Estimation of disease progression for ischemic heart disease using latent Markov with covariates

Article *in* Statistical Analysis and Data Mining · June 2022

DOI: 10.1002/sam.11589

Zarina Oflaz
KTO Karatay University
12 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Ceylan Yozgatligil
Middle East Technical University
41 PUBLICATIONS **569** CITATIONS
SEE PROFILE

Sevtap Kestel
Middle East Technical University
66 PUBLICATIONS **272** CITATIONS
SEE PROFILE

Received ; | Revised ; | Accepted

DOI: xxx/xxxx

RESEARCH ARTICLE

Estimation of Disease Progression for Ischemic Heart Disease using Latent Markov with covariates[†]

Zarina Oflaz*[1] | Ceylan Yozgatligil[2] | A. Sevtap Selcuk-Kestel[3]

[1]Department of Insurance and Social Security, KTO Karatay University, Konya, Turkey
[2]Department of Statistics, Middle East Technical University, Ankara, Turkey
[3]Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey

Correspondence
*Zarina Oflaz, Department of Insurance and Social Security, KTO Karatay University, Konya, Turkey. Email: zarina.oflaz@karatay.edu.tr

Abstract

Contemporaneous monitoring of disease progression, in addition to early diagnosis, is important for the treatment of patients with chronic conditions. Chronic disease-related factors are not easily tractable, and the existing data sets do not clearly reflect them, making diagnosis difficult. The primary issue is that databases maintained by health care, insurance, or governmental organizations typically do not contain clinical information and instead focus on patient appointments and demographic profiles. Due to the lack of thorough information on potential risk factors for a single patient, investigations on the nature of disease are imprecise. We suggest the use of a latent Markov model with variables in a latent process because it enables the panel analysis of many forms of data. The purpose of this study is to evaluate unobserved factors in ischemic heart disease (IHD) using longitudinal data from electronic health records. Based on the results we designate states as healthy, light, moderate, and severe to represent stages of disease progression. This study demonstrates that gender, patient age, and hospital visit frequency are all significant factors in the development of the disease. Females acquire IHD more rapidly than males, frequently developing from moderate and severe disease. Additionally, it demonstrates that individuals under the age of 20 bypass the light state of IHD and proceed directly to the moderate state.

KEYWORDS:
Ischemic Heart Disease, Latent Markov, Longitudinal data, Gender, Age, Disease progression, Electronic health records

1 | INTRODUCTION

Chronic diseases are the main cause of mortality and disability worldwide. In Turkey, 70% of all deaths caused by chronic diseases are ischemic heart disease (IHD), cerebrovascular, unipolar depressive, chronic obstructive pulmonary disease, and diabetes mellitus [33]. Along with the risk of morbidity and mortality, there are high costs involved in treating critical illnesses; thus, understanding of the contribution of factors leading to the development and progression of chronic disease is critical.

Along with early diagnosis, continuous monitoring of illness progression is critical for the management of patients with chronic disorders. A better understanding of disease progression would lead to the implementation of appropriate health care, improved cohort selection, and faster medication discovery. Clinical researchers frequently use quantitative models called disease progression models to define the course of disease development using longitudinal patient records. The rate of progression

[†]Ischemic Heart Disease progression with Latent Markov model

of chronic diseases differs substantially across patients due to various factors, including genetics, physiology, gender, habits, socioeconomic state, and behavior.

Chronic diseases typically have a multi-state nature that dynamically progresses from early to late stages and is influenced by a variety of internal and external risk factors. To investigate the structure of disease progression, multi-state models are increasingly utilized. Breast cancer progression is examined with non-homogeneous exponential regression Markov models [15]. Meenaxi and Singh [22] presented the use of the Markov process to demonstrate its efficacy in providing a survival analysis of a patient with chronic heart failure due to a reduced ejection fraction. The progression rate of type 2 diabetes was quantified using the Markov model [32]. Collaborative topic modeling and the Gaussian mixture method have been employed to study the distribution and progression of chronic disease in a population using information on human mobility patterns [34]. Luo et al. [20] proposes applying the continuous-time hidden Markov model to explore the progression of chronic obstructive pulmonary disease using longitudinal health records. A multistate continuous-time non-homogeneous Markov model was employed to the study disease progression of patients with decreased renal function [7]. Visual analytics with hidden Markov models have been employed to investigate disease progression pathways of chronic diseases [18]. The three-state Markov model and the Phase Type Law have been employed to investigate the morbidity and mortality rates of a chronic disease [1].

The rapid development of the electronic information system in insurance, health industries, and government departments has enabled studies on the characteristics of chronic diseases and monitoring of disease progression in a population. However, access to medical data of individuals is either limited or there is a lack of a centralized database for retrieving information on factors that contribute to or cause the development of diseases. The main disadvantage is that the available data usually do not include clinical information and contain only records on health institution appointments and patients' demographic profiles. The absence of detailed information on potential risk factors for a specific patient confines the precision of studies on the nature of disease. Moreover, time until exposure to the disease is not observed, therefore data on such cases usually include a non-zero claim, which makes it difficult to determine timing of first diagnosis. The observation set includes the ones who were sick and visited their doctors or practicians. Additionally, chronic disease patients do not use resources intensively during the early stages of disease; therefore, their records in health care claim databases may be limited. For this reason, unforeseen factors of disease onset and progression should be analyzed in more detail.

Motivated to develop an efficient way to retrieve information on background factors leading to or influencing disease onset and progression, we suggest to use the latent Markov model (LMM) framework [3, 35], which exposes underlying factors causing the onset of disease by including longitudinal factors. The considered framework defines information on the health status of each individual at each time point according to records of appointments at a health institution. Individuals' demographic data and frequency of hospital appointments are used as covariates influencing the latent process. The findings corroborate established medical facts. On the other hand, it enables researchers to investigate additional effects of covariates using the LMM approach. The findings are also beneficial for insurance companies, which base premiums on hidden risks associated with policyholders that may be concealed or not yet observed.

The LMM has the same model assumptions and estimation techniques as the well-known hidden Markov model (HMM). However, HMMs are mainly used in time-series data analyses, whereas LMMs are generally involved in longitudinal data analyses [3]. LMMs can be used as an extension of latent class analysis for longitudinal data and have been suggested in many studies for public health and medical research. In particular, an LMM is employed to study the performance of nursing homes [4] and their rankings [25], self-reported health statuses based on longitudinal data [2], the effect of health status on material hardship [11], the effect of population aging on future health services costs [36], the diagnostics for trachoma elimination [17], the effects of uncontrolled diabetes on health care consumption [14], and model smoking transitions [21].

