BMC Medical Research
Methodology

RESEARCH
Open Access

A discrete-time split-state framework for multi-state modeling with application to describing the course of heart disease

Ming Ding[1*], Haiyi Chen[2] and Feng-Chang Lin[2]

Abstract

In chronic disease epidemiology, the investigation of disease etiology has largely focused on an endpoint, while the course of chronic disease is understudied, representing a knowledge gap. Multi-state models can be used to describe the course of chronic disease, such as Markov models which assume that the future state depends only on the present state, and semi-Markov models which allow transition rates to depend on the duration in the current state. However, these models are unsuitable for chronic diseases that are largely non-memoryless. We propose a Discrete-Time Split-State Framework that generates a process of substates by conditioning on past disease history and estimates discrete-time transition rates between substates as a function of duration in a (sub)state. Specifically, as the substates are created by conditioning on past history, they satisfy the Markov assumption, regardless of whether the original disease process is Markovian; and the transition rates are approximated by competing risks in a short time interval estimated from cause-specific Cox models. In the simulation study, we simulated a Markov process with an exponential distribution, a semi-Markov process with a Weibull distribution, and a non-Markov process with an exponential distribution. The coverage rate of transition rates estimated using our framework was 94% for the Markov process and 93% for the non-Markov process. However, the estimated transition rates were under coverage (72%) for the semi-Markov process, which is likely due to the approximation of transition rates in discrete time. In the application, we applied the framework to describe the course of heart disease in a large cohort study. In summary, the framework we proposed can be applied to both Markov and non-Markov processes and has potential to be applied to semi-Markov processes. For future research, as substates created using our framework track past disease history, the transition rates between substates have the potential to be used to derive summary estimates that characterize the disease course.

Keywords Multi-state modeling, Course of chronic diseases, Heart disease

Introduction

In chronic disease epidemiology, the investigation of disease etiology has largely focused on a single endpoint. However, the development of chronic disease is a multi-state process, with each state playing a crucial role in affecting its progression. Understanding the course of disease progression allows us to gain new mechanistic insights into the disease at the population level. Despite its importance, this field remains understudied and represents a knowledge gap in epidemiology.

Multi-state models have been used to describe the course of chronic disease, including Markov models, which assume that the future state depends only on the present state [1, 2], and semi-Markov models, which allow transition rates to depend on the duration in the current state [3, 4]. However,

*Correspondence:
Ming Ding
ming_ding@med.unc.edu
[1] Department of Emergency Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2] Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

these models are unsuitable for chronic diseases where past states often interplay and affect future disease progression [5–7]. With only a few nonparametric methods proposed, non-Markov models remain a largely unexplored area [8–14]. Proposed for non-Markov processes, the landmark Aalen-Johansen (LMAJ) estimator estimates transition probabilities using landmarking, which divides participants into subgroups according to the state occupation probability at a certain time point [8]. However, it is a non-parametrical method and the subsampling of participants lowers the power of estimation [15, 16]. One non-Markov multi-state modeling approach has been proposed that uses logistic models to model transitions, allowing the transition to depend on the entire previous history [17].

In this paper, we propose a Discrete-Time Split-State Framework that generates a process of disease substates by conditioning on past disease history and estimates discrete-time transition rates as a function of duration in a substate. The substates created using our framework have two public health significances. First, regardless of whether the original process follows a Markov assumption, the substates are independent of past history and satisfy the Markov assumption. Thus, the Aalen-Johansen estimator can be used to estimate transition probabilities between substates based on the estimated transition rates. These transition parameters enhance our understanding of the dynamic process of chronic disease progression. Second, while the substates satisfy the Markov assumption, they track past disease history and are memorable. For future research, the transition rates between substates can be synthesized into summary estimates to characterize the disease course, which can then be used for the precision prevention and prediction of chronic diseases.

Methods: Development of the discrete-time split-state framework

Step 1. Split disease states into substates by conditioning on past history

Suppose a continuous-time process with $(M+1)$ multi-states S_0, S_1, S_2, … S_M. We use s_m (coded as 0 or 1) to indicate whether the participant is at S_m, where *m*is between 0 and M. For a forward transition Markov process, where a state can only transition to future states but not past states [18], the transition rate depends only on the present state and not on the previous states. Equation 1 establishes the memoryless property of the Markov process for the transition rate to state S_{m+1} from S_m at time *t*, given any combination of past history s_0, s_1, …., s_{m-1}:

$$\lambda_{S_{m+1}=1|S_m=1}(t) = \lambda_{S_{m+1}=1|S_0=s_0,\ S_1=s_1,\ldots,S_{m-1}=s_{m-1},\ S_m=1}(t) \tag{1}$$

However, for a non-Markov process, the transition rate depends on the present and past states. $\lambda_{S_{m+1}=1|S_m=1}(t)$ cannot be estimated using information only at S_m, as it also depends on the values of s_0, s_1, …., and s_{m-1}. As Eq. 1 does not establish for a non-Markov process, we condition on the past history and divide it into substates, $S_m\_J_{0\_m-1}$, where J_{0_m-1} is a joint indicator of past states S_0 to S_{m-1} and has $2^{\ m-1}$ combinations. Taking $m=3$ for example, J_{0_2} is a joint indicator of past states S_0, S_1 and S_2, and has 4 categories, "0", "0_1", "0_2", "0_1_2", which show disease path ($s_0=1$, $s_1=0$, and $s_2=0$), ($s_0=1$, $s_1=0$, and $s_2=1$), ($s_0=1$, $s_1=1$, and $s_2=0$), and ($s_0=1$, $s_1=1$, and $s_2=1$), respectively. Correspondingly, S_3 can be divided into four substates, S_{3_0}, $S_{3\_0\_1}$, $S_{3\_0\_2}$, and $S_{3\_0\_1\_2}$. By conditioning on the path of past history, the substates at state S_m are independent of past states. Thus, as Eq. 2 shows, the transition from substate $S_m\_J_{0\_m-1}$ to S_{m+1} is a Markov process, regardless of whether the original process is Markovian.

$$\lambda_{S_{m+1}=1|S_m\_J_{0\_m-1}=1}(t) = \lambda_{S_{m+1}=1|S_0=s_0,S_1=s_1,\ldots,\ S_{m-1}=s_{m-1},S_m\_J_{0\_m-1}=1}(t) \tag{2}$$

It turns out that transition rates between substates, $\lambda_{S_{m+1}=1|S_m\_J_{0\_m-1}=1}(t)$, can be estimated by including the joint indicator of past states, J_{0_m-1}, as a covariate into the model that estimates transition rate between disease states, $\lambda_{S_{m+1}=1|S_m=1}$ (Eq. 3).

$$\begin{aligned}\lambda_{S_{m+1}=1|S_m\_J_{0\_m-1}=1}(t) &= \lambda_{S_{m+1}=1|S_0=s_0,S_1=s_1,\ldots,\ S_{m-1}=s_{m-1},S_m\_J_{0\_m-1}=1}(t)\\ &= \lambda_{S_{m+1}=1|S_0=s_0,S_1=s_1,\ldots,\ S_{m-1}=s_{m-1},S_m=1}(t)\\ &= \lambda_{S_{m+1}=1|J_{0\_m-1},S_m=1}(t)\\ &\equiv \lambda_{S_{m+1}=1|S_m=1}\left(t,J_{0\_m-1}\right)\end{aligned} \tag{3}$$

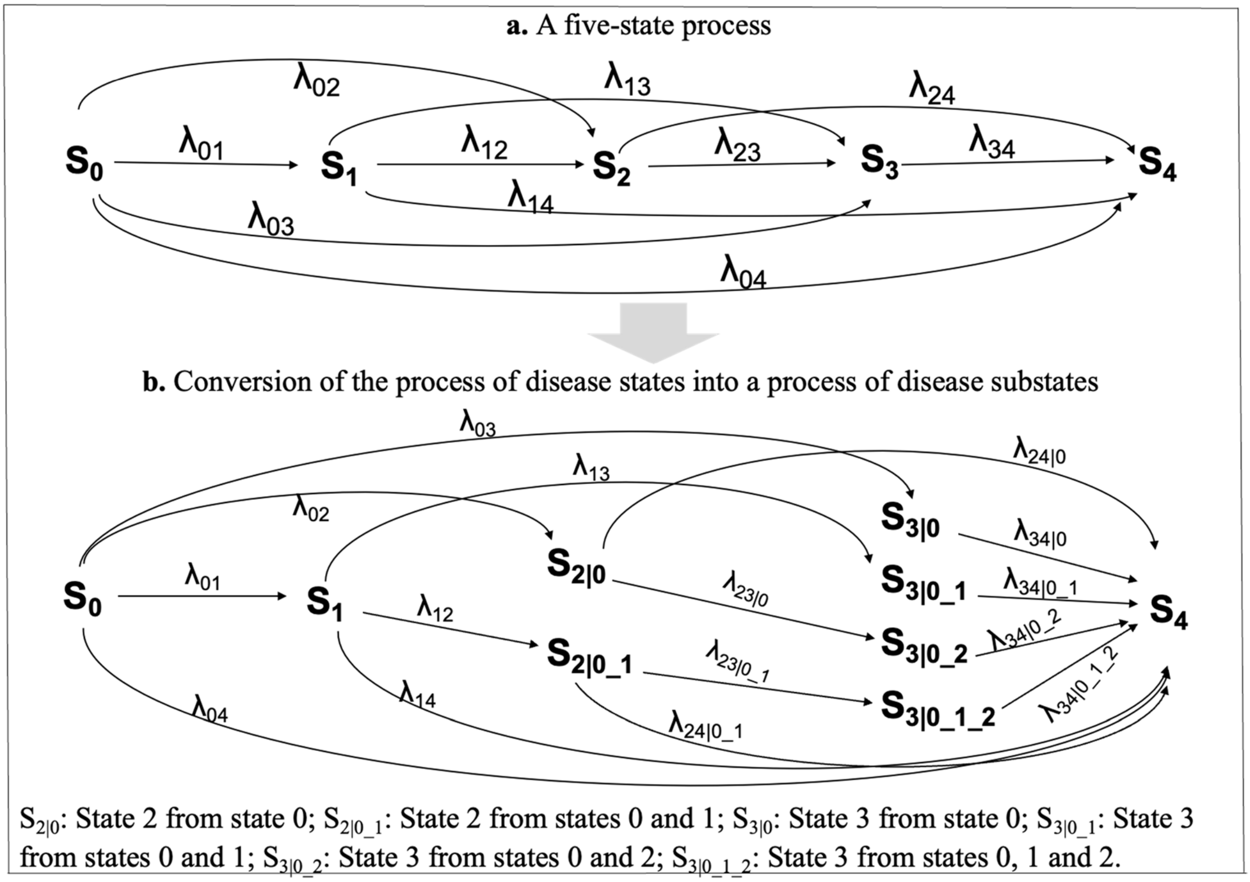
Construct and visualize the process of disease substates

