JOURNAL OF FUNDAMENTAL MATHEMATICS
AND APPLICATIONS (JFMA) VOL. 8 NO. 2 (2025)
Available online at www.jfma.math.fsm.undip.ac.id

# REGRESSION ANALYSIS FOR MULTISTATE MODELS USING TIME DISCRETIZATION WITH APPLICATIONS TO PATIENTS' HEALTH STATUS

Rianti Siswi Utami[1,*], Adhitya Ronnie Effendie[2], Danardono Danardono[3]

[1,2,3]*Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia*
Email : [1]riantisiswi.u@ugm.ac.id, [2]adhityaronnie@ugm.ac.id [3]danardono@ugm.ac.id
*Corresponding Author

**Abstract.** This paper addresses the estimation of multistate models in discrete time, which are widely used to describe complex event histories involving transitions between multiple health states. Accurate estimation of transition intensities and probabilities is essential for understanding disease progression and evaluating the impact of covariates. However, conventional estimators such as the Nelson–Aalen estimator often produce rough estimates, especially in sparse data settings. To improve estimation, we apply kernel smoothing to Nelson–Aalen estimators of transition intensities. Transition probabilities are then derived via product-integrals of the smoothed intensities. Covariate effects on transition intensities are modeled using the Cox proportional hazards model. Rather than modeling covariate effects on transition probabilities indirectly through their influence on transition intensities, we model them directly using pseudo-values of state occupation probabilities obtained through a jackknife procedure. These pseudo-values are treated as outcome variables in a Generalized Estimating Equation (GEE) framework. The proposed methodology is applied to patient visit data from a clinic in West Java, Indonesia, where it successfully captures both the progression dynamics across health states and the influence of key covariates.
**Keywords:** Multistate model; Transition intensities; Transition probabilities; State occupation probabilities; Generalized Estimating Equation (GEE).

## I. INTRODUCTION

In event history analysis, individuals are followed over time and the occurrence of events is recorded, where events may be transient (such as disease onset or recovery) or terminal (such as death). Multistate models provide a flexible and powerful framework for analyzing such complex longitudinal processes by representing individual trajectories as transitions between a finite number of discrete states [1,2]. These models have been widely applied in medical [3], epidemiological [4], and actuarial studies [5] to describe disease progression, treatment response, and health dynamics over time.

Multistate models are typically illustrated using state-transition diagrams, where states are represented by boxes and transitions between states are indicated by arrows. These diagrams provide an intuitive representation of the underlying stochastic process and help clarify the structure and assumptions of the model. For example, Figure 1 presents a three-state illness–death model consisting of the states: 1 (healthy), 2 (diseased), and 3 (dead). The complexity of a multistate model depends on both the number of states and the set of allowable transitions among them.

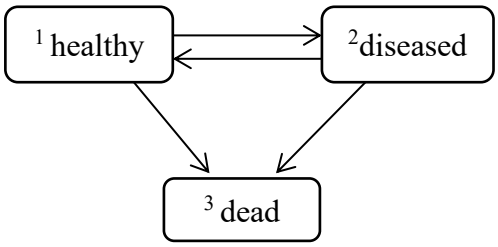


**Figure 1.** The illness-death model

Classical inference in multistate models is commonly based on continuous-time Markov processes, where transition intensities are estimated nonparametrically using the Nelson–Aalen estimator and transition probabilities are subsequently obtained via product-integrals [6]. Although this framework is theoretically well established, practical applications often encounter challenges when event times are sparse, irregular, or observed only at discrete visit times [1]. In such situations, the resulting step-function estimates of transition intensities can be unstable and difficult to interpret, which may propagate uncertainty into the estimated transition probabilities [7,8].

Recent developments in multistate modeling have focused on improving estimation efficiency [9], relaxing model assumptions [10], and enhancing interpretability of covariate effects [11]. Several studies published in the past few years have extended multistate models to incorporate flexible hazard structures [12], time-dependent covariates [13], and complex censoring mechanisms [14]. However, most of these approaches still rely on continuous-time formulations and model covariate effects on transition probabilities indirectly through transition intensities. This indirect relationship often leads to highly nonlinear and model-dependent interpretations, particularly when interest lies in marginal state occupation probabilities rather than instantaneous transition risks [15].

An alternative line of research has introduced pseudo-value regression methods for directly modeling functionals of multistate processes, such as cumulative incidence functions or state occupation probabilities. Pseudo-values, typically constructed via jackknife procedures [15], allow these quantities to be treated as response variables within generalized estimating equation (GEE) frameworks [16]. While this approach offers improved interpretability and flexibility, its application has largely been restricted to continuous-time settings and has not been fully explored in discrete-time multistate models, especially in combination with smoothed estimators of transition intensities.

