for the states representing discrete clinical conditions or death. In dementia studies, the multistate processes are

typically characterized with four states, namely $X(t) \in \{0, 1, 2, 3\}$, where 0 denotes cognitively intact for age, 1 denotes cognitive impairment, 2 denotes clinical dementia, and 3 denotes death. The biologically plausible transitions between the four states are depicted in Fig. 1, where the transition between state 0 and state 1 is potentially reversible, the transition from state 0 to 2 is assumed to proceed through MCI, and state 3 is the absorbing state that can be reached from all other states.

Assume that $X(t)$ follows a conditional Markov multistate model with four states, characterized by the conditional transition probabilities,

$$p_{kl}(t_1, t_2 | \mathbf{Z}) = P\{X(t_2) = l | X(t_1) = k, \mathbf{Z}\},$$

where $k, l \in \{0, 1, 2, 3\}$, $t_1 \leq t_2$ and \mathbf{Z} is a vector of risk factors. Then, the conditional transition intensity from k to l , $q_{kl}(t | \mathbf{Z})$, is

$$q_{kl}(t | \mathbf{Z}) = \lim_{\Delta t \rightarrow 0^+} \frac{P(X(t^- + \Delta t) = l | X(t^-) = k, \mathbf{Z})}{\Delta t},$$

for $k \neq l$ and $q_{kk}(t | \mathbf{Z}) = -\sum_{l \neq k} q_{kl}(t | \mathbf{Z})$ (Cox and Miller 1965). Let $\mathbf{P}(t_1, t_2)$ denote the transition probability matrix whose (k, l) entry is $p_{kl}(t_1, t_2 | \mathbf{Z})$ and $\mathbf{Q}(t)$ denote the transition intensity matrix whose (k, l) entry is $q_{kl}(t | \mathbf{Z})$ and $k, l \in \{0, 1, 2, 3\}$. With the possible transitions shown in Fig. 1,

$$\mathbf{P}(t_1, t_2) = \begin{pmatrix} p_{00}(t_1, t_2 | \mathbf{Z}) & p_{01}(t_1, t_2 | \mathbf{Z}) & p_{02}(t_1, t_2 | \mathbf{Z}) & p_{03}(t_1, t_2 | \mathbf{Z}) \\ p_{10}(t_1, t_2 | \mathbf{Z}) & p_{11}(t_1, t_2 | \mathbf{Z}) & p_{12}(t_1, t_2 | \mathbf{Z}) & p_{13}(t_1, t_2 | \mathbf{Z}) \\ 0 & 0 & p_{22}(t_1, t_2 | \mathbf{Z}) & p_{23}(t_1, t_2 | \mathbf{Z}) \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and}$$

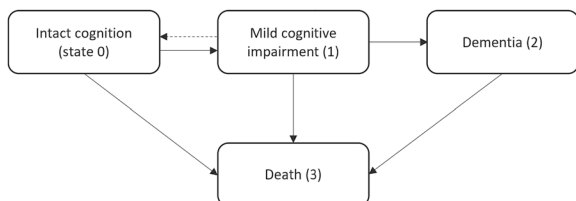
$$\mathbf{Q}(t) = \begin{pmatrix} q_{00}(t | \mathbf{Z}) & q_{01}(t | \mathbf{Z}) & 0 & q_{03}(t | \mathbf{Z}) \\ q_{10}(t | \mathbf{Z}) & q_{11}(t | \mathbf{Z}) & q_{12}(t | \mathbf{Z}) & q_{13}(t | \mathbf{Z}) \\ 0 & 0 & q_{22}(t | \mathbf{Z}) & q_{23}(t | \mathbf{Z}) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The transition probabilities for a continuous-time Markov process can be obtained by solving the following Kolmogorov forward equation,

$$\frac{d\mathbf{P}(t_0, t)}{dt} = \mathbf{P}(t_0, t)\mathbf{Q}(t), \tag{1}$$

where $\mathbf{P}(t_0, t_0) = \mathbf{I}$ is the identity matrix. For a nonhomogeneous Markov model, however, the solution to equation (1) is intractable with non-trivial forms of $\mathbf{Q}(t)$, and it is computationally intensive to get the solution with covariates (Titman 2011).

Fig. 1 Transition diagram with four states for dementia: state 0 for intact cognition, state 1 for mild cognitive impairment (MCI), state 2 for dementia and state 3 for death



In some situations, an analytic solution to equation (1) is available; the first example is a progressive process with a small number of states (Pak et al. 2017, 2019), and the second example is a time transformation model that assumes the transition intensity matrix is a form of $\mathbf{Q}(t) = \mathbf{Q}_0 dh(t)/dt$ with the operation time $h(t)$, where \mathbf{Q}_0 is the transition intensity matrix for a homogeneous Markov process and $h(t)$ is a non-negative function (Kalbfleisch and Lawless 1985; Omar et al. 1995; Hubbard et al. 2008). We adapt the time transformation model when constructing the likelihood function for event-history data.

We assume the multiplicative intensity model for each transition that is

$$q_{kl}(t|\mathbf{Z}) = q_{kl}^{(0)}(t) \exp(\mathbf{Z}'\boldsymbol{\beta}_{kl}), \quad k \neq l \tag{2}$$

where $q_{kl}^{(0)}(t)$ is the k -to- l baseline transition intensity and $\boldsymbol{\beta}_{kl}$ is the regression coefficient vector for the transition from state k to state l . We use the Weibull distribution to model the baseline transition intensities under the assumption of the same time dependency across transitions, namely, $q_{kl}^{(0)}(t) = \lambda_{kl}\gamma t^{\gamma-1}$ with unknown positive parameters, λ_{kl} and γ . Then, the transition intensity matrix $\mathbf{Q}(t)$ can be expressed as $\mathbf{Q}_0 dh(t)/dt$, where \mathbf{Q}_0 is the transition intensity matrix whose (k, l) entry is $\lambda_{kl} \exp(\mathbf{Z}'\boldsymbol{\beta}_{kl})$ and $dh(t)/dt = \gamma t^{\gamma-1}$. This implies that there exists a new time scale, $h(t)$, which leads to the homogeneous process with the transition intensity matrix \mathbf{Q}_0 (Hubbard et al. 2008). Thus, we have $\mathbf{P}(t_1, t_2) = \exp[\mathbf{Q}_0\{h(t_2) - h(t_1)\}]$, where $h(t) = \int_0^t \gamma s^{\gamma-1} ds$.

