New Method Utilizing Continuous Time Markov Chains to Analyze Evolution in the Nine States Model of Non-Alcoholic Fatty Liver Disease

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

09-11-2021 / 12-11-2021

CITATION

Mohammed Attia Abd-Elkhalik Abo Elreesh, Iman (2021): New Method Utilizing Continuous Time Markov Chains to Analyze Evolution in the Nine States Model of Non-Alcoholic Fatty Liver Disease. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.16967485.v1>

DOI

[10.36227/techrxiv.16967485.v1](https://doi.org/10.36227/techrxiv.16967485.v1)

New Method Utilizing Continuous Time Markov Chains to Analyze Evolution in the Nine States Model of Non-Alcoholic Fatty Liver Disease

Iman M. Attia *

Imanattiathesis1972@gmail.com, imanattia1972@gmail.com

**Department of Mathematical Statistics , Faculty of Graduate Studies for Statistical Research , Cairo University , Egypt*

Abstract

Global increase in prevalence of obesity and type 2 Diabetes is strongly connected to increase prevalence of non-alcoholic fatty liver disease (NAFLD) worldwide. In the present paper, progression of (NAFLD) process is modeled by Continuous time Markov chains (CTMC) with 9 states. Maximum likelihood is used to estimate the transition intensities among the states. Once the transition intensities are obtained the mean sojourn time and its variance are estimated as well as the state probability distribution and its asymptotic covariance matrix. A hypothetical example based on longitudinal study assessing patients with NAFLD in various stages is discussed. The mean time to absorption is estimated as well as the other above mentioned statistical indices are examined. In this paper MLE is utilized in a new approach to compensate for the missing values in the follow up period of patients evaluated in longitudinal studies. A matlab code is provided in the appendix for estimation of the rate transition matrix and probability transition matrix

Index Terms: Continuous time Markov chains, Life expectancy, Maximum Likelihood estimation, Mean Sojourn Time, Non-Alcoholic Fatty Liver Disease, Panel Data.

Introduction

CTMC is commonly used to model data obtained from longitudinal studies in medical research and to investigate the evolution and progression of the diseases over time. Estes et al. [1] used multistate Markov chains to model the epidemic of nonalcoholic fatty liver disease. Younossi et al. [2] used the multistate Markov chains to demonstrate the

economic and clinical burden of nonalcoholic fatty liver disease in United States and Europe.

According to American Association for Study of Liver Disease , American College of Gastroenterology, and the American Gastroenterological Association, NAFLD to be defined requires (a) there is evidence of hepatic steatosis (HS) either by imaging or by histology and (b)there are no causes for secondary hepatic fat accumulation such as significant alcohol consumption, use of steatogenic medications or hereditary disorders [3].This is the same definition established by European Association for the Study of the Liver (EASL),European Association for the Study of Diabetes (EASD)and European Association for the Study of Obesity(EASO)[4]. NAFLD can be categorized histologically into nonalcoholic fatty liver (NAFL) or nonalcoholic steato-hepatitis (NASH). NALF is defined as the presence of $\geq 5\%$ (HS) without evidence of hepatocellular injury in the form of hepatocyte ballooning .NASH is defined as the presence of $\geq 5\%$ HS and inflammation with hepatocyte injury (ballooning), with or without any fibrosis.

NAFLD is a multistage disease process consisting of 9 stages as depicted in figure 1[2]. As shown in the figure; the patient can move across the stages of the disease process. While the remission rates are allowed from stage 4 (compensated liver cirrhosis) to the earlier stages, patient progresses to HCC and liver transplantation once he arrives to stage 5(decompensated liver cirrhosis) and remission rates are not allowed. Death state can be reached from any state. The patient can move from the first 5 stages to stage 8 (HCC) with higher rate of progression from stage 4(CC) or stage 5 (DCC) to stage 8(HCC) compared to first 3 stages. A brief definition of each stage is illustrated below the figure (1).

Figure 1: disease model structure :

NAFLD-NO FB = nonalcoholic fatty liver disease with no fibrosis (stage 1). NASH-NO FB = nonalcoholic steato-hepatitis with no fibrosis (stage 2). NASH-FB = nonalcoholic steato-hepatitis with fibrosis (stage 3). CC= compensated cirrhosis (stage 4). DCC= de-compensated cirrhosis (stage 5). LT= liver transplant(stage 6). PLT =post liver transplant (stage 7). HCC =hepato-cellular carcinoma (stage 8). EM= extra-mortality (stage 9). [2]

NAFLD stages are modeled as time homogenous CTMC , that is to mean $P_{ij}(\Delta t)$ depends on Δt and not on t , with constant transition intensities λ_{ij} over time, exponentially distributed time spent within each state and patients' events follow Poisson distribution. The states are finite and can be defined or identified based on various aspects such as clinical symptoms and invasive or noninvasive investigations. The gold standard method for classification of histopathological changes in the liver is the invasive liver biopsy. It is presently the most trustworthy procedure for diagnosing the presence of steatohepatitis (HS) and fibrosis in NAFLD patients[5]. The limitations of this procedure are cost, sampling error, and procedure-related morbidity and mortality. MR imaging, by spectroscopy[6] or by proton density fat fraction[7], is an excellent noninvasive technique for quantifying HS and is being widely used in NAFLD clinical trials[8]. The use of transient elastography (TE) to obtain continuous attenuation parameters is a promising tool for quantifying hepatic fat in an ambulatory setting[9]. However, quantifying noninvasively HS in patients with NAFLD is limited in routine clinical care. Also one of the most recent biological markers are the keratin(K18) and its caspase-cleaved fragments(cK18). There are many scoring systems that can identify the stages of the disease process [10]. NAFLD has higher prevalence rate in individuals with risk factors such as visceral obesity, type 2 diabetes mellitus (T2DM), dyslipidemia, older age , male sex and being of Hispanic ethnicity[11].

