Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
1–21

DOI: 10.1177/09622802231155100
journals.sagepub.com/home/smm
SAGE

Modeling comorbidity of chronic diseases using coupled hidden Markov model with bivariate discrete copula

Zarina Oflaz[1], Ceylan Yozgatligil[2] and A Sevtap Selcuk-Kestel[3]

Abstract

A range of chronic diseases have a significant influence on each other and share common risk factors. Comorbidity, which shows the existence of two or more diseases interacting or triggering each other, is an important measure for actuarial valuations. The main proposal of the study is to model parallel interacting processes describing two or more chronic diseases by a combination of hidden Markov theory and copula function. This study introduces a coupled hidden Markov model with the bivariate discrete copula function in the hidden process. To estimate the parameters of the model and deal with the numerical intractability of the log-likelihood, we use a variational expectation maximization algorithm. To perform the variational expectation maximization algorithm, a lower bound of the model's log-likelihood is defined, and estimators of the parameters are computed in the M-part. A possible numerical underflow occurring in the computation of forward–backward probabilities is solved. The simulation study is conducted for two different levels of association to assess the performance of the proposed model, resulting in satisfactory findings. The proposed model was applied to hospital appointment data from a private hospital. The model defines the dependency structure of unobserved disease data and its dynamics. The application results demonstrate that the model is useful for investigating disease comorbidity when only population dynamics over time and no clinical data are available.

Keywords

Coupled hidden Markov model, discrete copula, dependency in hidden process, comorbidity, chronic diseases

1 Introduction

Comorbidity is a medical term that refers to when a patient has two or more diseases concurrently. A range of chronic diseases have a significant influence on each other and share common risk factors. Morbidity is more likely to result from complications caused by comorbid conditions than from the primary illness itself.[1] It can refer to a variety of chronic diseases that are highly interconnected and share risk factors.[2,3]

Statistical models have been used to identify patterns of co-occurrence of diseases.[4–6] These studies, however, address the question of which comorbidities frequently co-occur but do not model their progression. Comorbidities are represented as networks using dynamic structural equation models.[7,8] or deep diffusion processes.[9] The studies provide information on which diseases co-occur but not on the dynamics of diseases. Bayesian networks have been used to examine comorbidities and their temporal relationships.[10–12] These works, which are based on a variety of different clinical and demographic data about patients, provide only a brief understanding of the underlying disease states representing the illness development. The interaction among diabetes and chronic liver disease under Metformin treatment is modeled by a coupled hidden Markov model (CHMM) with a personalized, non-homogeneous transition mechanism.[13] The model has a fixed number of hidden states but a greater number of states may be more informative about the progression of comorbidity.

[1]Department of Industrial Engineering, 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

Corresponding author:
Zarina Oflaz, Department of Industrial Engineering, KTO Karatay University, Akabe Mah. Alaaddin Kap Cad. No:130 42020 Karatay/Konya, Turkey.
Email: zarina.oflaz@karatay.edu.tr

Also, the model is applied to clinical and demographic data, however, the application of the model to restricted administrative data may be limited.

Comorbidity is used in studies employing hidden Markov model (HMM) and its extensions.[14–16] However, Markov models with memoryless properties imply that a patient's current state is distinct from their future trajectory. As a result, HMM-based models are unable to adequately explain the variability in individuals' progression trajectories, which are frequently caused by their varied clinical histories or chronologies (order and timing) of clinical events. This is a critical shortcoming in models of survival analysis for complex chronic diseases with numerous morbidities.[17]

Our aim is to investigate these unknown shared factors that contribute to the development of these diseases. We are particularly interested in estimating the probability of chronic disease comorbidity, or how the presence of one disease may affect the likelihood of being exposed to another.

The theoretical set up in this study is mainly to combine hidden Markov and copula theories describing the probabilistic joint behavior of two or more chronic diseases. We use a CHMM as the fundamental model in this study to capture the comorbidity phenomenon by combining the hidden processes underlying chronic diseases. The causal nature of disease-disease interactions can be represented using dynamic networks, in which the strength of the edges connecting nodes varies over time in accordance to the joint copula distribution. The underlying network dynamics are frequently unknown, and what we perceive are sequences of observed events propagating across the network. To infer latent network dynamics from observed sequences, one must consider both when and what events occurred in the past, as both provide insight into the mechanisms behind disease generation and progression.

We propose a novel CHMM with copula accounting for interaction between diseases in the latent space. Since the hidden process of the CHMM is defined on a discrete state space, the probability of joint hidden states is modeled by bivariate discrete copula proposed by Geenens.[18] Sklar's theorem states that the copula of a discrete random vector is not completely identifiable, resulting in serious inconsistencies. Therefore, Geenens[18] develops the rejuvenating approach of copula modeling for discrete data based on Yule's,[19] Goodman and Kruskal's,[20] and Mosteller's[21] conceptions.

Several research have used copula theory to examine competing risks and comorbidity. While competing risks are investigated using a variety of copula-based approaches,[22–26] there is only one study examining chronic disease comorbidity using a copula-based approach.[27] To our knowledge, no study has been conducted that uses a combination of any type of HMM and copula function to investigate comorbidity or competing risks.

The proposed model is an incomplete data model for which the expectation maximization (EM) algorithm[28] is the most often used probability maximization technique. However, the model's exact inference creates a number of computing problems. In this study, we define a probabilistic model in general form when the number of underlying hidden chains exceeds two, resulting in a large number of parameters. We employ variational approximation to make the expectation (E) step of the EM algorithm tractable in terms of computation. The resulting variational EM (VEM) seeks to maximize the lower bound on the log-likelihood. While VEM is initially formally described in machine learning applications such as Saul and Jordan,[29] it is now frequently deployed and generalized in a variety of ways. Comprehensive summary of studies can be found in the works of Jaakkola,[30] Blei et al.,[31] Wainwright and Jordan.[32] Ormerod and Wand[33] gives also an explanation of VEM in statistical terms.

In this study, we present theoretical advances for defining a probabilistic model and developing the necessary inference to implement the model. In particular, we compute the complete data log-likelihood (CDLL) and its lower bound, derive forward–backward probabilities and conditional expectations required for the E-step, and derive estimators for the model parameters required for the maximization (M) step. We propose an approximate inference algorithm based on a variational approach. A simulation study is performed to assess the performance of the proposed method, and an application to the detection of comorbidity levels in heart diseases and hypertension is presented. In the study by Oflaz et al.,[34] a latent Markov model with covariates influencing hidden states is used to examine the progression of ischemic heart disease (IHD). Extracted information on background factors leading to or influencing disease onset or progression is helpful for identifying hidden comorbidity levels of IHD and hypertension.

The proposed model takes into account the time influence of the disease improvement on the patient in both univariate and bivariate forms; exposes the factors having an impact on the diseases both in univariate and bivariate forms; controls the dependence structure in bivariate dimensions.

Interaction among diseases is implemented on limited administrative time series data not clinical data. It is critical to note that the majority of studies on competing risks rely on data from detailed clinical observations. However, a lack of data, particularly clinical data on individuals, may make classical statistical models of competing risks difficult to use. This could be due to restricted access to hospital data or a dearth of information on a particular cause or geographic location. As a result, it is necessary to develop alternative statistical models for sparse data.

Comorbidity of diseases have various etiological models, where risk factors have an influential role.[35] For example, in the associated risk factors model, risk factors for one disease are correlated with risk factors for another, increasing the

likelihood of the diseases occurring concurrently. On the other hand, in the heterogeneity model, disease risk factors are not correlated, but each is capable of causing diseases associated with the other risk factor among diseases. The proposed model can be applicable to study various pathways to comorbidity, assuming that these interactions happen in an unobserved process.

1.1 Extensions of HMM

A strength of the HMM is that it can be readily extended to accommodate a variety of applications. In the literature, a linear combination of the prior estimate of the transition matrix and the empirical transition matrix is established by Siu et al.,[36] Monte Carlo Markov chain is used to perform Bayesian inference and evaluate the posterior distribution of transition matrix.[37] The bivariate HMM have been developed to study the dependency between discrete and continuous observations.[38]

There are established approaches to model interaction of several processes by combination of two or more HMMs. Hierarchical HMM is a model where each hidden state is an HMM as well, where children states depend on parent states.[39] Factorial HMM splits the hidden state into multiple variables that are merged at output, and each state has its own transition matrix.[40] Event-specific HMMs developed by Kristjansson et al.,[41] aims to model a class of weakly connected time series in which only the onset of events are coupled in time. The representation capacity of event-coupled HMMs is clearly constrained by the restrictive structure, which is designed for a relatively narrow class of applications. Coupled HMM factor HMM to many chains in which the present state of each chain is dependent on the prior state of all chains.[42] Kwon and Murphy[43] models traffic velocities by using CHMM. Clearly, the completely coupled architecture developed by Brand et al.[42] is the most powerful in terms of representing interactions between many sequences. This framework can be used to naturally model a wide variety of applications.[44]

Our model is distinguished by the fact that we combine interacting processes via a copula function, implying that joint hidden states follow a predefined joint probability. Other coupled HMMs with a variety of architectures combine hidden chains through the use of conditional probabilities; that is, hidden states are connected to preceding states following Markov property, and their distribution is defined solely by transition state probabilities. Assuming that hidden states are dependent on one another and have a joint probability distribution, and as we cannot observe hidden states directly, using a copula to represent the dependence structure is the optimal choice. Copulas have evolved into one of the most widely used statistical tools for describing, analyzing, and modeling the dependence of random variables.

There are several studies on integrating copula with an HMM. To construct a dependency between the intensity levels of the various modalities, a Gaussian copula is utilized, that is, marginal distributions of those are linked by a copula.[45] Another way of merging two theories is importing hidden Markov chain in the copula parameter.[46] Instead of the assumption of conditional independence between observed variables and hidden states researches suggest studying the dependence of observed values on unknown states via copula.[47] Also, there is a copula-based HMM of cylindrical time series, where a mixture of copula-based cylindrical densities approximates the distribution of cylindrical data, the parameters of which rely on the development of a latent Markov chain.[48] A copula is used to construct the dependence structure of the Markov process by providing copula representation of the Markov property.[49] To our best knowledge, there is no study that combines hidden chains using the copula function in the context of a coupled HMM.

2 Copula modeling for discrete random variables

Copulas are utilized to extract the dependency structure of a multivariate distribution. We can construct any multivariate distribution by providing the marginal distributions, F and G, and its copula separately. Copulas naturally arise in statistical modeling as a result of the well-known Sklar's theorem.[50] When both F and G are continuous, Sklar's theorem indicates that the various objects presented in definition of copula[50] coincide and the set of "sub-copulas," $\mathcal{A}$, consists of the unique copula associated with $\mathbb{P}(X \leq x,\ Y \leq y)$.[51] When F and G are not continuous, their inverses exhibit plateaus; also, a copula representation for discrete functions exists in $\mathcal{A}$, but it is no longer unique, resulting in an identifiability issue.

Due to issues with the copula function's unidentifiability for discrete random variables, Geenens[18] proposes the bivariate discrete copula for count data. The existence and uniqueness of the proposed copula probability mass function is established.