We introduce additional notation to model the prevalent samples. The prevalent cohort includes participants who are alive but with dementia at the time of recruitment. Let U denote individual's age at study enrollment. Further, we let A^o be the age of onset for clinical dementia and A^d be the age at death. Then, the prevalent cohort is formed by the individuals whose age at enrollment U is between A^o and A^d . Let $\tilde{W} = U - A^o$ denote the time from dementia onset to study enrollment and $\tilde{T} = A^d - A^o$ denote the time from dementia onset to death within a prevalent population of patients with dementia. Let (W, T) be a pair of (\tilde{W}, \tilde{T}) for the observed samples in the prevalent cohort among those who are eligible to be sampled at the time of recruitment (i.e., $\tilde{T} > \tilde{W}$). Note that (W, T) is not exactly observed. This is because the exact age at dementia onset is generally unavailable from the prevalent cohort. We let $T = W + V$, where V is the observed time from study recruitment to death and C denote a residual censoring time from study recruitment to loss of follow-up. The diagram in Fig. 2 depicts these notations for the prevalent sampling.

2.2 Likelihood and estimation

Consider a random sample of m independent subjects that are combined from both the incident and prevalent cohorts with respective sample sizes of m_1 and m_2 , where $m = m_1 + m_2$. Suppose that those in the incident cohort are labeled $1, \dots, m_1, m_1 < m$. Let $X_i(t)$ be the state of the health condition at time t for the i -th subject ($i = 1, \dots, m$). For the i -th subject, we let $s_i = (s_{i1}, \dots, s_{i n_i})'$ denote the states, consecutively observed by n_i observations with the corresponding time points $t_i = (t_{i1}, \dots, t_{i n_i})'$, and \mathbf{z}_i denote a

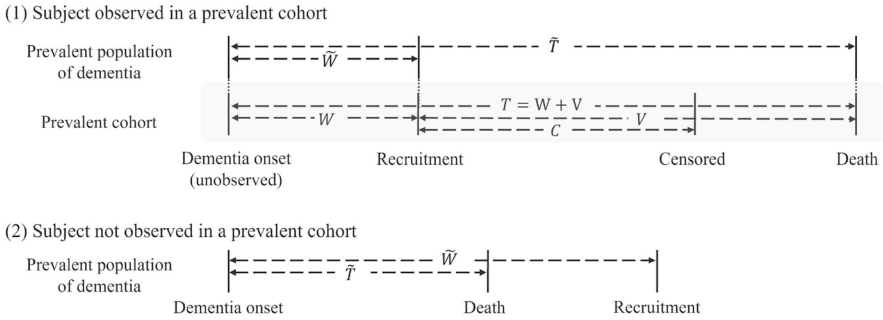


Fig. 2 The diagram for left truncation in a prevalent cohort study: Case (1) is for a subject sampled in the cohort, and Case (2) is for a subject not sampled in the cohort

vector of baseline risk factors. In the Nun Study, the time to death was observed exactly when a subject died during the study. Let $\delta_i = 1$ if the i -th subject died before the end of study and $\delta_i = 0$ otherwise. Then, $s_{in_i} = 3$ for $\delta_i = 1$ and $s_{in_i} \in (0, 1, 2)$ for $\delta_i = 0$.

The observed event history data from the incident cohort consist of $\mathcal{O}^{(inc)} \equiv (\mathcal{O}_1^{(inc)}, \dots, \mathcal{O}_{m_1}^{(inc)})$, where $\mathcal{O}_i^{(inc)} = (t_i, s_i, \delta_i, \mathbf{z}_i)$ and $i = 1 \dots, m_1$. We assume that the disease progression is a Markov process and $\mathcal{O}_i^{(inc)}$ is independent across subjects. Denote the vector of all model parameters by θ . The likelihood for the incident cohort of the i -th subject who was alive at the end of the study, denoted by $\mathcal{L}_i^{(inc)}(\theta; t_i, s_i, \delta_i = 0, \mathbf{z}_i)$, can be expressed as the product of the transition probabilities, which is

$$\mathcal{L}_i^{(inc)}(\theta; t_i, s_i, \delta_i = 0, \mathbf{z}_i) = P\{X(t_{i1}) = s_{i1} | \mathbf{z}_i\} \left\{ \prod_{r=2}^{n_i} P_{s_{i(r-1)}s_{ir}}(t_{i(r-1)}, t_{ir} | \mathbf{z}_i) \right\}, \quad (3)$$

where $P\{X(t_{i1}) = s_{i1} | \mathbf{z}_i\}$ can be approximated by $p_{0s_{i1}}(t_0, t_{i1} | \mathbf{z}_i)$ with a possible time point t_0 of being in intact cognition, such as the age at which the disease initiated.

In the Nun Study, a subject's cognitive state at the moment of death was unknown unless they were diagnosed with dementia before death. Multiple possible risks of death may exist after the last cognitive assessment of the subject. For example, if the last assessment of a subject was the MCI state, there were three possible risks of death: (1) cognitive function was improved and she died in the cognitive intact state, (2) dementia was not developed before death and she died in the MCI state, and (3) dementia was developed and she died in the dementia state. The likelihood for the i -th subject of the incident cohort who died during the study (i.e., the subject with $s_{n_i} = 3$ for $i = 1, \dots, m_1$) is then

$$\begin{aligned} \mathcal{L}_i^{(inc)}(\theta; t_i, s_i, \delta_i = 1, \mathbf{z}_i) = & P\{X(t_{i1}) = s_{i1} | \mathbf{z}_i\} \left\{ \prod_{r=2}^{n_i-1} P_{s_{i(r-1)}s_{ir}}(t_{i(r-1)}, t_{ir} | \mathbf{z}_i) \right\} \\ & \times \sum_{k \neq 3} P_{s_{i(n_i-1)k}(t_{i(n_i-1)}, t_{in_i}^- | \mathbf{z}_i) q_{k3}(t_{in_i} | \mathbf{z}_i). \end{aligned} \quad (4)$$

