IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
https://doi.org/10.1080/24725579.2020.1866715

Modeling the length-of-stay of patients with geriatric diseases or alcohol use disorder using phase-type distributions with covariates

Wanlu Gu[a], Neng Fan[a], and Haitao Liao[b]

[a]Department of Systems & Industrial Engineering, University of Arizona, Tucson, AZ, USA; [b]Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA

ABSTRACT

The hospital length-of-stay (LOS), as an important measure of the effectiveness of healthcare, represents the level of medical requirement and is highly related to the treatment costs. As the human life expectancy has being increased rapidly in the past few decades, there is a pressing need to improve health systems for geriatric patients. Similarly, the alcohol use disorder (AUD), as a chronic relapsing brain disease related to severe problem drinking, has caused negative impacts to society and put patients' health and safety at risk. In both cases, more efficient hospital management is in demand due to increasing requirements for long-term hospital treatment and the continuously rising medical cost. In order to improve the healthcare efficiency, an accurate modeling of the LOS data and the further analysis of potential influencing factors are necessary. In this paper, we utilize the Coxian Phase-Type (PH) distribution and apply Maximum Likelihood Estimation (MLE) to fit the patient flow information of both geriatric patients and AUD patients collected in a hospital. The influences of the covariates of age, gender, admission type, admit source, and financial class on LOS are assessed and compared through Expectation-Maximization (EM) algorithms. The results show that the LOS data of both types of patients can be modeled well, and the differences with respect to covariates can be accurately identified by the proposed methods. Using the fitted Coxian PH distribution and the estimated coefficients of covariates will provide a guide for better decision-making in healthcare service and resource allocation.

KEYWORDS

Phase-type distribution; healthcare quality; length-of-stay; covariate; EM algorithms

1. Introduction

Nowadays, population ageing is much faster than ever before. According to World Health Organization Fact Sheets (WHO, 2015), the proportion of the world's population over 60 years old will be nearly doubled from 12% to 22% between 2015 and 2050. Facing fast increasing proportions of elder populations, all countries need to improve the planning and management of their health and service systems to handle this demographic shift.

According to the 2019 National Survey on Drug Use and Health, 14.1 million adults at age of 18 and older (5.6% of this age group) had alcohol use disorder (AUD) in U.S. (Alcohol Facts and Statistics, 2020). Following the ICD-10 (2016) code, the alcohol related disorders include alcohol abuse, alcohol dependence, alcohol use, and unspecified. Among the alcohol related disorders, alcohol abuse is a heterogeneous set of behaviors that include any pattern of ethyl alcohol intake that causes medical and social complications (Cloninger et al., 1981). It also influences children in both genetic and environmental aspects.

Both geriatric diseases and AUD are common seen diagnosis in heathcare, and always cause negative effects on patients' life. The symptoms of geriatric diseases and AUD are difficult to be completed treated, which leads to repeated admissions or even death. Therefore, an efficient treatment and a thorough understanding of the factors related to them are necessary.

The length-of-stay (LOS), the time difference between the admission time and discharge time, is a significant measure of healthcare efficiency. Analyzing LOS helps improving hospital efficiency for geriatric diseases (Turgeman et al., 2015). The importance of LOS as an indicator of severity and a measure of treatment in investigating alcohol related diseases are studied in both Finney et al. (1981) and Long et al. (1998).

A lot of previous research has been performed on analyzing the LOS information of geriatric or AUD patients. El-Darzi et al. (1998) and Faddy and McClean (2007) apply a multi-state model to classify geriatric patients to different LOS groups, while Toh et al. (2017) and Kwok et al. (2017) study the factors that influence the LOS of geriatric patients. However, the research on the LOS of patients with AUD has not been always consistent. In Gottheil et al. (1992), the relationships between LOS and outcome for patients grouped by severity are examined. In Saitz et al. (1997), it is observed that having an alcohol-related diagnosis is associated with more use of intensive care, longer inpatient stays, and higher hospital charges. J. H. Park et al. (2018) also find that the alcohol use is associated with increased emergency

CONTACT Neng Fan nfan@arizona.edu Department of Systems & Industrial Engineering, University of Arizona, Tucson, AZ, USA

department LOS, and a multivariate quantile regression model is applied to include information, such as age, gender, consciousness status, severity of injury, emergency medical service use, etc, for analysis.

In the literature, lots of studies on modeling and evaluating LOS of patients with various diseases have been conducted (Gu et al., 2019; Xie et al., 2005; Zhang et al., 2013). Among these methods of modeling LOS data, the phase-type (PH) distributions have been applied in the healthcare field increasingly over time to interpret healthcare systems and to improve the healthcare efficiency. A PH distribution describes the absorption time of an evanescent finite-state Continuous Time Markov Chain (CTMC) (Fackrell, 2003). The Coxian PH distributions, as a special type of PH distributions, are often used in modeling and investigating the influences of covariates on LOS data.

As stated in Faddy et al. (2009), understanding how age, gender, comorbid conditions, and iatrogenic events influence LOS will aid program evaluation and handle difficult tasks in managing hospital systems. These factors are called *covariates*, which are variables that are possibly predictive of the outcome under study. As a result, the covariates are sometimes referred as predictor variables.

Previous study of LOS for both geriatric patients and AUD mostly focuses on studying the effects of LOS on the treatment outcomes or a simple comparison between those alcohol users and non-alcohol users. It lacks of investigating simultaneously the pattern recognition of LOS and identifying important factors that significantly influence the LOS itself.

For simple PH distributions without the covariates, the EMpht-programme is developed based on the Expectation-Maximization (EM) algorithms proposed in Asmussen et al. (1996) and has been widely used in various fields. For models with covariates, Tang et al. (2012), McGrory et al. (2009) and Faddy et al. (2009) apply either the maximum likelihood estimation (MLE) or the Bayesian method to estimate Coxian PH models and incorporate covariates to explain the differences in distinct LOS groups. Gardiner (2012), Gardiner et al. (2014), and Zhu et al. (2018) apply an order restriction on Coxian PH models and reform it into a finite mixture of parametric distributions that can be easily interpreted.

Since there is no application of EM algorithms in PH distributions with covariates before, in this paper, we extend the original methods in Asmussen et al. (1996) to allow modeling the LOS with consideration of the influences of covariates. Further analysis on the estimated results of our modeling the patient flow information is also presented. The characteristics of patients on admission will be considered as the covariates in a LOS study from a more general viewpoint. It is aimed to verify the efficiency of our extended EM algorithms in fitting the Coxian PH distributions and capturing the impacts of covariates.

To this end, the flow information of patients with either geriatric diseases or AUD, and the demographic information collected in Banner University Medical Center Tucson - Main Campus and South Campus from 2012 to 2017 are preprocessed and analyzed. Because of limitations and availability of the patient information, the most common and basic factors that are available for all patients in this data source are identified, and thus the covariates studied in this research include age, gender, admission type, admit source, and financial class for the payment of medical care. Based on the collected data, the effects of these covariates on LOS will be studied through the PH distribution, and their coefficients in expressing LOS will be estimated through the extended EM algorithms. The top predictors of LOS and a series of useful comparisons among covariates are also presented. The fitting results provide a reference for patient cluster for both geriatric patients and AUD patients. It also helps to understand how the LOS is related to a set of given covariates and to better classify and identify the hidden pattern inside the LOS of patients.

The remainder of this paper is organized as follows. In Section 2, the properties of Coxian PH distributions are introduced, and the algorithms to fit Coxian PH distributions with covariates are proposed. In Section 3, we present how to collect and manipulate data of patients. In Section 4, the fitting results and further analysis are presented. Finally, Section 5 concludes the paper.

2. Model fitting

2.1. Coxian PH distributions

A continuous PH distribution is the distribution of the time from the initial state until absorption in the absorbing state in a CTMC (Neuts, 1981). Consider a CTMC $\{J_u\}_{u\geq 0}$ on a finite discrete state space $S = \{0, 1, 2, ..., m\}$, where state 0 is the absorbing state and states $1, ..., m$ are transient states. In PH distributions, the sojourn w_i in transient state i follows an exponential distribution with rate λ_i, i.e. $f(w_i|\lambda_i) = \lambda_i \exp(-\lambda_i w_i)$, $w_i \geq 0$. It is understandable to treat the time between transitions as the time spent in each previous state and the parameter λ as transition density or the transition rate (Cox, 1955).