For simplicity, all individuals are assumed to enter the disease process at stage one and they are all followed up with the same length of time interval between measurements.

The paper is divided into 7 sections. In section 1 the transition probabilities and transition rates are thoroughly discussed. In section 2 mean sojourn time and its variance are reviewed. In section 3 state probability distribution and its covariance matrix are discussed. While in section 4 the life expectancy of the patients are considered. In section 5 expected numbers of patients in each state is obtain .A hypothetical numerical example is used in section 6 to illustrate the above concepts. Lastly a brief summary is comprehended in section 7. In supplementary materials, appendix A provides the Matlab code for rate matrix estimation while appendix B provides Matlab code for probability matrix estimation, utilizing Matlab program version 14.

1. Transition Rates And Probabilities

NAFLD is modeled by a multistate Markov chains which define a stochastic process

$[X(t), t \in T]$ over a finite state space $S = \{1, 2, \dots, 9\}$ and $T = [0, t]$ and $t < \infty$

The transitions can occur at any point in time and hence called continuous time Markov chains in contrast to the discrete time Markov chains in which transitions occur at fixed points in time. The rates at which these transitions occur are constant over time and thus are independent of t that is to say the transition of patient from state i at time t to state j at $t = t + s$ where $s = \Delta t$ depends on difference between two consecutive time points. And it's defined as $\theta_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{P_{ij}(\Delta t) - I}{\Delta t}$ or the Q matrix.

For the above multistate Markov model demonstrating the NAFLD disease process; the forward Kolomogrov differential equations are the following:

$$\frac{d}{dt}P_{ij}(t) = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} & P_{15} & P_{16} & P_{17} & P_{18} & P_{19} \\ P_{21} & P_{22} & P_{23} & P_{24} & P_{25} & P_{26} & P_{27} & P_{18} & P_{29} \\ P_{31} & P_{32} & P_{33} & P_{34} & P_{35} & P_{36} & P_{37} & P_{18} & P_{39} \\ P_{41} & P_{42} & P_{43} & P_{44} & P_{45} & P_{46} & P_{47} & P_{18} & P_{49} \\ 0 & 0 & 0 & 0 & P_{55} & P_{56} & P_{57} & P_{18} & P_{59} \\ 0 & 0 & 0 & 0 & 0 & P_{66} & P_{67} & 0 & P_{69} \\ 0 & 0 & 0 & 0 & 0 & 0 & P_{77} & 0 & P_{79} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_{18} & P_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_{99} \end{bmatrix} \begin{bmatrix} -\gamma_1 & \lambda_{12} & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{18} & \lambda_{19} \\ \mu_{21} & -\gamma_2 & \lambda_{23} & 0 & 0 & 0 & 0 & 0 & \lambda_{28} & \lambda_{29} \\ 0 & \mu_{32} & -\gamma_3 & \lambda_{34} & 0 & 0 & 0 & 0 & \lambda_{38} & \lambda_{39} \\ 0 & 0 & \mu_{43} & -\gamma_4 & \lambda_{45} & 0 & 0 & 0 & \lambda_{48} & \lambda_{49} \\ 0 & 0 & 0 & 0 & -\gamma_5 & \lambda_{56} & 0 & 0 & \lambda_{58} & \lambda_{59} \\ 0 & 0 & 0 & 0 & 0 & -\gamma_6 & \lambda_{67} & 0 & \lambda_{69} \\ 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{79} & 0 & \lambda_{79} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{89} & \lambda_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Solving the Kolmogorov differential equations will give the transition probability matrix $P_{ij}(t)$ (see suppl. Info. section 1)

$P_{ij}(t)$ satisfies the following properties :

1. $P_{ij}(t+s) = \sum_{l \in S} P_{il}(t)P_{lj}(s)$, $\forall t \geq 0, s \geq 0, i, j, l \in S$; obeying kolmogorov equations
2. $\sum_S P_{ij}(t) = 1$
3. $P_{ij}(t) \geq 0$, $\forall t \geq 0$ and $i, j \in S$

While the Q matrix satisfies the following conditions:

1. $\sum_S q_{ij}(t) = 0$
2. $q_{ij}(t) \geq 0$, $i \neq j$
3. $-\sum_S q_{ij}(t) = q_{ii}$, $i = j$

Where the q_{ij} is the (i, j) th entry in the Q matrix emphasizing that the P_{ij} depends only on the interval between t_1 and t_2 not on t_1 .

1.1 Maximum Likelihood Estimation of the Q Matrix

Let n_{ijr} be the number of individuals in state i at t_{r-1} and in state j at time t_r . Conditioning on the distribution of individuals among states at t_0 , then the likelihood function for θ is

$$L(\theta) = \prod_{r=1}^w \left\{ \prod_{i,j=1}^k [P_{ij}(t_{r-1}, t_r)]^{n_{ijr}} \right\},$$

where k is the index of the number of states

$$\log L(\theta) = \sum_{r=1}^{\tau} \sum_{i,j=1}^k n_{ijr} \log P_{ij}(t_{r-1}, t_r)$$

where $\tau = (t_r - t_{r-1})$

According to Kalbfleisch and Lawless [12], applying Quasi-Newton method to estimate the rates mandates calculating the score function which is a vector -valued function for the required rates and it's the first derivative of the probability transition function with respect to θ . The second derivative is assumed to be zero.

$$S(\theta) = \frac{\partial}{\partial \theta_h} \log L(\theta) = \sum_{r=1}^{\tau} \sum_{i,j=1}^{k=9} n_{ijr} \frac{\partial P_{ij}(\tau) / \partial \theta_h}{P_{ij}(\tau)},$$