In this study, the probability of joint hidden states is modeled by bivariate discrete copula, in particular we use bivariate binomial copula. With the constant trial numbers n, the dependency structure of a bivariate binomial is determined solely by the odds ratio parameter ω, and the associated Binomial(n)-copula is a one-parameter model. It is a $((n+1)\times(n+1))$-discrete distribution with uniform margins.

For example, if $n = 2$, the bivariate Binomial copula distribution is defined as follows:

$$\bar{\mathbf{p}} = \frac{1}{3}\begin{pmatrix} \frac{\omega(\omega+1)}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} & \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} & \frac{\omega+1}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} \\ \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} & \frac{\omega^2+4\omega+1-2\sqrt{\omega(\omega+2)(2\omega+1)}}{(\omega-1)^2} & \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} \\ \frac{\omega+1}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} & \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} & \frac{\omega(\omega+1)}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} \end{pmatrix}, \quad (1)$$

where ω is the odds ratio of the initial bivariate Bernoulli defined as

$$\omega = \frac{\mathbb{P}(X=0,\, Y=0)\mathbb{P}(X=1,\, Y=1)}{\mathbb{P}(X=1,\, Y=0)\mathbb{P}(X=0,\, Y=1)}. \quad (2)$$

For $\omega \neq 1$. For $\omega = 1$, $\bar{\mathbf{p}}$ is the (3×3)-independence copula p.m.f. One also has

$$\Upsilon = \frac{\omega^2 - 1}{\omega^2 + \omega + 1 + \sqrt{\omega(\omega+2)(2\omega+1)}}$$

as Yule's coefficient for this copula p.m.f., which is $\Upsilon = -1$ for $\omega = 0$ and $\Upsilon = 1$ for $\omega = \infty$. Thus, this family of Binomial copulas is exhaustive since it accommodates all values of Yule's coefficients between -1 and 1.

3 Coupled HMM with copula

To be able to explain the joint behavior of two discrete time series variables, we combine two hidden Markov chains by copula function. For three or more time series variables, we suppose that there are identical number of hidden chains and we combine each pair of hidden chains by copula function. The proposed CHMM with discrete bivariate copula function accounts for dependency between diseases, it is a system of multiple interacting processes. The interaction between variables is considered in the hidden space rather than the observation space.

A series of observations $X_i = (X_{i,t})$ is supposed to be the total number of patients with disease i, $(i = 1, \ldots, I)$, observed at time t, $t = 1, \ldots, T$. We denote hidden process for disease i as $(\mathbf{S}_i) = (S_{i,1}, S_{i,2}, \ldots, S_{i,T})$, where $S_{i,t}$ takes Q_i different values, $Q_i \in \mathbb{N}$. In this setting, the joint hidden process, denoted as $(S_t)_t$ or $(\mathbf{S}_t)$, with $S_t = (S_{1,t}, \ldots, S_{I,t})$, consists in $\prod_{i=1}^{I} Q_i$ possible values.

Emission distribution or state-dependent conditional distribution $\Phi_{i,q}(X_{i,t})$ might be an arbitrary discrete distribution

$$\Phi_{i,q}(X_{i,t}) = \mathbb{P}(X_{i,t}|S_{i,t} = q).$$

(a) Binomial emission distribution: We assume that the observed variable $X_{i,t}$ conditional on state $S_{i,t}$ follows Binomial distribution with probability of success, $p_{i,q}$, that is, the probability of having disease i conditional on state q. Therefore, $\Phi_{i,q}(X_{i,t})$ is defined for each disease i as follows:

$$\Phi_{i,q}(X_{i,t}) = \mathbb{P}(X_{i,t} = x_{i,t}|S_{i,t} = q) = \binom{n_i}{x_{i,t}} p_{i,q}^{x_{i,t}}(1-p_{i,q})^{n_i - x_{i,t}}, \quad x_{i,t} = 0, 1, 2, \ldots, n_i$$

where n_i is the total number of trials for disease i.

(b) Poisson emission distribution: We assume that the observed variable $X_{i,t}$ conditional on hidden state $S_{i,t}$ follows Poisson distribution with the rate parameter, $\lambda_{i,q}$. Therefore,

$$\Phi_{i,q}(X_{i,t}) = \mathbb{P}(X_{i,t} = x_{i,t}|S_{i,t} = q) = \frac{\lambda_{i,q}^{x_{i,t}} e^{-\lambda_{i,q}}}{x_{i,t}!}, \quad x_{i,t} = 0, 1, 2, \ldots,$$

where $\lambda_{i,q}$ is the parameter representing average number of patients with disease i conditional on state q.

3.1 Hidden Markov chain

We assume that the joint hidden process $(\mathbf{S}_t)$ fulfills a Markov property. Hidden dependency structure for two diseases is represented in Figure 1. The set of state of all diseases $(S_{i,t})_i$ is a Markov chain and the edges between the state of all diseases at a given time t for all individuals allows us to consider disease dependence.

We assume that the transition probabilities of the joint hidden process $(\mathbf{S}_t)$ arise from the product of two terms (both supposed to be constant along time): one accounting for the transitions within each disease and one accounting for the

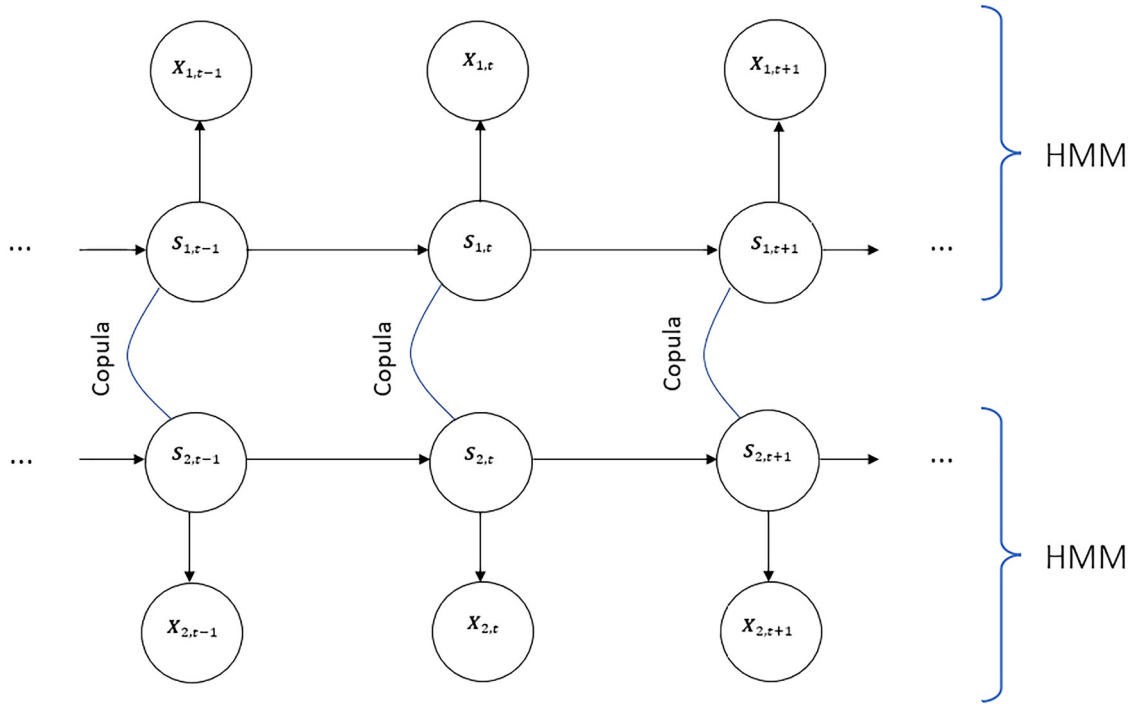

Figure 1. Directed graph of coupled hidden Markov model (CHMM) with copula function with two underlying Markov chains.

dependency between diseases by the discrete bivariate copula function defined as follows:

$$\mathbb{P}(\mathbf{S}_t = r|\mathbf{S}_{t-1} = q) =: P_{qr} \propto \pi_{q,r} \prod_i \prod_{j \neq i} \max\left(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)\right), \tag{3}$$

where

(i) π is a $Q \times Q$ transition matrix (each row sums to one) and q (resp. r) indicates the joint hidden state of the comorbidity of diseases;
(ii) the dependency relationship among the diseases is determined by the copula function, in particular, $\bar{\mathbf{p}} = \bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)$ is defined by equation (1) for Binomial copula.

In this model, π stands for the transitions within the comorbidity levels of diseases, while copula function introduces the dependency between diseases.

We further assume that the initial joint hidden process $\mathbf{S}_1 = (S_{i,1})_i$ has distribution

$$\mathbb{P}(\mathbf{S}_1 = q) \propto m_q \prod_i \prod_{j \neq i} \max\left(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)\right)$$

where m_q is an initial distribution of the states $1 \leq q \leq Q$, whose probability does not change across diseases. However, the model can be extended to include the assumption of varying transition probabilities across diseases; consequently, each hidden chain will have its own transition probability matrix and initial state probabilities.

The proposed method models the dependency of hidden chains by joint distribution of hidden states represented by discrete bivariate copula, whereas CHMM proposed by Brand et al.[42] couples chains by modeling the causal relationships between their hidden state variables with matrices of conditional probabilities.

The distribution of the hidden process $\mathbf{S}$ is defined as

$$\begin{aligned}\mathbb{P}(\mathbf{S}) = &\frac{1}{Z}\prod_{i,q}\prod_{j\neq i}[m_q \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))]^{\chi^q_{i,1}}\times \\ &\times \prod_{t\geq 2,q,r}\prod_{i}\pi_{q,r}^{\chi^q_{i,t-1}\chi^r_{i,t}}\prod_{j\neq i}\max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))^{\chi^q_{i,t-1}\chi^r_{i,t}},\end{aligned} \tag{4}$$

where Z is a normalizing constant and $\chi^r_{i,t}$ is an indicator function, $\chi^r_{i,t} = \mathbf{1}_{\{S_{i,t}=r\}}$.

Therefore, the joint probability for the sequence of states and observations is defined as follows:

$$\begin{aligned}\mathbb{P}(\mathbf{X}, \mathbf{S}) = &\frac{1}{Z}\prod_{i,q}\prod_{j\neq i}[m_q \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))]^{\chi^q_{i,1}}\times \\ &\times \prod_{t\geq 2,q,r}\prod_{i}\pi_{q,r}^{\chi^q_{i,t-1}\chi^r_{i,t}}\prod_{j\neq i}\max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))^{\chi^q_{i,t-1}\chi^r_{i,t}}\times \\ &\times \prod_{i,r,t}\Phi_{i,r}(X_{i,t})^{\chi^r_{i,t}}.\end{aligned} \tag{5}$$

It follows that the CDLL function is

$$\begin{aligned}\log\mathbb{P}(\mathbf{X}, \mathbf{S}) = &\sum_{i,q}\chi^q_{i,1}\log m_q + \sum_{i}\sum_{t\geq 2,q,r}\chi^q_{i,t-1}\chi^r_{i,t}\log\pi_{q,r} \\ &+ \sum_{i,r,t}\chi^r_{i,t}\sum_{j\neq i}\log\max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)) + \sum_{i,r,t}\chi^r_{i,t}\log\Phi_{i,r}(X_{i,t}) - \log Z.\end{aligned} \tag{6}$$

4 Variational EM algorithm

In the case of CHMM, where several hidden sequences are supposed to be dependent, the complexity of E-step becomes numerically inconvenient when number of parameters is too large. When both the number of chains and the number of states are small, that is, when $K = Q^I$ is less than a few tens, the global model can be viewed as a single HMM and the E-step can be performed via forward–backward recursion with complexity $\mathcal{O}(TK^2)$.[52] Therefore, we use a variational approximation of the E-step of the EM algorithm. As in HMM, the M-step for CHMM is straightforward and tractable.