In this paper, we study IHD morbidity rates by employing an LMM with covariates in the latent process. This approach makes it possible to include unobserved factors, which may be influenced by individual covariates; hence, the model assumptions account for unobserved individual heterogeneity. Examining the morbidity rates of IHD using the LMM scheme contributes to the literature on quantification of the likelihood of critical diseases while using incomplete information.

2 | LATENT MARKOV MODEL

In the multivariate scenario, we observe a vector of J response variables $\boldsymbol{Y}_i = (Y_{i1}^{(1)}, ..., Y_{iJ}^{(T)})$ for each subject i and time t. The LMM is provided in this paper for the special case of a single response variable per time period. Let $\boldsymbol{Y}_i = (Y_i^{(1)}, ..., Y_i^{(T)})$, $i = 1, ..., n$, be a vector of T categorical variables observed over n individuals. We denote c as the number of their categories, from 0 to $(c-1)$. An LMM is determined by the observed process $\boldsymbol{Y}_i$ depending on the latent process $\boldsymbol{U}_i = (U_i^{(1)}, ..., U_i^{(T)})$. The latent process is defined as the first-order Markov chain with state space $\{1, ..., k\}$, where k is the number of latent states. Therefore, [3]

$$P(U_i^{(t)}|U_i^{(t-1)}, \dots, U_i^{(1)}) = P(U_i^{(t)}|U_i^{(t-1)}), \quad t = 2, \dots, T$$
$$P(Y_i^{(t)}|Y_i^{(t-1)}, \dots, Y_i^{(1)}, U_i^{(t)}, \dots, U_i^{(1)}) = P(Y_i^{(t)}|U_i^{(t)}), \quad t = 1, \dots, T.$$

An HMM and an LMM utilize the same model assumptions and estimation techniques. The main differences between the two models are the asymptotic properties of time-series and longitudinal data. In the context of time series, the time points tend to infinity, whereas in the longitudinal data, the sample size tends to infinity [3].

To construct a basic LMM over a sequence of observations, we need to define an initial distribution of latent states, π_u; a matrix of transition probabilities, $\pi_{v|u}^{(t)}$; and conditional response probabilities, $\phi_{y|u}^{(t)}$. Transition probability denotes the probability of moving from state u to state v at time t:

$$\pi_{v|u}^{(t)} = P(U_i^{(t)} = v|U_i^{(t-1)} = u) \quad t = 2, ..., T; u, v = 1, ..., k.$$

Then the the matrix of transition probabilities is defined as follows,

$$\Gamma = \begin{bmatrix} \pi_{1|1}^{(t)} & \pi_{2|1}^{(t)} & \dots & \pi_{k|1}^{(t)} \\ \vdots & & \ddots & \vdots \\ \pi_{1|k}^{(t)} & \pi_{2|k}^{(t)} & \dots & \pi_{k|k}^{(t)} \end{bmatrix}.$$

The initial distribution of the Markov chain which specifies the starting state is defined as

$$\pi_u = P(U_i^{(1)} = u), \quad u = 1, \dots, k.$$

The conditional response probabilities, given as

$$\phi_{y|u}^{(t)} = P(Y_i^{(t)} = y|U_i^{(t)} = u), \quad t = 1, \dots, T; u = 1, \dots, k; y = 0, \dots, c-1.$$

define the relationship between observation and an unobserved state. In this study, we assume conditional response probabilities are time-invariant, $\phi_{y|u}$ is not dependent on t.

2.1 | Latent Markov model with covariates in a latent process

We suppose that there is an effect of p covariates, the column vector $\boldsymbol{X}_i^{(t)} = (X_{1i}^{(t)}, X_{2i}^{(t)}, \dots, X_{pi}^{(t)})$, on the health condition of an individual i at time t. The health condition is not directly observable and thus, represented by the latent process. We include the covariates in the latent model by parameterization of the initial state probabilities and transition probabilities employing multinomial logits. Figure 1 illustrates the structure of an LMM with individual covariates in the latent process. The state-dependent process $\boldsymbol{Y}_i$ is expressed in terms of their individual factors ($\boldsymbol{X}_i^{(t)}$) through the latent parameter process ($\boldsymbol{U}_i$).

We assume a time-heterogeneous Markov chain in which both the transitions between latent states and the arrival of observations may occur at arbitrary times. It is, therefore, appropriate for irregularly sampled time data such as clinical observations.

Having that the first state is taken as the reference, then multinomial logit for the initial probability of U is,

$$\log \frac{P(U_i^{(1)} = u|\boldsymbol{X}_i^{(1)} = \boldsymbol{x})}{P(U_i^{(1)} = 1|\boldsymbol{X}_i^{(1)} = \boldsymbol{x})} = \log \frac{\pi_{u|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}}} = \beta_{0u} + \boldsymbol{x}^\top \boldsymbol{\beta}_{1u}, \quad u = 2, ..., k,$$

where $\boldsymbol{\beta}_{1u}$ is the vector of the regression coefficients for the observed covariates and β_{0u} is the intercept parameter.

In order to parameterize the transition probabilities, we set the remaining in the latent state v as the reference transition and construct a multinomial logistic model as follows:

$$\log \frac{P(U_i^{(t)} = u|U_i^{(t-1)} = v, \boldsymbol{X}_i^{(t)} = \boldsymbol{x})}{P(U_i^{(t)} = v|U_i^{(t-1)} = v, \boldsymbol{X}_i^{(t)} = \boldsymbol{x})} = \log \frac{\pi_{u|v\boldsymbol{x}}^{(t)}}{\pi_{v|v\boldsymbol{x}}^{(t)}} = \gamma_{0vu} + \boldsymbol{x}^\top \boldsymbol{\gamma}_{1vu},$$
$$t = 2, ..., T; u \neq v,$$

where γ_{0vu} and $\boldsymbol{\gamma}_{1vu}$ are parameter vectors to be estimated.