Despite the increasing availability of longitudinal data collected at discrete and irregular observation times—such as clinical visit data—there remains a lack of integrated methodology that simultaneously addresses instability in transition intensity estimation and enables direct regression modeling of state occupation probabilities in discrete-time multistate models. This gap is particularly relevant in applied health studies, where marginal probabilities of occupying specific health states are often of primary interest to clinicians and policymakers.

To address these limitations, this study proposes a discrete-time multistate modeling framework that combines kernel-smoothed Nelson–Aalen estimators for transition intensities [17] with direct regression modeling of state occupation probabilities using pseudo-values and GEE. Kernel smoothing is applied to stabilize and improve the interpretability of transition intensity estimates derived from sparse discrete-time data. Covariate effects are incorporated

using two complementary approaches: Cox proportional hazards models for transition intensities [18] and a pseudo-value-based GEE framework for direct inference on state occupation probabilities [15,16]. This dual strategy allows for both instantaneous and marginal interpretations of covariate effects.

The proposed methodology is applied to patient morbidity data from a clinic in West Java, Indonesia. The application demonstrates how the approach captures dynamic health state progression while providing interpretable covariate effects on both transition risks and state occupation probabilities. By integrating smoothing techniques and pseudo-value regression within a discrete-time multistate framework, this study contributes a practical and interpretable modeling strategy for analyzing complex longitudinal health data.

## II. MODEL AND METHODS

### 2.1. Multistate Models

A multistate process is a stochastic process $(X(t), t \in T)$ with a finite state space $S = \{1, \dots, N\}$. Here, $T = [0, \tau]$, where $\tau < \infty$, represents a bounded time interval. As the process evolves over time, it generates a history $H_t-$, which consists of observations of the process over the interval $[0, t)$, including the states visited and the times of transitions [8]. The multi-state process is fully characterized through transition probabilities between states $h$ and $j$

$$P_{hj}(s,t) = \text{pr}(X(t) = j | X(s) = h, H_s-) \tag{1}$$

where $h, j \in S,\ s, t \in T$ and $s \leq t$ or through transition intensities,

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t, t + \Delta t)}{\Delta t} \tag{2}$$

which represent the instantaneous hazard of transitioning to state $j$ given that the process is in state $h$ at time $t$. In most applications, a Markovian structure is assumed, meaning that $P_{hj}(s,t)$ depends on the history only through the covariates and the state $X(s) = h$ occupied at time $s$.

We are also interested in modeling and inference for the state occupation probabilities, defined as $Q_h(t) = \text{pr}[X(t) = h]$, where $h \in S$. Given the initial state probabilities $Q_j(0)$, the state occupation probabilities at time $t$ can be expressed as,

$$Q_h(t) = \sum_{j \in S} Q_j(0) P_{jh}(0, t). \tag{3}$$

### 2.2 Estimation in Multistate Models with no Covariates

When a Markov assumption is made, the transition probabilities can be estimated as product–integrals of the Nelson–Aalen estimators of the transition intensities [6]. Let $N_{hj}(t)$ denote the total number of transitions from state $h$ to state $j$ in the interval $[0, t]$, and let $Y_h(t)$

represent the number of individuals in state $h$ at time $t^-$. The estimated cumulative transition intensity from state $h$ to state $j$ is given by:

$$\hat{A}_{hj}(t) = \int_0^t \frac{I[Y_h(s) \neq 0]}{Y_h(s)} \mathrm{d}N_{hj}(s), \tag{4}$$

where $I(\cdot)$ is the indicator function. The cumulative transition intensities for all possible transitions between states can be organized into a matrix $\hat{A}(t)$,

$$\hat{A}(t) = \begin{bmatrix} \hat{A}_{11}(t) & \hat{A}_{12}(t) & \cdots & \hat{A}_{1N}(t) \\ \hat{A}_{21}(t) & \hat{A}_{22}(t) & \cdots & \hat{A}_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{A}_{N1}(t) & \hat{A}_{N2}(t) & \cdots & \hat{A}_{NN}(t) \end{bmatrix} \tag{5}$$

whose rows sum to 0, so that the diagonal entries are $\hat{A}_{hh}(t) = -\sum_{j \neq h} \hat{A}_{hj}(t)$. The transition probability matrix is estimated as the product-integrals of Nelson-Aalen estimator,

$$\hat{P}[s,t] = \prod_{(s,t]} \left[I + \mathrm{d}\hat{A}(\mathrm{u})\right] \tag{6}$$

$$\hat{P}[s,t] = \begin{bmatrix} \hat{P}_{11}(t) & \hat{P}_{12}(t) & \cdots & \hat{P}_{1N}(t) \\ \hat{P}_{21}(t) & \hat{P}_{22}(t) & \cdots & \hat{P}_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{P}_{N1}(t) & \hat{P}_{N2}(t) & \cdots & \hat{P}_{NN}(t) \end{bmatrix} \tag{7}$$

where $I$ is the $N \times N$ identity matrix and $\prod$ is the product–integral [10].