The generation of disease substates can be visualized in Fig. 1, which shows the conversion from a process of disease states to a process of substates using five states as an example. In Fig. 1a, we assume a forward transition process and allow the transition from the current state to any future states. By conditioning on past states, a disease state is divided into substates (Fig. 1b). As substates memorize past disease history, transition only occurs between substates that match the past disease states.

Step 2. Estimate discrete-time transition rates

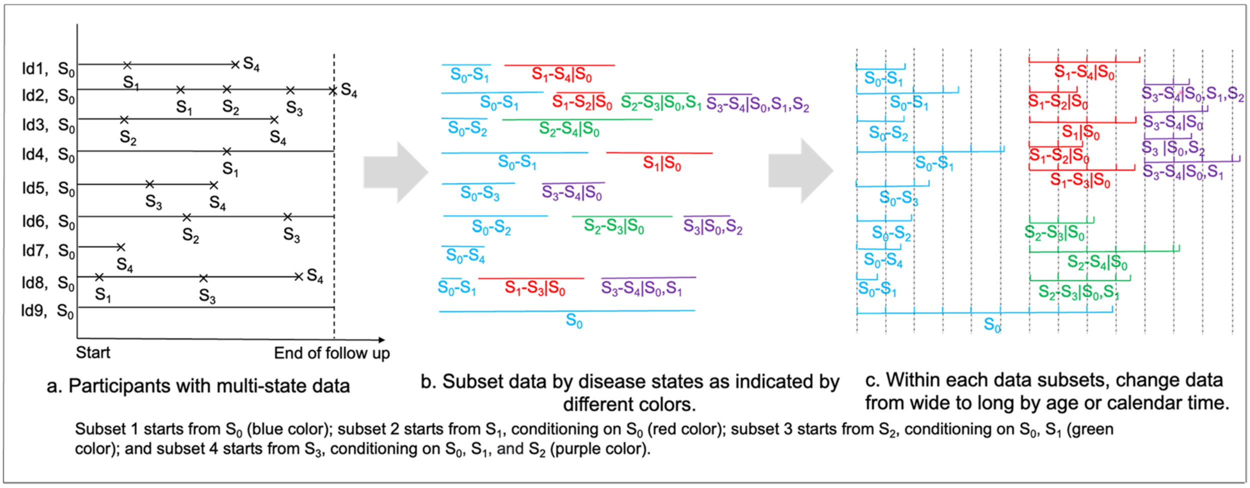
Fit cause-specific Cox models

Suppose we have survival data with disease states ascertained, with all participants starting from S_0 (Fig. 2a). We split each person's data into subsets 1–4 by disease state. Observations are classified into four data subsets: Subset 1 starts from S_0 (blue); Subset 2 starts from S_1, conditioning on S_0 (red); Subset 3 starts from S_2, conditioning on S_0, S_1 (green); and subset 4 starts from S_3, conditioning on S_0, S_1, and S_2 (purple) (Fig. 2b). We use



$S_{2|0}$: State 2 from state 0; $S_{2|0\_1}$: State 2 from states 0 and 1; $S_{3|0}$: State 3 from state 0; $S_{3|0\_1}$: State 3 from states 0 and 1; $S_{3|0\_2}$: State 3 from states 0 and 2; $S_{3|0\_1\_2}$: State 3 from states 0, 1 and 2.

Fig. 1 Construction of the disease substate process by conditioning on past states



Subset 1 starts from S_0 (blue color); subset 2 starts from S_1, conditioning on S_0 (red color); subset 3 starts from S_2, conditioning on S_0, S_1 (green color); and subset 4 starts from S_3, conditioning on S_0, S_1, and S_2 (purple color).

Fig. 2 Data preparation for applying the Discrete-time Split-State Framework

a discrete-time approach to fit cause-specific Cox (CSC) models to each data subset. As an extension of the Cox model, the CSC models assume different associations of each exposure with each specific event type [19]. However, the Cox model (including CSC models) is a semi-parametric model where the baseline hazard is arbitrary, as the baseline hazard is canceled out in the conditional likelihood estimation [20]. Although the Cox model can theoretically estimate the hazard nonparametrically, we have not yet found an R package that directly outputs the hazard (there are packages that output cumulative incidence, but not the hazard). Thus, to estimate how hazard changes with time, we use survival risk in a short time interval to approximate discrete-time hazards [21]. Specifically, within each data subset, we discretize duration in a state into small intervals, change the data structure from wide to long form, and model the duration t (t = 1, 2, 3, …) as a counting variable (Fig. 2c).

We fit four CSC models (Models 1–4) to data subsets 1–4. Duration in a (sub)state t, is included as a covariate in the CSC models, where f(t) is a function of duration in the current state. In Models 3 and 4, past disease history is included as a covariate, which allows past states to affect the transition rates of the current state, and the interaction between past states and time can be further included as covariates, which allows past states to affect the relationship of transition rates with duration in the current state.

Test Markov vs. non-Markov assumption

Whether past states affect the current state can be tested using a likelihood ratio test. Take Model 3 as an example, there are three ways to fit the model: A) including only f (t), B) further including past disease history J_{0_1}, and C) further including an interaction term $J_{0\_1} \times f(t)$. A likelihood ratio test can be used to compare models B and C with model A. A better fit for models B or C compared to model A indicates that the Markov assumption is violated.

Model 1 in subset 1 starting from S_0 state: $h_{(Sm=1|S0=1)}(\tau, t) = h_{0m}(\tau)\exp(f(t))$,

where m can be 1–4, indicating the development of S_1, S_2, S_3, or S_4 states, respectively. $h_{0m}(\tau)$ is the baseline hazard for developing state m in the first interval ($t = 1$). τ is time to event in each time interval and ranges from [0, u], where u is the length of the time interval.

Model 2 in subset 2 starting from S_1 state: $h_{(Sm=1|S1=1,S0=1)}(\tau, t) = h_{0m}(\tau)\exp(f(t))$,

where m can be 2–4, indicating the development of S_2, S_3, or S_4 states. There is no need to account for past states in Model 2, as we assumed all participants started from the $S_0 = 1$ state.

Model 3 in subset 3 starting from S_2 state:

$$h_{(Sm=1|J0\_1,S2=1)}(\tau, t, J_{0\_1}) = h_{0m}(\tau)\exp(f(t) + J_{0\_1} + J_{0\_1} \times f(t)),$$

where m can be 3 and 4, indicating the development of S_3 or S_4 states. J_{0_1} is a joint indicator of past states S_0 and S_1, and has two categories, 0 and 0_1, which show disease path ($s_0 = 1$ and $s_1 = 0$) and ($s_0 = 1$ and $s_1 = 1$), respectively. $J_{0\_1} \times f$ (t) is an interaction term between f(t) and J_{0_1}.

Model 4 in subset 4 starting from S_3 state:

$$h_{(Sm=1|J0\_2,S3=1)}(\tau, t, J_{0\_2}) = h_{0m}(\tau)\exp(f(t) + J_{0\_2} + J_{0\_2} \times f(t)),$$

where m indicates the development of S_4 state. J_{0_2} is a joint indicator of past states S_0, S_1 and S_2, and has 4 categories, 0, 0_1, 0_2, 0_1_2, which track disease path ($s_0 = 1$, $s_1 = 0$, and $s_2 = 0$), ($s_0 = 1$, $s_1 = 0$, and $s_2 = 1$), ($s_0 = 1$, $s_1 = 1$, and $s_2 = 0$), and ($s_0 = 1$, $s_1 = 1$, and $s_2 = 1$), respectively.

Estimate discrete-time transition rates

Λ(t). From each model, the survival probability at the end of each time interval can be estimated using the 'predict' command from the 'riskRegression' package [22]. Discrete-time transition rates at time t can be approximated as $\frac{1-\text{survival probability at t}}{\text{length of the time interval}}$. In particular, when age is used as a time scale, the predicted risk at the end of each time interval approximates the hazard in discrete time [21]. A conceptual illustration of the estimation of discrete-time transition rates is shown in Fig. 3. As the discrete-time transition rates can be estimated as a function of duration in a state, our framework has the potential to be applied to a semi-Markov process.