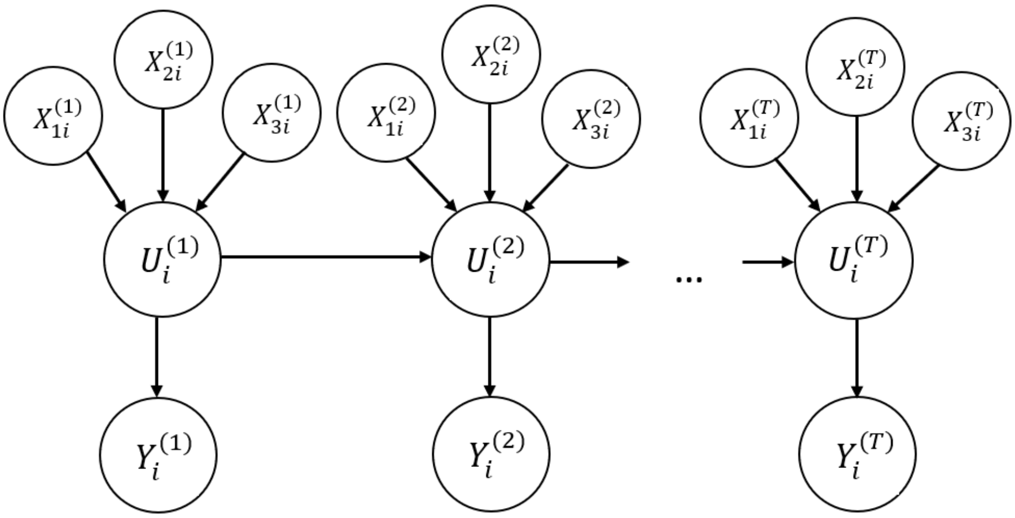


FIGURE 1 The directed graph of an LMM with individual covariates in the latent process.

2.2 | Parameter Estimation in an LMM

To construct an LMM, transition probabilities, initial probability, and parameters of the conditional response probabilities must be estimated. Regarding the complexity of the likelihood function for an analytical solution, an expectation–maximization (EM) algorithm is employed. The EM, which performs the maximum likelihood estimation of parameters in the presence of missing values in the data, functions well in an LMM estimation, since the latent states are treated as missing information [10] [19].

The complete-data log-likelihood of an LMM with covariates in latent process forms the expression [3]

$$\begin{aligned} l^*(\boldsymbol{\theta}) = & \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y=0}^{c-1} a_{uxy}^{(t)} \log \phi_{y|u\boldsymbol{x}} \\ & + \sum_{u=1}^{k}\sum_{\boldsymbol{x}} b_{u\boldsymbol{x}}^{(1)} \log \pi_{u|\boldsymbol{x}} + \sum_{t=2}^{T}\sum_{v=1}^{k}\sum_{u=1}^{k}\sum_{\boldsymbol{x}} b_{vu\boldsymbol{x}}^{(t)} \log \pi_{u|v\boldsymbol{x}}^{(t)}. \end{aligned} \tag{1}$$

Here, $a_{uxy}^{(t)}$ is the number of subjects that are in latent state u and responding by y at time t shown as,

$$a_{uxy}^{(t)} = \sum_{i=1}^{n} \boldsymbol{I}(u_{it} = u, \boldsymbol{x}_t = \boldsymbol{x}, y_{it} = y),$$

and $b_{u\boldsymbol{x}}^{(t)}$ is the frequency of subjects in u with covariate configuration $\boldsymbol{x}$ at time t, and $b_{vu\boldsymbol{x}}^{(t)}$ is the number of transitions from v to u at time t explained as

$$b_{u\boldsymbol{x}}^{(t)} = \sum_{i=1}^{n} \boldsymbol{I}(u_{it} = u, \boldsymbol{x}_t = \boldsymbol{x}),$$

and

$$b_{vu\boldsymbol{x}}^{(t)} = \sum_{i=1}^{n} \boldsymbol{I}(u_{it-1} = v, u_{it} = u, \boldsymbol{x}_t = \boldsymbol{x}),$$

respectively. u_{it} is the unknown latent state indicator, $i = 1, \ldots, n$, $t = 1, \ldots, T$.

The EM algorithm alternates between two phases. E-step involves computing the posterior expected value of each frequency in Equation 1 using appropriate forward-backward recursions [6]. M-step maximizes the complete data log-likelihood expressed as in Equation 1, substituting the expected value for each frequency. How to maximize this function is model-specific and, in particular, depends on whether covariates are included in the measurement or the latent model [6, 19, 4].

Iterations are repeated until a prescribed convergence is satisfied. The relative log-likelihood difference is used to assess the convergence of the EM algorithm, that is,

$$\frac{l(\boldsymbol{\theta}^{(s)}) - l(\boldsymbol{\theta}^{(s-1)})}{|l(\boldsymbol{\theta}^{(s)})|} < \epsilon$$

where $\boldsymbol{\theta}^{(s)}$ is the parameter estimate calculated at the s-th M-step, ϵ is an appropriate tolerance level.

3 | CASE STUDY: LMM ON TURKISH IHD PATIENTS

The considered approach is illustrated in real data collected from the Turkish health system, whose information can be obtained from the Social Security Institution (SSI) of Turkey. Before the complete digitalization of data for health branches in 2015, a centralized database with identification (ID)- based information was established under SSI, which covered the recordings of approximately 90% of the population and hence constituted a reliable basis from which to generalize the results of the proposed analysis. Around 2.5 million of the registrations in the data set are determined to be related to IHD, which is the data set used in the LMM methodology. The data set contains information on patients who were diagnosed and treated with IHD upon their registration to any health institute in Turkey, within the period 2007 to 2009. Trailing years are not available due to a change in SSI policy on data sharing. The ID number, age, gender, place of birth, date of appointment at a health institution, location of the health institute, code of patient diagnosis are the available variables.

The monthly re-modified data is structured with respect to individuals' frequency of repeated visits within a month. Monthly observations between June 2007 and July 2009 concerning age, gender, and frequency of hospital visits as explanatory variables have been employed to determine morbidity in terms of hidden states, $T = 26$ for all patients. If a patient is admitted to a hospital more than once a month, the visits are aggregated. If a patient is not observed during a given month, the visit number for that month is set to 0. The majority of insurance data contains loss occurrences, which we do have in our data collection. The observation set includes those who were ill and visited their doctors or practicians. For this reason, we made some assumptions about their emergence and disappearance based on the characteristics of the disease we chose, which requires ongoing surveillance and has no cure. The simplifying assumptions are:

i An insured patient who does not appear in the SSI system in consecutive periods is assumed to be severely ill so that they could not show at the follow-up meetings, since the mortality rate of IHD is high and the IHD requires regular medical controls or medication prescriptions;

ii As the prehistory of individuals before 2007 was unknown, an insured patient not appearing in the system for the selected period was assumed to be "healthy."

Therefore, we define a categorical variable, Status, which yielded three cases: 0 as healthy, 1 as IHD diagnosis, and 2 as severe IHD. Individuals are assigned a status of 0 until the first occurrence in the data set, a status of 1 from the first to the last visit, and a status of 2 from the last to the end of the study.