The infinitesimal generator (transition rate matrix) for the CTMC mentioned above can be presented in the form of block-matrix $\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{q} & \mathbf{T} \end{pmatrix}$. Here, $\mathbf{0}$ is a $1 \times m$ all zero vector, demonstrating that the transition rates $\lambda_{0i}, i = 1, ..., m$ from the absorbing state to transient states are all zeros. The matrix $\mathbf{T}$ consists of the transition rates between transient states, where the transition rates $\lambda_{ij} \geq 0, i, j = 1, ..., m$. The $m \times 1$ vector $\mathbf{q}$ is composed of transition rates $\lambda_{i0} \geq 0, i = 1, ..., m$ from transient states to the absorbing state. Let the random variable Y be the time from the initial state until absorption to the absorbing state. Then, Y is said to have a (continuous) PH distribution (Neuts, 1981), and a phase corresponds to a specific state in the CTMC. Then the distribution and density function of variable Y can be expressed in terms of initial state distribution $\boldsymbol{\pi}$ and matrix $\mathbf{T}$. The pair $(\boldsymbol{\pi}, \mathbf{T})$ is also known as a representation of the PH distribution.

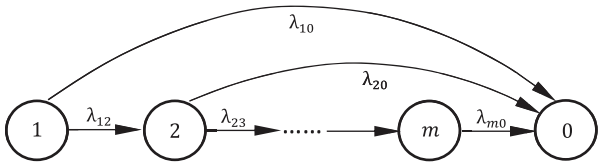


Figure 1. State transition diagram of a Coxian PH distribution.

One popular type of PH distributions is the Coxian PH distribution, shown in Figure 1, which ensures that the transient states of the model are ordered (Fackrell, 2009). In the Coxian PH distribution, the stochastic process begins from the first transient state and may either move sequentially or enter the absorbing state 0 directly. For the last transient state, it has only one direction which leads to the absorbing state. The time spent in each transient state i is exponentially distributed with parameter λ_i, which is also interpreted as the average rate moving out of state i. The transition rate from state i to state $i+1$ is $\lambda_{i,i+1}$, and the rate to absorbing state 0 is λ_{i0}. According to the special structure of Coxian PH distribution, we have the following relationships as $\lambda_i = \lambda_{i,i+1} + \lambda_{i0}$ for $i = 1, ..., m-1$ and $\lambda_m = \lambda_{m0}$. The initial state distribution for Coxian PH distribution is $\boldsymbol{\pi} = [1, 0, ..., 0]_{1\times m}$ and the probability density function (PDF), and cumulative distribution function (CDF) of absorbing time y can be expressed as

$$f(y) = -\mathbf{T}\boldsymbol{\pi}e^{\mathbf{T}y}\mathbf{e} = \mathbf{q}\boldsymbol{\pi}\exp(\mathbf{T}y),$$
$$F(y) = P(Y \le y) = 1 - \boldsymbol{\pi}\exp(\mathbf{T}y)\mathbf{e},$$

respectively, where $\mathbf{e} = (1, 1, ..., 1)^T_{m\times 1}$, and

$$\mathbf{T} = \begin{bmatrix} -\lambda_1 & \lambda_{12} & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_{23} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\lambda_{m-1} & \lambda_{m-1,m} \\ 0 & 0 & \cdots & 0 & -\lambda_m \end{bmatrix}_{m\times m}.$$

Recently, the Coxian PH distributions have been successfully applied to modeling patient LOS, corresponding to absorbing time y, in a hospital. The m states correspond to m phases in patients care processes in the hospital, which may be used to describe those steps or stages according to the LOS.

2.2. Incorporating covariates

Next, we will introduce how to incorporate the covariates information into a Coxian PH model. To study the LOS y_n $(n = 1, ..., N)$ for the nth patient, consider covariate information of all N observed patients as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, where the nth $(n = 1, ..., N)$ observation vector is $\mathbf{x}_n = (x_{n1}, x_{n2}, ..., x_{nc})^T$, and each x_{nj} corresponds to one categorical information of patient n (here c is the number of covariates to be considered). The coefficient vector $\mathbf{b} = (b_1, b_2, ..., b_c)$ contains the coefficients for all the covariates. Here vectors x_{nj} and b_j $(j = 1, 2, ..., c)$ have the same dimension, depending the number of categories in the jth covariates.

The early research on regression analysis of covariates can be traced back to Cox (1972), in which the Cox proportional hazards regression model is proposed. Li (1999) also recommends the Cox proportional hazards model to estimate the adjusted expected LOS using several factors that influence the outcome. The hazard rate usually refers to the rate of death or failure for an item at a certain time, which can be treated as the transition rate.

In order to assess the relationship between the probability distribution of absorbing time and covariates, the transition rate function is considered to be a function of the covariates. Specifically, we assume the hazard rate function as $\lambda(\mathbf{x}) = \lambda^0\exp(-\mathbf{b}\mathbf{x})$ in the Coxian PH distribution. The λ^0 is the baseline rate which should be the value when the covariates are not considered, and x is the vector for the information of covariates. Therefore, the adjusted transition rate matrix becomes $\tilde{\mathbf{T}} = \exp(-\mathbf{b}\mathbf{x})\mathbf{T}$, where $\mathbf{T}$ is the transition rate matrix if covariates $\mathbf{x} = 0$ or the so called baseline transition rate matrix.

In Section 2.1, it is already shown that the distribution of variable y depends on the pair $(\boldsymbol{\pi}, \mathbf{T})$. Then the parameters need to be estimated in model fitting without the covariates are all from $\mathbf{T}$, and there will be a total of $2m-1$ parameters. Additionally, with the coefficient vector $\mathbf{b}$, there will be $2m-1+|\mathbf{b}|$ parameters to be estimated in models that consider the covariates. Let $\mathbf{Y} = (y_1, y_2, ..., y_N)$ be an independent and identically distributed sample from a population with PDF $f(y|\Theta)$, where Θ is a vector consisting of $k = 2m-1+|\mathbf{b}|$ parameters to be estimated in $\mathbf{T}$ and $\mathbf{b}$.

For the models with covariates, the likelihood function becomes

$$L(\Theta|\mathbf{Y}) = \prod_{n=1}^{N} f(y_n|\Theta) = \prod_{n=1}^{N} \boldsymbol{\pi}\exp(\tilde{\mathbf{T}}_n y_n)\tilde{\mathbf{q}}_n.$$

Let $\lambda_i^n = \lambda_i(\mathbf{x}_n) = \lambda_i\exp(-\mathbf{b}\mathbf{x}_n)$ be the adjusted transition rate of the ith state with the influence of covariates on the nth observation. Then the adjusted transition matrix and adjusted absorbing matrix are expressed as

$$\tilde{\mathbf{T}}_n = \exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T} = \exp(-\mathbf{b}\mathbf{x}_n)\begin{bmatrix} -\lambda_1 & \lambda_{12} & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_{23} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\lambda_{m-1} & \lambda_{m-1,m} \\ 0 & 0 & \cdots & 0 & -\lambda_m \end{bmatrix}_{m\times m},$$

$$\tilde{\mathbf{q}}_n = \exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q} = \exp(-\mathbf{b}\mathbf{x}_n)\begin{bmatrix} \lambda_{10} \\ \lambda_{20} \\ \vdots \\ \lambda_{m-1,0} \\ \lambda_{m,0} \end{bmatrix}_{m\times 1}.$$

2.3. Constructing complete sample

With the method used in Asmussen et al. (1996) and the properties of the Coxian PH distributions, we consider the incomplete observation of N independent replications of

Markov process $J_u^1, ..., J_u^N$ as the time to absorption state, $\mathbf{Y} = (y_1, ..., y_N)$. Assuming there are p jumps until arriving absorbing state 0 within the embedded Markov chain $I_0, I_1, ..., I_{p-1}(I_p = 0)$, where the I_p represents the last state the Markov process visits and it equals 0 (state 0). The corresponding sojourn time in each state the Markov process visits are $S_0, S_1, ..., S_{p-1}(S_p = \infty)$, where S_p is the time it spends in state 0 and according to the property of absorbing state, $S_p = \infty$. Both the Markov chain and sojourn time list are unknown, and hence the complete sample set of N Markov processes can be represented by $\mathbf{C}$ which contains the information of N observations,

$$\mathbf{C}=(c_1,...,c_N)=(i_0^1,...,i_{p^1-1}^1,s_0^1,...,s_{p^1-1}^1,...,i_0^N,...,i_{p^N-1}^N,s_0^N,...,s_{p^N-1}^N).$$

Then the real observations becomes

$$\mathbf{Y} = (y_1, ..., y_N) = (s_0^1 + \cdots + s_{p^1-1}^1, ..., s_0^N + \cdots + s_{p^N-1}^N),$$

where p^n, $n = 1, ..., N$ is the total number of jumps the nth observation has.

