Use of Markov Chain Simulation in Long Term Care Insurance

Vladimír Mucha[1] | *University of Economics in Bratislava, Bratislava, Slovakia*
Ivana Faybíková[2] | *University of Economics in Bratislava, Bratislava, Slovakia*
Ingrid Krčová[3] | *University of Economics in Bratislava, Bratislava, Slovakia*

Received 13.4.2022, **Accepted** (reviewed) 27.6.2022, **Published** 16.12.2022

Abstract

The aim of this paper is to present the use of simulations of non-homogeneous Markov chains in discrete time in the context of the problem of long-term care delivery. The object of investigation is to model the distribution of clients into different states during specified time steps, then to estimate the average time a client stays in a given state, as well as to estimate the insurance premiums. Within the use of the Monte Carlo simulation method, the focus is on providing approaches that ensure more accurate results in the context of the number of simulations performed. Based on the statistical processing of the data obtained from the simulations, it is possible to obtain the information necessary for the provision of resources for the provision of health care and for the determination of the aforementioned premiums. For the implementation of the above techniques and their graphical presentation available packages such as markovchain, ggplot2 or custom code created using the R language were used.

Keywords	*DOI*	*JEL code*
Long-term care insurance, Markov Chains, multi-state models, simulations, Monte Carlo Method, markovchain package	*https://doi.org/10.54694/stat.2022.20*	*C63, G22*

INTRODUCTION

We currently see an increase in average life expectancy and we can assume that this trend will continue in the future. It is the older age group that suffers from various chronic illnesses or physical limitations, and it is the older age group that makes the most use of the services of healthcare providers. For this reason, health care institutions pay considerable attention to estimating the number of clients who will need health care later on. They focus particularly on the issue of Long-Term Care (LTC), which is provided to people who have reached a state of non-self-sufficiency. The increased number of people who become incapacitated due to illness also represents an increase in health care costs. Looking at the other side of the issue, people are also thinking about the capital that they would have available in the event that they need long-term care. Without long-term care insurance, the cost of providing these

[1] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: vladimir.mucha@euba.sk.

[2] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: ivana.faybikova@euba.sk.

[3] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: ingrid.krcova@euba.sk.

services can quickly deplete an individual's or family's savings. Markov chains are a popular tool used to make estimations as regards to the incidence of critical illnesses and the provision of long-term care. The issue of using these random processes to model the evolution of different illnesses in the context of multi-state models is currently being addressed by many authors. They use various software support to implement them, one suitable possibility being the R language with its available packages. For example, the available R language package markovchain (Spedicato, 2017) can be used to create Markov chain objects, the implementation of probabilistic operations using them, statistical analysis and simulation of homogeneous and non-homogeneous Markov chains in discrete time. Another tool for dealing with Markov chains in discrete time is the DTMCPack package (Nicholson, 2013). Functions to implement Markov Chain Monte Carlo (MCMC) using the Metropolis algorithm, for example, are contained in the package mcmc (Geyer and Johnson, 2013). The open-source software MARCH, which is a set of functions from the MATLAB programming environment (Berchtold, 2001), is also available to model Markov chains in discrete time. If necessary, custom code can be developed in the above programming languages to simulate Markov chains in discrete and continuous time according to algorithms available in various publications, e.g. (Janková, et al., 2014: 85–87). As mentioned before Markov chains are also used in the context of multi-state models for modelling in the field of life and non-life insurance, e.g. for estimating the costs associated with the provision of health care, estimating premiums, predicting the evolution of various illnesses, as well as for modelling the number of insured lives in a bonus -malus system in the framework of compulsory car insurance. The problem of planning financial resources for health care is presented by Garg et al. (2010), using non-homogeneous Markov chains in discrete time to model the number of patients, as well as by Diz and Query (2012). Methods for estimating transition probabilities or transition intensities, the use of Markov Chain Monte Carlo simulation and modelling with Markov Chains also in continuous time in the field of long-term care are discussed by other authors such as (Sato and Zouain, 2010; Esquível et al., 2021; Fleischmann, Hirz, Sirianni, 2021; Xie, Chaussalet, Millard, 2005). Modelling with Markov chains also allows estimation of the expected or average stay time of an individual in the healthy and sick states, respectively (Dudel and Myrskylä, 2020). Another area of interest in the implementation of Markov chains in critical illness modelling in the context of healthcare is critical illness insurance. A lump sum benefit will be paid to the insured in case of a critical illness diagnosis. The above issue is presented by Pasaribu et al. (2019), using the continuous-time Markov chain apparatus to estimate the premiums for specified age categories of insureds. Many authors use Markov chains to predict the evolution of various infectious diseases (Li, Dushoff and Bolker, 2018; Twumasi, Asiedu and Nortey, 2019). An alternative stochastic modeling approach that can be implemented in this area is represented by Hawkes processes (Maciak, Okhrin and Pešta, 2021; Unwin et al., 2021). In non-life insurance, homogeneous Markov chains in discrete time are used to model the distribution of the number of policyholders in a bonus-malus system (Fernandez-Morales, 2015). The present paper focuses on the simulation of trajectories of non-homogeneous Markov chains in discrete time using the R language. Based on the processing of the generated data, we will analyze the modelling of the distribution of the number of clients in each state during the specified years, as well as the estimation of the average time a client stays in a given state, and the estimation of the premiums in the case of long-term care insurance. To present the above techniques, we have used data obtained from the markovchain package, which were for the male population in Italy. In the three-state model, the sick state represents the state of unfitness into which the client has fallen due to, for example, Alzheimer's disease.

1 METHODS OF ANALYSIS

If the insurer has real data on the health status of insured lives, it can obtain transition probabilities between the different states, which can be used to model the evolution of the insured's state over the analysed time periods. Since these probabilities depend on the age of the insured in our dataset, we use non-homogeneous Markov chains and their simulations for this purpose.

1.1 Non-homogenous Markov chains in discrete time

A random chain $\{X_t\}_{t\in T}$ is called a Markov chain, if for each h = 0, 1, 2, ..., for all times $0 \le t_0 \le t_1$ ..., $\mathrm{t}_h \le \mathrm{t}_{h+1}$, t_0, t_1 ..., t_h, $t_{h+1} \in T$ and for all states $s \in S$ we have

$$P(X_{t_{h+1}} = s_{t_{h+1}} \mid X_{t_h} = s_{t_h}, \dots , X_{t_1} = s_{t_1}, X_{t_0} = s_{t_0}) = P(X_{t_{h+1}} = s_{t_{h+1}} \mid X_{t_h} = s_{t_h}), \tag{1}$$

assuming that the random variable $X_{t_{h+1}}$ is independent from $X_{t_0}, X_{t_1}, \dots , X_{t_h}$ (Janková, et al., 2014).