The application of LMM with covariates in the latent process for all 2,523,686 insured patients was analyzed through a random sample of 3,000 individuals using a stratified sampling technique in terms of male and female subgroups. The data set contains the recorded values of patients enrolled in hospitals in Turkey who are covered by social insurance. The ICD10 codes are the identifying characteristics on which we base our study. As a result, the data set includes both the selected critical disease and its occurrences during the selected period. As previously stated, we are unable to extend the data set to more recent years due to data privacy concerns. As a result, given the Turkish data and the time, this is the maximum sample size to consider. Moreover, the sample size is determined by computational time constraints. The proportions of males and females are 0.5038 and 0.4962, respectively. Afterwards, we allocate samples into training and test sets with 70%–30% split to control over the performance of the model and overcome overfitting. Therefore, the training set includes 1,059 males and 1,044 females, and the test set includes 452 males and 444 females.

Figures 2 a and 2 b show the patterns on the frequency visits over a month and in total. It should be mentioned that IHD is a chronic disease that obliges the patient to visit their practitioners or health institute periodically after they have received the IHD diagnosis. The number of visits in a month or in total varied by gender. We observe that the number of males with one, two, or three checkups per month was greater than the number of females (Figure 2 a). However, hospital visits became approximately equal among genders as the number of visits increased above three.In 26 months, 57% of patients visited the hospital only once, the majority of whom were female patients (Figure 2 b); 18.6% of patients visited the hospital twice during the study period; and the remaining 24% of patients visited the hospital regularly.

We apply the LMM with covariates in a latent process for males, females, and both genders by utilizing an optimization algorithm (lmestSearch) [5] using R. LMest combines deterministic and random initialization strategies to achieve the maximum global log-likelihood of the model via the EM algorithm. It uses one deterministic initialization and several random initializations that are proportional to the number of latent states. A final run is conducted, starting from the best solution obtained in the above initialization steps.

To evaluate the model's performance, we establish three data sets from the sampled individuals, which includes males, females, and both genders. For the male and female sets, we use age and number of hospital visits as covariates influencing the latent process, whereas for the aggregate set of both genders, we add gender to the covariates as well. The LMM model is applied to the data sets to determine the optimal number of latent states, k, which varies from 1 to 6 based on the log-likelihood function, $\hat{l}$, and performance indicators, AIC and BIC, whose results are summarized in Table 1 . We observe that the optimal model yields a 4-state for females, whereas 4 and 6 states are found to be optimal for male and aggregate data sets based on AIC and BIC values. The model for the aggregate set performed considerably different than for the male or female set, having the lowest AIC for 6 states and the lowest BIC for 4 states. It is noteworthy that the aggregate model deals with different data structures and includes an additional covariate compared to gender-base models. For this reason, we focus on the gender-based data sets in the implementation of the model. To assess the effect of the hospital visit frequency, we then exclude the covariate from the models whose fitting performance (Table 1) became significantly weaker, which indicate that the frequency of hospital visits is an important covariate to consider. Based on these results, we conclude that the 4-state model using age and hospital visits as covariates should be used for males and females.

In the absence of covariates, identification issues may develop in the presence of a singleton response variable [3].While the inclusion of covariates may reduce this issue in this case, further inquiry into identifiability may be necessary. The information matrix $\boldsymbol{J}\left(\hat{\theta}\right)$ is used for checking local identifiability at $\hat{\theta}$. The model is considered to be local identifiable if the matrix is of full rank [4, 3]. The matrix $\boldsymbol{J}\left(\hat{\theta}\right)$ is computed using standard error for parameter estimate. The matrices of estimated parameters from the selected 4-state models for males and females are full of rank. Therefore, the selected models do not have an identifiability problem.

Based on the train and the test set, we randomly draw 100 samples from the selected LMM models with individual covariates and estimated parameters. The mean accuracy of each generated sample is calculated. The mean accuracy of the selected model is found to be 80.74% for training and 81.21% for testing sets for males; for females, it is 73.08% and 72.52% for training and testing sets, respectively. The overfitting problem is not observed; furthermore, the prediction performance for the testing set is surpassed. It is important to emphasize that the primary goal of the study is to identify unobserved explanatory factors; hence, the prediction performance of primarily explanatory and not predictive models is sufficient.

3.1 | Results

After the determination of latent states, we depict the response probabilities and logit transition probabilities, whose outcomes are presented and interpreted in the framework of the LMM model.

The results of the EM algorithm indicate that the conditional response probabilities, i.e., the conditional probabilities of the categorical response variable given latent states, demonstrate that "healthy" should be in state 1, "illness" should be in states 2 and 3, and "severe illness" should be in state 4 with the probability of 1 for both genders.

To analyze latent state movements, we obtain initial and transition state probabilities for males and females at the mean age of the sample, which is 58 years old.

According to the dynamics of transitional and initial state probabilities, we name the 1st state as “healthy”, the 2nd state as “light”, the 3rd state as “moderate”, and the 4th state as “severe” to represent stages of disease progression. Initial state probabilities indicate that all individuals transitioned from a healthy state with a probability of 1. We observe that the movement dynamics of the latent states are different for males and females, which is shown in more detail in Table 2 . The other points to be mentioned are as follows:

i Males have a higher probability of remaining in a light state (0.9567), whereas females remain in that state only with a probability of 0.8817.

TABLE 1 The fitting performance of the LMM models for *k* state numbers and *r* parameters on IHD data set.

	Covariates	k	$\hat{l}$	r	AIC	BIC
Male	Age, Hospital visits	1	-27970	2	55945	55955
		2	-8886	13	17798	17863
		3	-2198	30	4455	4604
		4	-1129	53	2365	**2628**
		5	-1128	82	2420	2828
		6	-1052	117	**2339**	2920
Female	Age, Hospital visits	1	-26691	2	53385	53395
		2	-7938	13	15902	15967
		3	-2123	30	4306	4455
		4	-1454	53	**3014**	**3277**
		5	-1529	82	3221	3627
		6	-1789	117	3813	4392
Total	Gender, Age, Hospital visits	1	-54711	2	109425	109436
		2	-16892	16	33817	33907
		3	-4227	38	8531	8745
		4	-2825	68	5786	**6170**
		5	-2980	106	6173	6772
		6	-2513	152	**5330**	6189
Male	Age	1	-27970	2	55945	55955
		2	-12643	10	25307	25356
		3	-6228	22	12500	12609
		4	-5763	38	11603	11792
		5	-5557	58	11231	11519
		6	-5481	82	**11125**	**11532**
Female	Age	1	-26691	2	53385	53395
		2	-11651	10	23322	23371
		3	-5919	22	11882	11991
		4	-5422	38	10919	11107
		5	-5200	58	10516	**10803**
		6	-5121	82	**10406**	10812
Total	Gender, Age	1	-54711	2	109425	109436
		2	-24330	13	48687	48760
		3	-12148	30	24356	24525
		4	-11186	53	22478	22778
		5	-10760	82	21685	22148
		6	-10603	117	**21441**	**22102**

ii For both genders, the probability of moving directly from a healthy to moderate state is more than three times higher than the probability of moving from a healthy to light state.

iii The probability of moving directly from a healthy to severe state is insignificant for both genders. This result meets our expectation that chronic disease is more likely to grow stable rather than rapidly.

iv Women are more likely to transition directly from the light to severe stage of the disease than were men. Interestingly, females are more likely to move from the light state directly to severe (6.4% probability) rather than steadily to a moderate level (5.5% probability).