Since the sojourn in transient state i follows an exponential distribution with rate λ_i, the PDF of one complete observation c_n, $n = 1, ..., N$ is

$$\begin{aligned} f(c_n|\boldsymbol{\pi}, \tilde{\mathbf{T}}_n) = f(c_n|\boldsymbol{\pi}, \mathbf{T}, \mathbf{b}) = \pi_{i_0^n} \exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i_0^n} \exp\{-\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i_0^n} s_0^n\} p_{i_0^n i_1^n} \\ \times \exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i_1^n} \exp\{-\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i_1^n} s_1^n\} p_{i_1^n i_2^n} \\ \cdots \times \exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i_{p^n-1}^n} \exp\{-\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i_{p^n-1}^n} s_{p^n-1}^n\} p_{i_{p^n-1,0}^n}, \end{aligned}$$

where i_q^n, $q = 0, ..., p^n$ is the state the nth observation arrives after q jumps. Besides, we have the relationships that $p_{ij} = Pr(I_{n+1} = j|I_n = i) = \frac{\lambda_{ij}}{\lambda_i}$, $p_{i0} = Pr(I_{n+1} = 0|I_n = i) = \frac{\lambda_{i0}}{\lambda_i}$, $i, j = 1, ..., m$. Then the likelihood function can be expressed as follows

$$\begin{aligned} L(\boldsymbol{\Theta}|\mathbf{Y}) &= \prod_{n=1}^{N} f(c_n|\boldsymbol{\pi}, \mathbf{T}, \mathbf{b}) = f(\mathbf{C}|\boldsymbol{\pi}, \mathbf{T}, \mathbf{b}) \\ &= \prod_{n=1}^{N} \left\{ \prod_{i=1}^{m} \pi_i^{B_{in}} \prod_{i=1}^{m} \exp\{-\exp(-\mathbf{b}\mathbf{x}_n)\lambda_i Z_{in}\} \right\} \prod_{i=1}^{m} \prod_{j=0, j\neq i}^{m} \\ &\quad \times (\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{ij})^{N_{ijn}} \}, \end{aligned}$$

and by taking log of the likelihood function, we have

$$\begin{aligned} \log L(\boldsymbol{\Theta}|\mathbf{Y}) = \sum_{n=1}^{N} \Bigg\{ \sum_{i=1}^{m} B_{in} \log(\pi_i) + \sum_{i=1}^{m} (-\exp(-\mathbf{b}\mathbf{x}_n)\lambda_i Z_{in})) \\ + \sum_{i=1}^{m} \sum_{j=0, j\neq i}^{m} N_{ijn} \log(\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{ij}) \Bigg\}, \end{aligned} \tag{1}$$

where a set of variables are defined as follows:

- B_{in} is the frequency of the nth observation starting in state i, $i = 1, \ldots, m$;
- N_{i0n} is the probability of processes exiting from state i to the absorbing state for the nth observation;
- N_{ijn} is the probability of jumping from state i to state j for the nth observation;
- Z_{in} is the total time spent in state i prior to absorption for the nth observation.

To estimate $2m - 1 + |\mathbf{b}|$ parameters, it corresponds to estimate the values of $\mathbf{b}, N_{i0n}, N_{ijn}, Z_{in}$.

2.4. Fitting method

The EM algorithm is a broadly applicable approach to the iterative computation of MLE, and is useful in a variety of incomplete-data problems. Given observed data Y, missing values $\hat{Y}$, and thus the complete data can be denoted by $\tilde{Y} = (Y, \hat{Y})$. The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:

E-Step. Evaluate the conditional expectation of the log likelihood function of a parameter θ with respect to the current conditional distribution of complete data $\tilde{Y}$ given the information of Y and the current estimate of the parameter $\theta^{(s)}$, where s is the number of iteration

$$Q(\theta|\theta^{(s)}) = E_{\hat{Y}|Y,\theta^{(s)}}\left[\log L(\theta|\tilde{Y})\right];$$

M-Step. Find the parameter $\theta^{(s+1)} = \arg\max_\theta Q(\theta|\theta^{(s)})$ such that

$$Q(\theta^{(s+1)}|\theta^{(s)}) \geq Q(\theta|\theta^{(s)})$$

until the difference between the likelihood in two iterations $L(\theta^{(s+1)}) - L(\theta^{(s)})$ is small enough. Otherwise, let $s = s + 1$ and go to E-step.

In order to obtain the log likelihood function $\log L(\boldsymbol{\Theta}|\mathbf{y})$ in Equation (1), we calculate the expectation of B_{in}, N_{i0n}, N_{ijn}, and Z_{in}, which are related to the state i given the absorbing time y_n and covariates value $\mathbf{x}_n$. The formula of the expectations without the information of covariates are clearly presented in Asmussen et al. (1996), and here we derive the conditional expectations given the influences of covariates on each observation as follows

$$E\left[B_{in}|y_n,\mathbf{x}_n\right]=P(J_0=i|y_n,\mathbf{x}_n)=\frac{P(J_0=i|\mathbf{x}_n)P(y_n|J_0=i,\mathbf{x}_n)}{P(y_n|\mathbf{x}_n)}$$
$$=\frac{\pi_i\mathbf{e}_i^T\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}}{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}}$$
$$=\frac{\pi_i b_i(y_n|\boldsymbol{\pi},\mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_n|\mathbf{T})},\ i=1,...,m,$$
$$E\left[Z_{in}|y_n,\mathbf{x}_n\right]=E\left[\int_0^\infty 1_{\{J_u=i\}}du|y_n,\mathbf{x}_n\right]$$
$$=\int_0^\infty P(J_u=i|y_n,\mathbf{x}_n)du=\int_0^{y_n}P(J_u=i|y_n,\mathbf{x}_n)du$$
$$=\int_0^{y_n}\frac{P(J_u=i|\mathbf{x}_n)P(y_n|J_u=i,,\mathbf{x}_n)}{P(y_n|\mathbf{x}_n)}du$$
$$=\frac{1}{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}}$$
$$\times\int_0^{y_n}\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}u\}\mathbf{e}_i\mathbf{e}_i^T$$
$$\times\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}(y_n-u)\}\times\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}du$$
$$=\frac{c_{in}(y_n,i|\boldsymbol{\pi},\mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_n|\mathbf{T})},\ i=1,...,m.$$