The VEM algorithm maximizes a lower bound of the log-likelihood. In the study, we mainly rely on the approach of Wang et al.[52] and follow the lines of Jaakkola,[30] Wainwright and Jordan[32] to derive the variational approximation of the log-likelihood.

For any distribution $\tilde{P}$, we have

$$\begin{aligned}\log\mathbb{P}(\mathbf{X}) &\geq \log\mathbb{P}(\mathbf{X}) - KL[\tilde{\mathbb{P}}(\mathbf{S})|\mathbb{P}(\mathbf{S}|\mathbf{X})] \\ &= \tilde{E}\log\mathbb{P}(\mathbf{S}, \mathbf{X}) - \tilde{E}\log\tilde{\mathbb{P}}(\mathbf{S}) \\ &=: J(\mathbf{X}, \theta, \tilde{\mathbb{P}}),\end{aligned} \tag{7}$$

where $\tilde{E} = E_{\tilde{\mathbb{P}}}$ and KL denotes the Kullback-Leibler divergence. The maximization of CDLL turns into the maximization of the lower bound $J(\mathbf{X}, \theta, \tilde{\mathbb{P}})$ with respect to the parameter θ. As EM algorithm, VEM includes two steps:

VE-step: compute the approximate conditional distribution $\tilde{\mathbb{P}}$, given the observed data and the current value of the parameter θ^h, as

$$\tilde{\mathbb{P}}^{h+1} = \arg\max_{\tilde{\mathbb{P}}} J(\mathbf{X}, \theta^h, \tilde{\mathbb{P}}) = \arg\min_{\tilde{\mathbb{P}}} KL[\tilde{\mathbb{P}}(\mathbf{S})|\mathbb{P}(\mathbf{S}|\mathbf{X}; \theta^h)].$$

M-step: maximize the updated lower bound with respect to the set of parameters as

$$\theta^{h+1} = \arg\max_{\theta} J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^{h+1}).$$

The accuracy of this approximation depends mainly on the class of approximate distributions within which $\tilde{\mathbb{P}}$ is searched for. Here we employ the variational methods for graphical models proposed by Saul and Jordan[53] and modified for coupled

HMM by Ghahramani and Jordan.[54] The approximation bases on the setting $\tilde{P}$ to be a product of the independent Markov chains, that is,

$$\tilde{\mathbb{P}}(\mathbf{S}) = \prod_i \tilde{\mathbb{P}}(\mathbf{S}_i), \quad \text{where} \quad \tilde{\mathbb{P}}(\mathbf{S}_i) = \prod_i \tilde{\mathbb{P}}(S_{i,1}) \prod_{t\geq 2} \tilde{\mathbb{P}}(S_{i,1}|S_{i,t-1}). \tag{8}$$

Then we have

$$\tilde{\mathbb{P}}(\mathbf{S}_i) = \frac{1}{\tilde{Z}_i} \left(\prod_q (m_q h_{i,1}^q)^{\chi_{i,1}^q} \right) \prod_{t\geq 2} \left(\prod_{q,r} (\pi_{q,r} h_{i,t}^r)^{\chi_{i,t-1}^q \chi_{i,t}^r} \right), \tag{9}$$

where $\tilde{Z}_i$ is the normalizing constant that ensures that $\tilde{\mathbb{P}}(\mathbf{S}_i)$ equals one. The variational parameters $h_{i,t}^l$ can be thought of as correction terms for a Markov chain with parameters $(\mathbf{m}, \boldsymbol{\pi})$.

Let $\tau_{i,t}^r = \tilde{E}(\chi_{i,t}^r)$ and $\Lambda_{i,t}^{qr} = \tilde{E}(\chi_{i,t-1}^q \chi_{i,t}^r)$ denote the conditional expectations given the observations $\mathbf{X}_{i,t}$. According to these, we define the lower bound as shown in Theorem 1.

Theorem 1. Given observed random variable $X_{i,t}$, $X_{i,t} \in \mathbb{N}$, unobserved random variable $S_{i,t}$ taking Q different values, $Q \in \mathbb{N}$, $i, j = 1, \ldots, I$, $t = 1, \ldots, T$, $r = 1, \ldots, Q$ the lower bound $J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^h)$ of complete data log-likelihood $\log \mathbb{P}(\mathbf{X}, \mathbf{S})$ is defined as

$$\begin{aligned} J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^h) = \sum_{i,r,t} \tau_{i,t}^r \Big(\log \Phi_{i,r}(X_{i,t}) + \sum_{j\neq i} \log \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)) - \log h_{i,t}^r \Big) \\ + \sum_i \log \tilde{Z}_i - \log Z. \end{aligned} \tag{10}$$

Proof. Proof is provided in the Supplemental Material.

Based on these, we define analytically the optimal variation parameter in the E-part with the conditions for forward and backward algorithms.

Theorem 2. Given observed random variable $X_{i,t}$, $X_{i,t} \in \mathbb{N}$, unobserved random variable $S_{i,t}$ taking Q different values, $Q \in \mathbb{N}$, $i, j = 1, \ldots, I$, $t = 1, \ldots, T, r = 1, \ldots, Q$, the optimal value for the variation parameter $h_{i,t}^r$ of the lower bound $J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^h)$ is defined as

$$h_{i,t}^r = \Phi_{i,r}(X_{i,t}) \prod_{j\neq i} \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)).$$

Proof. Proof is provided in the Supplemental Material.

The conditional moment, which depend on the normalizing constant $\tilde{Z}_i$, are then computed using an independent forward–backward recursion for each disease.

Forward recursion: set $F_{i,1}^q \propto m_q h_{i,1}^q$ for $t \geq 2$ and compute

$$F_{i,t}^r \propto \sum_q F_{i,t-1}^q \pi_{q,r} h_{i,t}^r.$$

Backward recursion: $\tau_{i,T}^r = F_{i,T}^r$ holds and, for $1 \leq t \leq T-1$, compute

$$G_{i,t+1}^r = \sum_r F_{i,t}^q \pi_{q,r},$$

$$\Delta_{i,t}^{q,r} = \pi_{q,r} \frac{\tau_{i,t+1}^r}{G_{i,t+1}^r} F_{i,t}^q,$$

$$\tau_{i,t}^q = \sum_r \Delta_{i,t}^{q,r}.$$

The derivation of the M-part for transition and initial probabilities is done with respect to the proposed framework. In this respect, we define the M-part derivation under the assumptions of Binomial and Poisson distributions separately.

Theorem 3. Given observed random variable $X_{i,t}$, $X_{i,t} \in \mathbb{N}$, unobserved random variable $S_{i,t}$ taking Q different values, $Q \in \mathbb{N}$, $i, j = 1, \ldots, I$, $t = 1, \ldots, T$, $r = 1, \ldots, Q$, EM estimates of transition and initial probabilities are, respectively,

$$\hat{\pi}_{q,r} = \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}}, \tag{11}$$

$$\hat{m}_q = \frac{\sum_i \tau^r_{i,1}}{\sum_r \sum_i \tau^r_{i,1}}. \tag{12}$$

Proof. Proof is provided in the Supplemental Material.

Theorem 4. Given observed random variable $X_{i,t}$ following Binomial $(n; p_{i,q})$ distribution, $X_{i,t} \in \mathbb{N}$, unobserved random variable $S_{i,t}$ taking Q different values, $Q \in \mathbb{N}$, $i = 1, \ldots, I$, $t = 1, \ldots, T$,q, $r = 1, \ldots, Q$, the EM estimate of Binomial emission distribution parameter $p_{i,q}$ with fixed number of trials n is

$$\hat{p}_{i,q} = \frac{\sum_t \sum_r B^{q,r}_{i,t} X_{i,t}}{n \sum_t \sum_r B^{q,r}_{i,t}}, \tag{13}$$

where

$$B^{q,r}_{i,t} = \pi_{q,r} \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}.$$

Proof. Proof is provided in the Supplemental Material.

Theorem 5. Given observed random variable $X_{i,t}$ following Poisson $(\lambda_{i,q})$ distribution, $X_{i,t} \in \mathbb{N}$, unobserved random variable $S_{i,t}$ taking Q different values, $Q \in \mathbb{N}$, $i = 1, \ldots, I$, $t = 1, \ldots, T$, q, $r = 1, \ldots, Q$, the EM estimate of Poisson emission distribution parameter $\lambda_{i,q}$ is

$$\hat{\lambda}_{i,q} = \frac{\sum_t X_{i,t} \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}}{\sum_t \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}}. \tag{14}$$

Proof. Proof is provided in the Supplemental Material.

4.1 Estimation of the odds ratio

According to Wang et al.,[52] estimation of the dependency function within the M-step result in a poor estimation of the dependency function. The study suggests to use grid search of parameters and select the ones that minimize the weighted Residual Sum of Squares (RSS). Thus, we use the weighted RSS to select the optimal odds ratio, ω, for copula probabilities defined as follows:

$$RSS = \sum_{i,r,t} \tau^r_{i,t} (x_{i,t} - \mu_{i,r})^2, \tag{15}$$

where $\mu_{i,r} = \lambda_{i,r}$ and $\mu_{i,r} = np_{i,r}$ for Poisson and Binomial emission probabilities, respectively.

4.2 Algorithm and numerical stability

The proposed model is presented in algorithm to be more elaborated. The algorithm includes VE-part in steps 3 to 6, and the M-part in step 7.

The most frequently encountered problem in computing forward–backward probabilities, numerical underflow of the calculated probabilities, is resolved by transforming the functions used in the algorithm; additionally, scaling the observed values helps overcome numerical underflow. Precision is required when manipulating and executing arithmetic on small

Algorithm 1. VEM algorithm of CHMM with bivariate discrete copula.

1: Set initial values for initial state probabilities, m_r, transition probabilities, $\pi_{q,r}$, and probabilities of having disease i conditional on state r, $p_{i,r}$
2: Set or calculate ω for each pair of diseases
3: Calculate max($\bar{p}(S_i, S_j)$) for each pair of diseases. Calculate emission probabilities, $\Phi_{i,r}(X_{i,t})$, for each disease i. Then, for each disease i calculate

$$h_{i,t}^{r} = \Phi_{i,r}(X_{i,t}) \prod_{j \neq i} \max(\bar{p}(S_i, S_j))$$

4: Calculate forward probabilities: recursive calculation with initial $F_{i,1}^{q} \propto m_q h_{i,1}^{q}$ with normalized m_q
5: Calculate backward probabilities using obtained forward probabilities
6: Normalize obtained posterior probabilities and calculate the matrix $\Lambda_{i,t}^{qr}$ using posterior probabilities
7: Calculate/update parameters m_r, $\pi_{q,r}$, $p_{i,r}$ based on M-part formulas
8: Calculate weighted RSS value for each disease
9: Calculate criteria of convergence.
10: **if** criteria less than the threshold **then**
11: stop the algorithm, return the estimated parameters
12: **else**
13: update parameters and do steps 3-9
14: **end if**

numbers. When possible, it is recommended to work with logarithms of probabilities. In particular, the "log-sum-exp" trick is useful when dealing with numerical underflow,[55,31]

$$\log\left[\sum_i \exp(y_i)\right] = \alpha + \log\left[\sum_i \exp(y_i - \alpha)\right].$$

The constant α is typically set to $\max_i y_i$. This ensures that common computations in variational inference processes are numerically stable.