Therefore, the likelihood of the incident cohort data, $\mathcal{L}^{(inc)}$, is

$$\mathcal{L}^{(inc)}(\boldsymbol{\theta}; \mathcal{O}^{(inc)}) = \prod_{i=1}^{m_1} \mathcal{L}_i^{(inc)}(\boldsymbol{\theta}; \mathbf{t}_i, \mathbf{s}_i, \delta_i, \mathbf{z}_i). \tag{5}$$

Next, we construct the likelihood for the prevalent cohort after adjusting for left-truncation. Based on our multistate model, the probability density function of the age of dementia onset A^o given the covariates $\mathbf{Z} = \mathbf{z}$, denoted by $f_{A^o}(a|\mathbf{z})$, is

$$f_{A^o}(a|\mathbf{z}) = \frac{p_{01}(0, a|\mathbf{z})q_{12}(a|\mathbf{z})}{\int_0^\infty p_{01}(0, u|\mathbf{z})q_{12}(u|\mathbf{z})du}. \tag{6}$$

Then, the conditional density function of \tilde{T} given $A^o = a$ follows:

$$f_{\tilde{T}|A^o}(t|a, \mathbf{z}) = \frac{f_{23}(a + t|\mathbf{z})}{S_{23}(a|\mathbf{z})}, \tag{7}$$

where f_{23} and S_{23} are, respectively, the density function and the survival function for A^d given $\mathbf{Z} = \mathbf{z}$.

We assume that \tilde{T} is independent of \tilde{W} given A^o and $\mathbf{Z} = \mathbf{z}$. Note that the subjects from the prevalent cohort of the Nun Study are among those who have $\tilde{T} > \tilde{W}$ due to the prevalent sampling, and T is dependent of (W, V) due to their relationship in $T = W + V$. If a subject from the prevalent cohort dies during the study, the age of study enrollment U and the residual survival time V are observable. With (6) and (7), the joint density of (U, V) given $\mathbf{Z} = \mathbf{z}$ and the sampling constraint, denoted by $f(u, v|\mathbf{z})$, follows:

$$\begin{aligned} f(u, v|\mathbf{z}) &= P(U = u, V = v | \tilde{T} > \tilde{W}, \mathbf{Z} = \mathbf{z}) \\ &= \frac{\int_0^u P(\tilde{T} = u + v - a, \tilde{W} = u - a | A^o = a, \mathbf{Z} = \mathbf{z}) f_{A^o}(a|\mathbf{z}) da}{P(\tilde{T} > \tilde{W} | \mathbf{Z} = \mathbf{z})} \\ &= \frac{\int_0^u f_{\tilde{T}|A^o}(u + v - a|\mathbf{z}) g(u - a|\mathbf{z}) f_{A^o}(a|\mathbf{z}) da}{P(\tilde{T} > \tilde{W} | \mathbf{Z} = \mathbf{z})}, \end{aligned}$$

where $g(w|\mathbf{z})$ is the probability density function for \tilde{W} given $\mathbf{Z} = \mathbf{z}$, and

$$\begin{aligned} P(\tilde{T} > \tilde{W} | \mathbf{Z} = \mathbf{z}) &= \int_0^\infty P(\tilde{T} > \tilde{W}, A^o = a | \mathbf{Z} = \mathbf{z}) da \\ &= \int_0^\infty \int_0^\infty P(\tilde{T} > \tilde{W}, \tilde{W} = w | A^o = a, \mathbf{Z} = \mathbf{z}) f_{A^o}(a|\mathbf{z}) dadw \\ &= \int_0^\infty \int_0^\infty S_{\tilde{T}|A^o}(w|\mathbf{z}) g(w|\mathbf{z}) f_{A^o}(a|\mathbf{z}) dadw. \end{aligned}$$

A parametric family of distribution can be used for $g(w|\mathbf{z})$. Length-biased sampling can also be assumed by letting $g(w|\mathbf{z})$ be a uniform distribution over $(0, \tau)$, where τ is a constant that makes the probability mass $P(T = t | T > A) \rightarrow 0$ for $t > \tau$ (Shen et al. 2009).

Since the transition that is possible for a subject with dementia is the transition toward death, the information on study enrollment and last assessment suffices to describe the prevalent cohort data. Let u_i be the age at study enrollment and y_i be the observed time from study entry to the last assessment for the i -th subject. Then, the observed data for the i -th subject consist of $u_i, y_i = \min(V_i, C_i), \delta_i = I(V_i < C_i)$, and \mathbf{z}_i . We denote the prevalent cohort data as $\mathcal{O}^{(pre)} \equiv (\mathcal{O}_{m_1+1}^{(pre)}, \dots, \mathcal{O}_m^{(pre)})$, where $\mathcal{O}_i^{(pre)} = (u_i, y_i, \delta_i, \mathbf{z}_i)$ and $i = m_1 + 1, \dots, m$.

Assume that C is independent of $(\tilde{T}, \tilde{W}, V)$ given $\mathbf{Z} = \mathbf{z}$. Then, the likelihood for the prevalent cohort is proportional to

$$\mathcal{L}^{(pre)}(\boldsymbol{\theta}, \boldsymbol{\xi}; \mathcal{O}^{(pre)}) \propto \prod_{i=m_1+1}^{m_2} f(u_i, y_i | \mathbf{z}_i)^{\delta_i} S(u_i, y_i | \mathbf{z}_i)^{1-\delta_i}, \tag{8}$$

where $S(u, y | \mathbf{z}) = \int_y^\infty f(u, r | \mathbf{z}) dr$, $\boldsymbol{\theta}$ is a vector of the same model parameters for the target population of the incident cohort, and $\boldsymbol{\xi}$ is a vector of the parameters in $g(w | \mathbf{z})$.

The likelihood function for combined data from incident and prevalent cohorts can be expressed as the product of the likelihoods given in (5) and (8),

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\xi}; \mathcal{O}^{(inc)}, \mathcal{O}^{(pre)}) \propto \mathcal{L}^{(inc)}(\boldsymbol{\theta}; \mathcal{O}^{(inc)}) \times \mathcal{L}^{(pre)}(\boldsymbol{\theta}, \boldsymbol{\xi}; \mathcal{O}^{(pre)}). \tag{9}$$

We can estimate $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ by maximizing the logarithm of the likelihood in (9).