Here N_{ijn} denotes the probability of jumping from state i to state j given the covariates information of the nth observation. A set of discrete approximations of N_{ijn} as $N_{ijn}^{\epsilon}=1_{J_u=i,J_{(u+\epsilon)}=j},\ \epsilon>0,\ i\neq j$ are dominated by the $\sum_{i\neq j}^m N_{ijn}$ and converge to N_{ijn} as $\epsilon\downarrow 0$. Suppose the system is in state i at time $k\epsilon$, then

$$E\left[N_{ijn}^{\epsilon}|y_n,\mathbf{x}_n\right]=\sum_{k=0}^{[y/\epsilon]-1}P(J_{k\epsilon}=i,J_{(k+1)\epsilon}=j|y_n,\mathbf{x}_n)$$
$$=\sum_{k=0}^{[y/\epsilon]-1}\frac{P(y_n|J_{k\epsilon}=i,J_{(k+1)\epsilon}=j,\mathbf{x}_n)P(J_{k\epsilon}=i,J_{(k+1)\epsilon}=j|\mathbf{x}_n)}{P(y_n|\mathbf{x}_n)}$$
$$=\sum_{k=0}^{[y/\epsilon]-1}\frac{P(J_{k\epsilon}=i|\mathbf{x}_n)P(J_{(k+1)\epsilon}=j|J_{k\epsilon}=i,\mathbf{x}_n)P(y_n|J_{(k+1)\epsilon}=j,\mathbf{x}_n)}{P(y_n|\mathbf{x}_n)}$$
$$=\sum_{k=0}^{[y/\epsilon]-1}\frac{1}{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}}$$
$$\times\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}u\}\mathbf{e}_i\mathbf{e}_i^T\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}\epsilon\}\mathbf{e}_j\mathbf{e}_j^T$$
$$\times\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}(y_n-(u+\epsilon))\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}\}$$
$$\to\int_{u=0}^{y_n}\frac{1}{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}}$$
$$\times\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}u\}\mathbf{e}_i(\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{ij})\mathbf{e}_j^T$$
$$\times\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}(y_n-u)\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}\}du,$$

and thus

$$E\left[N_{ijn}|y_n,\mathbf{x}_n\right]=\frac{\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{ij}c_{jn}(y_n,i|\boldsymbol{\pi},\mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_n|\mathbf{T})},$$
$$i=1,...,m,\ j=1,...,m \text{ and } i\neq j.$$

Similarly, we can achieve

$$E\left[N_{i0n}^{\epsilon}|y_n,\mathbf{x}_n\right]$$
$$=\frac{P(J_{y_n-\epsilon}=i,y_n|\mathbf{x}_n)}{P(y_n|\mathbf{x}_n)}$$
$$=\frac{P(y_n|J_{y_n-\epsilon}=i,\mathbf{x}_n)P(J_{y_n-\epsilon}=i|\mathbf{x}_n)}{P(y_n|\mathbf{x}_n)}$$
$$=\frac{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}(y_n-\epsilon)\}\mathbf{e}_i\mathbf{e}_i^T\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}\epsilon\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}}{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}}$$
$$\to\frac{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\mathbf{e}_i\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i0}}{\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}},$$
$$E\left[N_{i0n}|y_n,\mathbf{x}_n\right]=\frac{\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{i0}a_{in}(y_n|\boldsymbol{\pi}\mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_n|\mathbf{T})},\ i=1,...,m,$$

where

$$b_i(y_n|\boldsymbol{\pi},\mathbf{T})=\mathbf{e}_i^T\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q},$$
$$\mathbf{b}(y_n|\mathbf{T})=\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q},$$
$$c_{in}(y_n,i|\boldsymbol{\pi},\mathbf{T})=\int_0^{y_n}\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}u\}\mathbf{e}_i\mathbf{e}_i^T\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}(y_n-u)\},$$
$$\times\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}du$$
$$a_{in}(y_n|\boldsymbol{\pi}\mathbf{T})=\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y_n\}\mathbf{e}_i,$$

and $\mathbf{e}_i$ is an all-one vector. The most difficult part of the E-Step is to calculate the terms of $c_{jn}(y_n,i|\boldsymbol{\pi},\mathbf{T})$ and $a_{in}(y_n|\boldsymbol{\pi}\mathbf{T})$, and since it is differentiable, we can apply the Runge-Kutta method (Asmussen et al., 1996) to approximate its value. By the Runge-Kutta method, we can obtain

$$a'_{in}(y|\boldsymbol{\pi},\mathbf{T})=\boldsymbol{\pi}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y\}\mathbf{e}_i$$
$$=\sum_{j=1}^m\exp(-\mathbf{b}\mathbf{x}_n)\lambda_{ji}a_{jn}(y|\boldsymbol{\pi},\mathbf{T}).$$

We can set the initial $y_0=0$, then

$$a_i(0|\boldsymbol{\pi},\mathbf{T})=\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}\times 0\}\mathbf{e}_i=\pi_i.$$

Similarly,

$$c_{in}(y,i|\boldsymbol{\pi},\mathbf{T})=\int_0^{y_n}\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}u\}\mathbf{e}_i\mathbf{e}_i^T$$
$$\times\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}(y_n-u)\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}du.$$

The initial function value is $c_{in}(0|\boldsymbol{\pi},\mathbf{T})=0$,

$$c'_{jn}(y,i|\boldsymbol{\pi},\mathbf{T})=\boldsymbol{\pi}\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}y\}\mathbf{e}_i\mathbf{e}_j^T$$
$$\exp\{\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{T}(y_n-y)\}\exp(-\mathbf{b}\mathbf{x}_n)\mathbf{q}$$
$$=\sum_{k=0}^m\lambda_{ik}c_{kn}(y,j|\boldsymbol{\pi},\mathbf{T}).$$

In the M-Step, the maximum likelihood estimators for λ_{ij} and λ_i are given by

$$\hat{\pi}_i^{(k+1)}=\frac{\sum_{n=1}^N B_{in}^{(k+1)}}{N},$$
$$\hat{\lambda}_{ij}^{(k+1)}=\frac{\sum_{n=1}^N N_{ijn}^{(k+1)}}{\sum_{n=1}^N Z_{in}^{(k+1)}},\ \hat{\lambda}_{i0}^{(k+1)}=\frac{\sum_{n=1}^N N_{i0n}^{(k+1)}}{\sum_{n=1}^N Z_{in}^{(k+1)}},$$

where N is the number of samples. In Coxian PH distribution, $\pi_1 = 1, \pi_i = 0, i = 2, ..., m$.

We then need to maximize $\log L(\Theta|\mathbf{Y})$ upon the coefficients $\mathbf{b}$, that is getting the root of $\frac{\partial \log L(\Theta|\mathbf{Y})}{\partial \mathbf{b}} = 0$. The Newton-Raphson method is used to obtain the new estimator of $\mathbf{b}$. The first and second derivative of the log-likelihood function over coefficient $\mathbf{b}$ are.

$$\frac{\partial \log L(\Theta|\mathbf{Y})}{\partial \mathbf{b}} = \sum_{n=1}^{N} \left\{ \sum_{i=1}^{m} \mathbf{x}_n \exp(-\mathbf{b}\mathbf{x}_n) \lambda_i Z_{in} + \sum_{i=1}^{m} \sum_{j=0, j\neq i}^{m} -\mathbf{x}_n N_{ijn} \right\}$$

$$\frac{\partial^2 \log L(\Theta|\mathbf{Y})}{\partial \mathbf{b}^2} = \sum_{n=1}^{N} \left\{ \sum_{i=1}^{m} -\mathbf{x}_n \mathbf{x}_n^T \exp(-\mathbf{b}\mathbf{x}_n) \lambda_i Z_{in} \right\}$$

Therefore the new estimator is given as $\mathbf{b}^{(s+1)} = \mathbf{b}^{(s)} - \frac{\partial \log L(\Theta|\mathbf{Y})}{\partial \mathbf{b}} \left[\frac{\partial^2 \log L(\Theta|\mathbf{Y})}{\partial \mathbf{b}^2}\right]^{-1}$.

2.5. Los groups

In this paper, the Akaike Information Criterion $AIC = 2k - 2\max_{\Theta} \log L(\Theta|\mathbf{Y})$ will be used to determine the number of states by taking into consideration the number k of parameters. The most appropriate number m of states in the Coxian PH distribution is obtained by minimizing the value of AIC.