$h = 1, \dots, 22$ while $P_{ij}(\tau) = \frac{n_{ijr}}{n_{i+}}$

$$\frac{n_{ijr}}{P_{ij}(\tau)} = n_{i+}, \text{ such that } n_{i+} = \sum_{j=1}^k n_{ijr}$$

$$S(\theta) = \tau e^{\Lambda \tau} d\Lambda$$

Λ is the eigenvalues for each Q matrix in each τ (see suppl. Info.)

$$\frac{\partial^2}{\partial \theta_g \partial \theta_h} \log L(\theta) =$$

$$\sum_{r=1}^{\tau} \sum_{i,j=1}^k n_{ijr} \left\{ \frac{\partial^2 P_{ij}(\tau) / \partial \theta_g \partial \theta_h}{P_{ij}(\tau)} - \frac{\partial P_{ij}(\tau) / \partial \theta_g \partial P_{ij}(\tau) / \partial \theta_h}{P_{ij}^2(\tau)} \right\}$$

Assuming the second derivative is zero and

$$\frac{n_{ijr}}{P_{ij}(\tau)} = n_{i+} \text{ then}$$

$$M_{ij}(\theta) = \frac{\partial^2}{\partial \theta_g \partial \theta_h} \log L(\theta)$$

$$= - \sum_{r=1}^{\tau} \sum_{i,j=1}^k n_{i+} \frac{\partial P_{ij}(\tau) / \partial \theta_g \partial P_{ij}(\tau) / \partial \theta_h}{P_{ij}(\tau)}$$

The Quasi-Newton formula is

$$\theta_1 = \theta_0 + [M(\theta_0)]^{-1} S(\theta_0)$$

According to Klotz and Sharples [13] the initial

$$\theta_0 = \frac{n_{ijr}}{n_{i+}} \text{ for } \Delta t = 1$$

For this NAFLD process (see suppl. info. section 1.1)

2. Mean Sojourn Time:

It is the mean time spent by a patient in a given state i of the process. It is calculated in relations to transition rates $\hat{\theta}$. These times are independent and exponentially distributed random variables with mean $\frac{1}{\lambda_i}$ where $\lambda_i = -\lambda_{ii}$; $i = 1, \dots, 8$.

According to Kalbfleisch and Lawless [12] the asymptotic variance of this time is calculated by applying multivariate delta method:

$$var(s_i) = \left[(q_{ii}(\hat{\theta}))^{-2} \right]^2 \sum_{h=1}^{22} \sum_{g=1}^{22} \frac{\partial q_{ii}}{\partial \theta_g} \frac{\partial q_{ii}}{\partial \theta_h} [M(\theta)]^{-1} |_{\theta=\hat{\theta}}$$

For this NAFLD process (see suppl. info section 2)

3. State Probability Distribution

According to Cassandras and Lafortune [14], it is the probability distribution for each state at a specific time point given the initial probability distribution. Thus using the rule of total probability; a

solution describing the transient behavior of a chain characterized by Q and an initial condition $\pi(0)$ is obtained by direct substitution to solve:

$$\pi(t) = \pi(0)P(t)$$

Stationary probability distribution when t goes to infinity or in other words when the process does not depend on time is obtained by differentiating both sides of the following equation:

$$\pi(t) = \pi(0)P(t) = \pi(0)e^{Qt}$$

differentiate both sides to obtain $\frac{d}{dt}\pi(t)\Big|_{t=0} = \pi(0)Q$

$$\begin{aligned} \frac{d}{dt}\pi_i(t)\Big|_{t=0} &= [\pi_{0(1)} \ \pi_{0(2)} \ \pi_{0(3)} \ \pi_{0(4)} \ \pi_{0(5)} \ \pi_{0(6)} \ \pi_{0(7)} \ \pi_{0(8)} \ \pi_{0(9)}] \times \\ &\begin{bmatrix} -\gamma_1 & \lambda_{12} & 0 & 0 & 0 & 0 & 0 & \lambda_{18} & \lambda_{19} \\ \mu_{21} & -\gamma_2 & \lambda_{23} & 0 & 0 & 0 & 0 & \lambda_{28} & \lambda_{29} \\ 0 & \mu_{32} & -\gamma_3 & \lambda_{34} & 0 & 0 & 0 & \lambda_{38} & \lambda_{39} \\ 0 & 0 & \mu_{43} & -\gamma_4 & \lambda_{45} & 0 & 0 & \lambda_{48} & \lambda_{49} \\ 0 & 0 & 0 & 0 & -\gamma_5 & \lambda_{56} & 0 & \lambda_{58} & \lambda_{59} \\ 0 & 0 & 0 & 0 & 0 & -\gamma_6 & \lambda_{67} & 0 & \lambda_{69} \\ 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{79} & 0 & \lambda_{79} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{89} & \lambda_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

$[\pi_1 \ \pi_2 \ \pi_3 \ \pi_4 \ \pi_5 \ \pi_6 \ \pi_7 \ \pi_8 \ \pi_9]$ at specific time point is obtained by solving this system of differential equations. Solving these differential equations for this complex chain is not a trivial matter. If the limit of $\pi_z = \lim_{t \rightarrow \infty} \pi_z(t)$ exists, so there is a stationary or steady state distribution and as $t \rightarrow \infty$, the $\frac{d}{dt}\pi_j(t) = 0$, since $\pi_z(t)$ does not depend on time. Therefore $\frac{d}{dt}\pi(t) = \pi(t)Q$ will reduce to $\pi(t)Q = 0$. The stationary state probability distribution is obtained by solving $\pi Q = 0$ subject to $\sum_{all \ z} \pi_z = 1$

For this NAFLD process (see suppl.info section 3)

3.1 Asymptotic Covariance of the State Probability Distribution:

Multivariate delta method is applied to following function $Q'\pi = F(\theta_h, \pi_i) = 0$ to obtain asymptotic covariance matrix of the state probability distribution, as π is not a simple function of thetas.