This means that in the case of Markov chains, the probability of transition to the next state depends only on the current state and not on previous states, hence they are also called "memoryless" chains (Dobrow, 2016). We consider Markov chains in discrete time, so T is the set of natural numbers with zero. The values taken by the random variables X_t, $t \in T$, are called states, we denote their set by $S = \{s_1, s_2, \dots s_m \}$. We call the Markov chain $\{X_t\}_{t\in T}$ *non-homogenous* (in time), unless we have that as follows:

$$\forall i, j \in S, \forall k \in N: \; P(X_{t+k+1} = j \mid X_{t+k} = i) = P(X_{t+1} = j \mid X_t = i) . \tag{2}$$

Transition probabilities from state i to state j after one time step from the time t we denote by

$$p_{i,j}(t, t + 1) = P(X_{t+1} = j \mid X_t = i) , \tag{3}$$

and arrange them for a given t into the transition probability matrix

$$P(t; t + 1) = \big(p_{i,j}(t; t + 1)\big)_{i,j\in S} , \tag{4}$$

for which we have $\sum_{j\in S} p_{i,j}(t; t + 1) = 1$, which means, that each row of this matrix is a probability distribution, we call it a stochastic matrix (Jones and Smith, 2018).

Let $\{X_t\}_{t\ge 0}$ be a Markov chain. The probability distribution $\alpha = \{\alpha_k\}_{k\in S}$ such that $P(X_{t_0} = s_k) = \alpha_k = p_{s_k}(0)$ for $s_k \in S$, we call *the initial distribution of the chain* $\{X_t\}_{t\ge 0}$.

The vector $\mathbf{p}(0) = (p_{s_1}(0); \dots ; p_{s_k}(0); \dots ; p_{s_m}(0))$ will be called the *vector of initial probabilities*. The probability of transition from the initial state k to state j, $j \in S$ in h time steps from time 0, i.e. from the beginning of the Markov chain, is called *the absolute probability of the states of the Markov chain* and is denoted as follows

$$p_{k,j}(0,h) = p_j^{(k)}(h), \tag{5}$$

whereby we will call the vector $\mathbf{p}^{(k)}(h) = (p_j^{(k)}(h))_{j\in S}$ *the vector of absolute probabilities*.

We get the following expression for the vector of absolute probabilities $\mathbf{p}^{(k)}(h)$ using the *Chapman-Kolmogorov equality* (Fecenko, 2018).

$$\mathbf{p}^{(k)}(h) = \mathbf{p}^{(k)}(h - 1) \cdot P(h - 1; h) = \mathbf{p}(0) \cdot \dots \cdot P(h - 1; h). \tag{6}$$

1.2 Generating trajectories of a non-homogeneous Markov chain in discrete time in R

For a non-homogenous Markov chain with transition matrices $P(t; t + 1) = (p_{ij}(t; t + 1))_{i,j\in S}$ for $t \in T$, we define a random variable Z_r.

The values of the probability function $P(Z_r = j)$ represent in the corresponding matrix $P(t; t + 1)$ the values $p_{s_r j}(t; t + 1)$, $j \in S = \{s_1, s_2, \dots s_m\}$, which appear in its r-th row. We write this discrete distribution using the notation $p_{Z_r}(j) = \{p_{s_r j}(t; t + 1)\}_{j \in S}$. The algorithm for generating the random variable values Z_r by the inverse transformation method can be written as follows in the given context:

1. generate the value u of random variable $U \sim Unif(0; 1)$
2. transform the value u to the value of the random variable Z_r as follows

$$Z_r = s_1, \text{ if } u \le p_{Z_r}(s_1) \quad \text{or} \quad Z_r = j, \text{ if } \sum_{l=s_1}^{j-1} p_{Z_r}(l) < u \le \sum_{l=s_1}^{j} p_{Z_r}(l) \,. \tag{7}$$

We will simulate a non-homogeneous Markov chain with transition matrices $P(t; t + 1)$ and initial distribution $\alpha=\{\alpha_k\}_{k \in S}$ on a set of states $S = \{s_1, s_1, \dots s_m\}$ in discrete time, i.e. construct its trajectory, until time t_h using the following steps:

1. From the discrete initial distribution $\{\alpha_k\}_{k \in S}$ we generate the value $s_{t_0} = s_k$ of the random variable X_{t_0} at the initial point of time.
2. From the discrete distribution $\{p_{s_k j}(t_0;t_1)\}_{j \in S}$, i.e. from the k-th row of the transition matrix $P(t_0; t_1)$, we generate the value s_{t_1}, which represents a value of the random variable X_{t_1}.
3. If $t_c < t_h$ and we have generated the value of the random variable X_{t_c}, then from the distribution. $\{p_{s_{t_c},j}(t_c; t_{c+1})\}_{j \in S}$, i.e., from the row corresponding to the state s_{t_c} in the transition matrix $P(t_c; t_{c+1})$, we generate the value $s_{t_{c+1}}$, which represents the value of the random variable $X_{t_{c+1}}$.

If $t = t_h$, we stop the generation. The result will be the realisation of a set of h states $s_{t_0}, s_{t_1}, \dots , s_{t_h}$, which we get after h time steps. By repeating this algorithm n times, we get n trajectories of the Markov non-homogeneous chain in discrete time (Janková et al., 2014).

1.3 Accuracy of Monte Carlo estimation of the probability of an event occurring

To estimate the probability of occurrence of an event we use the law of large numbers, or Bernoulli's theorem, according to which as the number n of repeated independent trials increases, the relative frequency of occurrence of the observed event f_n approaches the theoretical probability p of occurrence of this event in each trial, which we can express as:

$$\lim_{n \to \infty} P(|f_n - p\,| < \varepsilon) = 1, \varepsilon > 0. \tag{8}$$

Thus, the number of occurrences of the observed event in a series of n independent simulation steps follows a binomial distribution $Y_n \sim B_i(n; p)$ with characteristics $E(Y_n) = n \cdot p$ and $D(Y_n) = n \cdot p \cdot q$. Using the Moivre – Laplace theorem we get:

$$P\left(\left|\frac{Y_n}{n} - p\right| < \varepsilon\right) \approx 2 \cdot \Phi\left(\varepsilon \cdot \sqrt{\frac{n}{p \cdot q}}\right) - 1, \tag{9}$$

and hence we can determine with probability $(1 - \alpha)$ the accuracy of the theoretical probability estimate p using the relative frequency $f_n = \frac{Y_n}{n}$ (Mucha and Páleš, 2018) by means of the confidence interval $(p - \varepsilon; p + \varepsilon)$, for which:

$$\frac{Y_n}{n} \in \left(p - u_{1-\frac{\alpha}{2}} \cdot \sigma; p + u_{1-\frac{\alpha}{2}} \cdot \sigma\right), \text{ where } \sigma = \sqrt{D\left(\frac{Y_n}{n}\right)} = \sqrt{\frac{p \cdot q}{n}} \,. \tag{10}$$

The accuracy, or error, ε therefore depends on the chosen level of confidence (1 – α) and from the standard deviation, the value of which can be bounded by the expression (Horáková and Mucha, 2002).

$$\sigma = \sqrt{\frac{p \cdot q}{n}} \leq \frac{1}{2} \cdot \sqrt{\frac{1}{n}} \ . \tag{11}$$

We can thus estimate more generally the maximum deviations of the simulated values $f_n = \frac{Y_n}{n}$ from the theoretical probability p for a given number of simulations from the equation:

$$\varepsilon = u_{1-\frac{\alpha}{2}} \cdot \frac{1}{2} \cdot \sqrt{\frac{1}{n}} \ . \tag{12}$$

Table 1 gives the calculated maximum errors ε for probability 1 – α = 0.9 and for different numbers of simulations n.