After the model structure is determined, the transition rates λ_{ij} and $\lambda_{i0}, i, j = 1, ..., m$ can be obtained from the estimated $\mathbf{T}$, and the coefficients for the covariates can be obtained from the estimated $\mathbf{b}$. Let P_i, $i = 1, ..., m$ be the probability of transferring from state i to absorbing state 0, and the time spent in state i follows $\exp(\lambda_i)$. According to the balance equation $\lambda_i = \lambda_{i,i+1} + \lambda_{i,0}$, $i = 1, ..., m-1$ and $\lambda_m = \lambda_{m0}$, the time spent in each transient state contributes to two directions, one transferring to the next neighboring state with rate $\lambda_{i,i+1}$ and another one for being absorbed to state 0 with rate $\lambda_{i,0}$. Then, the proportion of being absorbed to state 0 from state i directly is obtained using the formula below (for calculations, we refer to Gu et al. (2019)):

$$P_1 = \frac{\lambda_{10}}{\lambda_{10} + \lambda_{12}},$$

$$P_i = \frac{\lambda_{12}}{\lambda_{10} + \lambda_{12}} \times \frac{\lambda_{23}}{\lambda_{20} + \lambda_{23}} \times \cdots \times \frac{\lambda_{i,0}}{\lambda_{i,0} + \lambda_{i,i+1}}, \quad i = 2, ..., m-1,$$

$$P_m = \frac{\lambda_{12}}{\lambda_{10} + \lambda_{12}} \times \frac{\lambda_{23}}{\lambda_{20} + \lambda_{23}} \times \cdots \times \frac{\lambda_{m-1,m}}{\lambda_{m-1,0} + \lambda_{m-1,m}}.$$

To determine the LOS groups, we first sorted the LOS data in an ascending order. The first group has the shortest LOS, while the mth one has the longest LOS.

3. Data preprocessing

3.1. Data description and covariates

The proposed approaches are applied to the patient flow data collected from 2012 to 2017 in Banner University Medical Center Tucson - Main Campus and South Campus. Since most developed countries have accepted 65 years as a definition of "elderly" (WHO, 2002), we collected 3287 electronic medical records (EMRs) of 2183 patients with their age larger than or equal to 65 in the data. The diagnosis geriatric patients at admission have common seen geriatric diseases, like Alzheimer Disease, Heart failure, etc. From the same data source, 3586 records are collected with diagnosis code containing F10.1 (Alcohol abuse) in terms of ICD-10 codes, and of 2019 patients are identified among them.

Table 1. Statistical description of LOS.

	Geriatric	AUD
No. of records	3287	3586
Mean	4.96	0.8118
Min	0.01	0.0007
Max	93.49	32.3944
Median	1.47	0.3688
Mode	3.76	0.1361
Std	8.58	1.9652
Skewness	3.66	7.0761
Kurtosis	18.63	74.1356
25th percentile	0.25	0.2326
75th percentile	6.19	0.5521

The statistical description of the LOS information for the collected data is presented in Table 1. The range of the LOS of geriatric patients (from 0.01 day to 93.49 days) is much wider than that of AUD patients (from 0.0007 day to 32.3944 days), and so are the mean, the 25th percentile and the 75th percentile. The mean LOS of both diseases are larger than the mode, and then the median, indicating that the LOS data is right skewed and this can be confirmed by the positive-valued skewness. Furthermore, the LOS of two diseases has large values of kurtosis, especially that the LOS of AUD patients has kurtosis as 74.1356, and it shows that the LOS data has a heavy tail.

The covariate information of patients in the collected data includes: gender, age, admission type, admission source, and financial class for the payment of medical care. (i) For the *gender*, both (male and female) are recorded. (ii) When considering geriatric diseases, the covariate *age* is partitioned into two categories, which are Age$\leq$ 85 and Age $>$ 85. While for AUD, the age of the patients varies from 10 to 90. According to the minimum alcohol drinking age and the common-seen geriatric age, we divide the age variable into three categories as less than 21 years, 21 to 65 years, and older than 65 years. (iii) According to (CMS Manual System, 2018) "FL 14 - Type of Admission/Visit", the *admission types* are classified into four categories as Emergency, Urgent, Elective and Trauma Center. However, these four categories are similar to each other so that we may use some statistical tests to check whether the variables of admission types matter. (iv) In the same code list, *admit sources* are divided into nine categories, Ambulatory Surgery Center, Court/Law Enforcement, Discharge and Readmit, Emergency Room, Outside Hospital, Outside Healthcare Facility, Physician or Clinic Referral, Self Referral, and Skilled Nursing Facility. Different from admission types which refer to the circumstances under which the patient is admitted, the admit sources describe the origin of the patient's admission. (v) The *financial classes* for the payment of medical care include Commercial insurance, Medicaid, Medicare, Self-pay and Other.

Table 2. Kruskal-Wallis test of LOS of geriatric patients.

Covariates	df	h	*p* value
Gender	1	8.57	0.0034
Age	1	26.54	0.0000
Admission type	3	351.45	0.0000
Admit source	4	294.41	0.0000
Financial class	4	29.48	0.0000

Table 3. Kruskal-Wallis test of LOS of AUD patients.

Covariates	df	h	*p* value
Gender	1	1.24	0.2657
Age	2	47.80	0.0000
Admission type	3	108.65	0.0000
Admit source	4	83.77	0.0000
Financial class	4	37.26	0.0000

3.2. Statistical tests on covariates

Before incorporating these five type of covariates into our model, we firstly use Kolmogorov-Smirnov (KS) tests to check the normality of LOS in each category. The p values of KS tests of LOS in each category are close to 0, and the results reject the original hypothesis that the LOS in each category is normally distributed or follows a specific distribution. Thus, the parametric test of One-Way Analysis of variance (ANOVA), assuming the distribution of residuals are normal, may not be used here. In situations where the normality assumption is unjustified, an alternative procedure that does not depend on this assumption is needed. The Kruskal-Wallis (KW) test is used to test the null hypothesis that the LOS distribution affected by covariates are identical against the alternative hypothesis that at least one of the distributions is different from others. In general, the KW test is a nonparametric alternative to the usual ANOVA (Montgomery, 2017).

In this study, the significance level is chosen as $\alpha = 0.05$, and if the test statistic h is larger than $\chi^2_{\alpha,t-1}$ (t is the number of categories in each covariate), the null hypothesis will be rejected. The h, p value and degrees of freedom $(t-1)$ with the corresponding covariates categories are listed in Tables 2 and 3. One can see that *the effects of gender, age, admission type, admit source, financial class are significant in studying the geriatric patients based on the test result. While the all of the covariates except gender have significant effects on the LOS of AUD patients.*

The KW test is significant, and thus a post-hoc analysis can be performed to determine which categories of a certain covariate differ from others. One of the most popular test for this is the Dunn's test. Dunn's Multiple Comparison Test is a post-hoc non-parametric test, which means that it should run after the Kruskal-Wallis test and it is a "distribution free" test (Dunn, 1961). The null hypothesis for the test is that there is no difference between the effects of categories under one covariate on LOS distribution (categories can be equal or unequal in size). The alternative hypothesis for the test is that there is a difference between categories under one covariate. Since Dunn's test is appropriate for groups with unequal numbers of observations (Zar, 2010), it is applied to each covariate separately in this study to identify where the difference occurs.

Table 4. Distributions fitting restults of LOS of geriatric patients.

Transition rate		P_i	Group	Min	Max	Average	No. of records
λ_{10}	0.0000	0.0000	–	–	–	–	–
λ_{20}	3.4393	0.4047	G_1	0.0083	0.5139	0.2226	1331
λ_{30}	0.1267	0.5953	G_2	0.5153	93.4931	8.1789	1956

For a pair of categories under same covariate, if the p value is small enough, the null hypothesis that the pair has the same effect on the LOS distribution will be rejected.

We set the significance level as $\alpha = 0.05$ in this study and for *patients with geriatric diseases*, the category pairs under gender, age and admission type are all rejected, meaning that the LOS of patients in each category under same covariate follow different distributions.

Based on the Dunn's test results for the patients with geriatric diseases, we can classify admit source of Court/Law Enforcement, Discharge and Readmit, Outside Hospital, Outside Healthcare Facility, and Skilled Nursing Facility into one group. Besides, Physician or Clinic Referral, Self Referral, and Emergency Room can be grouped together since the pair of these three categories are significant. Thus, the admit source actually has 3 categories: Admit source 1 (Ambulatory Surgery Center), Admit source 2 (Court/Law Enforcement, Discharge and Readmit, Outside Hospital, Outside Healthcare Facility, Skilled Nursing Facility), and Admit source 3 (Physician or Clinic Referral, Self Referral, Emergency Room).