5 Comorbidity by simulation

We conduct a simulation study that requires specific modeling to generate the comorbidity to capture the proposed model. Data sets with odds ratios of 0.85 and 8, which represent different correlation structures between the variables, are simulated.

For an odds ratio of 1, hidden states are independent. This scenario is inapplicable to the proposed model, as it is intended to model dependencies between hidden states. Additionally, when $\omega = 1$, the joint probability of any two states is equal to $1/m$, where m is the number of states. In this situation, the CDLL of the model does not change to independent HMMs or classical CHMM, so there is no theoretical connection between the proposed model and known hidden Markov-based models.

When ω is negative, that is, outside of the range, the joint probabilities are outside of the [0, 1] range; consequently, further calculation steps are not possible, and the model cannot maximize CDLL.

Simulation set up for hidden states: Firstly, consider the Bernoulli distribution with a specified probability of success, 0 and 1 represent statuses of "no disease" and "having disease," respectively. Let the number of trials, n, be the number of medical check up or diagnostic analysis at the time t. We assume that $n = m - 1$, where m is the number of hidden states. For each time point, we generate vector of Bernoulli values (Figure 2, Step 1). Then, the outcomes are aggregated, so we get total number of successful trials, that is, number of diagnostic analyses resulting in having a disease, at the time t, $t = 1, \ldots, T$. Therefore, for two diseases, we generate two hidden chains with states following Binomial(n, p) distribution (Figure 2, Step 2). Thus, we obtain pair of generated values, (S_{1t}, S_{2t}), at time t. We suppose that (S_{1t}, S_{2t}) follows Binomial copula distribution with predefined odds ratio parameter ω. The number of joint states over the time period is arranged to be

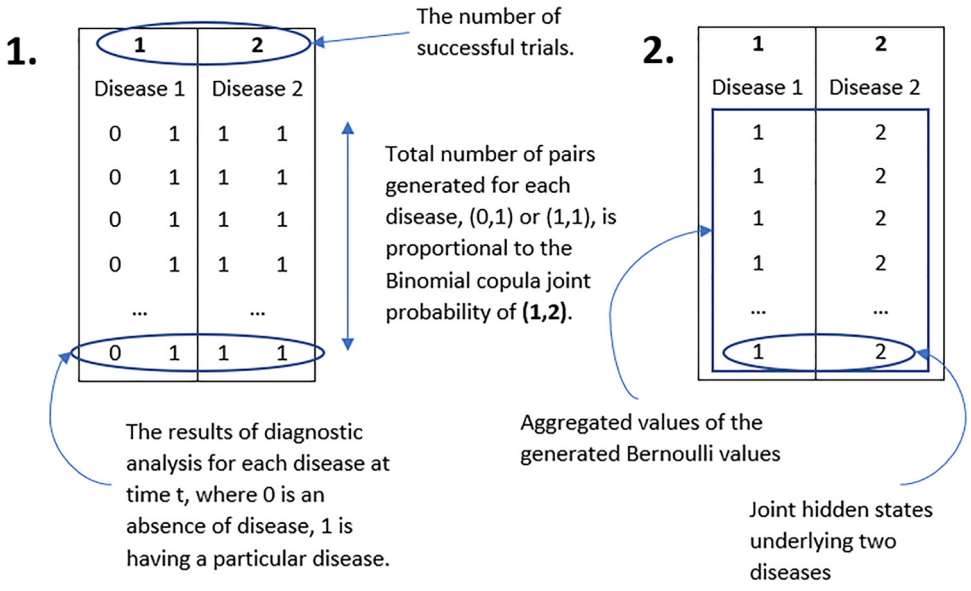

Figure 2. Simulation set up of hidden states for two diseases: generation and aggregation of Bernoulli values. For illustration purpose, only one out of eight possible pairs is displayed.

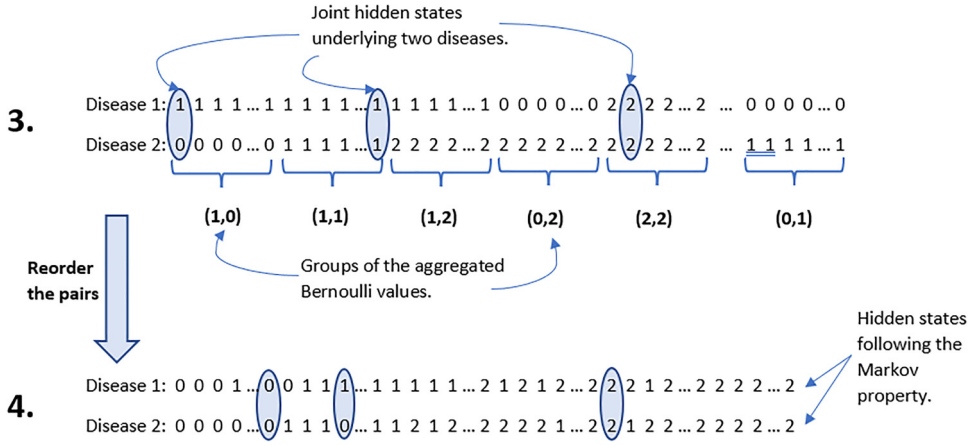

Figure 3. Simulation set up of hidden states for two diseases: reorder of the generated pairs.

proportional to the joint probability of states, defined in a Binomial copula matrix (equation (1)). Additionally, hidden chains should follow the Markov property, thus, to check whether the generated sequences follow the Markov property, chi-square based test proposed by Anderson and Goodman[56] is conducted. *verifyMarkovProperty* function from *markovchain* package[57] is used. We reorder the pairs of hidden states so that the sequences follow the Markov property (Figure 3, Steps 3 and 4). If the high proportion of sequences satisfies the Markov property, we do not drop non-Markov sequences and suppose that the established simulation is optimal.

We generate dependencies between the hidden chains using a bivariate Binomial copula, then observations dependent on hidden states are generated from a predefined distribution.

Simulation design steps for three state model:

S1: Calculate Binomial copula probabilities given a predefined odds ratio (equation (2)). For $n = 2$, Binomial copula distribution is a 3×3-matrix (equation (1)). The rows and columns of the matrix represent 0, 1, 2 number of successful trials for disease 1 and disease 2, respectively. The calculated Binomial copula probabilities represent the true distribution of the joint hidden states.

S2: Generate joint states as explained in simulation set up for hidden states. The frequency of the generated joint states and Binomial copula probabilities given in equation (1) are compared based on bias and mean square error (MSE),

$$Bias = \frac{\sum_{i=1}^{N}(x_i - y_i)}{N},$$

$$MSE = \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N},$$

where x_i is a true probability, y_i is a frequency of the generated joint states, $i = 1, \ldots, N$.

S3: Calculate transition probability matrices for each hidden chain. Aggregate probabilities of corresponding states and normalize by the total value in the corresponding row. Thus, the comorbidity transition matrix of the two diseases is obtained.

S4: Initial probability of the generated hidden states is defined as the frequency of the generated states at time $t = 1$.

S5: Finally, generate observations from Poisson distribution given λ_i parameter conditional on the hidden state for each disease i.

To test whether the CHMM copula model optimally fits the simulated data, estimated transition, and initial probabilities, λ_i parameters are compared with the parameters given in the simulation set up.

5.1 Model performance on the simulated data

The three state hidden Markov chains representing two diseases are generated from Binomial copula distribution according to the simulation design. The order of the generated Bernoulli trials for each disease is such that the joint hidden states follow the Binomial copula and the sequence of hidden states for each disease follows the Markov property. For each disease, observed values were generated using a Poisson distribution with hidden state-dependent rate parameters. For odds ratios of 0.85 and 8, two data sets were simulated 500 times. There are 120 time points for an odds ratio of 0.85, and 99 for an odds ratio of 8. To determine if the generated hidden states follow the Binomial copula distribution, bias and MSE were computed. The calculated error metrics are close to zero for both odds ratios. In addition, the Markov property was examined using a chi-square test.

According to the chi-square-based Markovian test, 98.6% of the generated hidden states for disease 1 and 99.4% for disease 2 satisfy the Markov property for an odds ratio of 0.85, while 87% of the generated hidden states for disease 1 and 93% for disease 2 satisfy the Markov property for an odds ratio of 8. According to the results of statistical tests, a large proportion of the 500 generated observations for both odds ratios and both diseases are stationary (Table 1).

To distinguish estimated values of the generated data set from the estimated values by model, we call them true parameters of the generated data, see Table 2.

To assess the performance of the proposed model, 50 simulated data sets with the same parameters were used. Since the estimation of the parameters is an iterative process, we need to decide on the optimal iteration number to capture the best estimated results. Models with identical starting parameter settings but a varied number of inner iterations (70, 100, 150, 200, and 250) were applied to the simulated data sets.

The true values of the simulation design parameters are compared to the model's estimated parameters using bias and MSE metrics. For clear demonstration of the model results on the simulated data, bias and MSE of the estimated parameters for each hidden state are calculated (see Tables 1 and 2 in the Supplemental Materials), and then mean of the values are displayed in Figure 4. Generally, for any number of iterations, the calculated bias and MSE between the estimated and true parameters of the simulation designs with an odds ratio of 0.85 and 8 are small, but for some cases we obtain moderately larger values. Having 16 parameters we aim to minimize the loss of the whole model while minimizing RSS, it is

Table 1. Proportion of stationary sequences among the 500 generated observations. Tests are conducted with significance level of 0.05. KPSS test for odds ratio of 8 also includes the results for significance level of 0.01.

	$\omega = 0.85$			$\omega = 8$			
	ADF	KPSS	PP	ADF	KPSS (0.05)	KPSS (0.01)	PP
Disease 1	99.2%	90.8%	100%	94%	61.4%	88%	100%
Disease 2	98.8%	94.4%	100%	93.8%	68.2%	88.6%	100%

KPSS: Kwiatkowski-Phillips-Schmidt-Shin; ADF: augmented Dickey-Fuller; PP: Phillips- Perron.

Table 2. True values of the generated hidden states and observations for odds ratios of 0.85 and 8.