The likelihood function can readily be extended to handle the event history data with interval-censored events in the prevalent cohort. Assume that the residual period of a prevalent subject is only known to be within a specific interval, i.e., $V \in [L, R)$, where $L \geq U$ and $R = \infty$ for right-censoring. The likelihood for the prevalent cohort can be generalized to

$$\begin{aligned} \mathcal{L}^{(pre)}(\boldsymbol{\theta}, \boldsymbol{\xi}; \mathcal{O}^{(pre)}) \propto & \prod_{i=m_1+1}^{m_2} \{f(u_i, y_i | \mathbf{z}_i)^{\delta_i} S(u_i, y_i | \mathbf{z}_i)^{1-\delta_i}\}^{1-\delta_{li}} \\ & \times \{S(u_i, l_i | \mathbf{z}_i) - S(u_i, r_i | \mathbf{z}_i)\}^{\delta_{li}}, \end{aligned} \tag{10}$$

where (l_i, r_i) is the interval of (L, R) for the i -th subject, and $\delta_{li} = I(R_i < \infty)$. Thus, the likelihood function for the combined data with interval-censored prevalent samples is proportional to the product of the likelihoods given in (5) and (10).

The maximum likelihood implementation requires calculating integrals over finite or infinite intervals, which may not have analytic solutions. In such a case, we obtain a numerical integral value by using the Gauss-Jacobi quadrature after transforming an integral over the unit interval $(0, 1)$. The statistical inference about $\boldsymbol{\eta} \equiv (\boldsymbol{\theta}', \boldsymbol{\xi}')'$ can be performed with the asymptotic distribution of $\hat{\boldsymbol{\eta}}$, which we approximate by $N(\hat{\boldsymbol{\eta}}, I_{obs}^{-1}(\hat{\boldsymbol{\eta}}))$, where I_{obs} is the observed information matrix.

3 Simulation studies

We conducted a series of simulation studies to assess finite sample properties of the proposed likelihood-based approach. Two different sets of sample sizes for the incident cohort and the prevalent cohort were considered: (a) $m = 600$ ($m_1 = 300$ and $m_2 = 300$) and (b) $m = 1200$ ($m_1 = 600$ and $m_2 = 600$). We generated two independent covariates from $Z_1 \sim \text{Bernoulli}(0.4)$ and $Z_2 \sim \text{Uniform}(0, 1)$, i.e., $\mathbf{Z} = (Z_1, Z_2)'$. In data generation of the life-history process for a subject, the Weibull baseline transition intensities were used with the set of parameters: $\lambda_{01} = 0.5$, $\lambda_{03} = 0.2$, $\lambda_{10} = 0.4$, $\lambda_{12} = 0.3$, $\lambda_{13} = 0.4$, $\lambda_{23} = 0.3$, and $\gamma = 1.5$. The coefficients of \mathbf{Z} for transitions, denoted by $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{01}, \boldsymbol{\beta}'_{03}, \boldsymbol{\beta}'_{10}, \boldsymbol{\beta}'_{12}, \boldsymbol{\beta}'_{23})'$, were chosen to be $\boldsymbol{\beta}'_{01} = (0.2, -0.2)$, $\boldsymbol{\beta}'_{03} = (0.1, -0.2)$, $\boldsymbol{\beta}'_{10} = (-0.2, 0.2)$, $\boldsymbol{\beta}'_{12} = (0.2, -0.3)$, $\boldsymbol{\beta}'_{13} = (0.2, 0.2)$, and $\boldsymbol{\beta}'_{23} = (0.3, -0.2)$. Ten cognitive assessment times for each subject were simulated to generate interval censored data, which were set to $t_1 = 0.2 + \text{Uniform}(0, 0.4)$ and $t_j = t_{j-1} + \text{Uniform}(0.2, 0.4)$ for $j = 2, \dots, 10$. With these assessment rules, about 50% of subjects were still alive at the last assessment. For the prevalent cohort, the distribution of \tilde{W} was assumed to follow the Weibull distribution that is $g(w) = 0.6w^{0.2} \exp(-0.5w^{1.2})$, and the residual censoring time C was generated from $\text{Uniform}(0, c)$, where c was set to achieve two censoring percentages, 15% and 30%. Lastly, 1000 replicates were generated in each set of sample sizes for Monte Carlo simulations.

Tables 1 and 2 show the simulation results for estimating $\boldsymbol{\beta}$ from the combined cohort data with the different censoring rates of a prevalent cohort, along with the results of fitting a multistate model to the incident cohort data only. The mean estimates from all simulated cohort data are close to the true parameter values. The empirical standard errors of the estimates are almost the same as the mean of the estimated asymptotic standard errors, and they decrease by almost $\sqrt{2}$ when the sample size doubles. The coverage probability for every estimator is also close to the nominal level. We further calculated the relative efficiency (RE) of the two approaches, defined as the ratio of the mean square error of the estimator from the combined cohort to that from the incident cohort. As expected, the approach using the combined cohort data has the highest efficiency gain in estimating $\boldsymbol{\beta}_{23}$ with the range of REs being about from 7.4 to 11.7 across all scenarios. The censoring rate of the prevalent samples is negatively related to the RE of $\boldsymbol{\beta}_{23}$. The efficiency gains in estimating the baseline intensity for the 2-to-3 transition are also depicted in Fig. 3; the 95% pointwise confidence intervals of the baseline intensities from the combined cohort data are narrower than those of the incident cohort data. Although the prevalent samples provide the information on the 2-to-3 transition only, REs of the parameters for other transitions also tend to be over one. This implies that combining the incident and prevalent cohorts affects the estimation of overall transition probabilities.