In the same way for the patients with geriatric diseases, financial classes of Medicaid, Medicare, Self-pay, and Other should be combined together. So the final categories in the Financial Class should be Noncommercial and Commercial. Above all, we can conclude that the LOS distributions of geriatric patients depend on covariates consisting of *gender, age, admission type, admit source, and financial class.*

Similarly, for *patients with AUD*, all of the categories under the covariates of age, admission type, financial class will be kept unchanged since we reject the null hypotheses for all category pairs. Four categories under Admit Source has been reassigned into one category as Admit source other (Physician or Clinic Referral, Outside Healthcare Facility, Other Banner Hospital, and Psych, Substance Abuse, or Rehab Hospital). Totally, there are 5 categories under Admit Source: Court/Law Enforcement, Discharge and Readmission, Emergency Room, Self Referral, Admit source other. Above all, we can conclude that the LOS distributions depend on patients age, admission type, admit source, and financial class.

In Section 4, a Coxian PH distribution with inclusion of covariates will be fitted to the LOS data and the influences of covariates will also be identified.

4. Numerical analysis

4.1. Geriatric diseases

As verified in the Section 3, five types of covariates, including gender, age, admission type, admit source, and financial class may influence the geriatric patients' LOS. There are two categories in gender (Female, Male), age ($\leq$85, $>$85),

Table 5. Statistical description of covariates for geriatric patients.

	Category	#	b_j	Mean
Gender	Female	1763	−0.0344	4.6185
	Male	1524	–	5.3489
Age	Age $>$ 85	485	−0.0940	4.8381
	Age $\leq$ 85	2802	–	4.9778
Admission type	Elective	294	−0.5423	3.0028
	Emergency	2678	−0.5243	4.4556
	Trauma	39	−0.4749	5.2776
	Urgent	276	–	11.8609
Admit source	Admit source 1	67	−0.5475	0.2097
	Admit source 2	240	0.2599	10.2069
	Admit source 3	2980	–	4.6411
Financial class	Noncommercial	3062	0.4367	5.0309
	Commercial	225	–	3.9536

financial class (Noncommercial, Commercial) respectively, four categories in admission type, including Emergency, Urgent, Elective and Trauma Center, and three categories in admit source (named Admit Source 1,2,3). Therefore, a total of thirteen categories should be considered in the model. Eight dummy binary variables, with value one indicating the presence of the category, are created to keep the independency of all categories.

We apply the EM algorithm introduced in Section 2.4 on Coxian PH distributions with various number of phases, and the least *AIC* value occurred to the three phases. Since the transition rate of $\lambda_{10} = 0$, which means that all patients in the first LOS group transfer to the second LOS group, the final model is a two-phase Coxian PH distribution.

The estimated transition rates, absorbing rates, and the proportion of each LOS group are shown in the Table 4. About 40.47% of the records of geriatric patients, that is a total number of 1331, should be classified into the first LOS group with minimum LOS as 0.0083 day, maximum LOS as 0.5139 day, and average LOS as 0.2226 day. In the meanwhile, the second LOS group consists of 1956 records, which occupies a proportion of 0.5953, and its LOS varies widely from 0.5153 day to 93.4931 days. The average LOS in the second group is 8.1789 days, which indicates severer situations and much more resources demands than the first group.

In addition to the transition matrix, the estimated coefficients of each category in five covariates are also obtained (Table 5). A numerical comparison of LOS in each category is also presented in Table 5, by which the efficiency of our model is verified. One of the categories for each covariate is designated as the base category, which will has no value of coefficient corresponding to it. As such, the coefficient for each non-base category indicated the effect of the presence of that category relative to the base category.

As shown in Table 5, female patients have less LOS than male patients, patients older than 85 leave earlier from hospital than younger patients, and patients with Noncommercial financial situations may stay in hospital for a longer period. Therefore, the factors, Female, and Age $>$ 85, have negative relationship with LOS, and this has been verified by the estimated results of our model. The potential reason is that patients older than 85 may discharge due to being transferred to other hospitals, to nursing homes, or fatality. Similarly, patients with noncommercial financial class have longer LOS than those with commercial insurance type. This relationship is supported by the positive coefficient of Noncommercial financial class.

Table 6. Distributions fitting restults of LOS of AUD patients.

Transition rate		P_i	Group	Min	Max	Average	No. of records
λ_{10}	0.0202	0.0045	G_1	0.0007	0.0306	0.0175	16
λ_{20}	4.7597	0.9209	G_2	0.0313	1.7368	0.3887	3302
λ_{30}	0.0154	0.0045	G_3	1.7500	1.9208	1.8329	16
λ_{40}	0.6156	0.0701	G_4	1.9306	32.3944	6.3418	252

Same findings can be achieved for other covariates. The top negative predictor of LOS is the Admit source 1 (Ambulatory Surgery Center), with which immediate attention for the care and treatment is required, and it is always expected to have less than 24 h of LOS. The estimation coincides with the finding in Table 5 that patients admitted from Ambulatory Surgery Center may have the least average LOS as 0.2097 day. The Admit Source 2, including sources like Discharge and Readmit, Outside Hospital, Skilled Nursing Facility, etc., positively impact the LOS relative to the Admit Source 3 and the average LOS in such category is 10.2069 days. The Admit Source 3, consisting of Physician or Clinic Referral, Self Referral, and Emergency Room, is designated as base category for admit source and has a relatively short LOS averagely at 4.6411 days.

The admission type of Urgent has the largest mean LOS, and it is intuitive that the patient with such admission type are in a severe situation and require more time and effort in the care and treatment although they have priority in the first available and suitable accommodation. This also explains the negativeness of the coefficients in admission type when the Urgent is chosen as base category.

With the Elective admission type, patients' conditions are usually not severe, leading to a shorter LOS and a negative coefficient of the covariate, which is also verified in Earnest et al. (2006). Furthermore, the admission types of Emergency and Trauma will both decrease the LOS compared with type of Urgent. The Emergency type contains the largest proportion of patients and they have a wide range LOS, which may lead to negative relationships with LOS in general.

4.2. AUD

In the same way of treating categorical variables, thirteen dummy variables are created and the four-phase Coxian PH distribution is then identified. In Table 6, majority of AUD patients are in the second LOS group with LOS ranging from 0.0313 day to 1.7368 days. All of the patients from the first three LOS groups discharge within 2 days, while the last group has LOS ranging from 1.9306 days to 32.3944 days.

In Table 7, the average LOS of patients older than 65 is the longest, the LOS of patients younger than 21 is the shortest, and the LOS of patients aged from 21 to 65 is in the middle. The relationship is then verified by the coefficients in Table 7 that both types of patients with Age $<$21 and 21$\leq$ Age $<$65 have negative coefficients if Age $\geq$ 65 is the base category. Similarly, patients admitted with urgent type have the largest LOS among all admission types and so

Table 7. Statistical description of covariates for AUD patients.

Covariates	Category	#	b_j	Mean
Age	Age $<$ 21	96	−0.2947	0.3795
	21 $\leq$ Age $<$ 65	3140	−0.0806	0.7852
	Age $\geq$ 65	350	–	1.1696
Admission type	Elective	10	−0.3613	0.3101
	Emergency	3385	−0.1941	0.7447
	Trauma	159	−0.1072	1.1931
	Urgent	32	–	6.1773
Admit source	Court/Law enforcement	47	−0.3292	0.3744
	Discharge and readmission	12	2.60448	5.3406
	Emergency room	2700	0.1110	0.7798
	Admit source other	115	0.1230	2.4435
	Self referral	712	–	0.6224
Financial class	Commercial	321	−0.0991	0.7602
	Medicaid	2205	−0.0560	0.7974
	Medicare	535	0.0751	1.1092
	Other	436	−0.1458	0.5403
	Self pay	89	–	0.8978

that the coefficients of Elective, Emergency, Trauma Center are all negative. The results in the admission type of AUD patients go the same way with the ones of geriatric patients, which not only verifies that our model has the ability of capturing the impacts of covariates, but also conveys the generality of the distribution of admission types.