Specifically, from Model 1, $\lambda_{01}(t)$, $\lambda_{02}(t)$, $\lambda_{03}(t)$, and $\lambda_{04}(t)$, which are the transition rates from state 0 to states 1–4, can be estimated as $\frac{1-CIF_{(Sm=1|S0=1)}(u,t)}{u}$, where m = 1, 2, 3, 4 for $\lambda_{01}(t)$, $\lambda_{02}(t)$, $\lambda_{03}(t)$, and $\lambda_{04}(t)$, respectively, u is the length of time interval, and $CIF_{Sm=1|S0=1}(u, t)$ is the cumulative incidence function (CIF) of developing S_m at the end of the interval u for interval t among participants starting from S_0. Of note, for CSC model, CIF estimates the incidence of an event while taking competing risk into account [23]. Similarly, Model 2 estimates $\lambda_{12}(t)$, $\lambda_{13}(t)$, and $\lambda_{14}(t)$, which are the transition rates from S_1 to S_2, S_3, or S_4. Model 3 estimates $\lambda_{23|0}(t)$ and $\lambda_{24|0}(t)$ which are the transition rates from substate $S_{2|0}$ (i.e., S_2 with no history of S_1) to S_3 or S_4, and $\lambda_{23|0\_1}(t)$ and $\lambda_{24|0\_1}(t)$ which are the transition rates from substate $S_{2|0\_1}$ (i.e., S_2 with a history of S_1) to S_3 or S_4. Model 4 estimates $\lambda_{34|0}(t)$, $\lambda_{34|0\_1}(t)$, $\lambda_{34|0\_2}(t)$, and $\lambda_{34|0\_1\_2}(t)$, which are the transition rates to S_4 from substates $S_{3|0}$, $S_{3|0\_1}$, $S_{3|0\_2}$, and $S_{3|0\_1\_2}$, respectively.

Step 3. Construct matrices of transition parameters between disease substates

Transition rate matrix $\boldsymbol{Q}(\boldsymbol{t})$

Transition rates between substates can be constructed as below.

$$\begin{array}{c|ccccccccc}
 & S_0\,(t) & S_1\,(t) & S_{2|0}\,(t) & S_{2|0\_1}\,(t) & S_{3|0}\,(t) & S_{3|0\_1}\,(t) & S_{3|0\_2}\,(t) & S_{3|0\_1\_2}\,(t) & S_4\,(t) \\
S_0\,(t) & -(\lambda_{01}+\lambda_{02}+\lambda_{03}+\lambda_{04}) & \lambda_{01} & \lambda_{02} & 0 & \lambda_{03} & 0 & 0 & 0 & \lambda_{04} \\
S_1\,(t) & 0 & -\lambda_{12}-\lambda_{13}-\lambda_{14} & 0 & \lambda_{12} & 0 & \lambda_{13} & 0 & 0 & \lambda_{14} \\
S_{2|0}\,(t) & 0 & 0 & -\lambda_{23|0}-\lambda_{24|0} & 0 & 0 & 0 & \lambda_{23|0} & 0 & \lambda_{24|0} \\
S_{2|0\_1}\,(t) & 0 & 0 & 0 & -\lambda_{23|0\_1}-\lambda_{24|0\_1} & 0 & 0 & 0 & \lambda_{23|0\_1} & \lambda_{24|0\_1} \\
S_{3|0}\,(t) & 0 & 0 & 0 & 0 & -\lambda_{34|0} & 0 & 0 & 0 & \lambda_{34|0} \\
S_{3|0\_1}\,(t) & 0 & 0 & 0 & 0 & 0 & -\lambda_{34|0\_1} & 0 & 0 & \lambda_{34|0\_1} \\
S_{3|0\_2}\,(t) & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{34|0\_2} & 0 & \lambda_{34|0\_2} \\
S_{3|0\_1\_2}\,(t) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{34|0\_1\_2} & \lambda_{34|0\_1\_2} \\
S_4\,(t) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

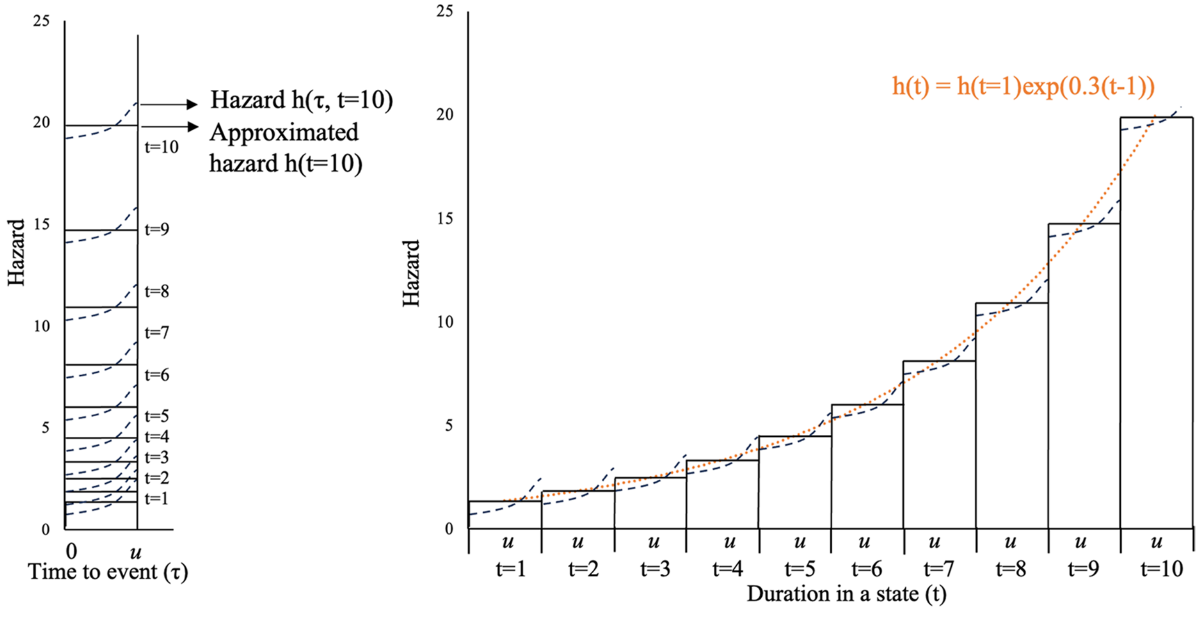
Each row and each column stand for a disease substate, and $\boldsymbol{Q}(\boldsymbol{t})[i,j]$ stands for the transition rate from the substate in row i to the substate in column j. If there is no transition between two substates, the corresponding value in the matrix is 0. The diagonal of $\boldsymbol{Q}(\boldsymbol{t})$ is the change in rate that participants remain in the state, which is the opposite of the sum of rates transited to other states. The sum of each row of matrix $\boldsymbol{Q}(\boldsymbol{t})$ is 0.

As the process of disease substates satisfies the Markov assumption, the Aalen-Johansen estimator can be used to estimate state occupation probabilities and transition probabilities based on transition rates between substates (i.e., the Aalen-Johansen estimators are systematically biased if directly used in a non-Markov process [12]).

State occupation probabilities $\boldsymbol{P}(\boldsymbol{t})$

$\boldsymbol{P}(\boldsymbol{t})$ Is defined as the proportion of participants occupied in each state at t and can be expressed as functions of transition rates.

$$\begin{aligned}
\mathbf{P}(t+1) &= \mathbf{P}(t) + \Delta\mathbf{P}(t) \\
&= \mathbf{P}(t) + \mathbf{P}(t) * \mathbf{Q}(t) \\
&= \mathbf{P}(t) * (I_9 + \mathbf{Q}(t)) \\
&= (\mathbf{P}(t-1) + \Delta\mathbf{P}(t-1)) * (I_9 + \mathbf{Q}(t)) \\
&= \mathbf{P}(t-1) * (I_9 + \mathbf{Q}(t-1)) * (I_9 + \mathbf{Q}(t)) \\
&= \mathbf{P}(t=0) * \textstyle\prod_{t=0}^{t=k}(I_9 + \mathbf{Q}(t))
\end{aligned}$$



a. Approximated hazard in long data format **b. Approximated hazard in wide data format**

Fig. 3 An illustration of estimating the discrete-time hazard using the Cox model. **a** shows how the hazard h(τ, t) changes with the time to event τ within each time interval t. We discretize the duration in a state (t) into small intervals (with the length of each interval denoted as *u*), change the data structure from wide to long format, and model the time interval t as a counting variable (t = 1, 2, 3, …). Suppose the Cox model is fitted as h(τ, t) = h(τ)exp(0.3t). The dashed lines (–) show how the hazard changes with τ within each time interval. As expected, the hazard between 0 and *u* is proportional across t. The solid lines (—) show that the hazards within each time interval can be approximated by cumulative incidence. **b** shows how the approximated discrete-time hazard h(t) changes with t

where $\mathbf{Q}(\mathrm{t})$ is the transition rate matrix, and I_9 is an identity matrix. Of note,$\mathbf{P}(\mathrm{t}) = \left(\mathrm{p}_0(\mathrm{t}), \mathrm{p}_1(\mathrm{t}), \mathrm{p}_{2|0}(\mathrm{t}), \mathrm{p}_{3|0}(\mathrm{t}), \mathrm{p}_{3|0\_1}(\mathrm{t}), \mathrm{p}_{3|0\_2}(\mathrm{t}), \mathrm{p}_{3|0\_1\_2}(\mathrm{t}),\ \mathrm{p}_4(\mathrm{t})\right)$, where p(t) is the probability of participants in a particular state. $\mathrm{p}_0(\mathrm{t}) + \mathrm{p}_1(\mathrm{t}) + \mathrm{p}_{2|0}(\mathrm{t}) + \mathrm{p}_{2|0\_1}(\mathrm{t}) + \mathrm{p}_{3|0}(\mathrm{t}) + \mathrm{p}_{3|0\_1}(\mathrm{t}) + \mathrm{p}_{3|0\_2}(\mathrm{t}) + \mathrm{p}_{3|0\_1\_2}(\mathrm{t}) +\ \mathrm{p}_4(\mathrm{t}) = 1$ · If all participants start from state 0, $\mathbf{P}(\mathrm{t} = 0) = (1, 0, 0, 0, 0, 0, 0, 0, 0)$.