In discrete time approach, the time interval is divided into subintervals, $s < t_1 < \cdots < t_{k-1} < t_k < \cdots < t_L \leq t$. The Nelson-Aalen estimator in Equation (4) becomes,

$$\hat{A}_{hj}(t) = \begin{cases} 0 & \text{if } t < t_1 \\ \sum_{t_k \leq t} \dfrac{\Delta N_{hj}(t_k)}{Y_h(t_k)} & \text{if } t_k \leq t \end{cases} \tag{8}$$

where $k = 1,2,\ldots,L$ and $\Delta N_{hj}(t_k)$ is the number of transitions from state $h$ to state $j$ at time $t_k$. The transition probability matrix is then estimated as,

$$\hat{P}[s,t] = \prod_{k=1}^{L} \left[I + \Delta\hat{A}(t_k)\right]. \tag{9}$$

The Nelson–Aalen estimator $\hat{A}_{hj}(t)$ provides an efficient nonparametric estimate of the cumulative transition intensity $A_{hj}(t)$. In discrete time, the change in the estimator at each time point $t_k$ is defined as $\Delta\hat{A}_{hj}(t_k) = \hat{A}_{hj}(t_k) - \hat{A}_{hj}(t_{k-1})$. In many applications, the

parameter of interest is not $A_{hj}(t)$, but rather its derivative $a_{hj}(t)$, which represents the transition intensity. The discrete increment $\Delta\hat{A}_{hj}(t_k)$ can be viewed as a crude estimate of $a_{hj}(t)$. However, when event times are sparse or unevenly distributed, this step function estimate can be unstable and difficult to interpret. To overcome this limitation, we apply kernel smoothing using Epanechnikov kernel, which assigns greater weight to time points closer to $t$. The smoothed estimator of the transition intensity is defined as,

$$\hat{\alpha}_{hj}(t) = b^{-1}\sum_{k=1}^{L} K\left(\frac{t-t_k}{b}\right)\Delta\hat{A}_{hj}(t_k) \tag{10}$$

for $b \le t \le t_{L-b}$, where $b$ is the bandwith, and $K(\cdot)$ is a kernel function. We use Epanechnikov kernel,

$$K(x) = 0.75(1-x^2) \tag{11}$$

for $-1 \le x \le 1$. When $t < b$, an asymmetric kernel is used to accommodate boundary effects [19].

### 2.3 Covariates Effect on Multistate Model

A common statistical approach for analyzing the effect of independent variables (covariates) on a dependent variable is regression analysis. In multistate modeling, covariate effects on state occupation or transition probabilities are typically examined indirectly by modeling the transition intensities. However, this approach often results in highly nonlinear and complex relationships between covariates and the resulting transition or occupation probabilities [15]. To address this, we analyze the effect of covariates on transition intensities and transition probabilities separately. Transition intensities are modeled using the Cox proportional hazards model, while transition probabilities are modeled directly through a regression framework based on pseudo-values.

We use the Cox regression model for modeling transition intensities, stratified by the type of transition,

$$a_{hji}(t|Z) = a_{hj,0}(t)\exp(\beta^{\mathrm{T}}Z_i) \tag{12}$$

where $i = 1,2,\dots,n$, $a_{hj0}(t)$ is the baseline transition intensity, which does not depend on covariate, $Z_i$ is the covariate vector, and $\beta$ is the corresponding vector of regression coefficients. The baseline transition intensity $a_{hj,0}(t)$ is estimated nonparametrically using the kernel-smoothed estimator $\hat{\alpha}_{hj}(t)$ in Equation (10) based on the discrete-time increments of the Nelson–Aalen estimator.

To analyze the effect of covariates on state occupation probabilities, we use pseudo-values as response variables in a Generalized Estimating Equation (GEE) framework. Let

$$\theta = \left(Q_h(t_1), Q_h(t_2), \dots, Q_h(t_L)\right) \tag{13}$$

be a vector of state occupation probability in state $h$ at discrete time points $t_1, \dots, t_L$. For individual $i$, let

$$\hat{\theta}_i = \left(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \dots, \hat{\theta}_{iL}\right) \tag{14}$$

be the vector of pseudo-values at time points $t_k, t_2, \dots, t_L$, where each $\hat{\theta}_{ik}$ represents the pseudo-value of being in particular state $h$ at time $t_k$. These pseudo-values are derived from a jackknife statistic [2]

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i} \tag{15}$$

where $\hat{\theta}$ is the full-sample estimator and $\hat{\theta}_{-i}$ is the estimator based on the sample of size with the $i$-th observation excluded. In the absence of censoring, the estimator $\hat{\theta}$ at time $t_k$ simplifies to the proportion of individuals in state $h$ at $t_k$. In such cases, the pseudo-value for subject $i$ at time $t_k$ is simply the indicator function that equals 1 if the subject is in state $h$ at time $t_k$, and 0 otherwise.