The top predictor in Admit Source is Discharge and Readmission, and it is intuitive that patients being readmitted to hospital due to AUD may have being through terrible addiction and severe situations. One can notice that the Discharge and Readmission is included in the Admit Source 2 for the geriatric patients, which also has a relatively large LOS.

The predictor in Financial Class with longest LOS is Medicare, with largest coefficient. Compared with other categories, the Medicare patients have relative worse healthcare condition. The Other Financial Class has the shortest LOS, which is consistent with the most negative coefficient.

Above all, the models for geriatric patients and AUD patients both coincide with the LOS distribution under each covariates. The priority of inpatient admission in terms of the service indicated by the admission type (CMS Manual System, 2018) is also confirmed by our model. We can conclude with effectiveness of our method of fitting Coxian PH distribution with consideration of covariates. Besides, patients admitted through discharge and readmission needs more attention since they may have larger LOS.

4.3. Further analysis

In Section 4.1 and Section 4.2, we can roughly catch the differences between the fitted distribution of the LOS of geriatric patients and AUD patients. In this section, we perform further analysis and comparison of the discharge destinations among different LOS groups and two types of patients.

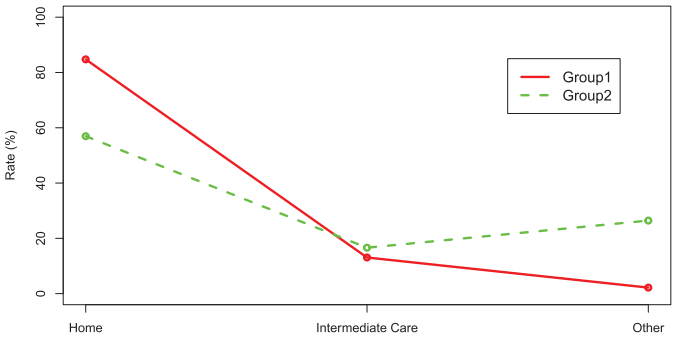
The Tables 8 and 9 and Figure 2 illustrate the percentage of each discharge destination and visual comparisons of it for both diseases. The home, as the most common seen discharge destination, always occupies the largest proportion among all discharge destinations no matter in which group or with which disease.

Table 8. Rates of discharge destinations for geriatric patients.

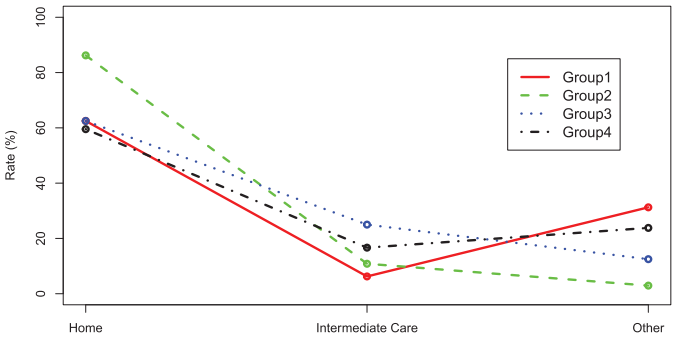
	Discharge to home	Discharge to intermediate care	Other
Group1	84.748310	13.072878	2.178813
Group2	56.952965	16.615542	26.431493

Table 9. Rates of discharge destinations for patients with AUD.

	Discharge to home	Discharge to intermediate care	Other
Group1	62.500000	6.250000	31.250000
Group2	86.220472	10.841914	2.937614
Group3	62.500000	25.000000	12.500000
Group4	59.523810	16.666667	23.809524



(a) geriatric patients



(b) patients with AUD

Figure 2. Plots of discharge destinations in each LOS group.

For geriatric patients (see Figure 2(a)), the first LOS group has higher rate of being discharged to home, and lower rate of going to both intermediate care and others than the second LOS group. That is due to the geriatric patients may require more treatment or care after discharging with the increase of LOS. Another fact is that the destination of others includes death, on which the second LOS group might has highest percentage. The proportion of patients going to intermediate care after discharge for two LOS groups are close to each other.

In Figure 3, we can notice an obvious rise in the proportion of the urgent admission type in the second LOS group, which explains the urgent status of patients at admission and the low rate of going back home after treatment.

For patients with AUD (see Figure 2(b)), the second LOS group has the largest number of patients, and this indicates that the LOS range in this group is the most reasonable for general AUD patients. Also, the second LOS group has the

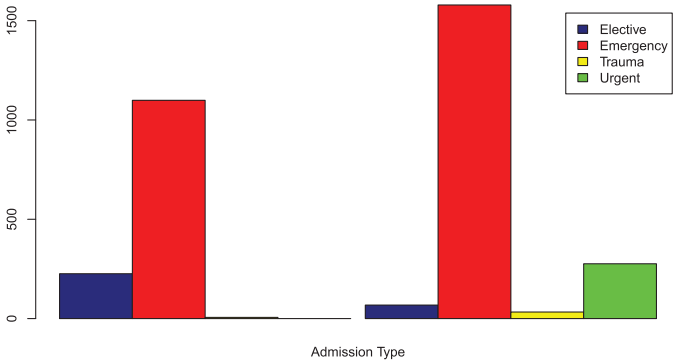


Figure 3. The distribution of admission type in each LOS group for geriatric patients.

highest percentage of home discharge, while the other three LOS groups share similar proportion. The third LOS group has the highest rate of being transferred to intermediate care.

5. Conclusions

This paper investigates the flow information of both geriatric patients and patients with AUD, based on the data collected in a medical center from 2012 to 2017. Several descriptive statistical modeling methods along with the EM algorithm for Coxian PH distributions are applied to analyze the LOS data with the information of covariates. In the descriptive statistical modeling part, we first present statistical descriptions of LOS in terms of each category according to the corresponding covariate. Then, several statistical hypothesis tests are conducted in order to better identify the significant covariates which indeed impact the LOS distributions. The KW test is chosen to check whether different covariates lead to different LOS distributions. A post hoc Dunn's test is then conducted to confirm the significance of the covariates by testing the existence of difference among the categories in pairs for each covariate.

The hypothesis tests verifies that the five covariates, including gender, age, admission type, admit source, and financial class are significantly impact the LOS of geriatric patients. The covariate of admit source is needed to be regrouped for both geriatric patients and AUD patients in order to keep the independency of each category.

In fitting the Coxian PH distribution to the LOS data considering the influences of covariates, we develop the EM algorithm by generating the expression of density of the complete observations and updating the sufficient statistics iteratively. Several numerical analysis methods are used to approximate the formulas and to obtain the MLE of parameters. The distribution of LOS groups and influences of each category under covariates on the LOS are obtained. The longest LOS group of geriatric patients stay in hospital from 0.5153 day to 93.4931 days, and the one of patients with AUD has LOS ranging from 1.9306 days to 32.3944 days.

Although the LOS distributions are different for the geriatric patients and AUD patients, there exists the same pattern of impacts of the admission type on LOS, which is captured by our model. The coincidence between the LOS in each category and the corresponding estimated coefficients also verifies the efficiency of our model. The pattern of the fitting results and further discharge destination analysis are then compared between geriatric patients and AUD patients. Patients in the same LOS group or with the same covariate may have some characteristics in common.

For geriatric patients, the top negative predictor of LOS is the admit source of Ambulatory Surgery Center, for which urgent requests for treatment are always the case. While for AUD patients, the most impactive factor is the admit source of Discharge and Readmission, which reveals the repeat of such a disease highly related to addiction. The geriatric patients with longer LOS are less likely to return home directly after treatment, which is due to more request of treatment and a higher probability of death during treatment with the increase of LOS. More urgent status for geriatric patients staying longer in hospital also explains the difference in the distribution of discharge destinations between LOS groups.