Table 1 Accuracy of the probability estimate *p* for a given number of simulations *n* with confidence 1 – *α* = 0.9

n	$\varepsilon_{0.9}$
1 000	0.0260
10 000	0.0082
100 000	0.0026

Source: Own construction

It should be noted that if the theoretical probability is close to p = 0.5, for a given number of simulations, the estimation error would be close to the values given in Table 1. Therefore, to obtain more accurate results, it is advisable to perform the order of tens or hundreds of thousands of simulations when estimating the probability of an event using relative frequency.

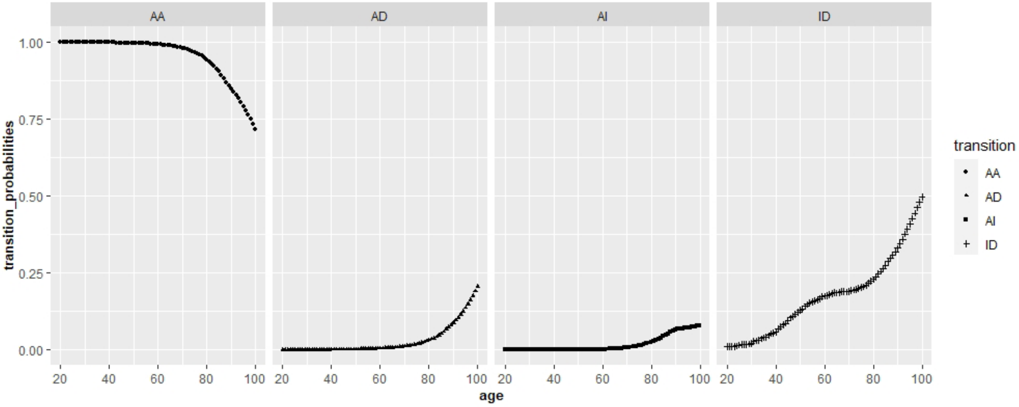
2 DATA DESCRIPTION AND MODEL BUILDING

We will use a multi-state model to solve the problem and focus on a unidirectional model with three states: healthy/active A, (terminally) ill I and dead D. From the healthy state it is possible to transition to the ill state and to the dead state. After leaving the healthy state, it is not possible to return to it again. From the ill state it is only possible to transition to the absorbing dead state. We can use this to model the situation of an illness for which there is no cure. This is also called the permanent disability model (Škrovánková and Simonka, 2021).

2.1 Data description

The dataset that we use to present the possibilities of using discrete-time simulation of non-homogeneous Markov chains in long-term care insurance was obtained in a text file from the *Markovchain* package, available in R. These data, in the form of transition probabilities between states depending on the age of the insured, were obtained from *Assicurazioni Sulla Salute*: *Caratteristiche, modelli attuariali e basi tecniche* by Paolo de Angelis and Luigi di Falco (2016). The data presented refer to the male population in Italy, whereby the status of ill I is considered, according to the author of the mentioned package (Spedicato, 2017), as a disability leading to the insured life's incapacity to work, for example Alzheimer's disease. We display graphically the obtained transition probabilities in Figure 1 using the *ggplot2* package in the R language environment (Wickham, 2016). This allows us to visually analyse the individual transition probabilities depending on the age of the insured. We have plotted their values for the age interval from 20 to 100 years.

Figure 1 Transition probabilities $p_{i,j}(t, t+1)$, $i,j \in \{A, I, D\}$ by age t of males in Italy



Source: Own construction, customized in R

2.2 Model building

We will consider the model as a system of generated trajectories of non-homogeneous Markov chains in discrete time, from which we obtain the desired results based on their statistical processing. Since the algorithm for simulating Markov chains uses transition matrices, we created them in the context of the rules of the given three-state model using the *Markovchain* package (Spedicato et al., 2017). We display the transition probability matrix in general for age t of the insured life:

$$P(t; t+1) = \begin{pmatrix} p_{A;A}(t; t+1) & p_{A;I}(t; t+1) & p_{A;D}(t; t+1) \\ 0 & p_{I;I}(t; t+1) & p_{I;D}(t; t+1) \\ 0 & 0 & 1 \end{pmatrix}. \tag{13}$$

In this way, we have modified the original data into the format of individual transition matrices $P(t; t+1)$, which we will use to create the final model for solving the presented problem. By simulating non-homogeneous Markov chains, we will create a model through the generation of their trajectories, which will mimic the real evolution of the states of the insured during h time steps. The results can be written into n rows and h columns of the matrix ${}^{(z)}M = [m_{ij}]_{n \times h}$, $z \in \{A, I\}$, where z represents the initial status of the insured life. For practical reasons, we will consider only the initial states healthy and ill. The elements of this matrix will be of interest to us in the context of carrying out analyses in the area of long-term care insurance.

Based on the statistical processing of a sufficient amount of n generated data in the h-th column of the matrix ${}^{(z)}M$ it is possible to estimate the percentage distribution of insured lives in each state $g \in \{A, I, D\}$ after h time steps according to the equation:

$$perc_g^{(z)}(h) = p_g^{(z)}(h) \cdot 100 \approx \frac{\sum_{i=1}^{n} I[m_{ih}=g]}{n} \cdot 100,\ g \in \{A, I, D\},\ z \in \{A, I\}, \tag{14}$$

where $p_g^{(z)}(h)$ represents a particular component of the absolute probability vector, which we estimate from the generated values in the h-th column of the matrix ${}^{(z)}M$.

In the presented model, we consider a portfolio composed of K insured lives, where we denote the initial number of insured lives in the healthy state by K_A and the number of insured lives in the ill state by K_I, thus:

$$K = K_A + K_I \,. \tag{15}$$

The absolute distribution of the number of insured in each state after h time steps can be written in the form of a vector **k**, which is a linear combination of absolute probability vectors $\mathbf{p}^{(A)}(h)$ and $\mathbf{p}^{(I)}(h)$, which we write as:

$$\mathbf{k} = K_A \cdot \mathbf{p}^{(A)}(h) + K_I \cdot \mathbf{p}^{(I)}(h) \,. \tag{16}$$

By substituting the mentioned vectors into the equation for vector **k**, we get the following expression in the considered three-state model:

$$\mathbf{k} = (k_1; k_2; k_3) = (K_A \cdot p_A{}^{(A)}(h); K_A \cdot p_I{}^{(A)}(h) + K_I \cdot p_I{}^{(I)}(h); K_A \cdot p_D{}^{(A)}(h) + K_I \cdot p_D{}^{(I)}(h)), \tag{17}$$

where k_1 represents the number of insured lives in the healthy state, k_2 represents the number of insured lives in the ill state and k_3 the number of insured lives in the dead state in the considered portfolio after h time steps. The above distribution of the number of policyholders into the different states makes sense, given Bernoulli's law of large numbers, if the numbers of policyholders K_A and K_I are large enough, i.e., in the order of tens of thousands or hundreds of thousands. We estimate the individual probabilities $p_A{}^{(A)}(h)$, $p_I{}^{(A)}(h)$, $p_I{}^{(I)}(h)$, $p_D{}^{(A)}(h)$, $p_D{}^{(I)}(h)$ using the relative frequencies from the generated matrices ${}^{(z)}M = [m_{ij}]_{n \times h}$, $z \in \{A, I\}$.