Although $\hat{\theta}_i$ represents a pseudo-value derived from estimated probabilities, its value is not restricted to the interval [0,1]. This is because the jackknife formula in Equation (15) can yield values slightly below 0 or above 1, especially in small samples or when an individual observation has a large influence on the overall estimate. This behavior is well-recognized and does not pose a problem for statistical inference. The pseudo-values remain asymptotically unbiased and are suitable as outcome variables in the GEE framework. If necessary, link functions such as the logit or probit may be applied in the modeling step to ensure that fitted values remain within the probability bounds [2].

To model the relationship between covariates and these pseudo-values, we specify a marginal model

$$g(\theta_{ik}) = \gamma^{\mathrm{T}} Z_{ik} \tag{16}$$

$$\theta_{ik} = g^{-1}(\gamma^{\mathrm{T}} Z_{ik}) \tag{17}$$

where $g(\cdot)$ is a suitable link function, $\gamma$ is a vector of regression coefficients, and $Z_{ik}$ is the covariate vector for individual $i$ at time $t_k$, which includes both time-specific indicators and individual-level covariates. which includes both time-specific indicators and individual-level covariates. This structure allows the model to capture both time-varying and subject-specific effects on the state occupation probabilities. In matrix form, the model for all individuals and time points becomes

$$g(\theta) = \begin{bmatrix} g(\theta_{11}) & g(\theta_{12}) & \cdots & g(\theta_{1L}) \\ g(\theta_{21}) & g(\theta_{22}) & \cdots & g(\theta_{2L}) \\ \vdots & \vdots & \ddots & \vdots \\ g(\theta_{n1}) & g(\theta_{n2}) & \cdots & g(\theta_{nL}) \end{bmatrix} \tag{18}$$

$$= \begin{bmatrix} \gamma^{\mathrm{T}} Z_{11} & \gamma^{\mathrm{T}} Z_{12} & \cdots & \gamma^{\mathrm{T}} Z_{1L} \\ \gamma^{\mathrm{T}} Z_{21} & \gamma^{\mathrm{T}} Z_{22} & \cdots & \gamma^{\mathrm{T}} Z_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^{\mathrm{T}} Z_{n1} & \gamma^{\mathrm{T}} Z_{n1} & \cdots & \gamma^{\mathrm{T}} Z_{nL} \end{bmatrix}.$$

The regression coefficients $\gamma$ are estimated by solving the unbiased estimating equations,

$$\sum_i \left( \frac{\partial}{\partial \gamma} g^{-1}(\gamma^{\mathrm{T}} Z_i) \right)^{\mathrm{T}} V_i^{-1} \left( \hat{\theta}_i - g^{-1}(\gamma^{\mathrm{T}} Z_i) \right) = \sum_i U_i(\gamma) = U(\gamma) = 0 \qquad (19)$$

where $V_i$ is a working covariance matrix, and $g^{-1}(\gamma^{\mathrm{T}} Z_i)$ represents the vector of fitted values on the probability scale for individual $i$, obtained by applying the inverse link function to the linear predictors at each time point.

## III. DATA APPLICATION

### 3.1 Analysis of Patients Morbidity Data

This method is applied to estimate a multistate model for patients' morbidity at Cihideung Clinic, located in Garut, West Java, Indonesia. A sample of 100 patients was randomly selected. The data were collected from patient visits between January 1, 2002 and January 6, 2007. The collected information includes patient ID, visit time, age (in years), systolic blood pressure (mmHg), diastolic blood pressure (mmHg), *Koch Pulmonum*[1] (KP, where 1 = positive and 0 = negative), sex (1 = male, 0 = female), and medical cost (in IDR).

State of patients are divided into four groups based on the cost, namely, Patient states are classified into four categories based on medical cost:

- State 1: Healthy
- State 2: Diseased type 1 (cost ≤ 25,000 IDR)
- State 3: Diseased type 2 (25,000 < cost ≤ 50,000 IDR)
- State 4: Diseased type 3 (cost > 50,000 IDR).

A higher medical cost is assumed to reflect greater disease severity, indicating that patients in higher-numbered states experience more severe health conditions.

The multistate model used is based on the structure proposed by Effendie [20], as illustrated in Figure 2.

[1]*Koch Pulmonum* also known as pulmonary tuberculosis (TB), is a bacterial infection of the lungs caused by *Mycobacterium tuberculosis*. It primarily affects the respiratory system and is characterized by symptoms such as chronic cough, chest pain, and difficulty breathing.