	$\omega = 0.85$	$\omega = 8$
Transition matrix	$\begin{pmatrix} 0.354 & 0.464 & 0.182 \\ 0.257 & 0.469 & 0.273 \\ 0.171 & 0.455 & 0.374 \end{pmatrix}$	$\begin{pmatrix} 0.858 & 0.124 & 0.018 \\ 0.090 & 0.794 & 0.116 \\ 0.018 & 0.097 & 0.885 \end{pmatrix}$
Initial state probabilities	$\begin{pmatrix} 0.5 & 0.4 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 0.95 & 0.05 & 0 \end{pmatrix}$
Emission distribution rate parameter for disease 1	$\begin{pmatrix} 8 & 9 & 9.5 \end{pmatrix}$	$\begin{pmatrix} 8 & 9 & 9.5 \end{pmatrix}$
Emission distribution rate parameter for disease 2	$\begin{pmatrix} 7 & 8 & 8.5 \end{pmatrix}$	$\begin{pmatrix} 7 & 8 & 8.5 \end{pmatrix}$
Binomial copula p.m.f.	$\begin{pmatrix} 0.10 & 0.11 & 0.12 \\ 0.11 & 0.11 & 0.11 \\ 0.12 & 0.11 & 0.10 \end{pmatrix}$	$\begin{pmatrix} 0.218 & 0.088 & 0.027 \\ 0.088 & 0.158 & 0.088 \\ 0.027 & 0.088 & 0.218 \end{pmatrix}$

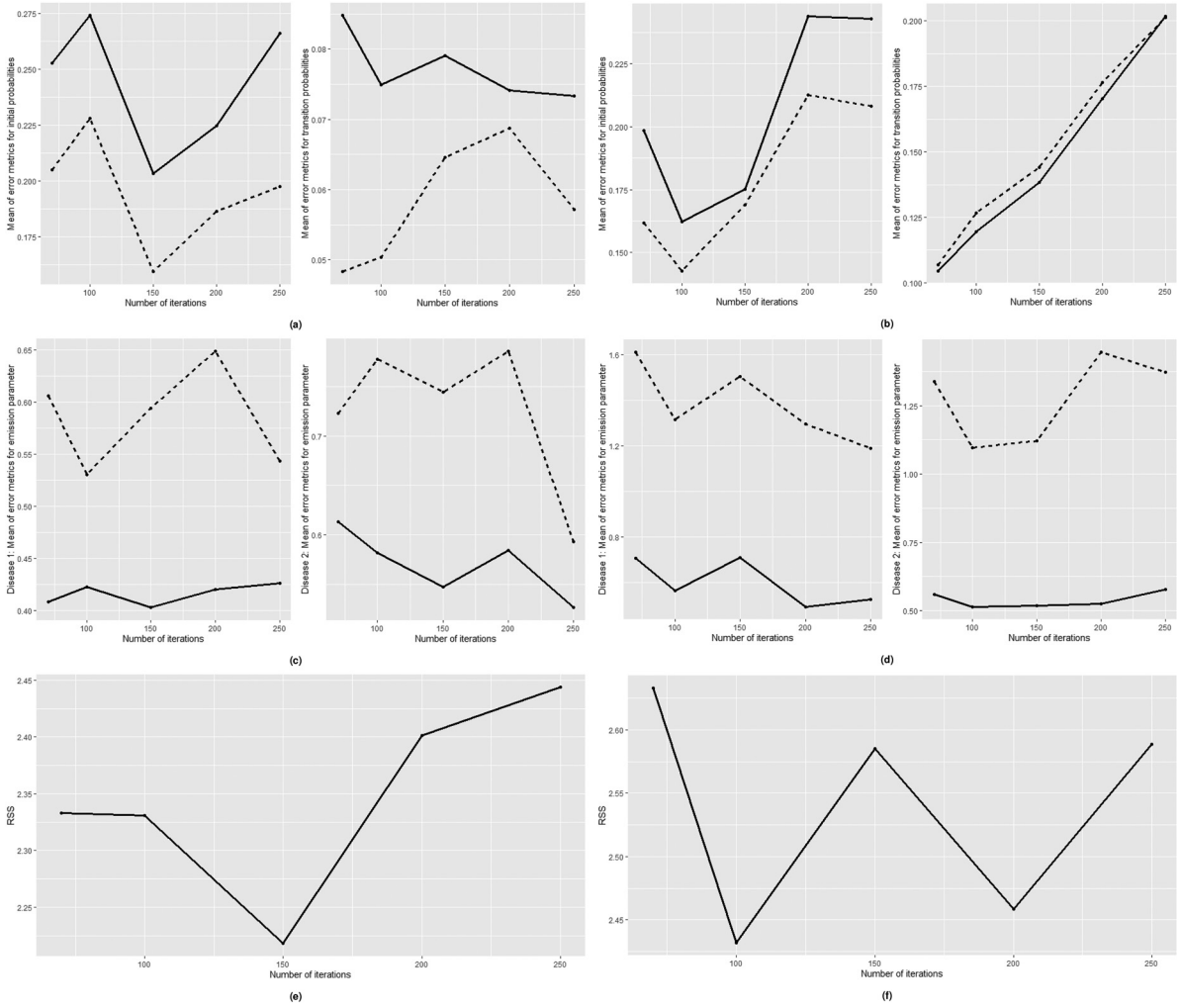

Figure 4. Mean of bias (dashed line) and MSE (solid line) of initial probabilities (left) and transition probabilities (right) for odds ratio of: a) 0.85 b) 8; mean of bias (dashed line) and MSE (solid line) of emission parameters for disease 1 (left) and disease 2 (right) for odds ratio of: c) 0.85 d) 8; weighted RSS values for odds ratio of: e) 0.85 f) 8. MSE: mean square error; RSS: Residual Sum of Squares.

challenging to minimize the bias of all parameters. We assume that the general results of the simulation study are acceptable; however, it is difficult to find well-defined initial parameter settings that produce unbiased results due to a large number of parameters. The mean of MSE of the estimated and true rate parameters for both diseases of the simulated

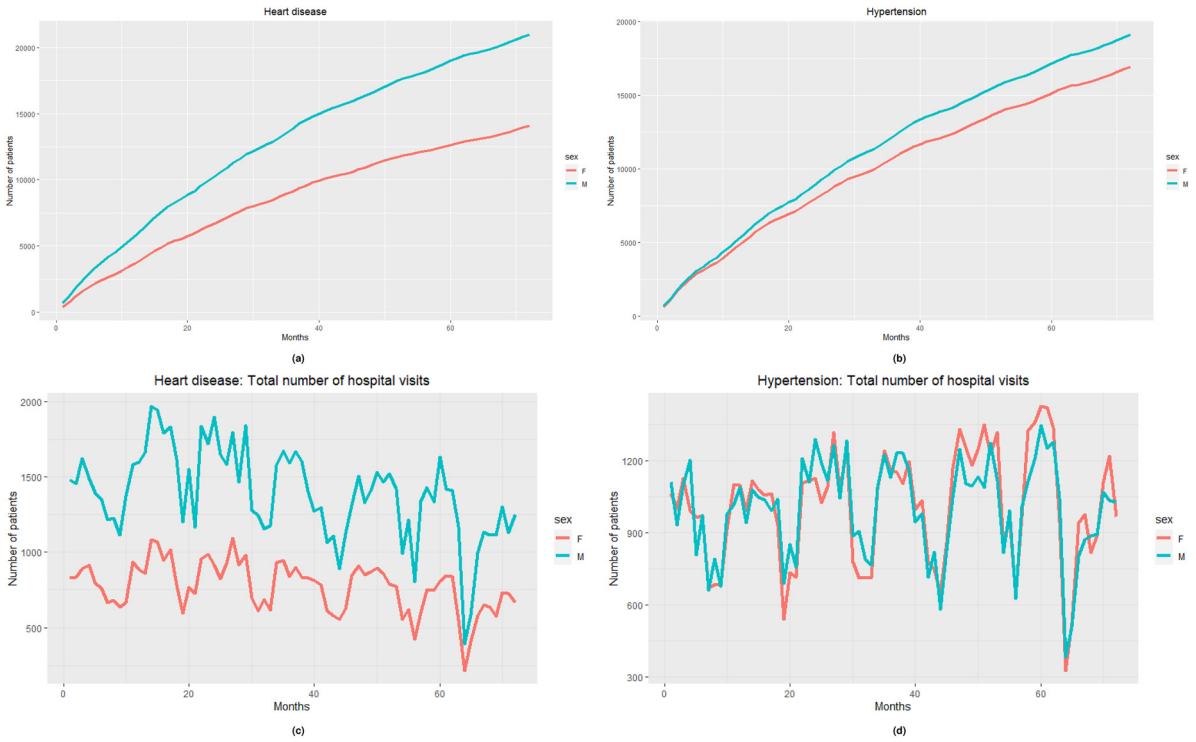

Figure 5. The number of patients by gender from January 2015 to December 2020 with: (a) heart disease, (b) hypertension. The number of hospital visits by gender from January 2015 to December 2020: (c) heart disease, and (d) hypertension.

data with an odds ratio of 8 occasionally exceeds one. The values of error metrics vary depending on the initial and transition probabilities, as well as the emission distribution rate parameters.

The proposed model with 100 or 150 inner iterations number display the lowest bias and MSE values for parameters of the simulated data with an odds ratio of 0.85 (Figure 4(a) and (c)), the model with 100 iterations is the most optimal for data with an odds ratio of 8 (Figure 4(b) and (d)).

Additionally, the weighted RSS is used to evaluate the performance of models with varying iteration numbers. The weighted RSS exhibits the lowest value after 150 and 100 iterations for odds ratios of 0.85 and 8, respectively (Figure 4(e) and (f)). The weighted RSS values support the most optimal number of iterations for acquiring parameters close to their true values on average.

For simulated data with an odds ratio of 0.85 indicating a negative and close to zero Yule's coefficient of association of -0.054, as well as for data with an odds ratio of 8 indicating a positive Yule's coefficient of association of 0.57, CHMM with discrete copula exhibits a satisfactory goodness of fit.

6 Real life application

In this section of the study, we present the applicability of the proposed method onto the real data set. The data set includes daily information relating to the patients with heart disease and the hypertension collected from 49,713 patients from January 2015 to December 2020. Data set is authorized by a private hospital in Turkey, the observations are provided with confidentiality condition.

The illness codes of raw data set were classified into two categories: heart disease and hypertension. The repeated rows have been removed. The data contains information on integer age and gender, indicator variables for ever-diagnosed heart disease and hypertension since 2015, whether an individual was admitted to a hospital in a given month, and the frequency with which the patient presented to the hospital with heart disease or hypertension. Once a patient has been diagnosed with a disease, the value for that disease remains 1 for the remainder of the column. As the prehistory of individuals before 2015 is unknown, it is assumed that patients who do not appear in the system for the chosen period did not have a disease until their first hospital visit. After cleaning and transforming panel data, it was aggregated to create time series data by counting patients with heart disease or hypertension in each month from January 2015 to December 2020.

The total number of patients with heart disease or/and hypertension was analyzed using time series analysis and exploratory data analysis. Figure 5(a) and (b) depicts the time-series plots of the total number of patients with a specific disease divided by gender. Both diseases have an increasing trend due to assumptions that if a patient is reported to have a particular disease, he will be assumed to have that disease until the study is completed. For both diseases, male patients are increasing faster than female patients over time.

Figure 5(c) and (d) shows the total number of hospital appointments for both diseases by gender. Hospital admissions for heart disease or/and hypertension follow a seasonal pattern. Females visit the hospital less frequently than males (check descriptive statistics), while hospital visits for patients with hypertension are comparable for both genders.

As illustrated in Figure 6(a) and (b), the number of patients with heart disease and hypertension, as well as patient hospital admissions, are highly linearly correlated. For both diseases, the total number of patients and their hospital visits follow a left-skewed distribution.