By generating the trajectories of non-homogeneous Markov chains, it is also possible to determine the percentage distribution of the number of insured lives K into the different states after h time steps, which we write using the vector:

$$\mathbf{perc} = (perc_1; perc_2; perc_3), \tag{18}$$

where $perc_1$ represents the percentage of insured lives in the healthy state, $perc_2$ represents the percentage of insured lives in the ill state and $perc_3$ the number of insured lives in the dead state in the considered portfolio after h time steps.

To determine the individual components $perc_i$, $i = 1, 2, 3$ in the considered three-state model we used a weighted arithmetic average with weights $w_1 = K_A$, $w_2 = K_I$, whereby:

$$perc_1 = \frac{K_A}{K} \cdot perc_A{}^{(A)}(h) + \frac{K_I}{K} \cdot perc_A{}^{(I)}(h) = \frac{K_A}{K} \cdot perc_A{}^{(A)}(h) \,, \tag{19}$$

$$perc_2 = \frac{K_A}{K} \cdot perc_I{}^{(A)}(h) + \frac{K_I}{K} \cdot perc_I{}^{(I)}(h) \,, \tag{20}$$

$$perc_3 = \frac{K_A}{K} \cdot perc_D{}^{(A)}(h) + \frac{K_I}{K} \cdot perc_D{}^{(I)}(h) \,, \tag{21}$$

If we do not consider a specific portfolio of insured lives, but the population in general, the above condition of a sufficiently large number of K_A and K_I is automatically satisfied and the predicted absolute and percentage distributions can be considered relevant without verification.

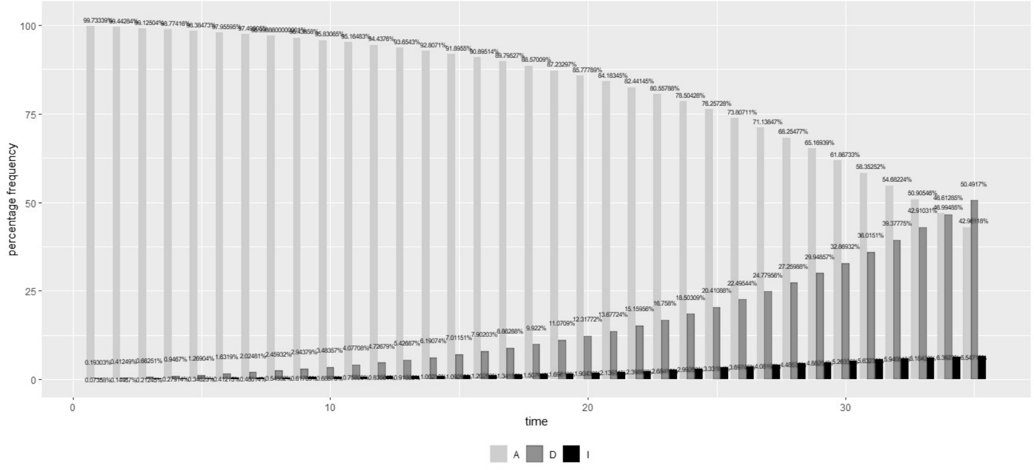
3 RESULTS AND DISCUSSION

In this part of the paper, we will use the described data set to model and analyse the development for a particular critical illness (for example Alzheimer's disease), which requires long-term care in the event of its occurrence. We will use the simulation of the trajectories of non-homogeneous Markov chains, which we will implement using the R language. Using the statistical data in the generated matrices ${}^{(z)}M = [m_{ij}]_{n\times h}$, $z \in \{A, I\}$, the insurance company can obtain information necessary for the provision of health care and for premium calculation.

3.1 Estimation of the distribution of insured lives in the separate states

Due to the nature of the database, we will focus on predicting the percentage distribution of the number of insured lives in the separate states on a yearly basis for a certain number of years. If the data were recorded differently, for example monthly, we could use that as our time interval for modelling purposes. First, we will show the evolution of the percentage distribution of the number of initially healthy insured lives aged 50, which we illustrate graphically in Figure 2 using the R language package *ggplot2* (Wickham, 2016).

Figure 2 Percentage distribution of initially healthy lives aged 50 in the different states A, I, D over time



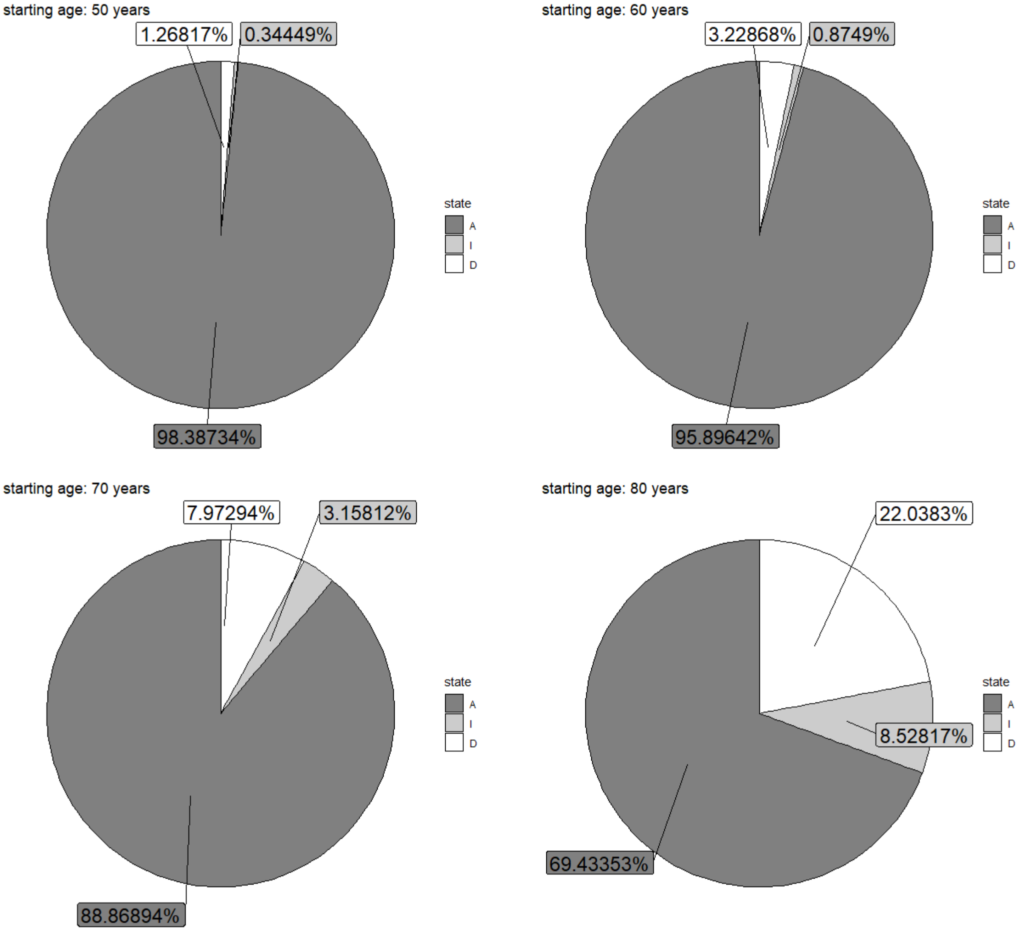
Source: Own construction, customized in R

The graphical presentation above shows the trends in the percentages of the separate states over time. To achieve more accurate results, we have carried out n = 100 000 simulations, whereby we repeated this scenario using the R language 100 times and calculated the results of the percentage distribution in each year as the arithmetic average.