Differentiating $F(\theta_h, \pi_i)$ implicitly with respect to θ_h is used in the following manner:

$$\begin{aligned} \frac{\partial}{\partial \theta_h} F(\theta_h, \pi_i) &= \frac{\partial}{\partial \theta_h} Q'\pi_i = 0 \\ \frac{\partial}{\partial \theta_h} Q'\pi_i &= [Q'] \left[\frac{\partial}{\partial \theta_h} \pi_i \right] + \pi_i \left[\frac{\partial}{\partial \theta_h} Q' \right]^T = 0 \\ \text{let's call } \pi_i \left[\frac{\partial}{\partial \theta_h} Q' \right]^T &= C(\theta) \\ [Q'] \left[\frac{\partial}{\partial \theta_h} \pi_i \right] + C(\theta) &= 0 \\ \text{solving for } \left[\frac{\partial}{\partial \theta_h} \pi_i \right] & \\ \left[\frac{\partial}{\partial \theta_h} \pi_i \right] &= -[Q']^{-1}C(\theta) \\ \text{Let } \left[\frac{\partial}{\partial \theta_h} \pi_i \right] &= A(\theta) \end{aligned}$$

By multivariate delta method

$$var(\pi) = A(\theta)var(\theta)A(\theta)' \quad , \quad \text{where } var(\theta) = [M(\theta)]^{-1}$$

For this NAFLD process: (see suppl.info section 3.1)

4. Life Expectancy of Patient in NAFLD Disease Process:

The disease process is composed of 8 transient states and one absorbing (death state). So the Q matrix is partitioned into 4 sets:

$$Q = \begin{bmatrix} B & A \\ 0 & 0 \end{bmatrix}, \quad \text{where } B = \begin{bmatrix} -\gamma_1 & \lambda_{12} & 0 & 0 & 0 & 0 & 0 & \lambda_{18} \\ \mu_{21} & -\gamma_2 & \lambda_{23} & 0 & 0 & 0 & 0 & \lambda_{28} \\ 0 & \mu_{32} & -\gamma_3 & \lambda_{34} & 0 & 0 & 0 & \lambda_{38} \\ 0 & 0 & \mu_{43} & -\gamma_4 & \lambda_{45} & 0 & 0 & \lambda_{48} \\ 0 & 0 & 0 & 0 & -\gamma_5 & \lambda_{56} & 0 & \lambda_{58} \\ 0 & 0 & 0 & 0 & 0 & -\gamma_6 & \lambda_{67} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{79} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_{89} \end{bmatrix}$$

Also the differential equations can be partitioned into the following:

$$\begin{bmatrix} \dot{P}(t) & \dot{P}_k(t) \end{bmatrix} = \begin{bmatrix} P(t) & P_k(t) \end{bmatrix} \begin{bmatrix} B & A \\ 0 & 0 \end{bmatrix}$$

B is the transition rate matrix among the transient states and the column vector A is the transition rate from each transient state to the absorbing (death) state.

$$A = -B1^T \quad \text{such that } 1^T \text{ is a column vector of } (k-1) \times 1 \text{ with all its elements equal to one}$$

$$\begin{bmatrix} \dot{P}(t) & \dot{P}_k(t) \end{bmatrix} = \begin{bmatrix} P(t) & P_k(t) \end{bmatrix} \begin{bmatrix} B & A \\ 0 & 0 \end{bmatrix} \quad \text{can be written as}$$

$$\dot{P}(t) = P(t)B \quad \text{and} \quad \dot{P}_k(t) = P(t)A$$

The solution to $\dot{P}(t) = P(t)B$ is

$$P(t) = P(0)e^{Bt} \quad \text{then} \quad \dot{P}_k(t) = P(0)e^{Bt}A$$

$$\text{and} \quad e^{Bt} = 1 + Bt + \frac{(Bt)^2}{2!} + \frac{(Bt)^3}{3!} + \frac{(Bt)^4}{4!} + \dots = \sum_{j=0}^{\infty} \frac{(Bt)^j}{j!}$$

If τ_k is the time taken from state i to reach the absorbing death state from the initial time

$$F_k(t) = pr[\tau_k \leq t] = pr[X(t) = k] = P_k(t) = 1 - P(t)1^t = 1 - P(0)e^{Bt}1^t$$

The moment theory for Laplace transform can be used to obtain the mean of the time which has the above cumulative distribution function. CTMC can be written in a Laplace transform such that:

$$[sP^*(s) - P(0) \quad sP_k^*(s)] = [P^*(s) \quad P_k^*(s)] \begin{bmatrix} B & A \\ 0 & 0 \end{bmatrix}$$

$$\therefore sP^*(s) - P(0) = P^*(s)B \quad \text{and} \quad sP_k^*(s) = P^*(s)A$$

Rearrange:

$$\therefore sP^*(s) - P^*(s)B = P(0)$$

$$P^*(s)[sI - B] = P(0) \rightarrow P^*(s) = P(0)[sI - B]^{-1}$$

$$\therefore sP_k^*(s) = P^*(s)A \rightarrow P_k^*(s) = \frac{1}{s} P^*(s)A = \frac{1}{s} P(0)[sI - B]^{-1}A$$

$$F_k^*(s) = \frac{1}{s} P(0)[sI - B]^{-1}A$$

$$f_k^*(s) = sF_k^*(s) = P(0)[sI - B]^{-1}A \quad ; \quad \text{where } A = -B1^T$$

Mean time to absorption:

$$E(\tau_k) = (-1) \frac{df_k^*(s)}{ds} \Big|_{s=0} = (-1)P(0)[sI - B]^{-2}A \Big|_{s=0} = P(0)[-B]^{-1}1^T$$