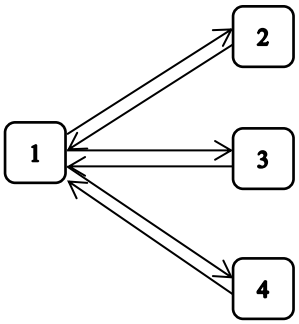


**Figure 2.** The multistate model for morbidity data

The time unit is measured in days. Data are only observed during clinic visits, so the exact time of transition to the healthy state is interval-censored—i.e., known only to lie between two clinic visits. It is assumed that if a patient does not visit the clinic within one week after their last recorded visit, the patient is considered healthy. The process is assumed to follow the Markov property, allowing direct transitions between healthy and diseased states. If a patient revisits the clinic within one week, the associated costs are accumulated, and the new state is determined based on the total cost. For modeling purposes, each patient's first visit since January 1, 2002, is treated as time $t_1$.

### 3.2 Estimation with no Covariate

Based on the model structure in Figure 2, the Nelson–Aalen estimator matrix can be written as

$$\hat{A}(t) = \begin{bmatrix} \hat{A}_{11}(t) & \hat{A}_{12}(t) & \hat{A}_{13}(t) & \hat{A}_{14}(t) \\ \hat{A}_{21}(t) & \hat{A}_{22}(t) & 0 & 0 \\ \hat{A}_{31}(t) & 0 & \hat{A}_{33}(t) & 0 \\ \hat{A}_{41}(t) & 0 & 0 & \hat{A}_{44}(t) \end{bmatrix}.$$

Before computing the Nelson–Aalen estimators, we determine the state evolution of patients from time $t_0$ to $T$. Table 1 presents the state evolution for 5 sample patients. State 0 indicates that the patient no longer visits the clinic. For this study, we observe patient trajectories up to 2 years ($T = 730$).

**Table 1.** State evolution of 5 patients

| Patient | Time | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 8 | 9 | 10 | … | 1430 | 1525 | 1532 | 1598 | 1831 |
| 1 | 1 | 3 | 1 | 1 | 1 | … | 1 | 2 | 1 | 2 | 0 |
| 2 | 1 | 3 | 3 | 3 | 3 | … | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 3 | 3 | 3 | 3 | … | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 2 | 1 | 1 | 1 | … | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 2 | 1 | 1 | 1 | … | 0 | 0 | 0 | 0 | 0 |

Using these state trajectories, we compute $Y_h(t_k)$ and $N_{hj}(t_k)$, then estimate the Nelson–Aalen increments $\Delta\hat{A}(t_k)$. Below are estimators at times 0, 1, and 8,

$$\Delta\hat{A}(0) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\Delta\hat{A}(1) = \begin{bmatrix} -1 & 0.55 & 0.29 & 0.16 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\Delta\hat{A}(8) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0.517 & 0 & -0.517 & 0 \\ 0.125 & 0 & 0 & -0.125 \end{bmatrix}.$$

To smooth the estimators, we use the Epanechnikov kernel with a bandwidth of 100. The smoothed estimators are then used to compute cumulative transition probabilities. The focus on transitions from State 1 arises because all patients begin in the healthy state at their first recorded visit, making it the natural starting point for estimating transition dynamics. Figure 3 presents the cumulative transition probabilities from State 1 to other states, which can be interpreted as occupation probabilities conditional on the initial state being 1. The figure shows that the probability of remaining in the healthy state remains high throughout the observation period. Transitions to more severe disease states (States 2, 3, and 4) are relatively rare and occur with decreasing frequency as severity increases, indicating that most patients experience mild or no deterioration in health over time.

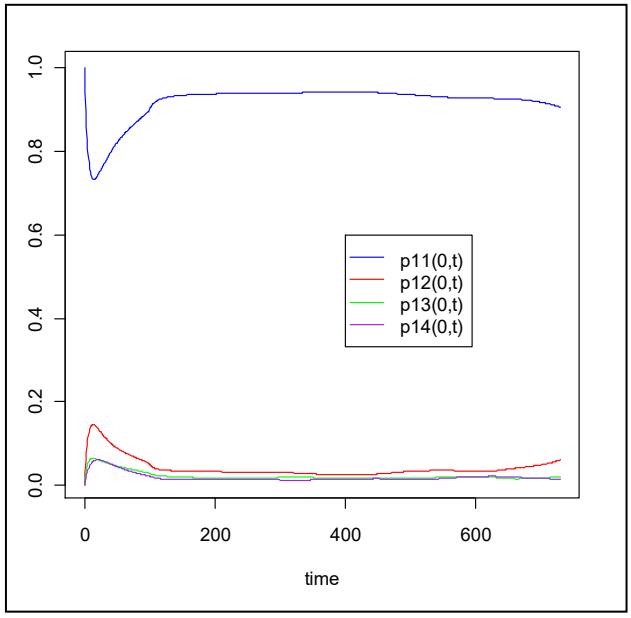


**Figure 3.** Cumulative transition probabilites from kernel smoothing with bandwith 100

### 3.3 Estimation Based on Covariates

Covariate effects on transition intensities are estimated using the `coxph` function from the `survival` package in R. Table 2 presents the estimation results. The resulting regression model for transition intensity is,