Transition probabilities $TP(t_1, t_2)$

$\boldsymbol{TP}(\boldsymbol{t}_1, \boldsymbol{t}_2)$ Is defined as the transition probability from each state to the following states from time t_1 to t_2. Specifically, at the start of t_1, $\mathbf{TP}\ (\mathrm{t} = t_1) = I_9$, which assumes the probability starting from each state is 1; and at the end of t_1, $\mathbf{TP}(t_1, t_1 + 1) = I_9 + \mathbf{Q}(\mathrm{t} = t_1)$. From t_1 to t_2, $\mathbf{TP}(t_1, t_2) = \prod_{\mathrm{t}=\boldsymbol{t}_1}^{\mathrm{t}=\boldsymbol{t}_2} (\mathrm{I}_9 + \mathbf{Q}(\mathrm{t}))$, where $\boldsymbol{TP}(t_1, t_2)[i, j]$ describes the proportion of participants transited from the state in row i to the state in column j. The sum of each row of matrix $\boldsymbol{TP}(\boldsymbol{t}_1, \boldsymbol{t}_2)$ is 1.

Use bootstrap to obtain 95% Cis

A parametric bootstrapping approach can be adopted to randomly draw samples with replacement to create a new sample that are the same size as the original population. Transition rates, transition probabilities, and state occupation probability are estimated from each bootstrap sample. Statistically, the resulting transition parameter estimates across all bootstrap samples represent the empirical distribution and can be summarized with median values and a 95% confidence interval (CI) defined by 2.5th to 97.5th percentiles of the empirical distribution.

Simulation study

Simulation methods

We performed a simulation study to validate the utility of the Discrete-time Split-state framework in estimating transition rates. Data were simulated across five states, and three scenarios were considered: Markov and non-Markov processes with an exponential distribution, and semi-Markov process with a Weibull distribution. For the Markov process with an exponential distribution, transition rates were simulated as constant. For semi-Markov process with a Weibull distribution, transition rates were simulated as either increasing or decreasing over time. For the non-Markov process, the parameters of transition rates differed based on past states. We simulated survival data starting from each disease state using the R package 'crisk.sim', and with the parameters for the simulation study shown in Table S1. The processes with exponential distributions were simulated as special cases of Weibull distributions, with the ancillary parameters fixed at 1. The datasets were simulated 500 times, each with a sample size of 2000.

For each scenario, we evaluated the performance of our framework in estimating the transition rates of substates. When converting the data from wide to long format, we divided time into intervals of 0.05. However, if the Cox model did not converge due to a lack of cases within some intervals, we increased the length of interval to 0.1. Models 5–8 were used to estimate the transition rates. Specifically, t referred to the duration in the current state (*e.g.*, for Model 6, t is the duration since entering the S_1 state), and restricted cubic splines were used to model the time intervals with high flexibly. For the non-Markov scenario, we included past disease states as covariates in Models 7 and 8. To estimate the 95% CI of the transition rates, bootstrap was conducted 1000 times, with the same sample size as the original population for each bootstrap.

Model 5 in subset 1 starting from S_0 state: $\mathrm{h}_{(\mathrm{Sm}=1|\mathrm{S0}=1)}(\tau, \mathrm{t}) = \mathrm{h}_{0\mathrm{m}}(\tau) \exp(\mathrm{spline}(\mathrm{t}))$

Model 6 in subset 2 starting from S_1 state: $\mathrm{h}_{(\mathrm{Sm}=1|\mathrm{S1}=1,\mathrm{S0}=1)}(\tau, \mathrm{t}) = \mathrm{h}_{0\mathrm{m}}(\tau) \exp(\mathrm{spline}(\mathrm{t}))$

Model 7 in subset 3 starting from S_2 state: $\mathrm{h}_{(\mathrm{Sm}=1|\mathrm{J0\_1},\mathrm{S2}=1)}(\tau, \mathrm{t}, \mathrm{J}_{0\_1}) = \mathrm{h}_{0\mathrm{m}}(\tau) \exp(\mathrm{spline}(\mathrm{t}) + \mathrm{J}_{0\_1})$

Model 8 in subset 4 starting from S_3 state: $\mathrm{h}_{(\mathrm{Sm}=1|\mathrm{J0\_2},\mathrm{S3}=1)}(\tau, \mathrm{t}, \mathrm{J}_{0\_2}) = \mathrm{h}_{0\mathrm{m}}(\tau) \exp(\mathrm{spline}(\mathrm{t}) + \mathrm{J}_{0\_2})$

Survival probability at the end of each interval was obtained using the CIF function, and the hazard was estimated as (1-survival probability)/length of time interval. Transition rates were estimated at time points 0.3, 0.6, 0.9, 1.2, and 1.5, when most of the simulated participants had developed the endpoint.

The estimated transition rates were compared to theoretical rates, calculated as

$$\frac{1/\ \mathrm{beta0}}{(e^{\mathrm{ancillary\ parameters}})^{(1/\ \mathrm{beta0})}} \times t^{(1/\ \mathrm{beta0}-1)}$$

The mean squared error (MSE), coverage, and width of the 95% CI were obtained. MSE is the average squared difference between the estimate and the true value. Coverage is defined as the proportion of 95% CIs that cover the true value across all simulated datasets. We compared the coverage to determine which method was closer to the nominal 95% coverage rate. The width of the 95% CI is the difference between the upper and lower bounds. To compare transition rates across models in a summarized manner, we calculated the mean MSE, coverage, and width of the 95% CI across all states and time points.

Simulation results

The coverage rates of transition rates were 94% for the Markov process an exponential distribution and 93% for the non-Markov processes with an exponential distribution (Table 1). However, the transition rates showed under-coverage (72%) for the semi-Markov process with a Weibull distribution. The coverage rates, width of the 95% CI, and MSE of transition rates between disease substates are shown in Table 2. For each simulated dataset, the average runtime for the Markov, non-Markov, and semi-Markov models was 25 min, 29 min, and 41 min, respectively.

Application study

Study population

The Atherosclerosis Risk in Communities Study (ARIC) study was designed to investigate the causes of atherosclerosis and its clinical outcomes, as well as variations in cardiovascular risk factors and disease by sex and race [24]. The enrolled participants in ARIC underwent a phone interview and clinic visit at baseline and were followed up by telephone calls and re-examinations. Participants were contacted periodically by phone and interviewed about interim hospital admissions, cardiovascular outpatient diagnoses, and deaths. Participants who reported CVD-related events were asked to provide medical records that were reviewed by physicians. For each event, the month and year of diagnosis were recorded as the diagnosis date. Heart failure was ascertained by surveillance calls or clinic visits and was verified using death certificates, medical records, and outpatient diagnoses. Deaths were identified from systematic searches of vital records in states and the National Death Index, supplemented by reports from family members and postal authorities [25]. We obtained ARIC data through the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC), an open repository created by the National Heart, Lung, and Blood Institute (NHLBI), with the Institutional Review Board (IRB) approval deemed exempt by the University of North Carolina (UNC)-Chapel Hill Review Board.

Methods

We modeled heart disease progression in five states: Healthy (S_0), at metabolic risk (S_1), coronary heart disease (CHD) (S_2), heart failure (S_3), and mortality (S_4). The at-metabolic-risk state was defined as the development of hypertension, hyperlipidemia, or diabetes. Hypertension was defined as blood pressure ≥ 140/90 mmHg or a history of hypertension or use of blood pressure medications [26]. Hyperlipidemia included primary hypertriglyceridemia (≥ 175 mg/dL) or primary hypercholesterolemia (LDL-c 160–189 mg/dL, and/or non-HDL-c 190–219 mg/dL) [27]. Diabetes was defined as fasting glucose ≥ 7.7 mmol/L [28]. We calculated the time of incidence of the at-metabolic-risk state as the earliest time that any of the risk factors were documented.

We assume a forward model of heart disease progression while recognizing that it can also occur backward (*e.g.*, some participants may develop CHD before the at-metabolic-risk state). Our model can accommodate this reverse scenario by treating the transition from CHD to the at-metabolic-risk state as a new transition. However, given the low proportion of participants who develop CHD before the at-metabolic-risk state, including this reverse transition may lower study power. Thus, we classified those following this reverse transition to a forward path (*e.g.*, “healthy → CHD → at risk → mortality” was classified as the path “healthy → CHD → mortality”).