For this NAFLD process: (see suppl.info section 4)

5. Expected Number of Patients in Each State

Let $u(0)$ be the size of patients in a specific state at specific time $t = 0$. The initial size of patients $u(0) = \sum_{j=1}^8 u_j(0)$, as there are 8 transient states and 1 absorbing states, where $u_j(0)$ is the initial size or number of patients in state j at time $t = 0$ given that $u_9(0) = 0$ i.e initial size of patients in state 9 (absorbing death state) is zero at initial time point $t = 0$. As the transition or the movement of the patients among states are independent so at the end of the whole time interval $(0, t)$ and according to Chiang [15], there will be $u_j(t)$ patients in the transient states at time t , also there will be $u_9(t)$ patients in state 9 (death state) at time t .

$$E[u_j(t)|u_j(0)] = \sum_{j=1, i=1}^9 u_i(0)P_{ij}(t), \quad i \& j = 1, \dots, 9$$

For this NAFLD process: (see suppl. info section 5)

6. Hypothetical Numerical Example :

To illustrate the above concepts and discussion, a hypothetical numerical example is introduced. It does not represent real data but it is for demonstrative purposes. (see suppl. Info. Section 6)

A study was conducted over 15 years on 1050 patients with risk factors for developing NAFLD such as type 2 diabetes mellitus, obesity, and hypertension acting alone or together as a metabolic syndrome. The patients were decided to be followed up every year by a liver biopsy to identify the NAFLD cases, but the actual observations were recorded as shown in the (see supplementary material).

The estimated transition rate matrix Q is:

$$\hat{Q} = \begin{bmatrix} -.397 & .39 & 0 & 0 & 0 & 0 & 0 & 0 & .007 \\ .02 & -.281 & .25 & 0 & 0 & 0 & 0 & 0 & .011 \\ 0 & .05 & -.365 & .225 & 0 & 0 & 0 & .047 & .043 \\ 0 & 0 & .041 & -.538 & .281 & 0 & 0 & .109 & .107 \\ 0 & 0 & 0 & 0 & -.348 & .19 & 0 & .059 & .099 \\ 0 & 0 & 0 & 0 & 0 & -.934 & .767 & 0 & .167 \\ 0 & 0 & 0 & 0 & 0 & 0 & -.421 & 0 & .421 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.745 & .745 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$var(\hat{\theta}) = 1 \times 10^{-13} \begin{bmatrix} v_1 & v_2 \\ v_3 & v_4 \end{bmatrix} \text{ where}$$

$$v_1 = \begin{bmatrix} .3292 & .0327 & .4414 & .0827 & .4004 & .0268 & .2540 & .1517 \\ .0327 & .0064 & .0428 & .0108 & .0385 & .0053 & .0268 & .0159 \\ .4414 & .0428 & .5922 & .1100 & .5373 & .0350 & .3400 & .2032 \\ .0827 & .0108 & .1100 & .0228 & .0995 & .0088 & .0650 & .0387 \\ .4004 & .0385 & .5373 & .0995 & .4876 & .0315 & .3082 & .1843 \\ .0268 & .0053 & .0350 & .0088 & .0315 & .0044 & .0219 & .0130 \\ .2540 & .0268 & .3400 & .0650 & .3082 & .0219 & .1967 & .1174 \\ .1517 & .0159 & .2032 & .0387 & .1843 & .0130 & .1174 & .0702 \end{bmatrix}$$

Expected count in the first interval:

v_2, v_3 and v_4 are all zero matrices of size (8 by 14), (14 by 8) and (14 by 14) respectively.

Transition probability matrix at 1 year:

$$P(1) = \begin{bmatrix} .6751 & .279 & .0345 & .0024 & 0 & 0 & .0006 & .0082 \\ .0143 & .7625 & .1819 & .0188 & .0019 & 0 & 0 & .0044 \\ .0004 & .0364 & .7017 & .1439 & .0206 & .0012 & .0002 & .0346 \\ 0 & .0007 & .0262 & .5868 & .1800 & .0145 & .0039 & .0616 \\ 0 & 0 & 0 & 0 & .7061 & .1015 & .0416 & .0344 \\ 0 & 0 & 0 & 0 & 0 & .3930 & .3938 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .6564 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .4747 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Mean time spent by the patient in state 1 is approximately 2 years and 6 months, in state 2 the mean sojourn time is approximately 3 years and 6 months, in state 3 it is approximately 2 years and 9 months, in state 4 it is approximately 1 years and 10 months, in state 5 it is approximately 2 years and 10 months, in state 6 it is approximately 1 years and 1 month, in state 7 it is approximately 2 years and 5 months and lastly in state 8 the mean sojourn time is approximately 1 years and 4 months.

If a cohort of 5000 NAFLD patients have initial distribution of

[.62 .22 .081 .03 .028 .005 .007 .009 0] and

initial counts of patients in each state

are [3100 1100 405 150 140 25 35 45 0] then at 1 year the state probability distribution

is [.4217 .3437 .1191 .035 .0274 .0054 .0079 .0111 .0287]

and the expected counts of patients are

[2109 1718 595 175 137 27 39 56 144].