$$\hat{\alpha}_{hji}(t|Z) = \hat{\alpha}_{hj,0}(t) \exp\begin{pmatrix} -0.0038\ \text{age}_i - 0.0106\ \text{systole}_i + 0.0131\ \text{diastole}_i \\ -0.1129\ \text{KP}_i + 0.1432\ \text{sex}_i \end{pmatrix} \quad (20)$$

where $\hat{\alpha}_{hj,0}(t)$ is the baseline hazard for transitions $h \to j \in \{1 \to 2, 1 \to 3, 1 \to 4, 2 \to 1, 3 \to 1, 4 \to 1\}$. The baseline used for the transition intensities is the smoothed kernel-based estimate with a bandwidth of 100. Figure 4 shows the cumulative transition intensities for the six transitions. The curves indicate that transitions from State 1 (healthy) to State 2 (mild disease) occur most frequently over time, as reflected by the steep rise in the yellow line. In contrast, transitions from more severe disease states (e.g., State 4 to State 1) show slower accumulation, suggesting less frequent recovery from severe conditions.

**Table 2.** Parameter estimation of regression model for transition intensities

| Variable | Estimation | SE | P_value |
|---|---|---|---|
| $\hat{\beta}_1$ (age) | -0.003818 | 0.001707 | 0.025322 |
| $\hat{\beta}_2$ (systole) | -0.010645 | 0.002828 | 0.000167 |
| $\hat{\beta}_3$ (diastole) | 0.013102 | 0.004287 | 0.002242 |
| $\hat{\beta}_4$ (KP) | -0.112900 | 0.088308 | 0.201081 |
| $\hat{\beta}_5$ (sex) | 0.143202 | 0.061984 | 0.020871 |

From Table 2, we observe that several covariates significantly affect the transition intensities between states:

a. Age has a negative coefficient (–0.0038, $p = 0.0253$), indicating that older patients are slightly less likely to transition between states, suggesting possibly lower clinic visit frequency or disease progression at older ages.
b. Systolic blood pressure also shows a significant negative effect (–0.0106, $p < 0.001$), meaning that higher systolic pressure is associated with lower transition intensity, which may reflect more stable health conditions or reduced likelihood of transitioning to a more severe disease state.
c. Diastolic blood pressure has a positive coefficient (0.0131, $p = 0.0022$), suggesting that patients with higher diastolic values are more likely to experience transitions, possibly due to underlying cardiovascular strain.
d. Sex has a positive and significant coefficient (0.1432, $p = 0.0209$), indicating that male patients have a higher rate of state transition compared to female patients, controlling for other covariates.
e. Although the KP variable is not statistically significant at the 5% level (-0.1129, $p = 0.2011$), it is retained in the model to account for clinical relevance and potential confounding.

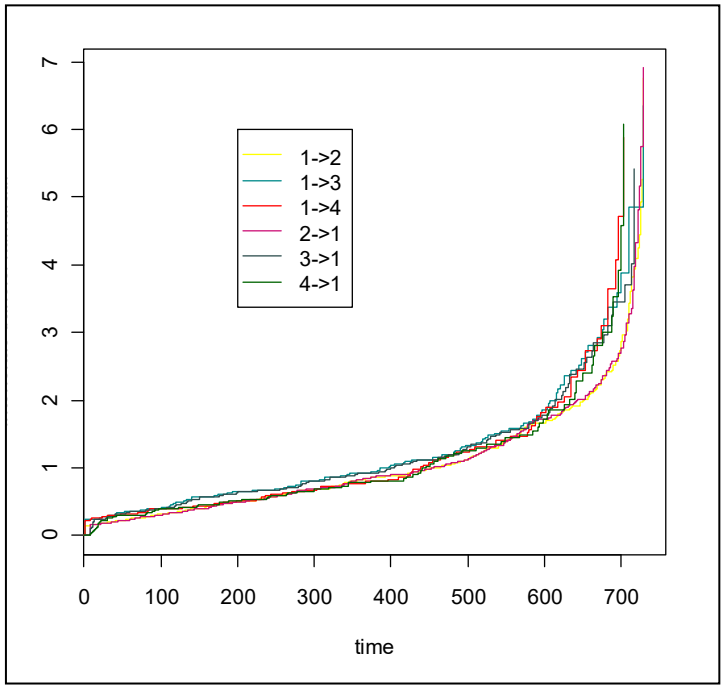


**Figure 4.** Cumulative baseline transition intensities

Let $Q_h^{(i)}(t_k)$ denote the state occupation probability of patient $i$ in state $h$ at time $t_k$. The analysis focuses on the state occupation probability of state 1 (healthy), $Q_1^{(i)}(t_k)$, as this state represents the clinically meaningful reference state and the initial state for all individuals in the study. From a practical perspective, the probability of remaining in or returning to the healthy state over time is of primary interest for assessing overall patient morbidity. In addition, state 1 is the most frequently observed state in the data, leading to more stable pseudo-value estimates compared with less frequently visited disease states.