From the presented results we see that, for example, after 20 years 85.77789% of the initially healthy insured lives aged 50 will still be in the healthy state, 1.90439% will be in the ill state and 12.31772% will be in the dead state. Of course, the above statement is in general only true if the initial portfolio of healthy males aged 50 was sufficiently large, i.e., in terms of the law of large numbers, it consists of the order of a few tens of thousands or hundreds of thousands of lives. In the case of the male population in Italy, this condition is of course met.

Figure 3 shows graphically the percentage distribution after 5 years for males initially aged 50, 60, 70 and 80 who were initially in the healthy state. In the case of males aged 70 and 80 who were initially healthy, there is a significant increase in the number in the ill state (illness requiring long-term care) after 5 years, compared to males aged 50 and 60. Given the nature of the illness (for example Alzheimer's disease), this increase is to be expected.

Figure 3 Distribution of the number of insured lives after 5 years as a percentage for males initially aged 50, 60,70 and 80 who started in the healthy state after running 100 000 simulations



Source: Own construction, customized in R

So far, we have been modelling assuming that the insured lives were in the healthy state at the start. We will now model the evolution of the number of insured lives for a specific portfolio that is composed of K_A in the healthy state and K_I in the ill state. Based on the simulation trajectories, we present in Tables 2 and 3 the distribution of the absolute and percentage number of insured lives initially aged 50 during a period of 10 years, where $K_A = 300\ 000$ and $K_I = 20\ 000$.

When expressing the number of insured lives as a percentage, based on the data generated in the matrices ${}^{(z)}M = [m_{ij}]_{n\times h}$, $z \in \{A, I\}$, it is not necessary to specify the absolute number of policyholders K_A and K_I. It is enough to enter the relative or percentage frequency of the considered states $\{A, I\}$ at the start of modelling.

Table 2 Distribution of the number of insured lives during 10 years for males aged 50, of which at the beginning $K_A = 300\ 000$ were in the healthy state and $K_I = 20\ 000$ in the ill state

	State			
Time	**healthy**	**ill**	**dead**	**Total**
1	299 211	17 649	3 140	320 000
2	298 339	15 551	6 110	320 000
3	297 386	13 628	8 986	320 000
4	296 329	11 900	11 771	320 000
5	295 166	10 407	14 427	320 000
6	293 895	9 140	16 965	320 000
7	292 519	8 072	19 409	320 000
8	291 014	7 185	21 801	320 000
9	289 364	6 460	24 176	320 000
10	287 541	5 871	26 588	320 000

Source: Own construction

Table 3 Percentage distribution of the number of insured lives during 10 years for males aged 50 of which $K_A = 300\ 000$ were initially in the healthy state and $K_I = 20\ 000$ in the ill state

	State			
Time	**healthy**	**ill**	**dead**	**Total**
1	93.503%	5.515%	0.982%	100%
2	93.231%	4.860%	1.909%	100%
3	92.933%	4.259%	2.808%	100%
4	92.603%	3.719%	3.678%	100%
5	92.239%	3.252%	4.509%	100%
6	91.842%	2.856%	5.302%	100%
7	91.412%	2.523%	6.065%	100%
8	90.942%	2.245%	6.813%	100%
9	90.426%	2.019%	7.555%	100%
10	89.857%	1.834%	8.309%	100%

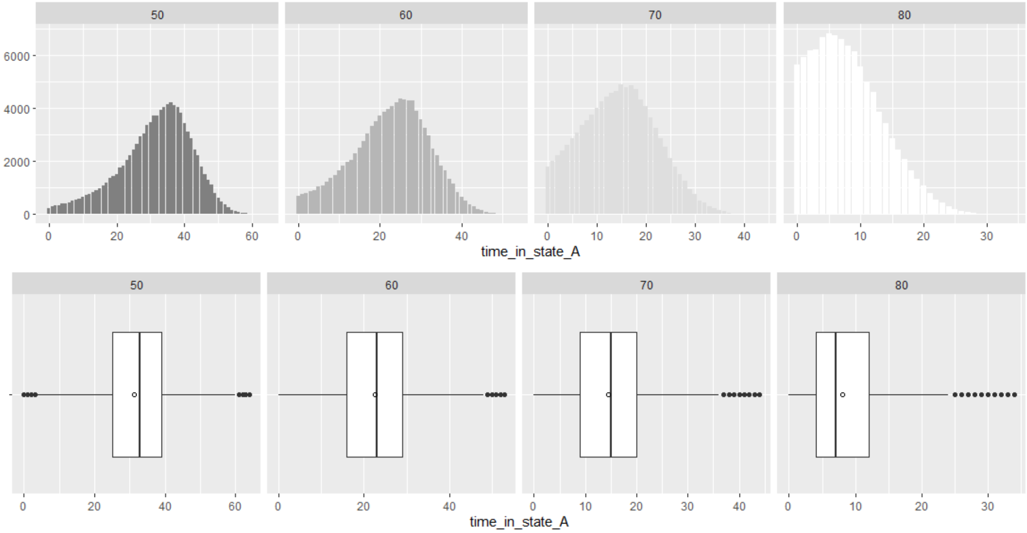
Source: Own construction

By using the simulation of non-homogeneous Markov chains in discrete time, it is possible to estimate the evolution of the representation in each of the separate states during the modelled years. By comparing the values obtained from the simulations with the values obtained from the absolute probability vectors, we can conclude that the presented simulation model provides relevant results for the described number of simulations. We will therefore use it to further model and obtain information that is relevant for the insurance of critical illnesses that require long-term care.

3.2 Estimation of time remaining healthy and remaining ill

In this part of the paper, we will analyse the estimation of the time during which the insured life remains in the healthy and ill state, respectively. We assume that the insured lives are healthy at ages 50, 60, 70, and 80 at the beginning of the modelling period. Using 1 000 simulations of the trajectories of non-homogeneous Markov chains in the matrix ${}^{(A)}M = [m_{ij}]_{n\times h}$ we recorded data on the number of years the insured life remained in the healthy state. We will use the full range of available transition probability matrices and model the states up to age 120. We present these data for each age category in the form of a bar plot and box plot in Figure 4. The circle in the box plot denotes the estimated mean value of the number of years the insured life remained in the healthy state and the line in the box denotes the median value.