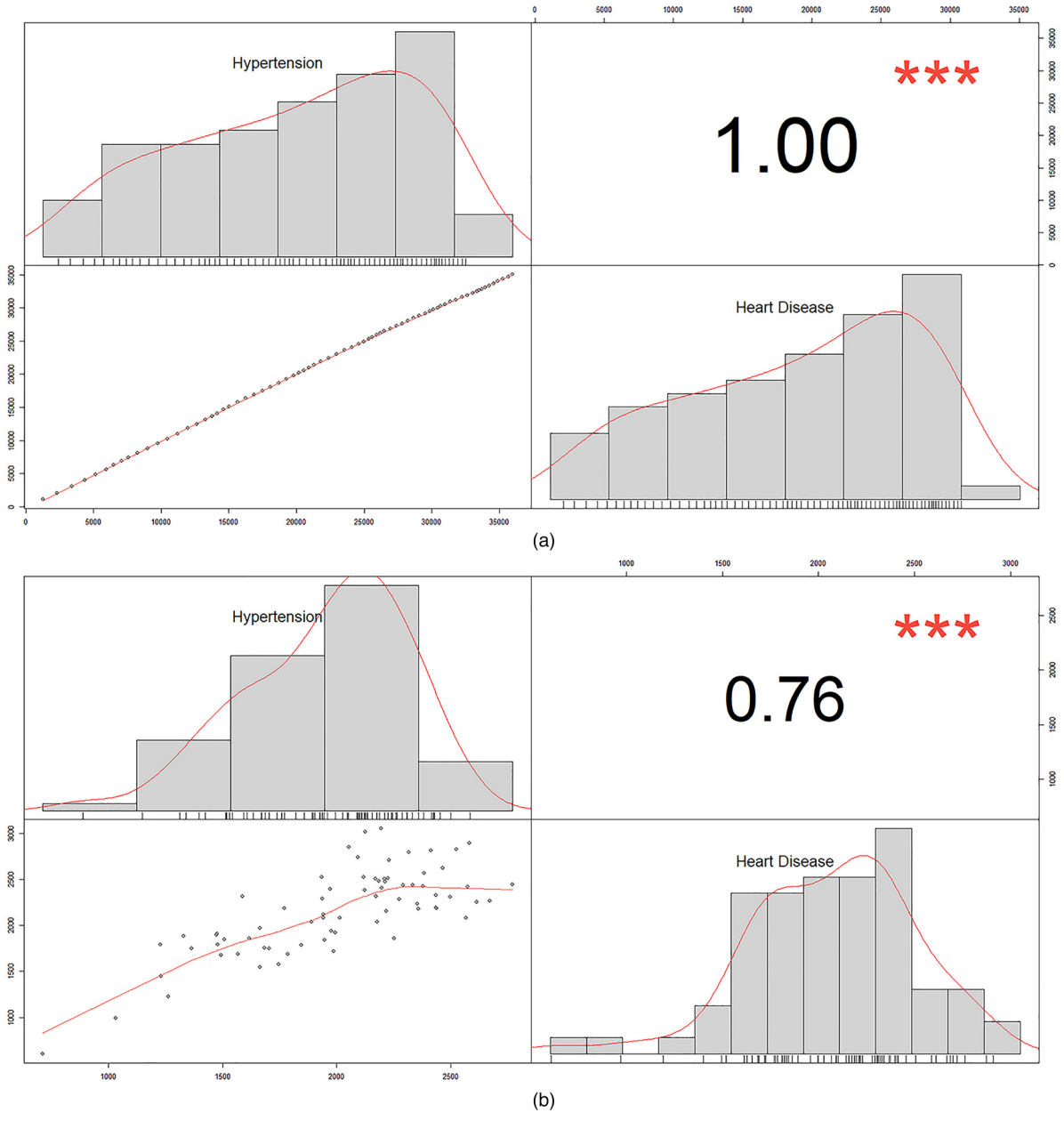

Figure 6. Heart disease and hypertension correlation plots and histograms: (a) patients number and (b) number of hospital visits.

We calculated the average time interval between disease occurrences. When hypertension occurs first, it takes an average of 12.96 months to be diagnosed with heart disease later; however, it takes an average of 18.88 months to be diagnosed with hypertension following heart disease. There are 4236 patients who develop heart disease after hypertension and 3108 patients who develop hypertension following heart disease.

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of the total number of patients over 72 months and those with the second-order difference for both diseases are represented in Figure 7. The time series of patients number for both diseases are nonstationary, have an increasing trend.

While HMMs are applicable to both stationary and nonstationary time series,[58] CHMMs, the proposed model's underlying model, do not include any theoretical information indicating whether this type of model can be used for nonstationary time series.[42] Additionally, application studies used stationary series or transformed nonstationary ones to work with CHMM or models based on CHMM.[59–62]

A Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is used to determine the number of differences required to make time-series stationary at the significance level of 0.05. Second-order differencing is performed on the time series of both diseases based on the test results. According to Figure 7, heart disease and hypertension series with second-order lag are stationary. The augmented Dickey-Fuller (ADF) test with p-values less than 0.01 for both diseases, the KPSS test with p-values greater than 0.01 for both diseases, and the Phillips-Perron (PP) test with p-value of 0.01 for heart disease and hypertension indicate that the series are stationary.

7 Application of the proposed model

The proposed model was applied to the number of patients with heart disease and hypertension resulting from the second-order differencing. Since the transformation produces negative values, the observations are increased by the minimum absolute value of the resulting observations, 270. Following that, the observed values are scaled by ten to obtain a stable VEM algorithm fit. These transformations does not affect the results because the aim is not to predict the number of patients but to predict the probability of co-occurrence. Initial setting for the model parameters are arranged as follows:

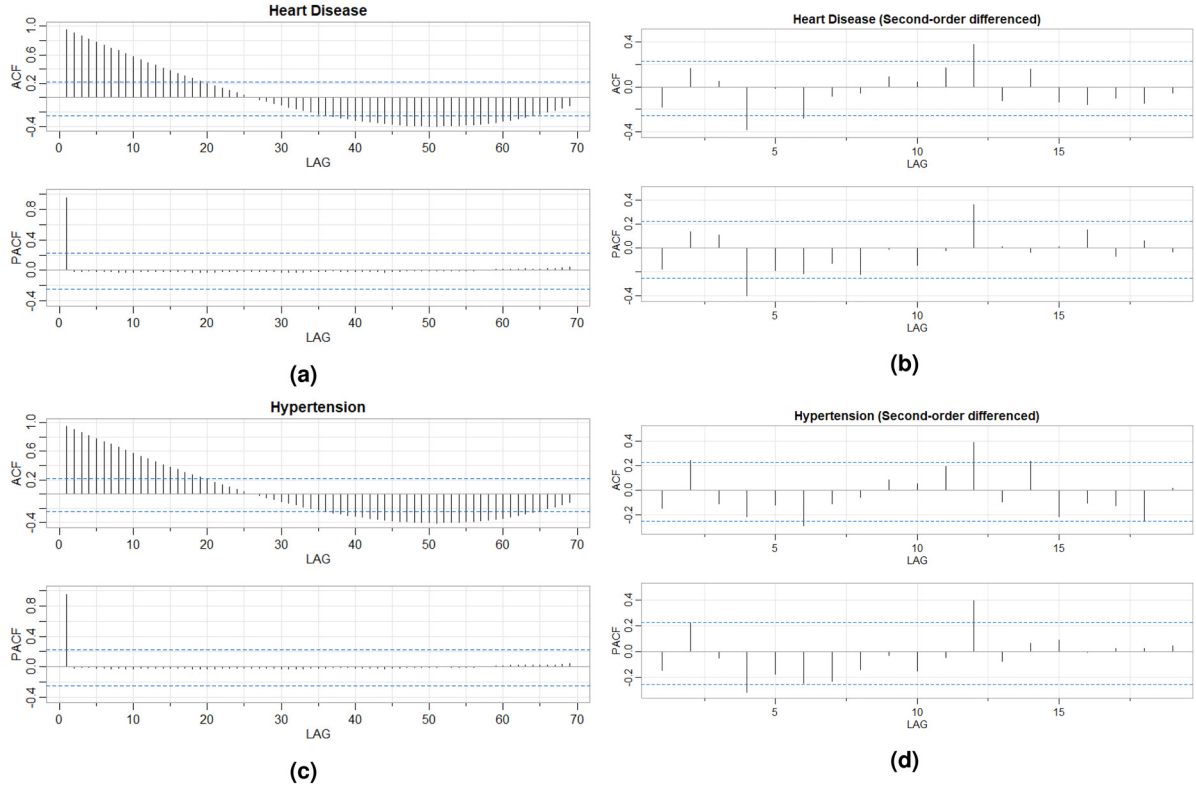

Figure 7. ACF and PACF plots of the number of patients over 72 months: (a) heart disease, (b) heart disease (second-order difference), (c) hypertension, and (d) hypertension (second-order difference). ACF: Autocorrelation Function; PACF: Partial Autocorrelation Function.

(i) transition probability matrix with the initial state probability, (0.9 0.1 0), is defined as

$$\begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix};$$

(ii) rate parameters for emission distribution are arranged according to the mean and median values of the observations, the initial values set to (22 26 30) for heart disease and (20 24 28) for hypertension.

To obtain the model with the optimum fit, the odds ratios between two hidden states of the diseases, (5, 15, 20, 25, 30, 50, 70, 100), were utilized. Due to the fact that we cannot observe the hidden states, in order to determine the odds ratio between the occurrences of the diseases' hidden states, we examined a range of odds ratio values and fitted the model for each odds ratio. The optimal odds ratio value is chosen based on the one with the lowest weighted RSS. Various inner iteration numbers were used, (150, 200, 250, 300, 350, 400); the optimal iteration number for each odds ratio is determined by the lowest weighted RSS value.

The estimated parameters of the model after 350 iterations for selected odds ratio of 70 are as follows:

- transition probability matrix is

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.000 & 0.000 & 1.000 \\ 0.023 & 0.972 & 0.005 \\ 0.000 & 0.074 & 0.926 \end{pmatrix} \end{array}$$

- initial state probabilities are (0.001 0.999 0);
- emission distribution parameters of the state-dependent observations for heart disease and hypertension are (25.23 26.08 26.46) and (24.38 27.68 28.70), respectively,

The weighted RSS value for the model is 10.38.

7.1 Interpretation of the findings

The Binomial copula distribution for estimated odds ratio of 70 is given as follows:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.2849 & 0.0443 & 0.0041 \\ 0.0443 & 0.2448 & 0.0443 \\ 0.0041 & 0.0443 & 0.2849 \end{pmatrix} \end{array}$$

The Yule's coefficient of association between hidden states representing heart disease and hypertension is 0.84; correlation analysis conducted on the data supports this result. High association of hidden chains representing heart disease and hypertension is supported by medical studies.[63,64] For example, pulmonary hypertension complicates the course of many adults with congenital heart diseases,[65] severity of pulmonary hypertension in patients with left heart diseases is studied.[66,67] According to the transition dynamic of hidden states and copula probabilities of joint states, we designate states as follows:

(i) State 2 represents no underlying risk factor;
(ii) State 1 corresponds to the presence of one risk factor;
(iii) State 3 corresponds to the presence of two risk factors.

Relationships between comorbid diseases are described by various etiological associations.[35] We suppose that the occurrence of diseases at the same or different states may recognize the level of comorbidity between diseases. For example, if a single risk factor affects two diseases, that is, both diseases are on state 1, then we suppose that these diseases have a light comorbidity. If both risk factors affect both diseases, that is, hypertension and IHD are on state 3, then severe comorbidity exists. Figure 8 lists all possible interactions between diseases and risk factors.