We applied our Discrete-time Split-state Framework to describe the course of heart disease. Briefly, we divided the longitudinal data into four subsets, “individuals starting from the healthy state (S_0) to the incidence of at-metabolic-risk, CHD, heart failure, or mortality, whichever outcome occurred first,” “individuals starting from the at-metabolic-risk state (S_1) to the incidence of CHD, heart failure, or mortality,” “individuals starting from the CHD state (S_2) to the incidence of heart failure or mortality,” and “individuals starting from the heart failure state (S_3) to the incidence of mortality.” Within each subset, we changed the data

Table 1 A simulation study evaluating the performance of the Discrete-time Split-state Framework under Markov, semi-Markov, and non-Markov scenarios

	Markov process with an exponential distribution	**Semi-Markov process with a Weibull distribution**	**Non-Markov process with an exponential distribution**
MSE (× 10^3)	2.60	4.52	3.04
Width of 95% CI	0.21	0.15	0.17
Coverage	94%	72%	93%

Mean squared error (MSE) was defined as the average squared difference between the estimated value and the true value. The width of the 95% confidence interval (CI) was defined as the difference between the 95% upper and lower bounds. Coverage was defined as the proportion of 95% CIs that include the true value across all simulated datasets

Table 2 A simulation study evaluating the performance of the Discrete-time Split-state Framework by disease substates under Markov, semi-Markov, and non-Markov scenarios

Markov process with exponential distribution															
	TR_{01}	TR_{02}	TR_{03}	TR_{04}	TR_{12}	TR_{13}	TR_{14}	$TR_{23\|0}$	$TR_{23\|0\_1}$	$TR_{24\|0}$	$TR_{24\|0\_1}$	$TR_{34\|0}$	$TR_{34\|0\_1}$	$TR_{34\|0\_2}$	$TR_{34\|0\_1\_2}$
MSE (× 10^3)	0.40	0.67	0.76	0.42	6.42	4.57	5.32	3.34		3.04		1.07			
Width of 95% CI	0.10	0.12	0.12	0.10	0.35	0.33	0.33	0.25		0.25		0.12			
Coverage (%)	94	91	92	93	95	95	95	94		95		92			
Semi-Markov process with Weibull distribution															
	TR_{01}	TR_{02}	TR_{03}	TR_{04}	TR_{12}	TR_{13}	TR_{14}	$TR_{23\|0}$	$TR_{23\|0\_1}$	$TR_{24\|0}$	$TR_{24\|0\_1}$	$TR_{34\|0}$	$TR_{34\|0\_1}$	$TR_{34\|0\_2}$	$TR_{34\|0\_1\_2}$
MSE (× 10^3)	0.98	1.10	1.52	0.82	10.58	4.71	7.85	3.87		3.89		9.92			
Width of 95% CI	0.08	0.09	0.09	0.08	0.20	0.18	0.19	0.28		0.28		0.08			
Coverage (%)	77	79	73	80	60	80	65	93		93		22			
Non-Markov process with exponential distribution															
	TR_{01}	TR_{02}	TR_{03}	TR_{04}	TR_{12}	TR_{13}	TR_{14}	$TR_{23\|0}$	$TR_{23\|0\_1}$	$TR_{24\|0}$	$TR_{24\|0\_1}$	$TR_{34\|0}$	$TR_{34\|0\_1}$	$TR_{34\|0\_2}$	$TR_{34\|0\_1\_2}$
MSE (× 10^3)	0.49	0.85	0.73	0.54	6.78	4.36	4.67	0.15	0.29	5.52	6.81	0.24	1.12	3.63	34.78
Width of 95% CI	0.10	0.12	0.12	0.10	0.35	0.32	0.32	0.05	0.07	0.26	0.28	0.06	0.13	0.20	0.37
Coverage (%)	94	91	93	92	95	96	97	94	95	91	92	95	94	92	80

TR transition rates. For example, $TR_{23|0}$ is the transition rate from the state 2 to state 3 with a history of state 0, and $TR_{34|0\_1\_2}$ is the transition rate from the state 3 to state 4 with a history of states 0, 1, and 2 into a long format by dividing them into small intervals by age. The fitted Models 9–12 are shown below.

To test whether the disease process is Markovian or non-Markovian, Models 11 and 12 were fitted in three ways: A) including only spline(age), B) further including past disease history J_{0_1}, and C) further including an interaction term, spline(age)×f(t). Likelihood ratio tests were conducted to choose the models with the best fit. It was found thatl, for Model 11, a non-Markov process without the interaction term best fit the data, and for Model 12, a Markov process best fit the data.

Model 9 in subset 1 starting from S_0 state: $h_{(Sm=1|S0=1)}(\tau, t) = h_{0m}(\tau)\exp(\text{spline}(\text{age}))$

Model 10 in subset 2 starting from S_1 state: $h_{(Sm=1|S1=1,S0=1)}(\tau, t) = h_{0m}(\tau)\exp(\text{spline}(\text{age}))$

Model 11 in subset 3 starting from S_2 state: $h_{(Sm=1|J0\_1,S2=1)}(\tau, t, J_{0\_1}) = h_{0m}(\tau)\exp(\text{spline}(\text{age}) + J_{0\_1})$

Model 12 in subset 4 starting from S_3 state:$h_{(Sm=1|J0\_2,S3=1)}(\tau, t, J_{0\_2}) = h_{0m}(\tau)\exp(\text{spline}(\text{age}))$

To estimate the 95% CI of transition rates, we conducted bootstrapping 1000 times, with the same sample size as the original population for each bootstrap.

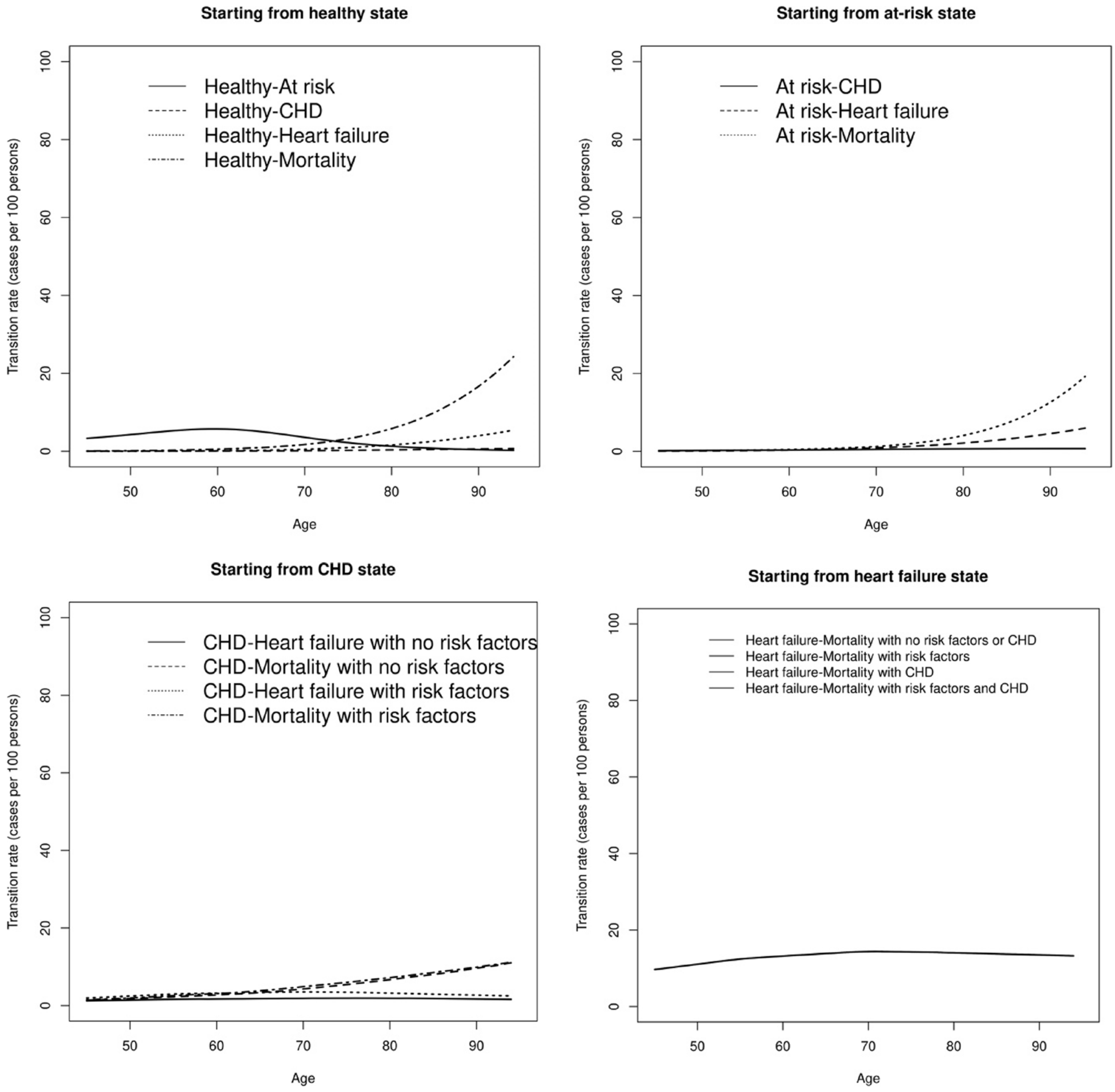


Fig. 4 Age-specific transition rates (cases per 100 persons) between the multi-states of heart disease in the Atherosclerosis Risk in Communities Study (ARIC) study

Table 3 Transition rates (cases per 100 persons) between the multi-states of heart disease in the Atherosclerosis Risk in Communities Study (ARIC) study