TABLE 2 Transition matrices for males and females at the average age 58.

	State	1	2	3	4
Male	1	0.9406	0.0133	0.0461	0.0000
	2	0.0000	0.9567	0.0433	0.0000
	3	0.0000	0.2259	0.0405	0.7336
	4	0.0000	0.0023	0.0392	0.9585
	State	1	2	3	4
Female	1	0.9088	0.0214	0.0698	0.0000
	2	0.0000	0.8817	0.0546	0.0637
	3	0.0055	0.1721	0.1597	0.6627
	4	0.0000	0.0000	0.0094	0.9906

v The probability of moving from the moderate to severe state is 0.7336 for males, which is 18 times higher than the probability of remaining in a moderate state. For comparison, females are less likely to move from moderate to the severe state than males, with a probability of 0.6627 and 0.7336, respectively.

vi Moreover, a moderate state is likely to return to a light state, with a probability of 0.2259 for males and 0.1721 for females. Thus, males are more likely to recover from a moderate state and transition to a light state than are females.

vii The most unstable state is the moderate state, both genders display a low probability of remaining in a moderate state, 0.0405 for males and 0.1597 for females. Remarkably, females are nearly four times more likely to remain in a moderate state than are males. Moreover, the moderate state have a high probability of transitioning to a severe state with a probability of 0.7336 for males and 0.6627 for females.

viii Females have a higher probability of staying in the severe state, with 99% chance, compared to males, with a 96% chance, which indicate that there is a lower chance of switching to a moderate state in comparison to males.

ix Noticeably, for both genders, there is an absolute or very close to 0 possibility of returning from any stage of the disease to the healthy state.

x Both genders have almost the same chance of moving from a light to moderate state.

The estimated LMM coefficients of the 4-state model for males and females are given in Table 3 . The parameters in γ_{1vu} refer to the transition from reference state, $logit = v$, to other states, $u, u \neq v$.

An increase in frequency of hospital visits has a positive impact on a transition probability from light to moderate state for both genders, which indicates a higher demand for hospital appointments when the disease advances. More frequent hospital visits increase the transition probability from a severe to moderate stage of the disease, $\gamma_{143} = 19.8407$ for males and $\gamma_{143} = 12.1964$ for females. Moreover, more frequent hospital appointments decrease the probability of moving from a moderate to severe state; males are more affected with $\gamma_{134} = -18.1044$ than females with $\gamma_{134} = -10.0981$.

The probability of moving directly from healthy to moderate is positively affected by the number hospital visits, $\gamma_{113} = 18.0062$, which is slightly higher than $\gamma_{112} = 15.0231$, which is the hospital visits parameter influencing transition probability from healthy to light state. A decrease in hospital visits number increases the probability of switching from a light or moderate state to a severe states for both genders, which indicates that if the patient does not receive treatment for their disease, it will negatively affect their health condition and prognosis. In general, age has a small effect on transition probabilities. However, age has a negative impact on the transition probability of moving from a moderate or light state to a healthy state for both genders. The estimated parameters indicate a negative effect of age on the recovery process. Interestingly, according to $\gamma_{224} = -0.1965$ for males and $\gamma_{224} = -0.3551$ for females, younger patients have a higher probability of transitioning from a light to severe state. The reason might be because the disease is more stable and progresses more steadily for older patients than for younger patients, as older patients are more likely to transition to a moderate and then severe state; however, young patients are more likely to switch directly from a light to severe state.

TABLE 3 Parameters affecting the logit for the transition probabilities for males and females.

Male	**logit = 1**	2	3	4
	Intercept	-5.8232 ***	-7.0043 ***	-12.2151 ***
	Hospital visits number	15.0231 ***	18.0062 ***	13.8145 ***
	Age	-0.0084	-0.0172	-0.0273
	logit = 2	1	3	4
	Intercept	-4.6820 ***	-15.2129 ***	-11.2939 ***
	Hospital visits number	-5.7964 ***	14.2246 ***	-7.4459 ***
	Age	-0.3989 ***	0.0355 ***	-0.1965 ***
	logit = 3	1	2	4
	Intercept	-5.1618 ***	9.4334 ***	12.6258 ***
	Hospital visits number	-4.1630 ***	-8.9555 ***	-18.1044 ***
	Age	-0.3421 ***	-0.0217	-0.0546 **
	logit = 4	1	2	3
	Intercept	-24.8644 ***	-20.0416 ***	-18.9757 ***
	Hospital visits number	8.6318 ***	17.8725 ***	19.8407 ***
	Age	0.0015 ***	0.0037 ***	-0.0011
Female	**logit = 1**	2	3	4
	Intercept	-5.3837 ***	-3.3264 ***	-9.9437 ***
	Hospital visits number	10.5458 ***	9.9898 ***	-3.4627 ***
	Age	-0.0010	-0.0070**	-0.0486
	logit = 2	1	3	4
	Intercept	-5.3768 ***	-7.1733 ***	-8.6754 ***
	Hospital visits number	-5.1672 ***	7.4613 ***	-8.2398 ***
	Age	-0.3869 ***	0.0179	-0.3551***
	logit = 3	1	2	4
	Intercept	1.2822	0.5600	4.1778 ***
	Hospital visits number	-8.5642 ***	-0.9666 ***	-10.0981 *
	Age	-0.0837 ***	-0.0047	-0.0432 ***
	logit = 4	1	2	3
	Intercept	-22.6497 ***	-21.1307 ***	-21.1368 ***
	Hospital visits number	9.3705 ***	8.7478 ***	12.1964 ***
	Age	-0.0124 ***	-0.0069 ***	-0.0112 ***

Significance: * (%10), ** (%5), *** (%1)

The global decoding algorithm calculates the most frequent latent state at a particular time at which the likelihood of the latent state for individuals of different age intervals and genders and their results for both genders (Figure 3) confirm the findings mentioned earlier.

i Patients younger than 20 years directly transition from a healthy to moderate state by skipping the light state. Males are more likely to progress to the severe state than are females.

ii Adult patients between the ages of 21 and 60 are more likely to proceed from a healthy to a severe state of IHD, with a lower likelihood of being in a light or moderate state.

iii In patients over the age of 60, IHD develops more steadily and frequently remains in a light or moderate state.

iv In general, IHD progresses faster and is more severe in younger people than in older patients.