The results for the baseline intensity parameters and the parameters in $g(w)$ are relegated to the Supplementary Information. They show similar trends to the regression parameters. The RE of λ_{23} ranges from 5.2 to 6.7 across all scenarios. A slight improvement in efficiency also was shown in the estimation of other

Table 1 Simulation results for estimating β with incident cohort data only ($m_1 = 300$) or with the combined cohort ($m_1 = 300$ and $m_2 = 300$) under different censoring rates for the prevalent cohort. The coefficient vector for k - l transition is denoted by $\beta_{kl} = (\beta_{kl}^{(1)}, \beta_{kl}^{(2)})'$, where $\beta_{kl}^{(1)}$ for Z_1 and $\beta_{kl}^{(2)}$ for Z_2

	$\beta_{01}^{(1)}$	$\beta_{01}^{(2)}$	$\beta_{03}^{(1)}$	$\beta_{03}^{(2)}$	$\beta_{10}^{(1)}$	$\beta_{10}^{(2)}$	$\beta_{12}^{(1)}$	$\beta_{12}^{(2)}$	$\beta_{13}^{(1)}$	$\beta_{13}^{(2)}$	$\beta_{23}^{(1)}$	$\beta_{23}^{(2)}$
Truth	0.20	-0.20	0.10	-0.20	-0.20	0.20	0.20	-0.30	0.20	0.20	0.30	-0.20
<i>Incident cohort only</i>												
Bias ^a	-0.01	-0.01	0.02	-0.02	-0.01	0.06	0.01	0.04	-0.01	-0.08	-0.01	-0.01
ESE ^b	0.15	0.67	0.27	1.04	0.33	1.32	0.28	1.22	0.27	1.09	0.36	1.62
SE ^c	0.15	0.65	0.25	1.05	0.32	1.31	0.27	1.17	0.26	1.10	0.36	1.57
CP ^d	0.95	0.95	0.96	0.96	0.95	0.95	0.96	0.94	0.95	0.96	0.95	0.95
<i>Combined cohorts with no censored prevalent samples</i>												
Bias	-0.01	-0.02	0.01	-0.02	-0.02	0.05	0.01	0.04	-0.01	-0.05	0.00	0.01
ESE	0.15	0.63	0.26	1.02	0.32	1.30	0.26	1.16	0.25	1.03	0.11	0.48
SE	0.15	0.62	0.25	1.02	0.32	1.29	0.26	1.11	0.24	1.03	0.11	0.46
CP	0.95	0.95	0.95	0.96	0.96	0.95	0.95	0.94	0.94	0.96	0.95	0.94
RE ^e	1.10	1.13	1.05	1.04	1.07	1.03	1.08	1.10	1.16	1.11	11.28	11.63
<i>Combined cohorts with 15% censored prevalent samples</i>												
Bias	0.01	-0.01	0.00	0.03	0.03	0.07	-0.01	-0.03	0.00	-0.05	-0.01	0.01
ESE	0.15	0.60	0.25	1.05	0.32	1.32	0.27	1.16	0.25	1.08	0.12	0.50
SE	0.15	0.62	0.25	1.03	0.32	1.30	0.26	1.11	0.24	1.03	0.12	0.49
CP	0.95	0.95	0.96	0.94	0.95	0.95	0.94	0.94	0.94	0.94	0.95	0.96
RE	1.07	1.24	1.14	0.99	1.05	1.00	1.04	1.10	1.16	1.02	9.25	10.50
<i>Combined cohorts with 30% censored prevalent samples</i>												
Bias	-0.01	0.02	0.01	0.03	0.01	0.08	-0.01	0.01	0.00	0.00	0.00	0.01
ESE	0.15	0.63	0.25	1.02	0.32	1.31	0.27	1.16	0.24	1.05	0.13	0.52
SE	0.15	0.62	0.25	1.03	0.32	1.29	0.26	1.11	0.24	1.03	0.12	0.53
CP	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.96
RE	1.11	1.12	1.13	1.03	1.06	1.01	1.08	1.11	1.29	1.08	8.12	9.86

^a the empirical bias

^b the empirical standard error of the parameter estimator

^c the average of the standard error estimator

^d the coverage of the 95% confidence interval with the Normal approximation

^e the relative efficiency obtained by the ratio of the mean-squared error from the combined cohort to that from the incident cohort

baseline intensity parameters, especially for γ , as their REs were greater than one. The parameters in $g(w)$ also were reasonably close to the true values with the coverage probabilities close to the nominal level.

We also performed a sensitivity analysis when the density of \tilde{W} was incorrectly specified under the current simulation settings (see Table S3 of the Supplementary Information). We considered two misspecified cases: \tilde{W} is assumed to follow (1) the exponential distribution and (2) the uniform distribution (i.e., the length-biased sampling). In summary, the parameter estimates from the proposed

Table 2 Simulation results for estimating β with incident cohort data only ($m_1 = 600$) or with the combined cohort ($m_1 = 600$ and $m_2 = 600$) under different censoring rates for the prevalent cohort. The coefficient vector for $k-l$ transition is denoted by $\beta_{kl} = (\beta_{kl}^{(1)}, \beta_{kl}^{(2)})'$, where $\beta_{kl}^{(1)}$ for Z_1 and $\beta_{kl}^{(2)}$ for Z_2