To calculate goodness of fit for multistate model used in this model, it is like the procedure used in contingency table, and it is calculated in each interval then sum up:

Step 1: H_0 = future state does not depend on the current state

H_1 = future state does depend on the current state

Step 2: calculate the $P_{ij}(\Delta t = 1)$

$$P_{ij}(\Delta t = 1) = \begin{bmatrix} .6751 & .279 & .0345 & .0024 & .0002 & .0000 & .0000 & .0006 & .0082 \\ .0143 & .7625 & .1819 & .0188 & .0019 & .0001 & .0000 & .0044 & .016 \\ .0004 & .0364 & .7017 & .1439 & .0206 & .0012 & .0002 & .0346 & .0604 \\ 0 & .0007 & .0262 & .5868 & .1800 & .0145 & .0039 & .0616 & .1215 \\ 0 & 0 & 0 & 0 & .7061 & .1015 & .0416 & .0344 & .1163 \\ 0 & 0 & 0 & 0 & 0 & .3930 & .3938 & 0 & .2132 \\ 0 & 0 & 0 & 0 & 0 & 0 & .6564 & 0 & .3436 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .4747 & .5253 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Step 3: calculate the expected counts in this interval by multiplying each row in the probability matrix with the corresponding total marginal counts in the observed transition counts matrix in the same interval to get the expected counts:

	State1	State2	State3	State4	State5	State6	State7	State8	State9	total
State1	1000.5	413.5	51.13	3.6	.3	0	0	.8892	12.15	1482
State2	7.4074	394.97	94.22	9.738	.9324	.0518	0	2.2792	8.288	517.89
State3	.08	7.28	140.34	28.18	4.12	.24	.04	6.92	12.08	199.28
State4	0	.0511	1.8761	40.953	12.9283	1.0585	.2847	4.4968	8.8695	70.518
State5	0	0	0	0	47.3087	6.8005	2.7872	2.3048	7.7921	66.9933
State6	0	0	0	0	0	3.93	3.938	0	2.132	10
State7	0	0	0	0	0	0	11.1588	0	5.8412	17
State8	0	0	0	0	0	0	0	9.494	10.506	20
State9	0	0	0	0	0	0	0	0	0	0

Step 4: apply

$$\sum_{i=1}^9 \sum_{j=1}^9 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 2226.362 \sim \chi_{(9-1)(9-1)(.05)}^2$$

The same steps are used for the observed transition counts in the $\Delta t = 2$ and $\Delta t = 3$ with the following results:

$$\text{Step2 : } P_{ij}(\Delta t = 2) = \begin{bmatrix} .4597 & .4023 & .0983 & .0131 & .0019 & .0001 & .0000 & .0032 & .0209 \\ .0206 & .5920 & .2673 & .0505 & .0097 & .0007 & .0002 & .0129 & .0441 \\ .001 & .0535 & .5026 & .1786 & .054 & .0054 & .0022 & .0503 & .1414 \\ 0 & .0018 & .0325 & .3183 & .2249 & .0318 & .0178 & .0708 & .2486 \\ 0 & 0 & 0 & 0 & .4986 & .1116 & .0967 & .0406 & .2525 \\ 0 & 0 & 0 & 0 & 0 & .1544 & .4133 & 0 & .4323 \\ 0 & 0 & 0 & 0 & 0 & 0 & .4308 & 0 & .5692 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .2254 & .7746 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Step 3: The expected counts in the second interval:

	State1	State2	State3	State4	State5	State6	State7	State8	State9	total
State1	272.6	238.56	58.292	7.7683	1.1267	.0593	0	1.8976	12.394	592.704
State2	4.2642	122.54	55.33	10.454	2.0079	.1449	.0414	2.6703	9.129	206.59
State3	.08	4.28	40.208	14.288	4.32	.432	.176	4.024	11.312	79.12
State4	0	.0522	.9425	9.231	6.5221	.922	.5162	2.0532	7.209	27.45
State5	0	0	0	0	13.462	3.013	2.6109	1.0962	6.818	27
State6	0	0	0	0	0	.6176	1.6532	0	1.729	4
State7	0	0	0	0	0	0	3.0156	0	3.984	7
State8	0	0	0	0	0	0	0	1.8032	6.197	8
State9	0	0	0	0	0	0	0	0	0	0

$$\text{step 4: } \sum_{i=1}^9 \sum_{j=1}^9 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 160.115 \sim \chi_{(9-1)(9-1)(.05)}^2$$

The same steps are used in $\Delta t = 3$ with the results:

$$\text{step 2: } P_{ij}(\Delta t = 3) = \begin{bmatrix} .3161 & .4386 & .1584 & .0299 & .0065 & .0006 & .0002 & .0078 & .0406 \\ .0225 & .4669 & .2973 & .0771 & .0223 & .0024 & .0011 & .0214 & .0842 \\ .0016 & .0595 & .367 & .172 & .0802 & .0108 & .0066 & .0544 & .2289 \\ .0001 & .0028 & .0313 & .1832 & .2159 & .0400 & .0348 & .0621 & .3655 \\ 0 & 0 & 0 & 0 & .352 & .0945 & .1282 & .0364 & .3889 \\ 0 & 0 & 0 & 0 & 0 & .0607 & .3321 & 0 & .6072 \\ 0 & 0 & 0 & 0 & 0 & 0 & .2828 & 0 & .7172 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .107 & .893 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Step 3: The expected counts in the third interval:

	State1	State2	State3	State4	State5	State6	State7	State8	State9	total
State1	46.783	64.913	23.443	4.4252	.962	.0888	.0296	1.1544	6.0088	147.81
State2	1.148	23.812	15.162	3.9321	1.1373	.1224	.0561	1.3914	4.2942	50.76
State3	.0288	1.071	6.606	3.096	1.4436	.1944	.1188	.9792	4.1202	17.66
State4	.0007	.0196	.2191	1.2824	1.5113	.28	.2436	.4347	2.5585	6.55
State5	0	0	0	0	2.464	.6615	.8974	.2548	2.7223	7
State6	0	0	0	0	0	.1214	.6642	0	1.2144	2
State7	0	0	0	0	0	0	.5656	0	1.4344	2
State8	0	0	0	0	0	0	0	.321	2.679	3
State9	0	0	0	0	0	0	0	0	0	0

$$\text{step 4 : } \sum_{i=1}^9 \sum_{j=1}^9 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 69.778 \sim \chi^2_{(9-1)(9-1)(.05)}$$

Step 5: sum up the above results to obtain:

$$\sum_{i=1}^9 \sum_{j=1}^9 \sum_{l=1}^{t=3} \frac{(O_{ijl} - E_{ijl})^2}{E_{ijl}} = 2456.255 \sim \chi^2_{(df=192)(.05)}$$

So from the above results the null hypothesis is rejected while the alternative hypothesis is accepted and the model fits the data that is to mean the future state depends on the current state with the estimated transition rate and probability matrices as obtained.