4 | DISCUSSION

In this paper, we suggest to adopt an LMM with individual-based covariates in the latent process to study chronic disease progression. We apply the LMM model to the real-world hospital appointment data for patients with an IHD diagnosis, the records for which are obtained from the SSI in Turkey.

The study has several limitations. First, the data on hospital visits is not a clinical data; therefore, the data does not provide detailed clinical information about the patient, as it includes only demographic information and hospital visit dates. Second, the study consists of a longitudinal data, including possible interdependence among observations of the same individual at different times and possible dependence among individuals [24]. Third, the data has non-zero claims, which does not provide information on the onset of disease.

The LMM framework successfully addresses possible dependence among individuals and times through Markovian dependence of latent states. The complex structure of the model allows us to investigate the unobserved information, which had limited sequences of occurrences and covariates. The selected 4-state LMMs for males and females demonstrate a more flexible classification of patients (latent states as healthy, light, moderate, and severe) than our assumption on health status with three levels: healthy, illness, and severe illness.

While LMMs are already extensively used for longitudinal data analyses, their applications are not prevalent in medical panel data analysis, especially for IHD. While well-known HMMs have been broadly used in time-series analyses in medical and public health studies, the LMM is more appropriate than the HMM for observations that must preserve the influence of longitudinal structure. This unique implementation of LMM enables us to capture the stages of IHD when there is limited information on the population.

The dynamic of IHD differs in different stages of the disease, which illustrates varying transition probabilities between stages of the diseases for different genders. IHD for females progresses more acutely than for males, as it switches from a light to severe state of disease with higher probability. Furthermore, females stay in a severe state with very high probability, whereas males return from a severe to moderate state with a probability of 4%. Moreover, females have less probability of recovering by returning from a state of disease to a preceding state. Garcia et al. [13] finds similar results. They showed that certain subcategories of women, particularly those under the age of 55 and ethnic minorities, continue to experience worse outcomes than age-matched men [13]. Traditional and novel risk factors promote IHD progression in both men and women; however, other risk factors are exclusive to women, such as pregnancy-related problems and menopause. After menopause, the prevalence of IHD rises. Menopause might not be an independent risk factor for IHD in women, as additional risk factors, such as diabetes, hypertension, and metabolic syndrome, increase after menopause [23]. Women who had a primary cesarean delivery, especially if it was for medical reasons, had a 30% higher risk of coronary heart disease than women who had a vaginal delivery [9].

The interpretation of risk factors and symptoms differs between men and women. When male-pattern IHD risk factors and presentations are applied to women, it leads to under-recognition, under-testing, and under-treatment of IHD in women [31]. Along with these risk factors, whose information was not available in our data set, we also find that the age of the patients is an important variable to explain the behavior of the disease from early to severe stages. It also reveals that patients younger than 20 years of age switch straight to a moderate state, thus skipping the light state of IHD. During childhood and adolescence, young people may develop congenital heart disease (CoHD). CoHD is one of the most common congenital malformations in newborns, affecting around 1% of all live births [16]. According to a cohort study in Sweden, CoHD patients are 16.5 times more likely to be hospitalized or die from IHD compared to the controls [12].

Adult patients under the age of 60 have a reduced likelihood of developing a light state than older patients, which indicates that IHD in younger patients progresses faster or is more severe than in older patients. According to the findings of a study on coronary heart disease in young adults, clinicians should be aware of the differences in risk factors, prognosis, and therapy between older and younger patients [30]. One of the causes of differentiated disease progression over ages might be genetic contribution to the development of coronary artery disease (CAD) [26], with genetic factors having a greater influence in those who develop CAD at a younger age [8]. Additionally, subclinical hypothyroidism, which is a prevalent risk factor for IHD, is linked to an increase in IHD mortality only in patients under the age of 65 [29]. Rawandi et al. [28] concludes that the frequency of uncommon causes of ischemic stroke is relatively higher in young patients. According to Özer et al. [27], risk factors, stroke etiology, stroke severity, and prognosis of ischemic heart disease are different for young patients compared with patients older than 45 years. Finally, the two studies mentioned above consider patients with ischemic stroke residing in Turkey. Further research on other populations is warranted before generalizing the findings of these studies to global populations.

The results of patient flow analysis using this model will improve the effectiveness of chronic disease management and control programs. This is very relevant in healthcare research, as access to clinical data is extremely limited and only demographic and patient attendance data has to be managed, as in this study. Even if specific epidemiological findings from the study are already well-known and confirmed in the literature, this indicates that the model is operating in the right direction, and more importantly, under conditions of limited information.

On the other hand, it enables researchers to use the LMM approach to investigate the additional effects of covariates. The findings are also beneficial for insurance companies, which set premiums based on concealed or unobserved risks associated with policyholders.

The outcomes of this ongoing research are expected to be utilized as an important and guiding indicator for experts in medicine, insurance companies, and investors in the health industry.

Conflict of interest

The authors declare no potential conflict of interests.

Data Availability Statement

Research data are not shared.

References

[1] F. Akat, A. S. Selcuk-Kestel, and F. Tank, *The estimation of adopted mortality and morbidity rates using model and the phase type law: the turkish case*, Commun. Stat. Simulat 48 (2019), 2552–2565.

[2] F. Bartolucci, S. Bacci, and F. Pennoni, *Longitudinal analysis of self-reported health status by mixture latent auto-regressive models*,J. Roy. Stat. Soc. C-App. 63 (2014), 267–288.

[3] F. Bartolucci, A. Farcomeni, and F. Pennoni, *Latent Markov models for longitudinal data*. Chapman and Hall/CRC, 2012.

[4] F. Bartolucci, M. Lupparelli, and G. E. Montanari, *Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes*, Ann. Appl. Stat. 3 (2009), 611–636.

[5] F. Bartolucci, S. Pandolfi, and F. Pennoni, *LMest: an R package for latent Markov models for longitudinal categorical data*, J. Stat. Softw. 81 (2017), 1-38.

[6] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*, Ann. Math. Stat. 41 (1970), 164-171.

[7] A. Begun et al., *Identification of a multistate continuous-time nonhomogeneous Markov chain model for patients with decreased renal function*, Med. Decis. Making 33 (2013), 298–306.

[8] R. A. Chaer, R. Billeh, and M. G. Massad, *Genetics and gene manipulation therapy of premature coronary artery disease*, Cardiology 101 (2004), 122–130.