	$\beta_{01}^{(1)}$	$\beta_{01}^{(2)}$	$\beta_{03}^{(1)}$	$\beta_{03}^{(2)}$	$\beta_{10}^{(1)}$	$\beta_{10}^{(2)}$	$\beta_{12}^{(1)}$	$\beta_{12}^{(2)}$	$\beta_{13}^{(1)}$	$\beta_{13}^{(2)}$	$\beta_{23}^{(1)}$	$\beta_{23}^{(2)}$
Truth	0.20	-0.20	0.10	-0.20	-0.20	0.20	0.20	-0.30	0.20	0.20	0.30	-0.20
<i>Incident cohort only</i>												
Bias	-0.00	-0.02	0.00	0.01	0.00	-0.05	-0.01	0.00	-0.01	0.01	-0.01	-0.04
ESE	0.11	0.45	0.18	0.73	0.22	0.91	0.196	0.80	0.18	0.753	0.25	1.12
SE	0.11	0.45	0.18	0.73	0.22	0.91	0.189	0.81	0.18	0.768	0.25	1.06
CP	0.95	0.95	0.94	0.96	0.96	0.95	0.945	0.95	0.95	0.957	0.95	0.94
<i>Combined cohorts with no censored prevalent samples</i>												
Bias	-0.00	-0.01	0.01	0.01	0.00	-0.05	-0.00	0.006	-0.01	0.01	0.00	0.00
ESE	0.10	0.43	0.18	0.71	0.22	0.90	0.189	0.76	0.171	0.71	0.08	0.32
SE	0.10	0.43	0.17	0.72	0.22	0.90	0.181	0.77	0.167	0.72	0.08	0.32
CP	0.95	0.96	0.94	0.96	0.95	0.95	0.934	0.95	0.944	0.96	0.96	0.97
RE	1.09	1.10	1.04	1.05	1.01	1.022	1.08	1.103	1.12	1.13	10.57	11.71
<i>Combined cohorts with 15% censored prevalent samples</i>												
Bias	-0.01	-0.02	0.01	-0.01	-0.00	0.01	0.01	-0.010	-0.01	0.00	0.00	0.02
ESE	0.10	0.42	0.18	0.74	0.23	0.91	0.186	0.78	0.173	0.71	0.08	0.34
SE	0.10	0.43	0.17	0.71	0.22	0.90	0.181	0.78	0.167	0.72	0.08	0.35
CP	0.94	0.96	0.95	0.95	0.95	0.95	0.954	0.95	0.943	0.95	0.96	0.95
RE	1.10	1.12	1.07	1.01	0.97	1.01	1.10	1.069	1.11	1.13	9.30	10.19
<i>Combined cohorts with 30% censored prevalent samples</i>												
Bias	-0.00	-0.01	-0.01	-0.03	0.02	-0.03	-0.00	0.009	0.00	0.02	0.00	0.01
ESE	0.10	0.46	0.18	0.71	0.217	0.95	0.19	0.78	0.170	0.73	0.09	0.39
SE	0.10	0.44	0.18	0.72	0.220	0.90	0.18	0.78	0.168	0.72	0.09	0.37
CP	0.95	0.94	0.95	0.94	0.955	0.94	0.94	0.95	0.946	0.95	0.94	0.93
RE	1.12	1.03	1.04	1.05	1.009	0.98	1.10	1.08	1.12	1.11	7.44	8.23

^a the empirical bias

^b the empirical standard error of the parameter estimator

^c the average of the standard error estimator

^d the coverage of the 95% confidence interval with the Normal approximation

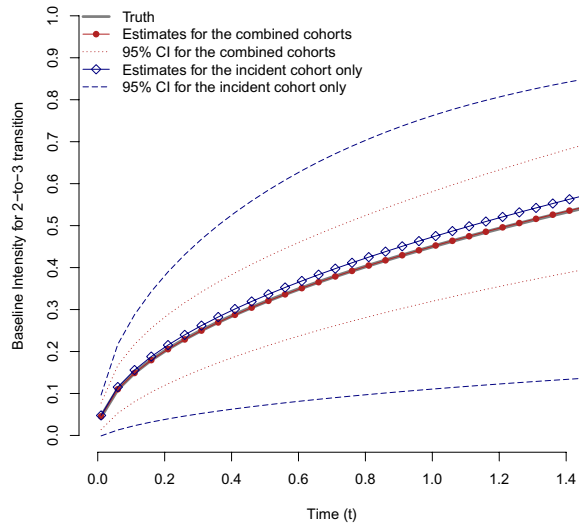
^e the relative efficiency obtained by the ratio of the mean-squared error from the combined cohort to that from the incident cohort

method were found to be reasonably robust against the misspecification of the form for $g(w)$.

4 Application

The Nun Study is a longitudinal study of aging and dementia with a cohort of 678 participants who were born before 1917 and recruited among members of the School Sisters of Notre Dame congregation in the United States. The cognitive status of

Fig. 3 The average of the estimated baseline intensities with the pointwise 95% confidence intervals, obtained from the combined cohort data (red) and the incident cohort only (blue) under the scenario with $m_1 = 300$, $m_2 = 300$, and 30% censoring rate for the prevalent cohort (Color figure online)



each participant was assessed annually for close to a decade and their awareness of deficits were classified into four stages according to severity: cognitively intact for age, cognitive deficit not affecting activities of daily living, cognitive deficit affecting one or more activities of daily living, and clinical dementia. Ages at their assessments and deaths were also recorded during the follow-up, along with two important covariates, the presence of at least one apolipoprotein E4 allele (APOE4) and their level of education (EDUCAT). To reduce dimensionality, we grouped the participants' status into four states and recorded them as follows: 0 for intact cognition, 1 for MCI, 2 for dementia, and 3 for death.

We applied the proposed approach to data from 501 subjects with complete baseline information. Among them, 424 were in the intact cognition or MCI state (i.e., incidence cases), while 77 were diagnosed with clinical dementia (i.e., prevalent cases) at their study recruitment. In the incident cohort, about 36% progressed to dementia and only 29% died with dementia during the follow-up. In the prevalent cohort, about 97% died during the follow-up. The flow chart for the samples of the incident cohort and the prevalent cohort from the Nun Study is available in the Supplementary Information. The average age at the first assessment was 82.53 years for the incident cohort and 85.89 years for the prevalent cohort. Table 3 shows the distribution of risk factors by cohort. The combined cohorts consist of more dementia cases in each group compared to the incident cohort. In the analysis, 75 years old, which is the earliest age at the intact cognition state in the incident cohort, is set as the initial time point of disease progression. The exponential distribution is used for the density of \tilde{W} , which implies that the period between dementia onset and study entry does not depend on the dementia onset itself.