7. Conclusion and Summary

Continuous time Markov chains are suitable mathematical and statistical tools to be used for analysis of disease evolution over time. CTMCs being a type of multistate models are utilized to study this evolution in NAFLD patients, with its main phenotypes NAFLD and NASH, as well as the associated presence of fibrosis and its stages. The prevalence of NAFLD is rapidly increasing worldwide, and parallels the epidemics of obesity and type 2 diabetes. Metabolic syndrome is a well-known risk factor.

In the present study, NAFLD is modeled in more elaborative expanded form, which includes nine states: the first eight states are the states of disease progression as time elapses; while, the ninth state is the death state. The importance of such analysis is that the health policy makers can predict the number of affected patients at each stage, the needed investigations and medications for each of them, and the costs and budgets that the medical insurance should assign to this disease burden. This analysis is of great value and benefit to the physicians, as they can conduct longitudinal studies to explore further investigations that better define each stage specifically and efficiently, as well as to explore further treatment needed for each stage. An example of the non-invasive diagnostic tools is the circulating level of cytokeratin-18 fragments, although promising it is not available in a clinical care setting and there is not an established cut-off value for identifying steato-hepatitis (NASH)[16]. A genetic polymorphism of patatin-like phospholipase domain-containing protein 3 gene variants (PNPLA-3) are associated with NASH and advanced fibrosis, however testing for these variants in routine clinical care is not supported and needs further studies.

The hypothetical examples of factitious non-real data is used to emphasize the attributes need to be estimated:

- Transition rate matrix among the various states.
- Transition probability matrix among states.
- Mean sojourn time in each state.
- Mean time to absorption (death state).
- Expected number of patients in each state.

- State probability distribution at specific time point in the future.

Such analysis may give better insights to physicians, especially when new drug classes will soon be released in the market. What drug classes are to be used first? How to monitor the disease throughout the journey of treatment? What investigations to be used in such monitoring? How to modify the drug treatment? What is the target that needs to be achieved and how to maintain this target? And what is more to be said; that is in late stage of the disease, when patients suffer from decompensated liver cirrhosis; liver transplantation is the treatment of choice to such patients, which increases the economic burden of NAFLD as was the disease course during treatment in early stages. Also, the load of what are the best economic noninvasive tests to be used in primary health care units for stratification and identification of high risk patients, whether to do genetic tests in health insurance setting, and when to refer for liver biopsy in secondary or special clinics. All these questions can be answered from such longitudinal studies conducted on susceptible individuals. Over and above, some of the recently investigated noninvasive scoring systems of fibrosis need further external validation so as to be generalized in ethnicities other than the one tested upon. There are some controversies of cutoff points of these scoring systems among countries, and among ethnicities within the same country. Although liver biopsy is considered the standard method for diagnosis of NAFLD and staging it; its limitations encourage the development of various noninvasive tests, which necessitate better correlation between the findings obtained from the biopsy and the results of these tests to minimize the misclassification errors, which hamper good diagnosis and prognosis of the patient. These tests should be easy, feasible, convenient and with high safety profile to be used repeatedly in patients for follow up in such longitudinal studies.

Multistate model represented by CTMC is a valuable statistical methodology, for longitudinal studies in medical researches to better comprehend and understand the pathophysiology, or the mechanism of the NAFLD process, and the interactions between the different modifiers either the external, or the internal modifiers. The external modifiers reside in bad dietary habits with excessive fat and carbohydrate ingestion, as well as sedentary life; while, the internal modifiers are represented in genetic factors affecting the metabolism of the food stuff (fat and carbohydrates) and other cellular functions such as risk factors for fibro-genesis (formation of fibrous tissue); as, fibrosis is a detrimental predictor factor for disease progression to liver cirrhosis and its complications. The importance of such understanding has a great impact to reveal the genes that must be

tested if ever needed, for whom to do such a test, and should it be in the utilities or services offered by the medical insurance. Moreover, should the degradation byproducts resulting from extracellular matrix destruction be used in routine clinical practice to mirror the fibrosis stages?

In Egypt, there are scarce data, or may be no available data, about the prevalence of NAFLD and its phenotypes. Guidelines for risk stratification and identification are also lacking. Thus, more longitudinal studies are needed to cover these issues.

Multistate models can also be used for analysis of competing risks to death in such patients, as the first and second most common causes of death in NAFLD patients are the cardiovascular diseases (CVD) and kidney diseases, while the liver-related mortality is the third common cause of death.

Some other statistical methodologies, like : semi Markov and hidden Markov chains can be used to model NAFLD, especially hidden Markov CTMC can be used to model misclassification errors encountered in studies conducted by time homogenous CTMC.