Figure 4 Analysis of the number of years the insured life remained healthy for the initial ages 50, 60, 70 and 80 using a bar plot and a box plot



Source: Own construction, customized in R

For the sake of illustration, we list the selected values in Table 4.

Table 4 Estimated values for the number of years the insured life has been in the healthy state

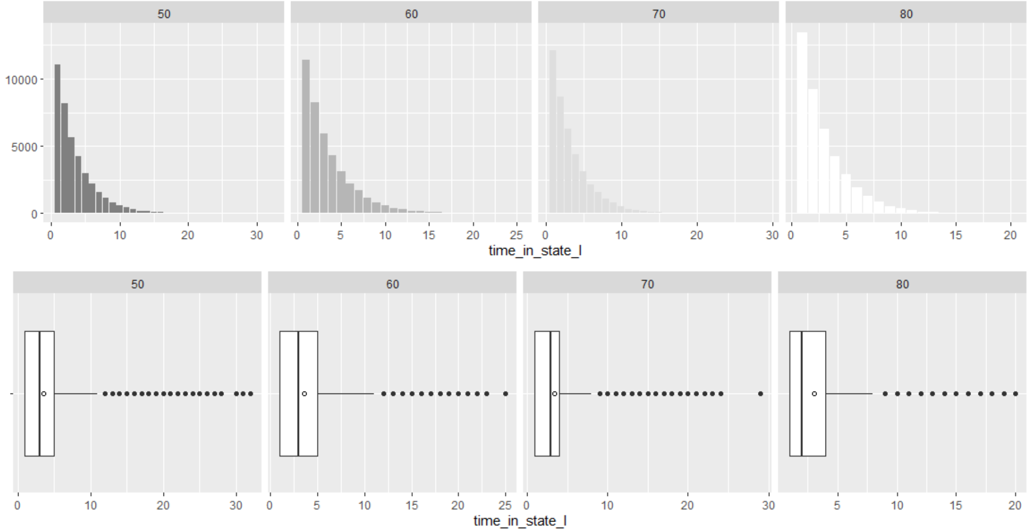
Age	$x_{0.25}$	Median	$x_{0.75}$	Mean
50	25	33	39	31.39801
60	16	23	29	22.53211
70	9	15	20	14.54286
80	4	7	12	8.10224

Source: Own construction

Thus, for example, in the case of healthy insured lives aged 60, 75% of them have the number of years they remain healthy less than or equal to 29, and 25% of them have the number of years they remain healthy greater than 29 years. The average number of years for a healthy insured life aged 50 is equal to 31.39801 years and for a 70 year old is equal to 14.54286 years. Given that the estimated mean value is an arithmetic mean, it is generally necessary to consider the dispersion of the values on the left and right sides of the mean. This may ultimately affect the relevance of the information thus obtained, despite a sufficiently large set of values. In this case, the median can be used for estimation.

Another important element for insurance calculations is the time during which the insured life remains ill. Again, we assume that the lives are healthy at ages 50, 60, 70 and 80 years at the start of the modelling period. Using 100 000 simulations of non-homogeneous Markov chains, the matrix ${}^{(A)}M = [m_{ij}]_{n \times h}$ records the data on the number of years the insured life remained ill. We present these data for each age category in the form of a bar plot and a box plot in Figure 5.

Figure 5 Analysis of the number of years during which the insured life remained ill for ages 50, 60, 70 and 80 using a bar plot and a box plot



Source: Own construction, customized in R

For the sake of illustration, we will list the selected values in Table 5.

Table 5 Estimated values for the number of years during which the insured life was in the ill state

Age	$x_{0.25}$	Median	$x_{0.75}$	Mean
50	1	3	5	3.599501
60	1	3	5	3.532215
70	1	3	4	3.365925
80	1	2	4	3.002761

Source: Own construction

Out of the 100 000 simulations in 59 754 cases an insured life aged 50 subsequently died whilst remaining in the healthy state. This means that, from the data available to us, he was not registered in the three-state model described above as an insured life in need of intensive long-term care because of illness. In the remaining 40 246 cases represented by the trajectory of the considered Markov chain we found that if an insured life entered the ill state, he stayed in this state 3.599501 years on average before moving to the dead state. This compares with a value of 3.365925 years for the 70 year old insured life as shown in Table 5.

3.3 Calculation of long-term care insurance premiums

Finally, we will deal with the determination of the single premium P, which a life aged x has to pay in order to receive an annual payment of C while in a state of non-self-sufficiency. We assume, of course, that the life is in the healthy state at the start of the policy. We use the generated trajectories of the non-homogeneous Markov chains, which we have written into the matrix ${}^{(A)}M = [m_{ij}]_{n\times h}$, to determine the above insurance premium, where $n = 100\ 000$ and $h = 120 - x$. For the purpose of determining the premium, we transform all elements of this matrix indicating the ill state I to the amount C and its other elements to zero values. We denote the resulting matrix by $M^C = [m^C_{ij}]_{n\times h}$. The single premium P is then determined using the equation:

$$P = M^C \cdot U\,, \tag{22}$$

where for the matrix elements $U = [u_{ij}]_{h\times 1}$ it holds that $u_{ij} = (1+u)^{-i}$, where u is the annual rate of interest.

The individual elements of the matrix $P = [p_{ij}]_{n\times 1}$ can be interpreted as representing the given premium determined for a particular modelled scenario of the insured life represented by the corresponding trajectory of the non-homogeneous Markov chain. The arithmetic average was then used to calculate the single premium P, to be paid by the life aged x. For more accurate results, we repeated this scenario using R 1 000 times and for the premium P we again used the arithmetic average as the estimated value. For example, a healthy life aged 50 would have to pay a premium of P = €12 583.42 at the interest rate used of $u = 0.01$, in order to receive an annual payment C = €12 000 at the beginning of each year if he falls ill. For comparison, we have also calculated the premium using a standard life insurance formula

$$P = \sum_{t=1}^{\omega-x} {}_{t-1}p_x^{AA} \cdot q^{AI}_{x+t-1} \cdot v^t \cdot \pi(\ddot{a}^{(I)}_{x+t}),\ t = 1,\ 2,\ \ldots,\ \omega\,, \tag{23}$$

ω – the highest age in the relevant mortality table,
${}_{t-1}p_x^{AA}$ – the probability that a life x years old remains healthy for $t - 1$ years,
q^{AI}_{x+t-1} – the probability that a life aged $x + t - 1$ years in the healthy state becomes ill within one year, i.e., at age $x + t$ is in the ill state,
v – is the discount factor, i.e. $v = \dfrac{1}{1+u}$, where u is the annual interest rate,
$\pi(\ddot{a}^{(I)}_{x+t})$ – the whole life annuity-due for a life aged $x + t$ years, if he is then in the ill state, for an annual payment of C payable in advance, i.e.