[9] C. W. Chuang et al., *Increased subsequent risk of coronary heart disease in primary cesarean delivery women: A population-based cohort study*, J. Womens Health 28 (2019), 323–330.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Stat. Soc. B Met. 39 (1977), 1-22.

[11] P. L. Donni, *The unobserved pattern of material hardship and health among older Americans*, J. Health Econ. 65 (2019), 31–42.

[12] M. Fedchenko et al., *Ischemic heart disease in children and young adults with congenital heart disease in Sweden*, Int. J. Cardiol. 248 (2017), 143-148.

[13] M. Garcia et al., *Cardiovascular disease in women: clinical perspectives*, Circ. Res. 118 (2016), 1273-1293.

[14] J. Gil, P. Li Donni, and E. Zucchelli, *Uncontrolled diabetes and health care utilisation: a bivariate latent Markov model approach*, Health Econ. 28 (2019), 1262–1276.

[15] H. J. Hsieh, T. H. H. Chen, and S. H. Chang, *Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan*, Stat. Med. 21 (2002), 3369–3382.

[16] B. Khoshnood et al., *Prevalence, timing of diagnosis and mortality of newborns with congenital heart defects: a population-based study*, Heart 98 (2012), 1667–1673.

[17] A. Koukounari et al., *Using a nonparametric multilevel latent Markov model to evaluate diagnostics for trachoma*, Am. J. Epidemiol. 177 (2013), 913–922.

[18] B. C. Kwon et al., *Dpvis: Visual analytics with hidden Markov models for disease progression pathways*, IEEE T. Vis. Comput. Gr. 27 (2021), 3685-3700.

[19] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, 2019.

[20] Y. Luo et al., *Bayesian latent multi-state modeling for nonequidistant longitudinal electronic health records*, Biometrics 77 (2021), 78–90.

[21] H. R. Mannan and J. J. Koval, *Latent mixed Markov modelling of smoking transitions using Monte Carlo bootstrapping*, Stat Methods Med Res 12 (2003), 125–146.

[22] D. S. Meenaxi and N. Singh, *A reliability model for the progression of chronic heart failure*, Int. J. Appl. Eng. Res. 13 (2018), 15351–15355.

[23] Mehta, P. K., Wei, J., and Wenger, N. K. (2015). Ischemic heart disease in women: a focus on risk factors. *Trends Cardiovas. Med.* **25**, 140–151.

[24] S. Menard, *Panel longitudinal studies*, Elsevier, 2005.

[25] G. E. Montanari and M. Doretti, *Ranking nursing homes' performances through a latent Markov model with fixed and random effects*, Soc. Indic. Res. 146 (2019), 307–326.

[26] M. A. Nordlie, L. E. Wold, and R. A. Kloner, *Genetic contributors toward increased risk for ischemic heart disease*, J. Mol. Cell. Cardiol. 39 (2005), 667–679.

[27] İ. Ş. Özer et al., *Genç İskemik İnmeli Hastaların Etiyolojik İnme Sebepleri, Risk Faktörleri Ve İzlemdeki Fonksiyonel Durumları*, Turk. J. Neurol. 21 (2015), 159-164.

[28] A. Rawandi et al., *Etiology And Risk Factors In The Young Patients With Ischemic Stroke*, Turk Beyin Damar Hast. Derg. 26 (2020), 126-132.

[29] S. Razvi et al., *The influence of age on the relationship between subclinical hypothyroidism and ischemic heart disease: a metaanalysis*, J. Clin. Endocr. Metab. 93 (2008), 2998–3007.

[30] J. B. Rubin and W. B. Borden, *Coronary heart disease in young adults*, Curr. Atheroscler. Rep. 14 (2012), 140–149.

[31] K. M. Schmidt et al., *Stable ischemic heart disease in women*, Curr. Treat. Options Cardiovasc. Med. 20 (2018), 1–13.

[32] H. C. Shih et al, *Estimation of progression of multi-state chronic disease using the Markov model and prevalence pool concept*, BMC Med. Inform. Decis. 7 (2007), 1–12.

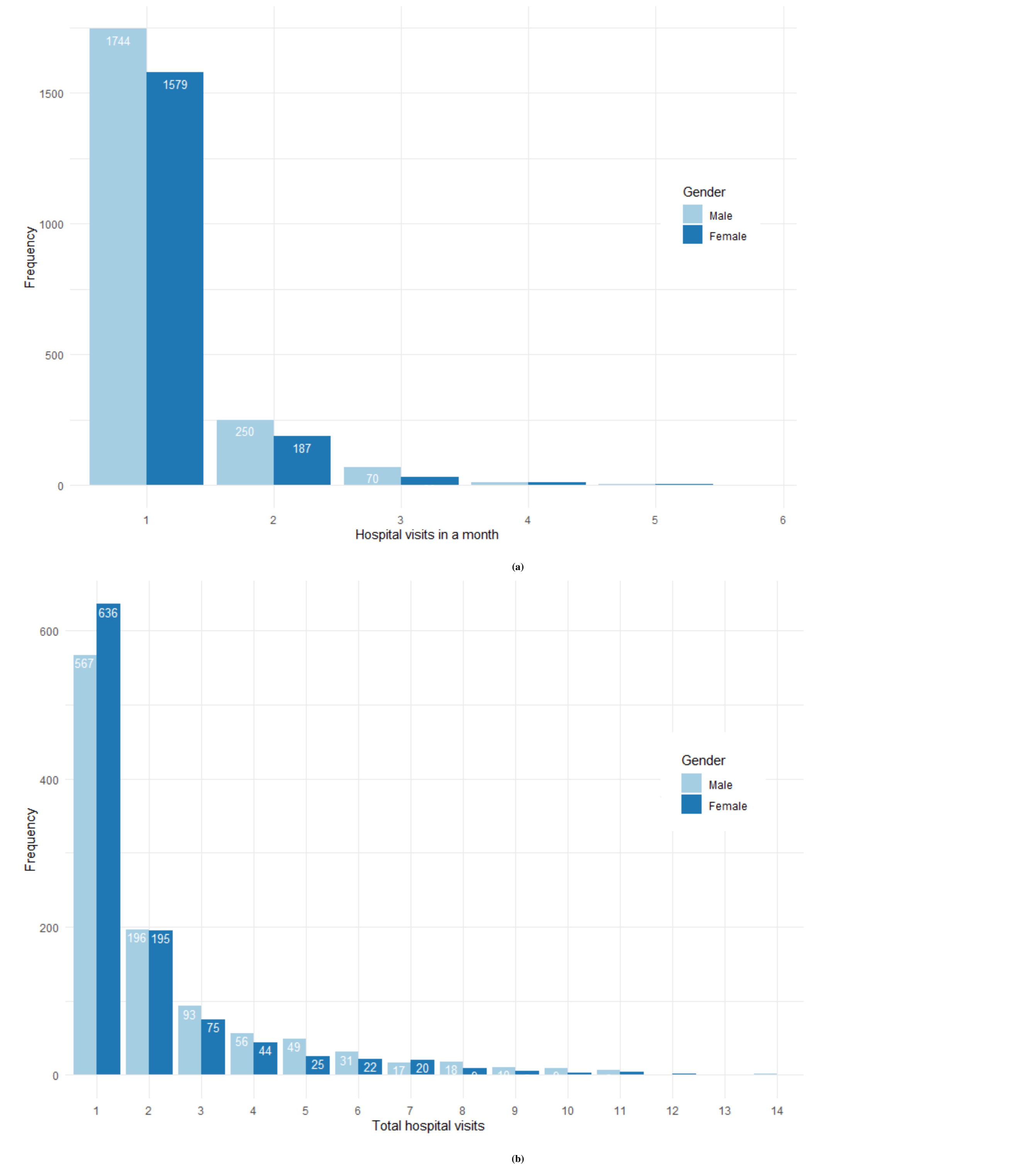
[33] B. Ünal et al., *Türkiye kronik hastalıklar ve risk faktörleri sıklığı çalışması*, Türkiye Halk Sağlığı Kurumu, 2013