	Age 70 years	Age 90 years
Transition from healthy state		
Healthy → at-risk	4.11 (3.88, 4.36)	0.62 (0.46, 0.82)
Healthy → CHD	0.19 (0.14, 0.24)	1.10 (0.57, 1.94)
Healthy → heart failure	0.54 (0.45, 0.64)	4.94 (3.68, 6.60)
Healthy → mortality	1.69 (1.53, 1.86)	17.53 (15.11, 20.00)
Transition from at-risk state		
At-risk → CHD	0.53 (0.49, 0.58)	0.86 (0.70, 1.05)
At-risk → heart failure	0.87 (0.81, 0.93)	4.75 (4.25, 5.25)
At-risk → mortality	1.19 (1.12, 1.26)	12.15 (11.30, 13.02)
Transition from CHD state with no risk factors		
CHD → heart failure	2.84 (1.86, 4.09)	3.40 (1.73, 6.16)
CHD → mortality	5.46 (4.09, 7.35)	14.69 (9.32, 22.00)
Transition from CHD state with risk factors		
CHD → heart failure	4.01 (3.49, 4.51)	4.82 (2.73, 7.76)
CHD → mortality	5.18 (4.64, 5.77)	13.85 (9.61, 19.90)
Transition from heart failure state with no risk factors or CHD		
Heart failure → mortality	15.07 (14.33, 15.82)	16.95 (13.58, 21.17)
Transition from heart failure state with risk factors		
Heart failure → mortality	15.07 (14.33, 15.82)	16.95 (13.58, 21.17)
Transition from heart failure state with CHD		
Heart failure → mortality	15.07 (14.33, 15.82)	16.95 (13.58, 21.17)
Transition from heart failure state with risk factors and CHD		
Heart failure → mortality	15.07 (14.33, 15.82)	16.95 (13.58, 21.17)

Results

Our study included 15,027 participants without CHD or heart failure at baseline. During a median follow-up of 27 years, our study documented 13,043 at-metabolic-risk participants, 2565 incident cases of CHD, 3283 incident cases of heart failure, and 7677 incident cases of mortality. The transition rates from healthy, at-metabolic-risk, or CHD states to subsequent states were low in mid-age and gradually increased with age (Fig. 4). There was a sharp increase in the transition rate from healthy or at-metabolic-risk status to mortality after the age of 80. The transition rate from heart failure to mortality was high across all ages, regardless of past history of risk factors or CHD. We present the transition rates between substates at ages 70 and 90 years in Table 3.

We estimated transition probabilities from each state to the following states, starting at age 45 (Fig. 5). Participants starting from a healthy state were very likely to develop the at-metabolic-risk state over follow-up. The transition probability from healthy to at-metabolic-risk state first increased and then decreased, which was due to participants' transition to subsequent states. The risk of mortality was higher for CHD participants with a history of risk factors compared to those without. Participants with heart failure had a high risk of mortality, regardless of past history of risk factors or CHD. For example, the risk of mortality was greater than 90% at age 70 and even higher beyond that. We present the transition probabilities between substates at ages 70 and 90 in Table 4.

We estimated the state occupation probability for participants starting from healthy states at age 45. The proportion of participants in the healthy state decreased over time, while the proportion in the mortality state increased (Fig. 6). The proportion of at-metabolic-risk participants first increased and then decreased, as these participants transitioned to the next states, such as CHD, heart failure, or mortality. At each age, the proportions of participants in CHD or heart failure states were low, either due to the low incidence from previous states or the high risks of transitioning to subsequent states. For example, the proportion of participants in the healthy, at-metabolic-risk, CHD, heart failure, and mortality states at age 90 were 1.64, 11.21, 1.57, 6.15, and 79.36%, respectively (Table 5). Most participants in the CHD or heart failure states had a history of risk factors: 82% of CHD participants had risk factors ($\frac{1.29\%}{1.57\%}$=82%), and 85% of heart failure participants had risk factors or a history of CHD (1-$\frac{0.90\%}{6.15\%}$=85%).

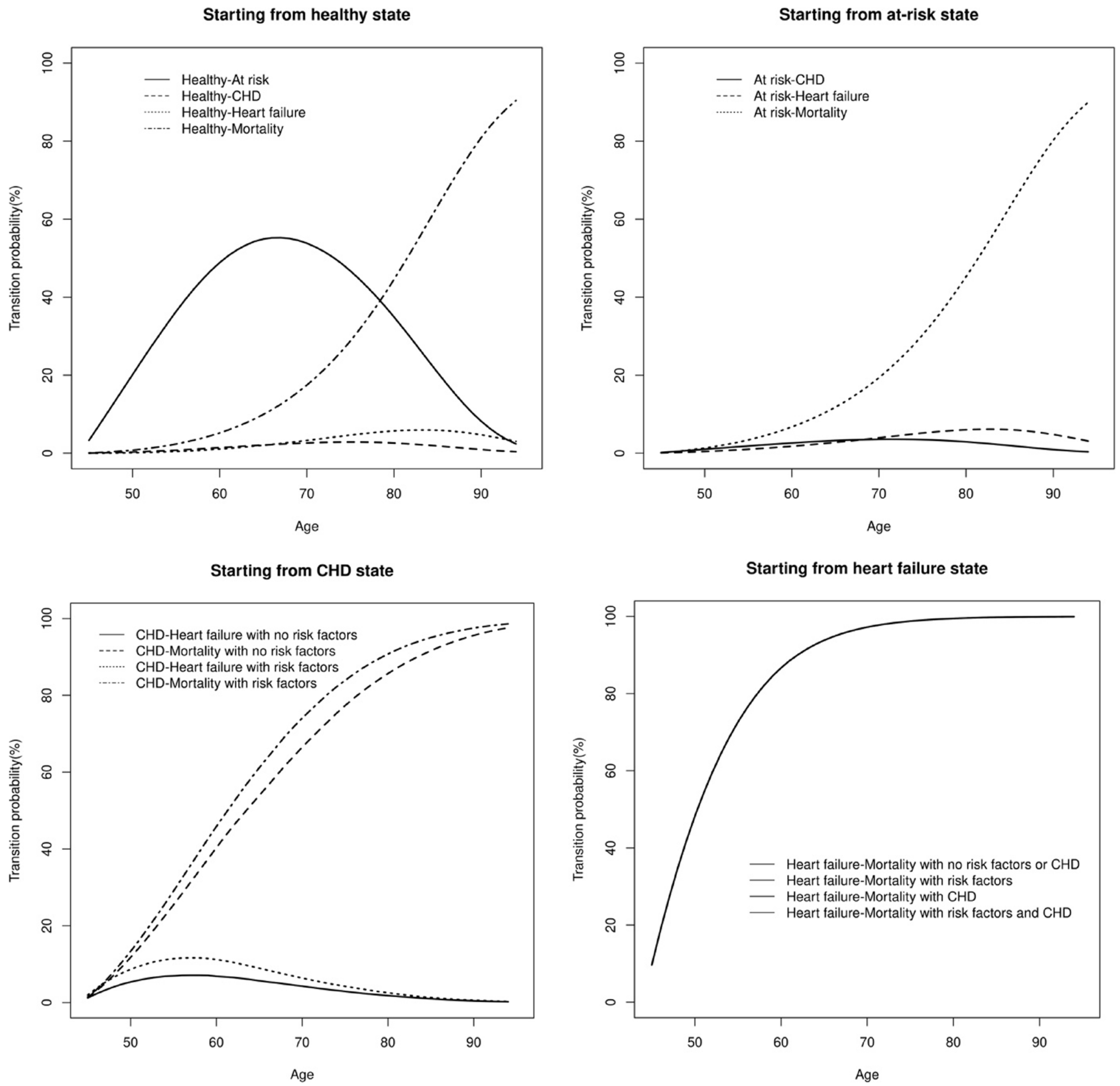


Fig. 5 Transition probabilities (%) between the multi-states of heart disease starting from age 45 years in the Atherosclerosis Risk in Communities Study (ARIC) study

Discussion

One feature of our Discrete-time Split-state Framework is the generation of a process of disease substates that satisfies the Markov assumption, regardless of whether the original disease process is Markovian. Thus, our framework can be applied to non-Markov processes where past states affect current and future states. By checking previous literature, a review paper on Markov models, published in 1999, mentioned a general form of the Markov process where the end state can be split into two states depending on past history [29]. However, this was mentioned briefly, and the author did not describe how to split the state. Additionally, the author explicitly stated that the approach was only applicable to progressive models (where each state, except the initial state, can only come from one previous state), but not to non-progressive models (where each state, except the initial state, can come from more than one previous state). In our paper, for the first time, we propose how to convert a process of disease states into a process of substates, as

Table 4 Transition probabilities (%) between the multi-states of heart disease starting from age 45 years in the Atherosclerosis Risk in Communities Study (ARIC) study

	Age 70 years	**Age 90 years**
Transition from healthy state		
Healthy → Healthy	21.48 (20.19, 22.72)	1.64 (1.34, 1.98)
Healthy → at risk	55.88 (54.57, 57.31)	11.21 (10.30, 12.20)
Healthy → CHD	2.82 (2.54, 3.13)	1.57 (1.13, 2.11)
Healthy → heart failure	3.21 (3.01, 3.44)	6.17 (5.07, 7.39)
Healthy → mortality	16.58 (15.69, 17.51)	79.36 (77.71, 80.98)
Transition from at-risk state		
At risk → At risk	73.66 (72.36, 74.72)	13.17 (12.10, 14.37)
At risk → CHD	3.88 (3.52, 4.23)	1.62 (1.12, 2.23)
At risk → heart failure	3.95 (3.71, 4.22)	6.28 (5.14, 7.58)
At risk → mortality	18.53 (17.59, 19.50)	78.90 (77.10, 80.65)
Transition from CHD state with no risk factors		
Remain in CHD state	21.01 (13.21, 29.17)	1.33 (0.36, 3.30)
CHD → heart failure	6.43 (4.56, 8.55)	0.99 (0.52, 1.83)
CHD → mortality	72.39 (64.10, 80.55)	97.62 (94.94, 99.04)
Transition from CHD state with risk factors		
Remain in CHD state	16.82 (14.27, 19.73)	0.89 (0.48, 1.56)
CHD → heart failure	7.89 (6.88, 9.01)	1.09 (0.70, 1.64)
CHD → mortality	75.24 (71.75, 78.54)	98.01 (97.11, 98.67)
Transition from heart failure state with no risk factors		
Remain in heart failure state	2.71 (2.24, 3.27)	0.08 (0.05, 0.14)
Heart failure → mortality	97.29 (96.73, 97.76)	99.92 (99.86, 99.95)
Transition from heart failure state with risk factors		
Remain in heart failure state	2.71 (2.24, 3.27)	0.08 (0.05, 0.14)
Heart failure → mortality	97.29 (96.73, 97.76)	99.92 (99.86, 99.95)
Transition from heart failure state with CHD		
Remain in heart failure state	2.71 (2.24, 3.27)	0.08 (0.05, 0.14)
Heart failure → mortality	97.29 (96.73, 97.76)	99.92 (99.86, 99.95)
Transition from heart failure state with risk factors and CHD		
Remain in heart failure state	2.71 (2.24, 3.27)	0.08 (0.05, 0.14)
Heart failure → mortality	97.29 (96.73, 97.76)	99.92 (99.86, 99.95)