The resulting parameter estimates and their standard errors, respectively obtained by analyzing the combined cohorts data and the incident cohort data only, are shown in Table 4. In model estimation, the proposed method was refitted to the

Table 3 Distribution of risk factors by cohort with the number of patients who experienced dementia before death during the study and the number of death at last follow-up

	Incident cohort only			Combined cohorts		
	$(m_1 = 424)$			$(m_1 = 424, m_2 = 77)$		
	Size	Dementia (%)	Death (%)	Size	Dementia (%)	Death (%)
APOE4						
Presence	82	39 (48%)	65 (79%)	111	68 (61%)	93 (84%)
Absence	342	114 (33%)	243 (71%)	390	162 (42%)	290 (74%)
EDUCAT						
College and higher	383	134 (35%)	274 (72%)	432	183 (42%)	323 (75%)
Others	41	19 (46%)	34 (83%)	69	47 (68%)	60 (87%)

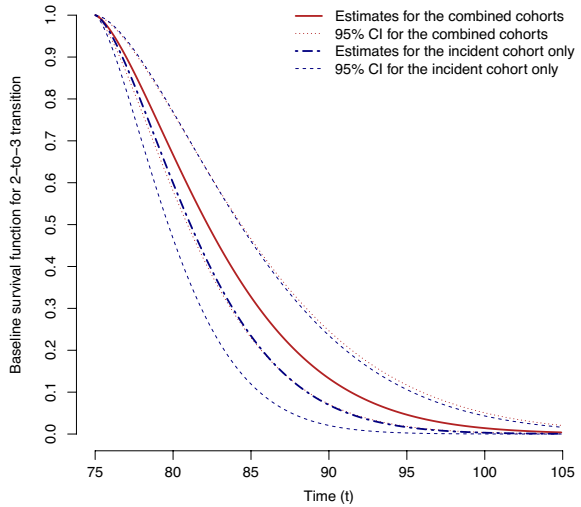
Table 4 Results of the parameter estimation in each transition using the combined cohorts and the incident cohort from the Nun Study

(Transition)	Combined cohorts					Incident cohort only				
	0-to-1	1-to-0	1-to-2	1-to-3	2-to-3	0-to-1	1-to-0	1-to-2	1-to-3	2-to-3
<i>APOE4 (presence=1, absence=0)</i>										
Estimate	0.28	-0.68	0.40	-0.09	-0.10	0.26	-0.71	0.39	-0.07	-0.19
SE	0.15	0.27	0.19	0.26	0.16	0.15	0.27	0.19	0.27	0.21
<i>EDCAT (college and higher=1, others=0)</i>										
Estimate	-0.89	0.62	0.17	0.59	0.34	-0.91	0.60	0.12	0.64	0.16
SE	0.30	0.43	0.24	0.34	0.17	0.31	0.44	0.25	0.39	0.26

data sets without the regression coefficients for the 0-to-3 transition because their estimates were close to zero. The results with both incident and prevalent cohorts of the Nun Study show that ApoE4 promotes the cognitive decline of a subject, in that an ApoE4 carrier is expected to have a relatively longer time of the transition from MCI back to intact cognition (p value = 0.011); a higher risk of being in MCI, which is marginally significant (p value = 0.054); and a shorter time of the transition from MCI to dementia (p value = 0.032), compared to a non-carrier. The education level of a subject is found to be significantly associated with the 0-to-1 and 2-to-3 transitions; the higher level of education (12 years or more) is found to decrease the risk of progression from intact cognition to MCI (p value = 0.003) and increase the risk of death with dementia (p value = 0.049). The results of incident cohort data show similar results with those of the combined cohort data except that education level was not significantly related to the 2-to-3 transition. The positive association between higher education and mortality with dementia also has been reported in other studies (Stern et al. 1995; Contador et al. 2017).

Figure 4 shows the estimated baseline intensity for the 2-to-3 transition with pointwise 95% confidence intervals. The analysis results of the combined data

Fig. 4 The estimated baseline survival functions with the pointwise 95% confidence intervals, obtained from the combined cohort (red) and the incident cohort (blue) of the Nun Study



from the incident and prevalent cohorts show narrower 95% pointwise confidence intervals than those of the incident cohort data only. The estimated density of the dementia onset in the prevalent population is presented in Fig. 5 as a byproduct of the analysis of the combined data. The median onset age of dementia was 86.43 (95% CI: 84.85 – 88.01) for an ApoE4 carrier with low education and 87.66 (95% CI: 86.57 – 88.76) for a non-carrier with low education. The estimated median survival after the onset of dementia was 3.13 years (95% CI: 2.27 – 4.00) for a patient with ApoE4 and a higher level of educational attained from the combined cohort data, while it was 3.04 years (95% CI: 1.97 – 4.11) from the incident cohort data.

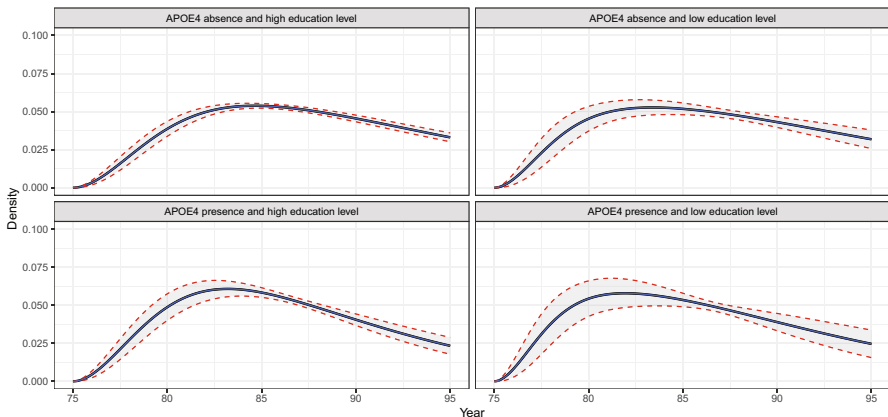


Fig. 5 The density of the dementia onset in the prevalent population by the presence of the apolipoprotein E4 allele and the level of education

5 Discussion

In this paper, we propose a multistate approach based on the likelihood-based inference to assess the effects of risk factors on the transitions that are related to the natural history of disease by using event history data that consists of both incident and prevalent samples. In the observational studies that are designed to follow the life course of a disease, prevalent samples are commonly available because patients receiving routine care in healthcare facilities are often within a study population. In the analysis with the prevalent samples, identifying the age of disease onset is essential because it plays a key role in addressing sampling bias induced by the prevalent sampling scheme. However, it is often difficult to retrospectively identify age of onset for many chronic diseases. The proposed method overcomes this challenge by using the complementary information that both incident cohort and prevalent cohort provide to each other, while accounting for interval censored observations of transition times between disease states.