[34] Y. Wang et al., *Predicting the spatio-temporal evolution of chronic diseases in population with human mobility data, Proceedings of 27th International Joint Conference on Artificial Intelligence*, 2018, International Joint Conferences on Artificial Intelligence, pp. 3578-3584.

[35] L. M. Wiggins,*Panel analysis: Latent probability models for attitude and behavior processes*. Elsevier, 1973.

[36] B. Wouterse et al., *The effect of trends in health and longevity on health services use by older adults*, BMC Health Serv. Res. 15 (2015), 1–14.

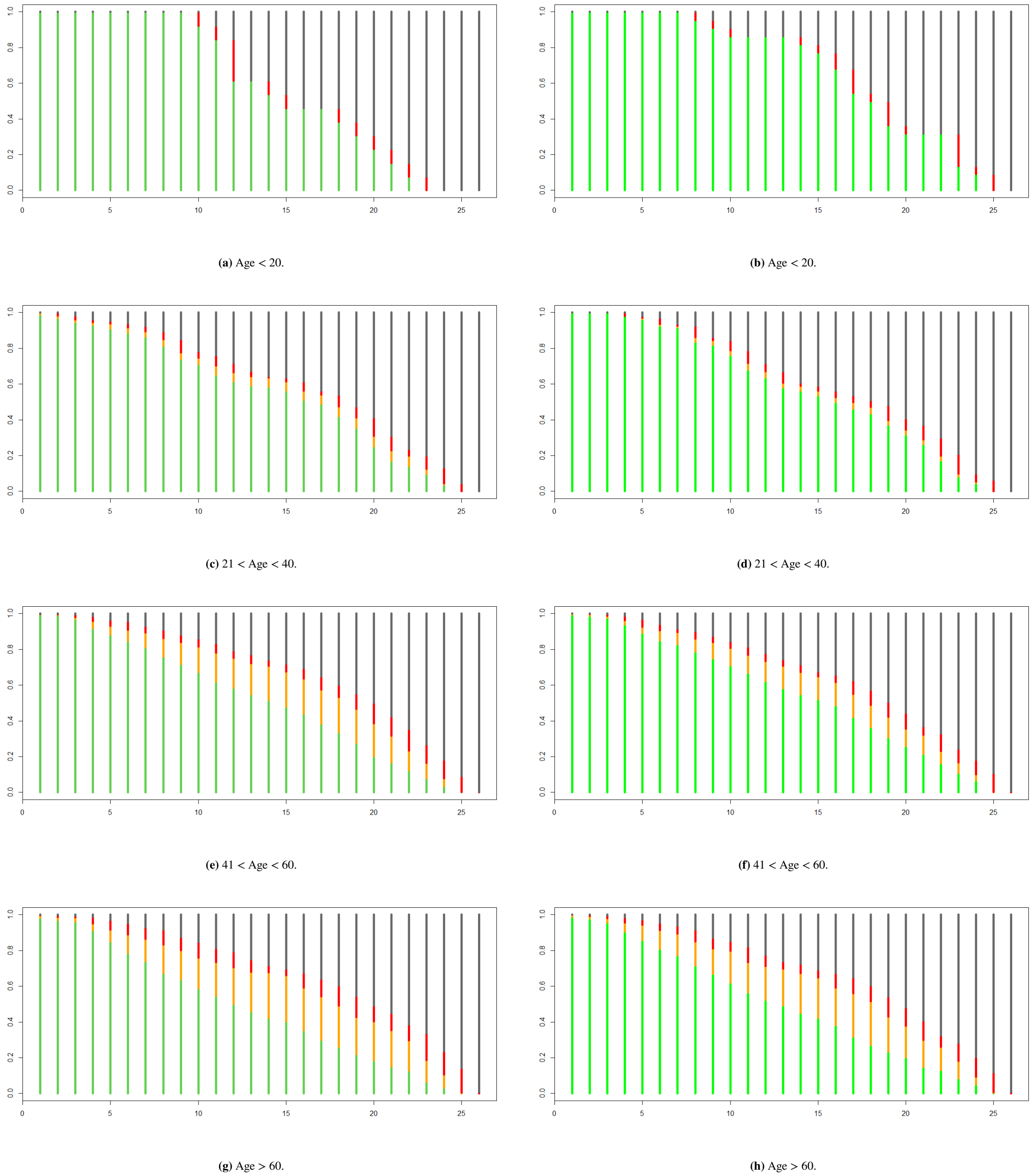
□



(a)

(b)

FIGURE 2 The number of visits to hospitals due to IHD per gender: (a) monthly, (b) daily between 2007 and 2009. This figure appears in color in the electronic version of this article, and color refers to that version.



(a) Age < 20.

(b) Age < 20.

(c) 21 < Age < 40.

(d) 21 < Age < 40.

(e) 41 < Age < 60.

(f) 41 < Age < 60.

(g) Age > 60.

(h) Age > 60.

FIGURE 3 Global decoding results. Horizontal axes are 26 months, vertical axes provide the probability of occurrence of the most frequent latent state over individuals. Healthy, light, moderate, and severe states are displayed by green, yellow, red, and black colors, respectively. Males are displayed on the left side, females are on the right side. This figure appears in color in the electronic version of this article, and color refers to that version.