shown in Fig. 1, and how to construct the transition rate matrix for the newly created process of substates. Moreover, our framework can be applied to both progressive and non-progressive models. In fact, the idea of dividing disease states into substates proposed in our paper was inspired by a fundamental concept in epidemiology: [30] Conditioning on a confounder controls for its effect on the relationship between the exposure and outcome. We creatively applied this concept to multi-state modeling and proposed the visualization of the process of disease substates in Fig. 1.

Another feature of our framework is the estimation of discrete-time transition rates as a function of duration in a state. Although Cox models can be applied to the process of newly created substates that satisfy the Markov assumption, the estimation of transition rates (or hazards) in our framework differs from Markov and semi-Markov Cox models due to two reasons. First, as the Cox model is semi-parametric, with the baseline hazard canceled out in conditional likelihood estimation [20]. Second, theoretically, although the Markov/semi-Markov Cox models can theoretically estimate the hazard nonparametrically [1, 2, 31–33], we have not yet found an R package that directly outputs the hazard. Thus, our framework estimates hazard parametrically by including the duration in a state as a covariate in cause-specific Cox models and using competing risk within a short interval to approximate hazard. Our framework has the potential to be applied to a semi-Markov process [21].

We recognize that the approximation of discrete-time hazards using competing risks may cause under coverage when modeling a continuous-time non-constant rate

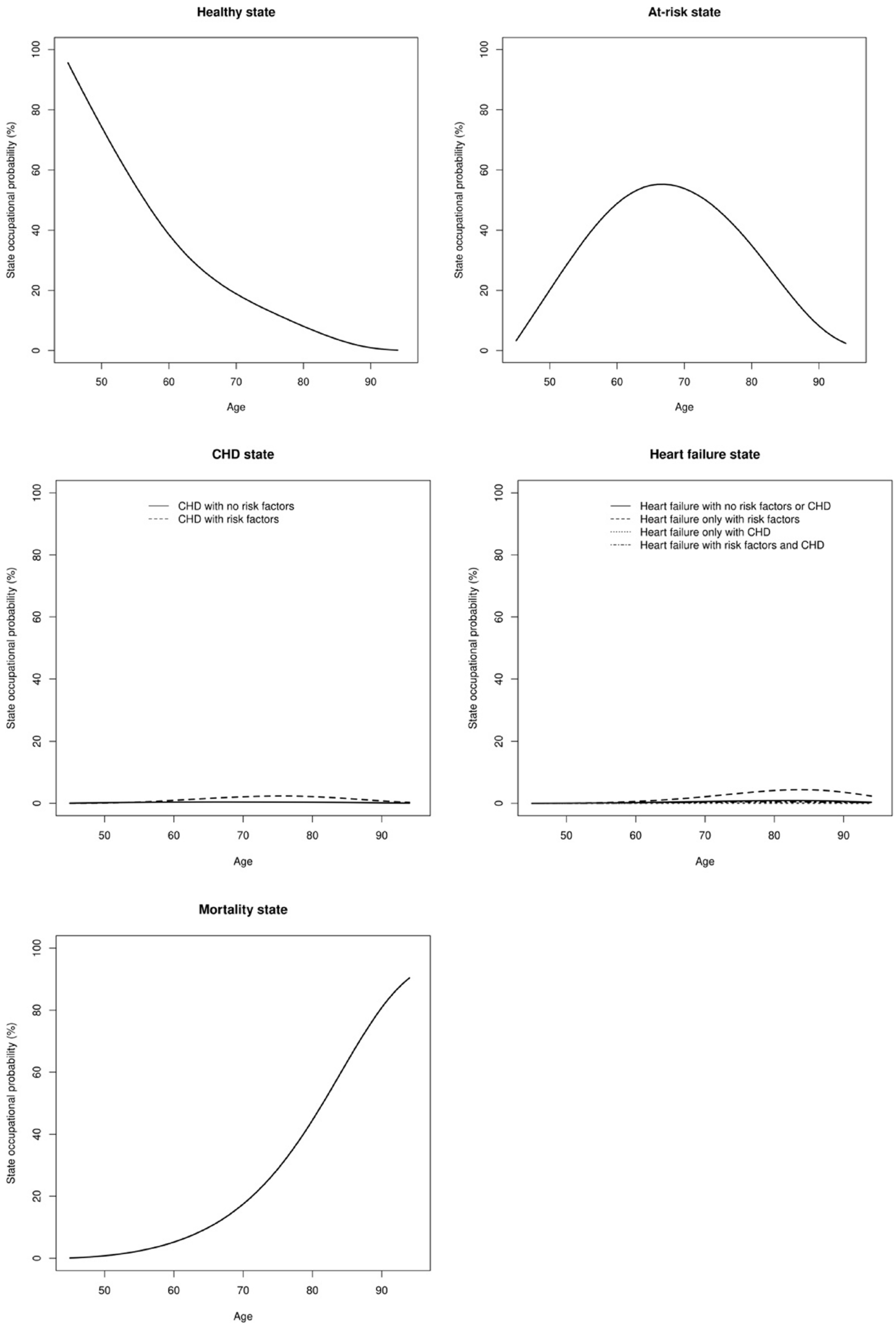


Fig. 6 State occupational probabilities (%) in each state of heart disease starting from age 45 years in the Atherosclerosis Risk in Communities Study (ARIC) study

Table 5 State occupational probabilities (%) in each state of heart disease starting from age 45 years in the Atherosclerosis Risk in Communities Study (ARIC) study

	Age 70 years	**Age 90 years**
Healthy	21.48 (20.19, 22.72)	1.64 (1.34, 1.98)
At-risk	55.88 (54.57, 57.31)	11.21 (10.30, 12.20)
CHD with no risk factors	0.62 (0.42, 0.86)	0.28 (0.16, 0.43)
CHD with risk factors	2.20 (2.00, 2.40)	1.29 (0.90, 1.75)
Heart failure with no risk factors or CHD	0.63 (0.53, 0.77)	0.90 (0.67, 1.17)
Heart failure with risk factors	2.10 (1.94, 2.25)	4.63 (3.81, 5.52)
Heart failure with CHD	0.11 (0.07, 0.17)	0.08 (0.04, 0.14)
Heart failure with risk factors and CHD	0.37 (0.32, 0.42)	0.54 (0.35, 0.80)
Mortality	16.58 (15.69, 17.51)	79.36 (77.71, 80.98)

[21]. This may explain the under coverage of the semi-Markov process with a Weibull distribution. Moreover, although we attempted to flexibly model duration in a disease state using restricted spline functions, modeling duration as counting numbers may cause model misspecification and reduce the coverage rate of the estimated transition rates in the simulation study.

The transition rates between substates estimated using our framework have three significances. First, our framework is suitable for modeling the course of chronic disease, where the Markov assumption is often violated [5–7]. The transition rates describe the dynamics of chronic disease progression and enhance our understanding of disease mechanisms. Second, regardless of whether the original process is Markovian, the generated disease substates follow the Markov assumption. Thus, the Aalen-Johansen estimator can be applied to estimate transition probabilities between substates based on transition rates (i.e., the Aalen-Johansen estimator is systematically biased when estimating transition probabilities between any two random time points in a non-Markov process) [10, 12, 34, 35]. Third, while the process of disease substates satisfies the Markov assumption, the substates are memorable—they track past disease states and inform multimorbidity. Thus, transition rates between substates can be used to derive summary metrics that characterize the course of chronic disease. These summary metrics can be applied to comparative effectiveness research and the prediction of disease course, which are interesting directions for future research.

We used a forward transition to illustrate our framework, as this is the case for most chronic diseases. However, our framework can also deal with backward transitions (i.e., a state can transition back to a previous state) by converting them to forward transitions. For example, for $S_0 \leftrightarrows S_1$, the S_0 state can be divided into two states: one where S_0 occurred before S_1, and one where S_0 occurred after S_1. In this way, a backward transition is transformed into a forward transition (S_0 before $S_1 \rightarrow S_1 \rightarrow S_0$ after S_1), allowing our framework to be applied.

In summary, we proposed a Discrete-time Split-state Framework that allows past states to influence the current state and enables transition rates to depend on the duration in a state. Our framework is well-suited for modeling the course of chronic diseases where the Markov assumption is often violated, and the estimated transition parameters may enhance our understanding of the mechanisms of disease progression. For future research, the transition rates between substates generated using our framework have the potential to be used for deriving summary estimates that characterize the disease course, which could stimulate new approaches for the precision prevention and prediction of chronic diseases.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02512-6.

Supplementary Material 1.