Pseudo-values of $Q_1^{(i)}(t_k)$ and subsequently treated as the dependent variable in the GEE analysis, assuming a binomial distribution with a logit link function. This specification allows the marginal expectation of the state occupation probability to be linked directly to covariates. The analysis is performed using the `geese` function from the `geepack` package in R. The estimation results are presented in Table 3, and the corresponding estimated regression model is given in Equation (21).

This approach differs fundamentally from conventional multi-state regression models, in which covariates are incorporated through transition-specific hazard models and their effects on state occupation probabilities are obtained only indirectly via nonlinear combinations of transition intensities (3,9). In contrast, the pseudo-value-based GEE model in Equation (21) provides marginal covariate effects on state occupation probabilities, leading to more transparent interpretation, particularly when interest lies in prevalence-type quantities rather than transition risks.

**Table 3.** Parameter estimation of the regression model for the state occupation probability of state 1

| Variable | Estimation | SE | P_value |
|---|---|---|---|
| Intercept | 1.9915100 | 0.4456713 | 0.0000079 |
| $\hat{\beta}_1$ (time) | 0.0013783 | 0.0003680 | 0.0001800 |
| $\hat{\beta}_2$ (age) | -0.0063219 | 0.0057964 | 0.2754225 |
| $\hat{\beta}_3$ (systole) | 0.0033701 | 0.0073332 | 0.6458322 |
| $\hat{\beta}_4$ (diastole) | -0.0000506 | 0.0104719 | 0.9961474 |
| $\hat{\beta}_5$ (KP) | -0.8851953 | 0.2127827 | 0.0000318 |
| $\hat{\beta}_6$ (sex) | 0.2081326 | 0.1584150 | 0.1888988 |

Based on the parameter estimates reported in Table 3, the fitted regression model for the state occupation probability of state 1 is given by:

$$\begin{aligned}\log\frac{\hat{Q}_1^{(i)}(t_k)}{1-\hat{Q}_1^{(i)}(t_k)} &= 1.9915 + 0.0014\,t_k - 0.0063\,\text{age}_i(t_k) \\ &\quad + 0.0034\,\text{systole}_i(t_k) - 0.00005\,\text{diastole}_i(t_k) \\ &\quad - 0.8852\,\text{KP}_i + 0.2081\,\text{sex}_i.\end{aligned} \tag{21}$$

As shown in Table 3, the estimated coefficients provide insight into covariate effects on the log-odds of occupying state 1.

a. The intercept is positive and highly significant (1.992, $p < 0.001$), representing the baseline log-odds of occupying state 1 when all covariates are set to zero.
b. Time has a positive and statistically significant coefficient (0.0014, $p = 0.00018$), indicating that the log-odds, and hence the probability, of occupying state 1 increases over time. This finding is consistent with the modeling assumption that patients are more likely to return to a healthy state if no clinic visits are recorded within a one-week period.
c. KP has a strong negative and statistically significant effect (–0.8852, $p < 0.001$), suggesting that patients diagnosed with KP have substantially lower odds of occupying the healthy state compared to patients without KP.
d. Although age, systolic blood pressure, diastolic blood pressure, and sex are not statistically significant at the 5% level, they are retained in the model to account for individual-level characteristics that may still contribute to variability in state occupation probabilities and to improve overall model adequacy.

## IV. CONCLUSIONS AND OUTLOOKS

This study proposes a discrete-time multistate modeling framework that combines kernel-smoothed Nelson–Aalen estimators for transition intensities with product-integrals for estimating transition probabilities. Covariate effects are incorporated using two complementary approaches: Cox regression for transition intensities and a pseudo-value-based GEE framework

for direct modeling of state occupation probabilities. By jointly analyzing instantaneous transition risks and marginal state occupation probabilities, the proposed approach provides improved interpretability of covariate effects in multi-state settings.

The application to patient morbidity data from a clinic in West Java, Indonesia, illustrates the practical relevance of the methodology for analyzing disease progression based on discretely observed clinical visit data. In particular, direct regression modeling of state occupation probabilities offers clinically meaningful insights that are not readily obtainable from intensity-based models alone.