Abbreviations :

CC: compensated cirrhosis (stage 4), CTMC: continuous time Markov chains, CVS: cardiovascular disease, DCC: de-compensated cirrhosis (stage 5),EASD: European Association for the Study of diabetes, EASL: European Association for the Study of liver, EASO: European Association for the Study of obesity, EM= extra-mortality (stage 9), HCC :hepato-cellular carcinoma (stage 8), HS: hepatic steatosis, LT= liver transplant (stage 6), NAFLD: non-alcoholic fatty liver disease, NAFLD-NO FB: nonalcoholic fatty liver disease with no fibrosis (stage 1), NASH: non-alcoholic steatohepatitis, NASH-NO FB : nonalcoholic steato-hepatitis with no fibrosis (stage 2), NASH-FB : nonalcoholic steato-hepatitis with fibrosis (stage 3), PLT : post liver transplant (stage7), PNPLA-3:patatin-like phospholipase domain-containing protein 3 gene variants, TE: transient elastography, T2DM: type 2 diabetes mellitus.

Declarations:

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and material

Not applicable. Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests

The author declares that I have no competing interests.

Funding

No funding resource. No funding roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript are declared

Authors' contribution

I am the author who has carried the mathematical analysis as well as applying these mathematical statistical concepts on the hypothetical example.

Acknowledgement

Not applicable.

Hint: A MATLAB code is edited to calculate the statistical indices in the hypothetical example. The code can be found published in the code ocean site with the following URL:

codeocean.com/capsule/7628018/tree/v1

Also the supplementary materials can be found in the same site.

References

- [1] C. Estes, H. Razavi, R. Loomba, Z. Younossi, and A. J. Sanyal, "Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease," *Hepatology*, vol. 67, no. 1, pp. 123–133, 2018. DOI: <https://doi.org/10.1002/hep.29466>
- [2] Z. M. Younossi *et al.*, "The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe," *Hepatology*, vol. 64, no. 5, pp. 1577–1586, 2016. DOI: <https://doi.org/10.1002/hep.28785>
- [3] N. Chalasani *et al.*, "The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases," *Hepatology*, vol. 67, no. 1, pp. 328–357, 2018, doi: 10.1002/hep.29367.
- [4] E. Association, E. Association, D. Eas, E. Association, and O. Easo, "Clinical Practice Guidelines EASL – EASD – EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease q," *J. Hepatol.*, vol. 64, no. 6, pp. 1388–1402, 2016, doi: 10.1016/j.jhep.2015.11.004.
- [5] N. Chalasani *et al.*, "Relationship of steatosis grade and zonal location to histological features of steatohepatitis in adult patients with non-alcoholic fatty liver disease," *J. Hepatol.*, vol. 48, no. 5, pp. 829–834, 2008. DOI: <https://doi.org/10.1016/j.jhep.2008.01.016>
- [6] S. B. Reeder, I. Cruite, G. Hamilton, and C. B. Sirlin, "Quantitative assessment of liver fat with magnetic resonance imaging and spectroscopy," *J. Magn. Reson. imaging*, vol. 34, no. 4, pp. 729–749, 2011. DOI: <https://doi.org/10.1002/jmri.22580>
- [7] I. S. Idilman *et al.*, "A comparison of liver fat

content as determined by magnetic resonance imaging-proton density fat fraction and MRS versus liver histology in non-alcoholic fatty liver disease," *Acta radiol.*, vol. 57, no. 3, pp. 271–278,2016.

DOI: <https://doi.org/10.1177/0284185115580488>

- [8] M. Nouredin *et al.*, "Utility of magnetic resonance imaging versus histology for quantifying changes in liver fat in nonalcoholic fatty liver disease trials," *Hepatology*, vol. 58, no. 6, pp. 1930–1940, 2013.
DOI: <https://doi.org/10.1002/hep.26455>
- [9] V. de Lédinghen *et al.*, "Controlled attenuation parameter for the diagnosis of steatosis in non-alcoholic fatty liver disease," *J. Gastroenterol. Hepatol.*, vol. 31, no.4,pp.848–855,2016.
DOI: <https://doi.org/10.1111/jgh.13219>
- [10] J. Paul, "Recent advances in non-invasive diagnosis and medical management of non-alcoholic fatty liver disease in adult," *Egypt. Liver J.*, vol. 10, no. 1, 2020, doi: 10.1186/s43066-020-00043-x.
- [11] P. Dongiovanni, Q. M Anstee, and L. Valenti, "Genetic predisposition in NAFLD and NASH: impact on severity of liver disease and response to treatment," *Curr. Pharm. Des.*, vol. 19, no. 29, pp. 5219–5238, 2013.
DOI: <https://doi.org/10.2174/13816128113199990381>
- [12] J. D. Kalbfleisch and J. F. Lawless, "The analysis of panel data under a Markov assumption," *J. Am. Stat. Assoc.*, vol. 80, no. 392,pp.863–871,1985.
DOI: <https://doi.org/10.2307/2288545>
- [13] J. H. Klotz and L. D. Sharples, "Estimation for a Markov heart transplant model," *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 43,no.3,pp.431–438,1994.
DOI: <https://doi.org/10.2307/2348579>
- [14] C. G. Cassandras and S. Lafortune, *Introduction to discrete event systems*. Springer Science & Business Media, 2009.
- [15] C. L. Chiang, "Introduction to stochastic processes in biostatistics," 1968.
- [16] G. Musso, R. Gambino, M. Cassader, and G. Pagano, "Meta-analysis: natural history of

non-alcoholic fatty liver disease (NAFLD) and diagnostic accuracy of non-invasive tests for liver disease severity," *Ann. Med.*, vol.43,no.8,pp.617–649,2011.

DOI: <https://doi.org/10.3109/07853890.2010.518623>

Iman Attia is currently M.Sc student in Faculty of Graduate Studies of Statistical Research, Mathematical Statistics Department, Cairo University. In 2001, she received M.Sc. degree in internal medicine, Cairo University. In 2015, she received Bachelor's degree in Computer Science and Information Technology from the Egyptian Electronic-learning University (EELU). In 2017, she received Diploma in statistics from Institute of Statistical studies and Research, Cairo University. Her research interests include biostatistics, bioinformatics and machine learning