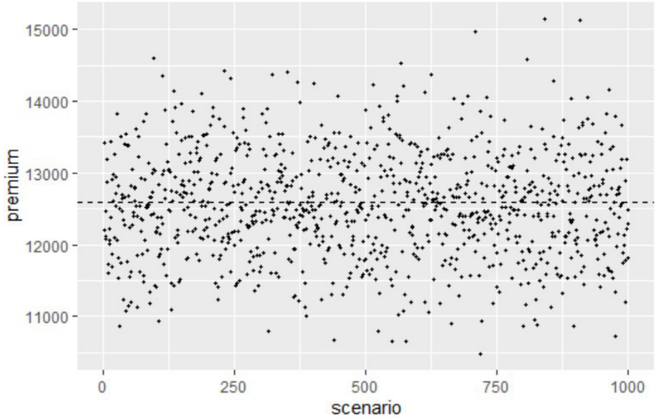
$$\pi(\ddot{a}^{(I)}_{x+t}) = \sum_{k=1}^{\omega-(x+t)} C \cdot {}_{k-1}p^{II}_{x+t} \cdot v^{k-1}, \tag{24}$$

where ${}_{k-1}p^{II}_{x+t}$ is the probability a life in the ill state at age $x + t$ remains in that state for a further $k - 1$ years (Dickson, Hardy and Waters, 2013).

Using this formula we calculated the value of the single premium for our male life aged 50 as P = €12 584.37, which is comparable to the amount obtained by using simulations. However, the ability to determine the premium based on the generation of the trajectory of the non-homogenous Markov chain, represents a more flexible and efficient approach.

We now review the importance of creating multiple scenarios and a sufficient number of simulations in the situation described in order to obtain relevant results for the insurance premium estimation. If we were to implement only one scenario in the form of $n = 1\ 000$ simulations, a sufficient accuracy of the results might not be achieved. We have therefore determined the premium as the average value of the 1 000 created premium scenarios whose variability can be seen in Figure 6.

Figure 6 Premium modelling based on the creation of 1 000 scenarios for 1 000 simulations of a non-homogenous Markov chain

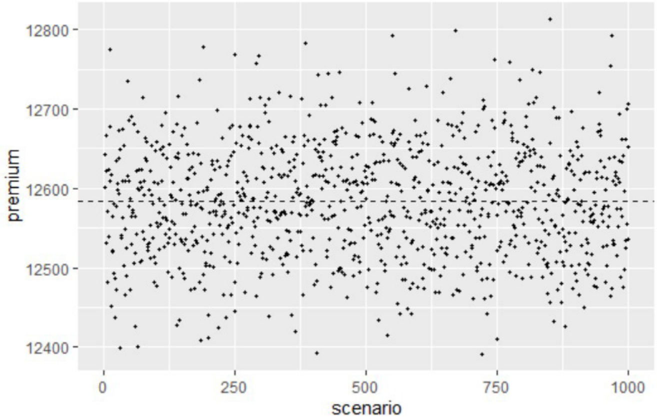


Source: Own construction, customized in R

For comparison, we have shown as a dashed line the value of the premium P = €12 584.37 as determined by the standard equation. The average premium calculated from the presented 1 000 values, each of which was itself calculated as the average from the 1 000 simulated trajectories of the insured life, is €12 552.58. If we choose P = €12 584.37 as a comparative premium value, then with a number of simulations n = 1 000 from the number of 1 000 created scenarios only 508 values of the premium are located in the interval (P – 500, P + 500). So, in 492 cases, the premium differed from the comparative value by more than €500.

If the number of simulations is increased to n = 100 000 the average premium is P = €12 583.42 and there is significantly less variability in the premiums as can be seen in Figure 7. For this number of simulations all 1 000 estimated premium values are in the interval (P – 500, P + 500).

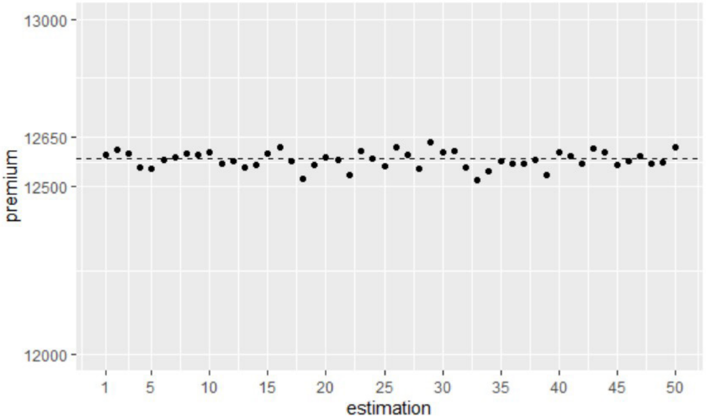
Figure 7 Modelling of the premiums based on 1 000 scenarios for 100 000 non-homogeneous Markov chain simulations



Source: Own construction, customized in R

Therefore, it is important for n = 1 000 simulations to estimate premiums as an average from values obtained from a sufficient number of created scenarios. Of course increasing the number of simulations will ensure more accurate results and we recommend implementing, for example 100 000 simulations. However, one needs to note that when creating 1 000 scenarios with n = 100 000 simulations the calculations using R were time-consuming. On the other hand they provide a sufficiently accurate result. If we were to carry out only one scenario with n = 1 000 simulations we could get an inaccurate estimate of the premium. The solution is to create enough scenarios for the given number of simulations. The result obtained in this way can then be considered as sufficiently accurate. For illustration, in Figure 8 we show 50 possible premium estimates for an alternative 1 000 scenarios with 1 000 simulations in comparison with the value P = €12 584.37.

Figure 8 50 premium estimates for 1 000 scenarios for 1 000 non-homogeneous Markov chain simulations



Source: Own construction, customized in R

It can be noted that the values presented in Figure 8 are comparable to the benchmark premium shown by the dashed line.

CONCLUSION

The use of Markov chains in the context of multi-state models is a frequently used tool for modelling the evolution of conditions in relation to disease incidence and long-term healthcare delivery. By simulating non-homogeneous Markov chains through the generation of their trajectories, we created a model that mimics the real evolution of insured lives' states over time. In the context of the Monte Carlo method, we also discussed in the paper the impact of the number of simulations on the accuracy of the obtained results. Due to the nature of the data in the context of recording a given disease, we performed our calculations in discrete time on an annual basis. The data presented here refer to the male part of the Italian population, where by the ill I(*ill*) state, according to the author of the markovchain package (Giorio Alfredo, Spedicato), we mean disability in the sense of the so-called non-self-sufficiency of the insured life, i.e., disability similar to that of Alzheimer's disease. Using the above modelling implemented using the R language, we have presented the absolute and percentage distribution of insured lives into different states over several years, based on the statistical processing of the generated data, and we have also described it by means of graphical and vector representations. We addressed the analysis of the illness state for the four selected ages, following the trend of its evolution over time. The results obtained could be used to estimate the costs of a health care institution. Another aspect of the use of the simulation model developed was the estimation of the average number of years that an insured life remains in the healthy state and the estimation of the time during which he/she remains in the ill state. This analysis was also carried out for selected ages, and the situation was presented graphically using bar plots and box plots. The information obtained may be important not only in the context of health care costing, but also in analyses for long-term care insurance contracts. In the last part of the paper we have dealt with the calculation of the premiums for such contracts, presenting in the context of simulations an approach that provides results at a sufficient level of accuracy. We pointed out that insufficient simulations in the premium calculation can provide inaccurate results. This shortcoming can be remedied by creating a sufficient number of scenarios and averaging the premium values we obtained from each scenario. The above analysis was also supported by a graphical presentation of the results of the individual simulation scenarios. The premium values obtained from the simulations were compared with those calculated using a standard life insurance formula and it can be concluded that they are comparable. However, the advantage of the simulation approach lies in greater computational flexibility and the possibility of interactive response when the parameters entering the premium calculation are changed. Modelling the evolution of states over time in the presented domain using Markov chain simulations represents a suitable and efficient solution tool.