The efficiency of our extended EM algorithms has been verified in this paper in fitting the Coxian PH distributions and capturing the impacts of covariates. Also, the consistence between the LOS distribution in each category and the corresponding estimated coefficients verifies the effectiveness of our model. Even though this paper only studies five covariates, the proposed approaches can be applied on flow information of patients with other type of disease, and distinct factors that influence the LOS. The analysis of both the distribution of LOS groups and influences of covariates on LOS will further offer variable insight into the improvement of healthcare service and resource allocation by reviewing patients information at admission. Patients with the characters that have larger impacts on LOS may have higher chances of requesting priority in treatment, resources assignments and staying longer in hospital.

Acknowledgements

We would appreciate the University of Arizona Center for Biomedical Informatics & Biostatistics Department of Biomedical Informatics Services for providing the data.

Funding

This material is based upon work supported by National Science Foundation Grants #1634282, #1635379 and #1740858.

ORCID

Neng Fan http://orcid.org/0000-0003-4333-3721

References

Alcohol Facts and Statistics. (2020). *Alcohol use disorder (AUD) in the United States*. National Institute on Alcohol Abuse and Alcoholism. Retrieved November 13, 2020, from https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics

Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, *23*(4), 419–441.

Cloninger, C. R., Bohman, M., & Sigvardsson, S. (1981). Inheritance of alcohol abuse. Cross-fostering analysis of adopted men. *Archives of General Psychiatry*, *38*(8), 861–868. https://doi.org/10.1001/archpsyc.1981.01780330019001

CMS Manual System. (2018). Retrieved September 10, 2018, from https://www.cms.gov

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220.

Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, *51*(2), 313–319. https://doi.org/10.1017/S0305004100030231

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Earnest, A., Chen, M. I., & Seow, E. (2006). Exploring if day and time of admission is associated with average length of stay among inpatients from a tertiary hospital in Singapore: An analytic study based on routine admission data. *BMC Health Services Research*, *6*(1), 6. https://doi.org/10.1186/1472-6963-6-6

El-Darzi, E., Vasilakis, C., Chaussalet, T., & Millard, P. H. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, *1*(2), 143–149.

Fackrell, M. W. (2003). *Characterization of matrix-exponential distributions* [PhD thesis]. School of Applied Mathematics, University of Adelaide.

Fackrell, M. W (2009). Modelling healthcare systems with phase-type distributions. *Health Care Management Science*, *12*(1), 11–26. https://doi.org/10.1007/s10729-008-9070-y

Faddy, M., Graves, N., & Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value in Health*, *12*(2), 309–314. https://doi.org/10.1111/j.1524-4733.2008.00421.x

Faddy, M. J., & McClean, S. I. (2007). Using a multi-state model to enhance understanding of geriatric patient care. *Australian Health Review*, *31*(1), 91–97. https://doi.org/10.1071/ah070091

Finney, J. W., Moos, R. H., & Chan, D. A. (1981). Length of stay and program component effects in the treatment of alcoholism: A comparison of two techniques for process analyses. *Journal of Consulting and Clinical Psychology*, *49*(1), 120–131. https://doi.org/10.1037/0022-006x.49.1.120

Gardiner, J. C. (2012). Modeling heavy-tailed distributions in healthcare utilization by parametric and Bayesian methods. *SAS Global Forum*. 418-2012, 1–15.

Gardiner, J. C., Luo, Z., Tang, X., & Ramamoorthi, R. V. (2014). Fitting heavy-tailed distributions to health care data by parametric and Bayesian methods. *Journal of Statistical Theory and Practice*, *8*(4), 619–652. https://doi.org/10.1080/15598608.2013.824823

Gottheil, E., McLellan, A. T., & Druley, K. A. (1992). Length of stay, patient severity and treatment outcome: Sample data from the field of alcoholism. *Journal of Studies on Alcohol*, *53*(1), 69–75. https://doi.org/10.15288/jsa.1992.53.69

Gu, W., Fan, N., & Liao, H. (2019). Evaluating readmission rates and discharge planning by analyzing the length-of-stay of patients. *Annals of Operations Research*, *276*(1-2), 89–108. https://doi.org/10.1007/s10479-018-2957-1

ICD-10. (2016). *International statistical classification of diseases and related health problems (ICD)* (10th Rev.). World Health Organization.

Kwok, C. L., Lee, C. K., Lo, W. T., & Yip, P. S. (2017). The contribution of ageing to hospitalisation days in Hong Kong: A decomposition analysis. *International Journal of Health Policy and Management*, *6*(3), 155–164. https://doi.org/10.15171/ijhpm.2016.108

Li, J. (1999). An application of lifetime models in estimation of expected length of stay of patients in hospital with complexity and age adjustment. *Statistics in Medicine*, *18*(23), 3337–3344. https://doi.org/10.1002/(SICI)1097-0258(19991215)18:23<3337::AID-SIM320>3.0.CO;2-5

Long, C. G., Williams, M., & Hollin, C. R. (1998). Treating alcohol problems: A study of programme effectiveness and cost effectiveness according to length and delivery of treatment. *Addiction (Abingdon, England)*, *93*(4), 561–571. https://doi.org/10.1046/j.1360-0443.1998.93456111.x

McGrory, C. A., Pettitt, A. N., & Faddy, M. J. (2009). A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis*, *53*(12), 4311–4321. https://doi.org/10.1016/j.csda.2009.05.021

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley and Sons.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models*. The Johns Hopkins University Press.

Park, J. H., Park, J. O., Ro, Y. S., & Do Shin, S. (2018). Effect of alcohol use on emergency department length of stay among minimally injured patients based on mechanism of injury: Multicenter observational study. *Clinical and Experimental Emergency Medicine*, *5*(1), 7–13. https://doi.org/10.15441/ceem.16.180

Saitz, R., Ghali, W. A., & Moskowitz, M. A. (1997). The impact of alcohol-related diagnoses on pneumonia outcomes. *Archives of Internal Medicine*, *157*(13), 1446–1452. https://doi.org/10.1001/archinte.1997.00440340078008

Tang, X., Luo, Z., & Gardiner, J. C. (2012). Modeling hospital length of stay by Coxian phase-type regression with heterogeneity. *Statistics in Medicine*, *31*(14), 1502–1516. https://doi.org/10.1002/sim.4490

Toh, H. J., Lim, Z. Y., Yap, P., & Tang, T. (2017). Factors associated with prolonged length of stay in older patients. *Singapore Medical Journal*, *58*(3), 134–138. https://doi.org/10.11622/smedj.2016158

Turgeman, L., May, J., Ketterer, A., Sciulli, R., & Vargas, D. (2015). Identification of readmission risk factors by analyzing the hospital-related state transitions of Congestive Heart Failure (CHF) patients. *IIE Transactions on Healthcare Systems Engineering*, *5*(4), 255–267. https://doi.org/10.1080/19488300.2015.1095823

World Health Organization (WHO). (2002). *Proposed working definition of an older person in Africa for the MDS project*. Retrieved from September 10, 2018, http://www.who.int/healthinfo/survey/ageingdefnolder/en

WHO. (2015). *World report on ageing and health 2015*. Retrieved from September 10, 2018, http://www.who.int/ageing/events/world-report-2015-launch/en

Xie, H., Chaussalet, T. J., & Millard, P. H. (2005). A continuous time Markov model for the length of stay of elderly people in institutional long-term care. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(1), 51–61. https://doi.org/10.1111/j.1467-985X.2004.00335.x

Zar, J. H. (2010). *Biostatistical analysis* (5th ed.). Pearson Prentice Hall.

Zhang, S., Payton, F. C., & Ivy, J. S. (2013). Characterizing the impact of mental disorders on HIV patient length of stay and total charges. *IIE Transactions on Healthcare Systems Engineering*, *3*(3), 139–146. https://doi.org/10.1080/19488300.2013.820238

Zhu, T., Luo, L., Zhang, X., & Shen, W. (2018). Modeling the length of stay of respiratory patients in emergency department using Coxian phase-type distributions with covariates. *IEEE Journal of Biomedical and Health Informatics*, *22*(3), 955–965. https://doi.org/10.1109/JBHI.2017.2701779