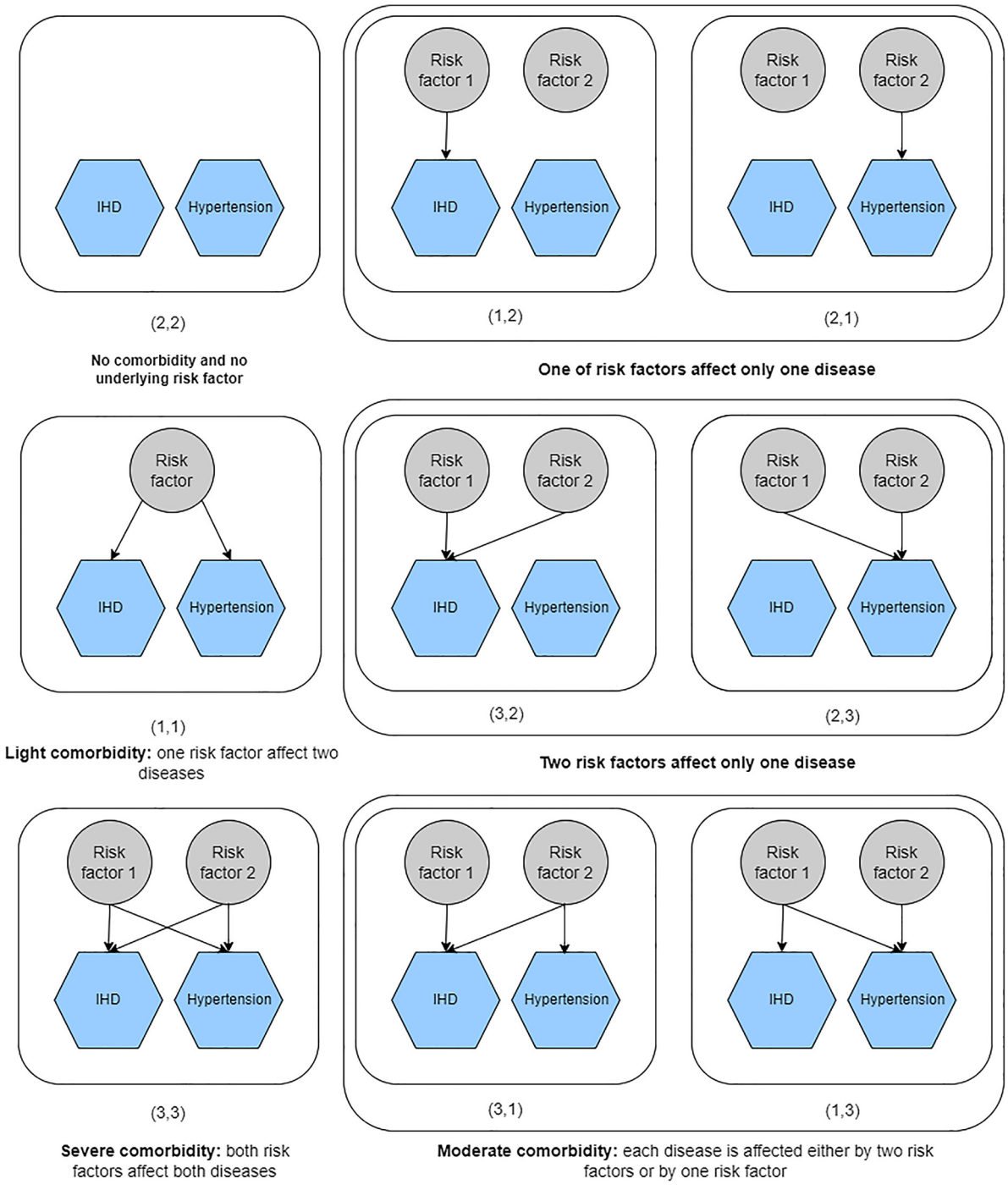

Figure 8. Interpretation of joint hidden states. Each structure relies on the interaction between diseases and risk factors.

According to the Binomial copula matrix, joint hidden states with different state numbers, such as (1, 2), (2, 1), (2, 3), (3, 2), (1, 3), and (3, 1), have a low probability of occurrence, implying that there is a low probability that two diseases occur on distinct states. For instance, there is a low possibility that heart disease is not affected by any risk factor, while hypertension is affected by two risk factors. On the other hand, the probability of a light comorbidity between diseases, as well as a severe comorbidity, is highest at 0.2849. The probability that both diseases are not affected by hidden risk factors is 0.2448.

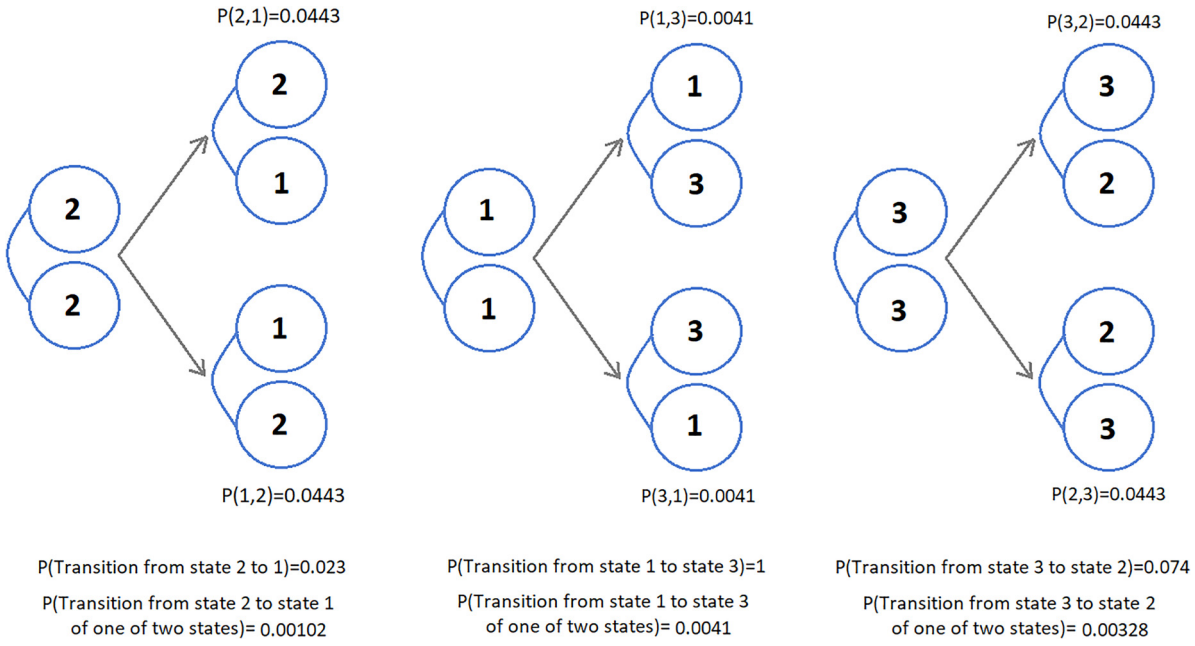

Figure 9. Example structure of joint hidden states with Binomial copula and transition probabilities.

Figure 9 displays the example schemes for transitioning between joint hidden states. By multiplying the joint state and transition probabilities, we calculate the probability of moving from "no risk factor" to "one of risk factors affect one disease" level, that is, $(2, 2) \rightarrow (2, 1)$ or $(2, 2) \rightarrow (1, 2)$. Thus,

$$P((2, 2) \rightarrow (2, 1)) = 0.023 * 0.0443 = 0.00102.$$

The probability of both diseases remaining in "severe comorbidity," that is, (3,3), is 0.2638. The probability of both diseases remaining in a "no risk factor," that is, (2,2), is 0.2379. The probability of both diseases remaining in a "light comorbidity," that is, (1,1), is 0.

The proposed model enables investigation of the joint behavior of hidden chains and hidden states, as well as their transition dynamics. Thus, we gain a better understanding of the dependency structure of unobserved knowledge regarding diseases with limited patient data based on hospital visits across time.

8 Concluding comments

The main focus of the proposition in this study is to use a combination of hidden Markov theory and copula function to model parallel interacting processes describing two or more chronic diseases.

We develop a novel coupled HMM in this study that incorporates a bivariate discrete copula function into the hidden process. We employ a novel type of discrete copula, the Binomial copula. We compute a CDLL and develop the necessary inference to implement the model. Due to the large number of parameters required to estimate parameters using the EM algorithm, even for two hidden chains, estimation becomes computationally intractable. As a result, we estimate the model's parameter using a VEM algorithm. Because the variational expectation part of the algorithm requires computing the lower bound approximation of CDLL, we use Kullback-Leibler divergence to derive the lower bound for the log-likelihood function based on the model. The VEM algorithm is carried out by computing conditional expectations based on forward-backward probabilities and estimators of parameters that maximize the CDLL.

Due to the unidentifiability of copula functions defined on discrete space, Geenens[18] developed a new bivariate discrete copula. Despite extensive theoretical development, statistical inference for parameter estimation of the joint copula function has not been developed yet. Thus, we designed a structure for joint states in such a way that they follow a predefined Binomial copula probability mass function with a given odds ratio while also satisfying the Markov property of hidden states in each chain. To verify and confirm the correctness of the designed structure, we calculated bias and MSE metrics and performed a Markovian test on 500 simulated data. The simulation study was conducted for two different odds ratio capturing weak and high association structure, and the developed model was applied. The findings of the simulation study are satisfactory.

The proposed model is applied to real data from a private hospital from January 2015 to December 2020, including information on hospital appointments. The observed variable is the total number of patients diagnosed with a particular disease in a given month. The purpose of this application is to define the dependency structure of unobserved clinical disease data. The application results demonstrate that the CHMM with discrete copula is useful for investigating disease comorbidity when only population dynamics over time are available and no clinical data are available.

During the application, one of the difficulties encountered is that the proposed model, like all HMM-based models, is sensitive to initial parameter settings. Because the developed model has a large number of parameters to estimate, the optimal fit is defined using combined information from weighted RSS. Additionally, the algorithm have been run multiple times with varying inner iteration counts in order to overcome a local optimum.

The model is applicable to the study of more than two diseases; more than two hidden chains can be included in the model. Additionally, the proposed model is general enough to account for any dependent events. The model can be extended to include covariates in both observed and hidden spaces. The proposed approach models interacting time-series observations; however, by combining CDLL and GLM functions, it is possible to extend this model to apply it to the longitudinal data.

Acknowledgements

This article is based on Ph.D. Thesis study by Oflaz Z.[68]

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Zarina Oflaz https://orcid.org/0000-0003-3234-3879

Supplemental material

Supplemental material for this article is available online.