The approach was illustrated by incorporating parametric transition intensities into a multistate model; however, it can be easily modified with other flexible forms such as a locally weighted smoother (Hubbard et al. 2008) and linear splines (Pak et al. 2019), although a substantial increase in the number of parameters to estimate is inevitable with many cases of possible transitions. By utilizing the commonly used temporal homogeneity model for transition intensity, we have a parsimonious model to capture the potential reversible transition between intact cognition and MCI in dementia progression. Alternatively, one could employ piecewise constant models to accommodate the more flexible temporal trend of the disease process, though it will require more transition events being observed to obtain stable estimates (Pérez-Ocón et al. 2001; Titman 2011).

As with other prevalent-data settings, it would be a challenge to incorporate time-dependent covariates into the analysis of combined data. If the time-dependent covariates are specified differently relative to each transition, one must know which transitions a prevalent individual went through before the development of dementia. Nevertheless, with extra information for the prevalent samples or modeling assumptions, one may incorporate time-dependent covariates when combining the incident and prevalent cohorts, which merits future research.

The key benefit of the analysis of the event history data is that one can simultaneously assess the effects of risk factors on every transition that represents the natural history of the disease. The proposed method can be applied to event history data from other disease studies. An example is with studies on the natural history of the coronavirus disease (COVID-19), where the information from travelers who are found to be infected upon arrival (i.e., prevalent samples) can be a complement to the study on the endpoints of death or recovery in periods of quarantine.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10985-023-09602-x>.

Acknowledgements This work was partially supported by the National Research Foundation of Korea (NRF) grant 2021R1G1A1009269 (DP), the National Cancer Institute grants R01CA269696 and

P30CA016672 (JN and YS) and the National Institute on Aging AG0386561 (RK). We also thank Jessica Swann for her editorial assistance.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Contador I, Stern Y, Bermejo-Pareja F, Sanchez-Ferro A, Benito-Leon J (2017) Is educational attainment associated with increased risk of mortality in people with dementia? A population-based study. *Curr Alzheimer Res* 14(5):571–576
- Cox D, Miller H (1965) The theory of stochastic processes. Chapman and Itall, New York
- Gorfine M, Keret N, Ben Arie A, Zucker D, Hsu L (2021) Marginalized frailty-based illness-death model: application to the UK-biobank survival data. *J Am Stat Assoc* 116(535):1155–1167
- Hubbard R, Inoue L, Fann J (2008) Modeling nonhomogeneous Markov processes via time transformation. *Biometrics* 64(3):843–850. <https://doi.org/10.1111/j.1541-0420.2007.00932.x>
- Kalbfleisch J, Lawless J (1985) The analysis of panel data under a Markov assumption. *J Am Stat Assoc* 80(392):863–871. <https://doi.org/10.1080/01621459.1985.10478195>
- Kulathinal S, Säävälä M, Auranen K, Saarela O (2020) Estimation of marriage incidence rates by combining two cross-sectional retrospective designs: Event history analysis of two dependent processes. arXiv preprint [arXiv:2009.01897](https://arxiv.org/abs/2009.01897)
- Lee C, Ning J, Kryscio R, Shen Y (2019) Analysis of combined incident and prevalent cohort data under a proportional mean residual life model. *Stat Med* 38(12):2103–2114. <https://doi.org/10.1002/sim.8098>
- McVittie J, Wolfson D, Stephens D, Addona V, Buckeridge D (2020) Parametric models for combined failure time data from an incident cohort study and a prevalent cohort study with follow-up. *Int J Biostatist*. <https://doi.org/10.1515/ijb-2020-0042>
- Omar R, Stallard N, Whitehead J (1995) A parametric multistate model for the analysis of carcinogenicity experiments. *Lifetime Data Anal* 1(4):327–346. <https://doi.org/10.1007/BF00985448>
- Pak D, Li C, Todem D, Sohn W (2017) A multistate model for correlated interval-censored life history data in caries research. *J R Stat Soc Ser C* 66(2):413–423. <https://doi.org/10.1111/rssc.12186>
- Pak D, Li C, Todem D (2019) Semiparametric analysis of correlated and interval-censored event-history data. *Stat Methods Med Res* 28(9):2754–2767. <https://doi.org/10.1177/0962280218788383>
- Pérez-Ocón R, Ruiz-Castro J, Gámiz-Pérez M (2001) Non-homogeneous Markov models in the analysis of survival after breast cancer. *J R Stat Soc Ser C* 50(1):111–124. <https://doi.org/10.1111/1467-9876.00223>
- Saarela O, Kulathinal S, Karvanen J (2009) Joint analysis of prevalence and incidence data using conditional likelihood. *Biostatistics* 10(3):575–587. <https://doi.org/10.1093/biostatistics/kxp013>
- Shen Y, Ning J, Qin J (2009) Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *J Am Stat Assoc* 104(487):1192–1202. <https://doi.org/10.1198/jasa.2009.tm08614>
- Stern Y, Tang M, Denaro J, Mayeux R (1995) Increased risk of mortality in Alzheimer's disease patients with more advanced educational and occupational attainment. *Ann Neurol Official J Am Neurol Assoc Child Neurol Soc* 37(5):590–595. <https://doi.org/10.1002/ana.410370508>
- Titman A (2011) Flexible nonhomogeneous Markov models for panel observed data. *Biometrics* 67(3):780–787. <https://doi.org/10.1111/j.1541-0420.2010.01550.x>
- Tyas S, Salazar J, Snowdon D, Desrosiers M, Riley K, Mendiondo M, Kryscio R (2007) Transitions to mild cognitive impairments, dementia, and death: findings from the nun study. *Am J Epidemiol* 165(11):1231–1238. <https://doi.org/10.1093/aje/kwm085>
- Wei S, Xu L, Kryscio R (2014) Markov transition model to dementia with death as a competing event. *Comput Stat Data Anal* 80:78–88. <https://doi.org/10.1016/j.csda.2014.06.014>
- Wolfson D, Best A, Addona V, Wolfson J, Gadalla S (2019) Benefits of combining prevalent and incident cohorts: an application to myotonic dystrophy. *Stat Methods Med Res* 28(10–11):3333–3345. <https://doi.org/10.1177/0962280218804275>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.