Log-Linear Model and Multistate Model to Assess the Rate of Fibrosis in Patients With NAFLD

*Iman M. Attia**

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

In this paper, the deleterious effects of obesity, type II diabetes, and insulin resistance, systolic and diastolic hypertension on the rate of progression of fibrosis in patients with non-alcoholic fatty liver disease (NAFLD) are illustrated using a new approach utilizing the Poisson regression to model the transition rate matrix. The observed counts in the transition count matrix are used as the response variables and the covariates are the risk factors for fatty liver. Then, the estimated counts from running the Poisson regression are used to estimate the transition rates using the continuous-time Markov chains (CTMCs) followed by exponentiation of the estimated rate matrix to obtain the transition probability matrix at specific time points. A depicted, hypothetical, observational, prospective longitudinal study of 150 participants followed up every year for a total of 29 years recording their demographic characteristics and their timeline follow-up is demonstrated. The findings revealed that insulin resistance expressed by HOMA2-IR had the most deleterious effects among other factors on increasing the rate of fibrosis progression from state 1 to state 2, from state 2 to state 3, and from state 3 to state 4. The higher the level of HOMA2-IR is, the more rapid the rate of progression is. This analysis helps the health policymakers and medical insurance managers to allocate the financial and human resources for investigating and treating high-risk patients with NAFLD. In addition, this analysis can be used by pharmaceutical companies to conduct longitudinal studies to assess the effectiveness of the newly emerging anti-fibrotic drugs.

Keywords: log-linear model, multistate model, non-alcoholic fatty liver disease, NAFLD, Poisson regression, continuous-time Markov chains, longitudinal studies, HOMA2-IR

OPEN ACCESS

Edited by:
Appanah Rao Appadu,
Nelson Mandela University,
South Africa

Reviewed by:
Ashish Awasthi,
NITc, India
Farai Julius Mhlanga,
University of Limpopo, South Africa

****Correspondence:***
Iman M. Attia
imanattiathesis1972@gmail.com
orcid.org/0000-0002-7333-9713

Specialty section:
This article was submitted to Mathematical Biology, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *18 March 2022*
Accepted: *12 April 2022*
Published: *16 June 2022*

Citation:
Attia IM (2022) Log-Linear Model and Multistate Model to Assess the Rate of Fibrosis in Patients With NAFLD. Front. Appl. Math. Stat. 8:899247. doi: 10.3389/fams.2022.899247

INTRODUCTION

Continuous-time Markov chains (CTMCs) are valuable mathematical and statistical tools. They are of great potential to evaluate the disease progression over time. NAFLD is an increasingly worldwide epidemic, paralleling the rise in the incidence of obesity and type II diabetes which are approaching a pandemic level. This emerging health problem is mainly due to sedentary life styles and western eating habits of ingesting high-fat and cholesterol diets. The pathological milestone for NAFLD is insulin resistance and hyperinsulinemia. This hyperinsulinemia will eventually result in type II diabetes with adverse complications like vascular diseases and fatty liver disease. On the other hand, NAFLD can cause type II diabetes, as the prevalence of diabetes in NAFLD ranges between 18 and 45%. Moreover, the prevalence of NAFLD in type II diabetic patients ranges between 49 and 75% [1].

Non-alcoholic fatty liver disease can be modeled using the simplest form for health, disease, and death model. It is composed of four states. One state is for susceptible individuals with risk factors like type II diabetes, dyslipidemia, obesity, and hypertension. The second state is the NAFLD phenotypes. The other two competing states for death are: one for liver-related mortality as a complication of NAFLD and the other state is the death causes unrelated to liver disease [2]. This model is shown in **Figure 1**.

In addition, NAFLD can be modeled in more elaborate expanded form which includes nine states [3]. The first eight states are the states of disease progression over time and the ninth state is the death state [2], as illustrated in **Figure 2**.

Moreover, fibrogenesis is a dynamic process that goes back and forth among the early stages of the expanded model. Stages of fibrous tissue formation are early seen in NAFLD process. Fibrosis progresses if the risk factors for its formation are not eliminated. Fibrosis is an ominous sign for loss of liver functions. When the fibrous tissue develops, a subset of the early states is used to relate these risk factors to the rates. Definition of each state is shown in **Figure 3** [4, 5]. F0 indicates that there is no fibrous tissue. F1 means that fibrous tissue is deposited due to non-alcoholic steatohepatitis (NASH) and not due to any other causes of liver disease; all other stages (F2 and F3) are maintained and are progressing over time by the presence of NASH till the liver cirrhosis (F4). If this NASH is well-treated by controlling the risk factors that induce it, the fibrous tissue formation and deposition will regress as shown in the **Figure 3**.