References

1. Koczwara B. *Cancer and Chronic Conditions*. Singapore: Springer, 2016.
2. Feinstein AR. The pre-therapeutic classification of co-morbidity in chronic disease. *J Chron Dis* 1970; **23**: 455–468.
3. Satariano WA. Comorbidities and cancer. In: Hunter CP, Johnson KA and Muss HB (eds) *Cancer in the Elderly*. 1st ed. CRC Press, 2000, pp. 486–508.
4. Guisado-Clavero M, Roso-Llorach A, López-Jimenez T, et al. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. *BMC Geriatr* 2018; **18**: 1–11.
5. Huang CF, Liu JC, Huang HC, et al. Longitudinal transition trajectory of gouty arthritis and its comorbidities: a population-based study. *Rheumatol Int* 2017; **37**: 313–322.
6. Violán C, Fernández-Bertolín S, Guisado-Clavero M, et al. Five-year trajectories of multimorbidity patterns in an elderly Mediterranean population using hidden Markov models. *Sci Rep-UK* 2020; **10**: 1–11.
7. Bringmann LF, Vissers N, Wichers M, et al. A network approach to psychopathology: new insights into clinical longitudinal data. *PLoS ONE* 2013; **8**: e60188.
8. Groen RN, Ryan O, Wigman JT, et al. Comorbidity between depression and anxiety: assessing the role of bridge mental states in dynamic psychological networks. *BMC Med* 2020; **18**: 1–17.
9. Qian Z, Alaa A, Bellot A, et al. Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. In: *23 rd International Conference on Artificial Intelligence and Statistics*, Online, 26-28 August 2020, pp. 3295–3305.
10. Faruqui SHA, Alaeddini A, Jaramillo CA, et al. Mining patterns of comorbidity evolution in patients with multiple chronic conditions using unsupervised multi-level temporal Bayesian network. *PLoS ONE* 2018; **13**: e0199768.
11. Lappenschaar M, Hommersom A, Lucas PJ, et al. Multilevel Bayesian networks for the analysis of hierarchical health care data. *Artif Intell Med* 2013; **57**: 171–183.
12. Lappenschaar M, Hommersom A, Lucas PJ, et al. Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity. *J Clin Epidemiol* 2013; **66**: 1405–1416.
13. Maag B, Feuerriegel S, Kraus M, et al. Modeling longitudinal dynamics of comorbidities. In: *Proceedings of the Conference on Health, Inference, and Learning* (ed Ghassemi M), Virtual USA, 8–10 April 2021, pp. 222–235. New York: Association for Computing Machinery.

14. Leiva-Murillo J, Rodrguez A and Baca-Garca E. Visualization and prediction of disease interactions with continuous-time hidden Markov models. In: *NIPS 2011 Workshop on Personalized Medicine* (ed Shawe-Taylor J and Zemel R), Granada, Spain, 16-17 December 2011.
15. Huang Z, Dong W, Wang F, et al. Medical inpatient journey modeling and clustering: a Bayesian hidden Markov model based approach. In: *AMIA Annual Symposium Proceedings*, San Francisco, USA, 14-18 November 2015, pp. 649–658, American Medical Informatics Association.
16. Powell G, Verma A, Luo Y, et al. Modeling chronic obstructive pulmonary disease progression using continuous-time hidden Markov models. *St Heal T* 2019; **264**: 920–924.
17. Lee C, Yoon J and Van Der Schaar M. Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE T Bio-Med Eng* 2019; **67**: 122–133.
18. Geenens G. Copula modeling for discrete random vectors. *Dependence Modeling* 2020; **8**: 417–440.
19. Yule GU. On the methods of measuring association between two attributes. *J R Stat Soc* 1912; **75**: 579–652.
20. Goodman LA and Kruskal WH. Measures of association for cross classifications. In: Goodman LA and Kruskal WH (eds) *Measures of association for cross classifications*. Springer, 1979, pp. 2–34.
21. Mosteller F. Association and estimation in contingency tables. *J Am Stat Assoc* 1968; **63**: 1–28.
22. Wang Y and Pham H. Modeling the dependent competing risks with multiple degradation processes and random shock using time-varying copulas. *IEEE T Reliab* 2011; **61**: 13–22.
23. Kaishev VK, Dimitrova DS and Haberman S. Modelling the joint distribution of competing risks survival times using copula functions. *Insur Math and Econ* 2007; **41**: 339–361.
24. Chen YH. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J Roy Stat Soc B Met* 2010; **72**: 235–251.
25. Escarela G and Carriere JF. Fitting competing risks with an assumed copula. *Stat Methods Med Res* 2003; **12**: 333–349.
26. Lo SM and Wilke RA. A copula model for dependent competing risks. *J Roy Stat Soc C-Appl* 2010; **59**: 359–376.
27. Stöber J, Hong HG, Czado C, et al. Comorbidity of chronic diseases in the elderly: patterns identified by a copula design for mixed responses. *Comput Stat Data An* 2015; **88**: 28–39.
28. Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Met* 1977; **39**: 1–22.
29. Saul LK, Jaakkola T and Jordan MI. Mean field theory for sigmoid belief networks. *J Artif Intell Res* 1996; **4**: 61–76.
30. Jaakkola T. Tutorial on variational approximation methods. In *Advanced mean field methods: theory and practice*, 2000, pp. 129–159. MIT Press.
31. Blei DM, Kucukelbir A and McAuliffe JD. Variational inference: a review for statisticians. *J Am Stat Assoc* 2017; **112**: 859–877.
32. Wainwright MJ and Jordan MI. Graphical models, exponential families, and variational inference. *Found Trends Mach Learn* 2008; **1**: 1–305.
33. Ormerod JT and Wand MP. Explaining variational approximations. *Am Stat* 2010; **64**: 140–153.
34. Oflaz Z, Yozgatligil C and Selcuk-Kestel AS. Estimation of disease progression for ischemic heart disease using latent Markov with covariates. *Stat Anal Data Min: ASA Data Sci J* 2022; **16**: 16–28.
35. Valderas JM, Starfield B, Sibbald B, et al. Defining comorbidity: implications for understanding health and health services. *Ann Fam Med* 2009; **7**: 357–363.
36. Siu TK, Ching WK, Fung SE, et al. On a multivariate Markov chain model for credit risk measurement. *Quant Financ* 2005; **5**: 543–556.
37. Pasanisi A, Fu S and Bousquet N. Estimating discrete Markov models from various incomplete data schemes. *Comput Stat Data An* 2012; **56**: 2609–2625.
38. Oflaz ZN, Yozgatligil C and Selcuk-Kestel AS. Aggregate claim estimation using bivariate hidden Markov model. *ASTIN Bull* 2019; **49**: 189–215.
39. Fine S, Singer Y and Tishby N. The hierarchical hidden Markov model: analysis and applications. *Mach Learn* 1998; **32**: 41–62.
40. Ghahramani Z and Jordan MI. Factorial hidden Markov models. In: *Advances in Neural Information Processing Systems 8* (ed. Touretzky D, Mozer MC and Hasselmo M), 1995, pp. 472–478.
41. Kristjansson TT, Frey BJ and Huang TS. Event-coupled hidden Markov models. In: *2000 IEEE International Conference on Multimedia and Expo*, New York, USA, 30 July–2 August 2000, volume 1, pp. 385-388, IEEE.
42. Brand M, Oliver N and Pentland A. Coupled hidden Markov models for complex action recognition. In: *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, San Juan, USA, 17–19 June 1997, pp. 994–999, IEEE.
43. Kwon J and Murphy K. *Modeling freeway traffic with coupled HMMs*. Report. USA: University of California, 2000.
44. Zhong S and Ghosh J. *A new formulation of coupled hidden Markov models*. Report. USA: Dept Elect Comput Eng, University of Austin, 2001.
45. Lapuyade-Lahorgue J, Xue JH and Ruan S. Segmenting multi-source images using hidden Markov fields with copula-based multivariate statistical distributions. *IEEE T Image Process* 2017; **26**: 3187–3195.
46. Hu X. *A Copula-based Quantile Risk Measure Approach to Hedging under Regime Switching*. Master Thesis, University of Waterloo, Canada, 2015.
47. Derrode S and Pieczynski W. Unsupervised classification using hidden Markov chain with unknown noise copulas and margins. *Signal Process* 2016; **128**: 8–17.

48. Lagona F. Copula-based segmentation of cylindrical time series. *Stat Probabil Lett* 2019; **144**: 16–22.
49. Sun F and Jiang Y. A hidden resource in wireless channel capacity: Dependence control in action. arXiv preprint arXiv:180500812 2018.
50. Sklar M. Fonctions de repartition an dimensions et leurs marges. *Publ inst statist univ Paris* 1959; **8**: 229–231.
51. Nelsen RB. An introduction to copulas, volume 139 of. Lecture Notes in Statistics 1999.
52. Wang X, Lebarbier E, Aubert J, et al. Variational inference for coupled hidden Markov models applied to the joint detection of copy number variations. *Int J Biostat* 2019; **15**: 20180023.
53. Saul L and Jordan M. Exploiting tractable substructures in intractable networks. *Adv Neur In* 1995; **8**: 486–492.
54. Ghahramani Z and Jordan MI. Factorial hidden Markov models. *Mach Learn* 1997; **29**: 245–273.
55. Nielsen F and Sun K. Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures using piecewise log-sum-exp inequalities. *arXiv preprint arXiv:1606.05850* 2016.
56. Anderson TW and Goodman LA. Statistical inference about Markov chains. *Ann Math Stat* 1957; **28**: 89–110.
57. Spedicato GA. Discrete Time Markov Chains with R. *R J* 2017.
58. Zucchini W, MacDonald IL and Langrock R. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2017.
59. Kubanek M, Bobulski J and Adrjanowicz L. Characteristics of the use of coupled hidden Markov models for audio-visual polish speech recognition. *B Pol Acad Sci-Tech* 2012; **60**: 307–316.
60. Ghosh S, Li J, Cao L, et al. Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. *J Biomed Inform* 2017; **66**: 19–31.
61. Darmanjian S, Kim SP, Nechyba MC, et al. Independently coupled HMM switching classifier for a bimodel brain–machine interface. In: *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, Maynooth, Ireland, 6–8 September 2006, pp. 379–384, Piscataway, NJ IEEE Service Center.
62. Abdelaziz AH, Charaf LA, Zeiler S, et al. On dynamic stream weight learning for coupled-hmm-based audio-visual speech recognition. In: *40th Annual German Congress on Acoustics*, Olderburg, Germany, 2014, pp. 531–532.
63. Wu J, Xun P, Tang Q, et al. Circulating magnesium levels and incidence of coronary heart diseases, hypertension, and type 2 diabetes mellitus: a meta-analysis of prospective cohort studies. *Nutr J* 2017; **16**: 1–13.
64. Schellevis FG, van der Velden J, van de Lisdonk E, et al. Comorbidity of chronic diseases in general practice. *J Clin Epidemiol* 1993; **46**: 469–473.
65. Beghetti M and Tissot C. Pulmonary arterial hypertension in congenital heart diseases. *Semin Respir Crit Care Med* 2009; **30**: 421–428.
66. Vachiéry JL, Adir Y, Barberà JA, et al. Pulmonary hypertension due to left heart diseases. *J Am Coll Cardiol* 2013; **62**: D100–D108.
67. Jin P, Gu W, Lai Y, et al. The circulating microRNA-206 level predicts the severity of pulmonary hypertension in patients with left heart diseases. *Cell Physiol Biochem* 2017; **41**: 2150–2160.
68. Oflaz Z. *Coupled Hidden Markov Model with Bivariate Discrete Copula to Study Comorbidity of Chronic Diseases*. PhD Thesis (Unpublished), Middle East Technical University, Turkey, 2022.