Acknowledgements

Dr. Ming Ding is supported by National Institutes of Health R21 HG012365, and Dr. Feng-Chang Lin is supported by National Center for Advancing Translational Sciences (NCATS) UM1TR004406.

Authors' contributions

M.D. generated the idea of the Discrete-time Split-state Framework, conducted statistical analyses, and wrote the manuscript. H.C. and F.L. provided highly constructive feedback and edited the manuscript. All authors reviewed the manuscript.

Funding

Dr. Ming Ding is supported by National Institutes of Health R21 HG012365, and Dr. Feng-Chang Lin is supported by National Center for Advancing Translational Sciences (NCATS) UM1TR004406.

Data availability
The R code for the Discrete-time Split-State Framework is available at the GitHub repository (https://github.com/mingding-hsph/multistate_framework).

Declarations

Ethics approval and consent to participate
We obtained ARIC data through BioLINCC, an open repository created by NHLBI, with the IRB deemed exempt by the University of North Carolina (UNC)-Chapel Hill Review Board.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 June 2024 Accepted: 18 February 2025
Published online: 28 February 2025

References
1. Meira-Machado L, de Una-Alvarez J, Cadarso-Suarez C, Andersen PK. Multi-state models for the analysis of time-to-event data. Stat Methods Med Res. 2009;18(2):195–222. https://doi.org/10.1177/0962280208092301. Epub 20080618 PubMed PMID: 18562394; PMCID: PMC2692556.
2. de Wreede LC, Fiocco M, Putter H. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. Comput Methods Programs Biomed. 2010;99(3):261–74. https://doi.org/10.1016/j.cmpb.2010.01.001. (Epub 20100315. PubMed PMID: 20227129).
3. Król A, Saint-Pierre P. SemiMarkov: An R Package for Parametric Estimation in Multi-State Semi-Markov Models. J Stat Softw. 2015;66(6):1–16.
4. Asanjarani A, Liquet B, Nazarathy Y. Estimation of semi-Markov multistate models: a comparison of the sojourn times and transition intensities approaches. Int J Biostat. 2021;18(1):243–62. https://doi.org/10.1515/ijb-2020-0083. (Epub 20210106 PubMed PMID: 35641138).
5. Drozd M, Relton SD, Walker AMN, Slater TA, Gierula J, Paton MF, Lowry J, Straw S, Koshy A, McGinlay M, Simms AD, Gatenby VK, Sapsford RJ, Witte KK, Kearney MT, Cubbon RM. Association of heart failure and its comorbidities with loss of life expectancy. Heart. 2021;107(17):1417–21. https://doi.org/10.1136/heartjnl-2020-317833. (Epub 20201105. PubMed PMID: 33153996; PMCID: PMC8372397).
6. Khan MS, Samman Tahhan A, Vaduganathan M, Greene SJ, Alrohaibani A, Anker SD, Vardeny O, Fonarow GC, Butler J. Trends in prevalence of comorbidities in heart failure clinical trials. Eur J Heart Fail. 2020;22(6):1032–42. https://doi.org/10.1002/ejhf.1818. (Epub 2020/04/16. PubMed PMID: 32293090; PMCID: PMC7906002).
7. Tran J, Norton R, Conrad N, Rahimian F, Canoy D, Nazarzadeh M, Rahimi K. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: A population-based cohort study. PLoS Med. 2018;15(3):e1002513. https://doi.org/10.1371/journal.pmed.1002513. (Epub 20180306 PubMed PMID: 29509757; PMCID: PMC5839540).
8. Putter H, Spitoni C. Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen-Johansen estimator. Stat Methods Med Res. 2018;27(7):2081–92. https://doi.org/10.1177/0962280216674497. (Epub 2018/05/31 PubMed PMID: 29846146).
9. de Una-Alvarez J, Meira-Machado L. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. Biometrics. 2015;71(2):364–75. https://doi.org/10.1111/biom.12288. (Epub 2015/03/05. PubMed PMID: 25735883).
10. Pepe MS. Inference for Events With Dependent Risks in Multiple Endpoint Studies. J Am Stat Assoc. 1991;86(415):770–8.
11. Andersen PK, Wandall ENS, Pohar Perme M. Inference for transition probabilities in non-Markov multi-state models. Lifetime Data Anal. 2022;28(4):585–604. https://doi.org/10.1007/s10985-022-09560-w. (Epub 20220628 PubMed PMID: 35764854).
12. Meira-Machado L, de Una-Alvarez J, Cadarso-Suarez C. Nonparametric estimation of transition probabilities in a non-Markov illness-death model. Lifetime Data Anal. 2006;12(3):325–44. https://doi.org/10.1007/s10985-006-9009-x. (Epub 2006/08/19. PubMed PMID: 16917736).
13. Glidden DV. Robust inference for event probabilities with non-Markov event data. Biometrics. 2002;58(2):361–8. https://doi.org/10.1111/j.0006-341x.2002.00361.x. (PubMed PMID: 12071409).
14. Meira-Machado L, Sestelo M. Estimation in the progressive illness-death model: A nonexhaustive review. Biom J. 2019;61(2):245–63. https://doi.org/10.1002/bimj.201700200. (Epub 20181120. PubMed PMID: 30457674).
15. Maltzahn N, Hoff R, Aalen OO, Mehlum IS, Putter H, Gran JM. A hybrid landmark Aalen-Johansen estimator for transition probabilities in partially non-Markov multi-state models. Lifetime Data Anal. 2021;27(4):737–60. https://doi.org/10.1007/s10985-021-09534-4. (Epub 20210930. PubMed PMID: 34595580; PMCID: PMC8536588).
16. Hoff R, Putter H, Mehlum IS, Gran JM. Landmark estimation of transition probabilities in non-Markov multi-state models with covariates. Lifetime Data Anal. 2019;25(4):660–80. https://doi.org/10.1007/s10985-019-09474-0. (Epub 20190417. PubMed PMID: 30997582).
17. Bacchetti P, Boylan RD, Terrault NA, Monto A, Berenguer M. Non-Markov multistate modeling using time-varying covariates, with application to progression of liver fibrosis due to hepatitis C following liver transplant. Int J Biostat. 2010;6(1):7. https://doi.org/10.2202/1557-4679.1213. (Epub 20100220. PubMed PMID: 20305705; PMCID: PMC2836212).
18. Jackson C. Multi-state models for panel data: the msm package for R. J Statist Software. 2011;38(8):1–28. https://doi.org/10.18637/jss.v038.i08.
19. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol. 2009;170(2):244–56. https://doi.org/10.1093/aje/kwp107. (Epub 20090603. PubMed PMID: 19494242; PMCID: PMC2732996).
20. Cox DR. Regression Models and Life-Tables. J Roy Stat Soc: Ser B (Methodol). 1972;34(2):187–220.
21. Hernán MA, Robins JM. Causal Inference: What If? Boca Raton: Chapman Hall/CRC; 2021. p. 221–34.
22. Gerds TA, Ohlendorff JS, Blanche P, Mortensen R, Wright M, Tollenaar N, Muschelli J, Mogensen UB. B. O. Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks. https://github.com/tagteam/riskregression. 2023.
23. Gray RG. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. Ann Stat. 1988;16(3):1141–54.
24. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am J Epidemiol. 1989;129(4):687–702. Epub 1989/04/01. PubMed PMID: 2646917.
25. Fung TT, Chiuve SE, McCullough ML, Rexrode KM, Logroscino G, Hu FB. Adherence to a DASH-style diet and risk of coronary heart disease and stroke in women. Arch Internal Med. 2008;168(7):713–20. https://doi.org/10.1001/archinte.168.7.713. (Epub 2008/04/17. PubMed PMID: 18413553).
26. Unger T, Borghi C, Charchar F, Khan NA, Poulter NR, Prabhakaran D, Ramirez A, Schlaich M, Stergiou GS, Tomaszewski M, Wainford RD, Williams B, Schutte AE. 2020 International Society of Hypertension Global Hypertension Practice Guidelines. Hypertension. 2020;75(6):1334–57. https://doi.org/10.1161/HYPERTENSIONAHA.120.15026. (Epub 20200506. PubMed PMID: 32370572).
27. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, Braun LT, de Ferranti S, Faiella-Tommasino J, Forman DE, Goldberg R, Heidenreich PA, Hlatky MA, Jones DW, Lloyd-Jones D, Lopez-Pajares N, Ndumele CE, Orringer CE, Peralta CA, Saseen JJ, Smith SC Jr, Sperling L, Virani SS, Yeboah J. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2019;139(25):e1046–81. https://doi.org/10.1161/CIR.0000000000000624. (Epub 20181110. PubMed PMID: 30565953).
28. American Diabetes A. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2010;33 Suppl 1(Suppl 1):S62–9. https://doi.org/10.2337/dc10-S062. PubMed PMID: 20042775; PMCID: PMC2797383.
29. Hougaard P. Multi-state Models: A Review. Lifetime Data Anal. 1999;5:239–64.

30. Rothman KJ, Greenland S, Lash T. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 5–31.
31. Aalen OO, Johansen S. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. Scand J Stat. 1978;5:141–50.
32. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc. 1958;53:457–81.
33. Strauss D, Shavelle R. An extended Kaplan-Meier estimator and its applications. Stat Med. 1998;17(9):971–82. https://doi.org/10.1002/(sici)1097-0258(19980515)17:9%3c971::aid-sim786%3e3.0.co;2-q. (PubMed PMID: 9612885).
34. Datta S, Satten GA. Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. Statist Probab Lett. 2001;55(4):403–11.
35. Dattaa S, Sattenb GA. Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. Stat Probability Lett. 2001;55:403–11.

Publisher's Note