Future research may extend this framework to continuous-time settings, allowing for finer temporal resolution in modeling transition dynamics. In addition, relaxing the Markov assumption—such as through semi-Markov or history-dependent multi-state models—would enable the incorporation of past disease trajectories into the modeling process, potentially yielding a more realistic representation of clinical progression in complex health processes.

## REFERENCES

[1] P. K. Andersen and N. Keiding, "Multi-state Models for Event History Analysis", *Stat Methods Med Res*., vol. 11, no. 2, pp. 91–115, 2002.

[2] P. K. Andersen, J. P. Klein, and S. RosthØj, "Generalised Linear Models for Correlated Pseudo-observations with Applications to Multi-state Models", *Biometrika*, vol. 90, no. 1, pp. 15–27, 2003.

[3] T. Ma, L. He, Y. Luo, D. Fu, J. Huang, G. Zhang, X. Cheng, and Y. Bai, "Frailty, An Independent Risk Factor in Progression Trajectory of Cardiometabolic Multimorbidity: A Prospective Study of UK Biobank", *The Journals of Gerontology: Series A*, vol. 78, no. 11, pp. 2127–2135, 2023.

[4] M. K. Lintu, K. M. Shreyas, and A. Kamath, "A Multi-state Model for Kidney Disease Progression", *Clinical Epidemiology and Global Health*, vol. 13, pp. 100946, 2022.

[5] Y. W. Sari, Gunardi, N. Y. Megawati, and S. H. Hutajulu, "Cox-Based Estimation Model for Critical Illness Insurance Policy for Breast Cancer Based on the Possible Transition of Status", *Malaysian Journal of Mathematical Sciences*, vol. 19, no. 1, pp. 35–31, 2025.

[6] O. O. Aalen and S. Johansen, "An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations", *Scand J Statist*, vol. 5, no. 3, pp. 141–150, 1978.

[7] H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in Biostatistics: Competing Risks and Multi-state Models", *Statistics in Medicine*, vol. 26, no. 11, pp. 2389–2430, 2007.

[8] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, P. K. Andersen, "Multi-state Models for The Analysis of Time-to-event Data", *Stat Methods Med Res.*, vol. 18, no. 2, pp. 195–222, 2009.

[9] A. S. Conlon, J. M. Taylor, and D. J. Sargent, "Improving Efficiency in Clinical Trials using Auxiliary Information: Application of A Multi-state Cure Model", *Biometrics*, vol. 71, no. 2, pp. 460–468, 2015.

[10] M. Hill, P. C. Lambert, M. J. Crowther, "Relaxing the Assumption of Constant Transition Rates in A Multi-state Model in Hospital Epidemiology", *BMC medical research methodology*, vol. 21, no. 1, pp. 16, 2021.

[11] M. Overgaard, P. K. Andersen, and E. T. Parner, "Pseudo-observations In A Multistate Setting, The Stata Journal, vol. 23, no. 2, pp. 491–517, 2023.

[12] J. G. Le-Rademacher, T. M. Therneau, and F. S. Ou, The Utility of Multistate Models: A Flexible Framework for Time-to-Event Data", *Curr Epidemiol Rep*., vol 9, pp. 183–189, 2022.

[13] L. Ferrer, V. Rondeau, J. Dignam, T. Pickles, H. Jacqmin-Gadda, and C. Proust-Lima, "Joint Modelling of Longitudinal and Multi-state Processes: Application to Clinical Progressions in Prostate Cancer", *Statistics in medicine*, vol. 35, no. 22, pp. 3933–3948, 2016.

[14] A. Nießl, A. Allignol, J. Beyersmann, and C. Mueller, "Statistical Inference for State Occupation and Transition Probabilities in Non-Markov Multi-state Models Subject to Both Random Left-truncation and Right-censoring", *Econometrics and Statistics*, vol. 25, pp.110–124, 2023.

[15] P. K. Andersen and J. P. Klein, "Regression Analysis for Multistate Models Based on A Pseudo-value Approach with Applications to Bone Marrow Transplantation Studies", *Scand J Statist*., vol. 34, no. 1, pp. 3–16, 2007.

[16] K. Y. Liang and S. L. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[17] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, "*Statistical Models Based on Counting Processes*", Springer-Verlag, 1993.

[18] T. M. Therneau and P. M. Grambsch, "*Modeling Survival Data: Extending the Cox Model*", Springer, 2000.

[19] J. P. Klein and M. L. Moeschberger, "*Survival Analysis Techniques for Censored and Truncated Data*", Springer-Verlag, New York, 2003.

[20] A. R. Effendie, "On Health Insurance Valuation Models with Dynamic Interest Rate Using Multistate Risk Approach", *Dissertation*, Universitas Gadjah Mada, 2011.