ACKNOWLEDGMENT

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA No 1/0431/22 *Implementation of innovative approaches of modeling and managing risks in internal models of insurance companies in accordance with the Solvency II.*

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA No 1/0410/22 *Analysis of insurance risks in relation to the economy of the life insurance company.*

References

BERCHTOLD, A. (2001). Markov Chain Computation for Homogeneous and Non-homogeneous Data: MARCH 1.1 Users Guide [online]. *Journal of Statistical Software*, 6(3): 1–81. <https://doi.org/10.18637/jss.v006.i03>.

DE ANGELIS, P., DI FALCO, L. (2016). *Assicurazioni sulla salute:caratteristiche, modelli attuariali e basi tecniche.* Il Mulino.

DICKSON, D. C. M., HARDY, M. R., WATERS, H. R. (2013). *Actuarial Mathematics for Life Contingent Risk.* New York: Cambridge University Press.

DIZ, E., QUERY, J. T. (2012). Applying a Markov model to a plan of social health provisions. *Insurance Markets and Companies*, 3(2): 27–34.

DOBROW, R. (2016). *Introduction to Stochastic Processes with R.* John Wiley & Sons.

DUDEL, C., MYRSKYLÄ, M. (2020). Estimating the number and length of episodes in disability using a Markov chain approach [online]. *Popul Health Metr.*, 18(1): 15. <https://doi.org/10.1186/s12963-020-00217-0>.

ESQUÍVEL, L. M., GUERREIRO, R. G., OLIVEIRA, C. M., REAL, C. P. (2021). Calibration of Transition Intensities for a Multistate Model: Application to Long-Term Care. Risks [online]. *MDPI*, 9(2): 1–17, <https://doi.org/10.3390/math9131496>.

FECENKO, J. (2018). *Teória pravdepodobnosti II v MAXIME.* Bratislava: Letra Edu.

FERNANDEZ-MORALES, A. (2015). Application of a Discrete-time Markov Chain Simulation in Insurance [online]. *International Journal of Recent Contributions from Engineering, Science´& IT (iJES)*, 3(3): 27–32. <https://doi.org/10.3991/ijes.v3i3.4929>.

FLEISCHMANN, A., HIRZ, J., SIRIANNI, D. (2021). A long-term care multi-state Markov model revisited: a Markov chain Monte Carlo approach [online]. *European Actuarial Journal.* <https://doi.org/10.1007/s13385-021-00285-y>.

GARG, L., MCCLEAN, S. et al. (2010). A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare systém [online]. *Health Care Manag Sci*, 13: 155–169. <https://doi.org/10.1007/s10729-009-9120-0>.

GEYER, C. J., JOHNSON, L. T. (2013). *mcmc:Markov Chain Monte Carlo* [online]. <http://CRAN.R-project.org/package=mcmc>.

HORÁKOVÁ, G., MUCHA, V. (2002). Určenie rozdelenia celkových škôd s využitím metódy Monte Carlo a jeho porovnanie s numericky presným výpočtom v danom portfóliu poistných zmlúv. *Managing and Modelling of Financial Risks*, VŠB – Technical University of Ostrava, 77–81.

JANKOVÁ, K., KILIANOVÁ, S., BRUNOVSKÝ, P., BOKES, P. (2014). *Markovove reťazce a ich aplikácie.* Bratislava: Epos.

JONES, W. P., SMITH, P. (2018). *Stochastic Processes. An Introduction.* Taylor & Francis Group.

LI, M., DUSHOFF, J., BOLKER, B. M. (2018). Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches [online]. *Statistical Methods in Medical Research*, 27(7): 1956–1967. <https://doi:10.1177/0962280217747054>.

MACIAK, M., OKHRIN, O., PEŠTA, M. (2021). Infinitely stochastic micro reserving [online]. *Insurance: Mathematics and Economics*, 100: 30–58. <https://doi.org/ 10.1016/j.insmatheco.2021.04.007>.

MUCHA, V., PÁLEŠ, M. (2018). *Teória pravdepodobnosti pre ekonómov. S podporou jazyka R.* Bratislava: Letra Edu.

NICHOLSON, W. (2013). *DTMCPack: Suite of functions related to discrete-time discrete-state Markov Chains* [online]. <https://CRAN.R-project.org/package=DTMCPack>.

PASARIBU, S. U., HUSNIAH, H., SARI, N. K. R., YANTI, R. (2019). Pricing Critical Illness Insurance Premiums Using Multiple State Continous Markov Chain Model [online]. *Journal of Physics*, 1366. <https://doi:10.1088/1742-6596/1366/1/012112>.

SATO, R., ZOUAIN, D. (2010). Markov Models in health care [online]. *Einstein*, 8(3): 376–379. <https://10.1590/S1679-45082010RB1567>.

SPEDICATO, A. G. (2017). Discrete Time Markov Chains with R [online]. *The R Journal*, 9(2): 84–104. <https://doi.org/10.32614/RJ-2017-036>.

SPEDICATO, A. G. et al. (2017). *The markovchain Package: a Package for Easily Handling Discrete Markov Chains in R* [online]. <https://cran.rproject.org/web/packages/markovchain/vignettes/an_introduction_to_markovchain_package.pdf>.

UNWIN, H. J. T., ROUTLEDGE, I., FLAXMAN, S., RIZOIU, M.-A., LAI, S., COHEN, J. et al. (2021). Using Hawkes Processes to model imported and local malaria cases in near-elimination settings [online]. *PLoS Comput Biol*, 17(4). <https://doi.org/10.1371/journal.pcbi.1008830>.

ŠKROVÁNKOVÁ, L., SIMONKA, Z. (2021). *Aktuárske metódy a modely v penzijnom, zdravotnom a nemocenskom poistení.* Brno: H.R.G. s.r.o.

TWUMASI, C., ASIEDU, L., NORTEY, E. (2019). Markov Chain Modeling of HIV, Tuberculosis, and Hepatitis B Transmission in Ghana [online]. *Interdisciplinary Perspectives on Infectious Diseases.* <https://doi.org/10.1155/2019/9362492>.

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 [online]. *Journal of the Royal Statistical Society: Series A*, 168(1): 51–61. <https://doi.org/10.1111/j.1467-985X.2004.00335.x>.

WICKHAM, H. (2016). *Ggplot2, Elegant Graphics for Data Analysis.* New York: Springer-Verlag.