Kalbfleisch and Lawless [6] related the instantaneous rate of transitions from state *i* to state *j* to covariates, by regression modeling of the Q transition rate matrix using log-linear model for the Markov rates.

The previous studies, as will be later mentioned in the discussion, mainly included the evaluation of 2 paired biopsies, initial and second biopsies, then grouping the patients according to the findings into stable, regressors, slow progressors, and rapid progressors without precise estimation of specific transition rates among states and without proper estimation of the predictive value of each variable on these specific rates. The rate of fibrosis progression was estimated by dividing the difference in fibrosis stage between biopsies by the time interval (in years), and this was performed to account for the time differences between the biopsies [7]. Additionally, either univariate or multivariate linear regression was used to relate the risk factors with the rate of progression. As will be later mentioned in the discussion, some studies utilized multivariate logistic regression instead of linear regression.

This depicted study differs from the previous studies in many aspects. First, it proposes recording multiple repeated observations over time. Second, it suggests running Poisson regression to relate the transition rates among states with the risk

Abbreviations: CC, compensated cirrhosis (stage 4); CTMC, continuous-time Markov chains; DCC, de-compensated cirrhosis (stage 5); EM, extramortality (stage 9); HCC, hepatocellular carcinoma (stage 8); LT, liver transplant (stage 6); NAFLD, non-alcoholic fatty liver disease; NAFL-NO FB, non-alcoholic fatty liver with no fibrosis (stage 1); NASH, non-alcoholic steatohepatitis; NASH-NO FB, non-alcoholic steatohepatitis with no fibrosis (stage 2); NASH-FB, non-alcoholic steatohepatitis with fibrosis (stage 3); PLT, post-liver transplant (stage 7); T2DM, type 2 diabetes mellitus.

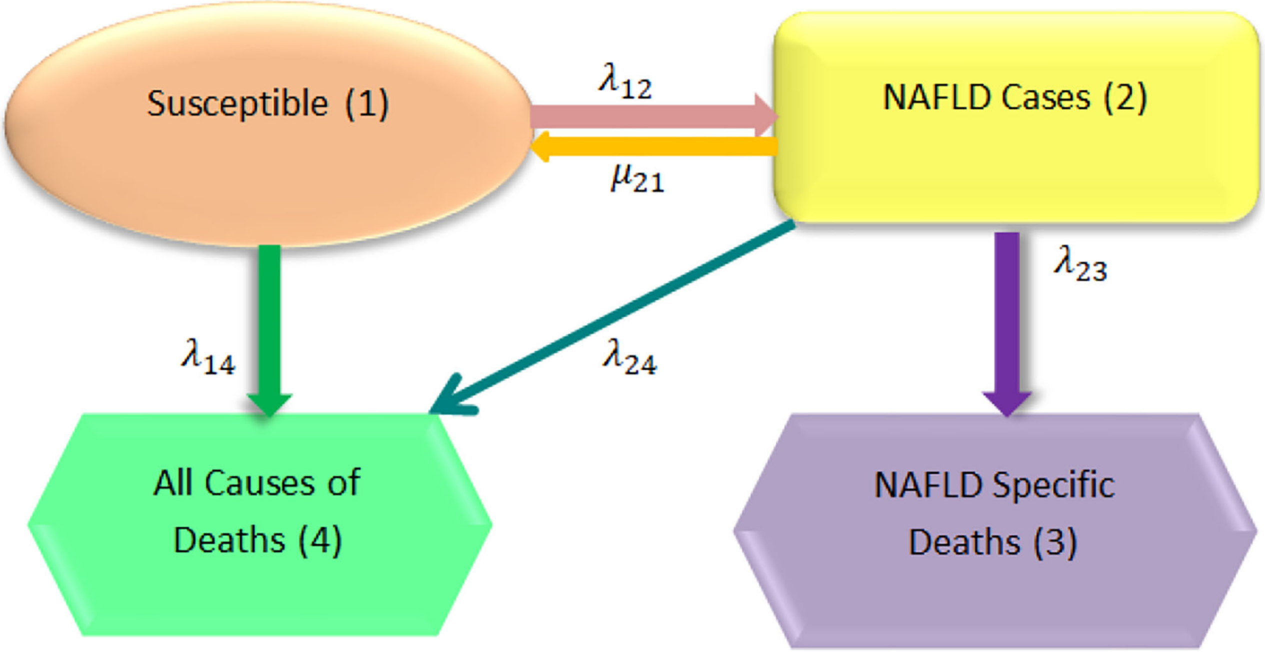

FIGURE 1 | General model structure [2].

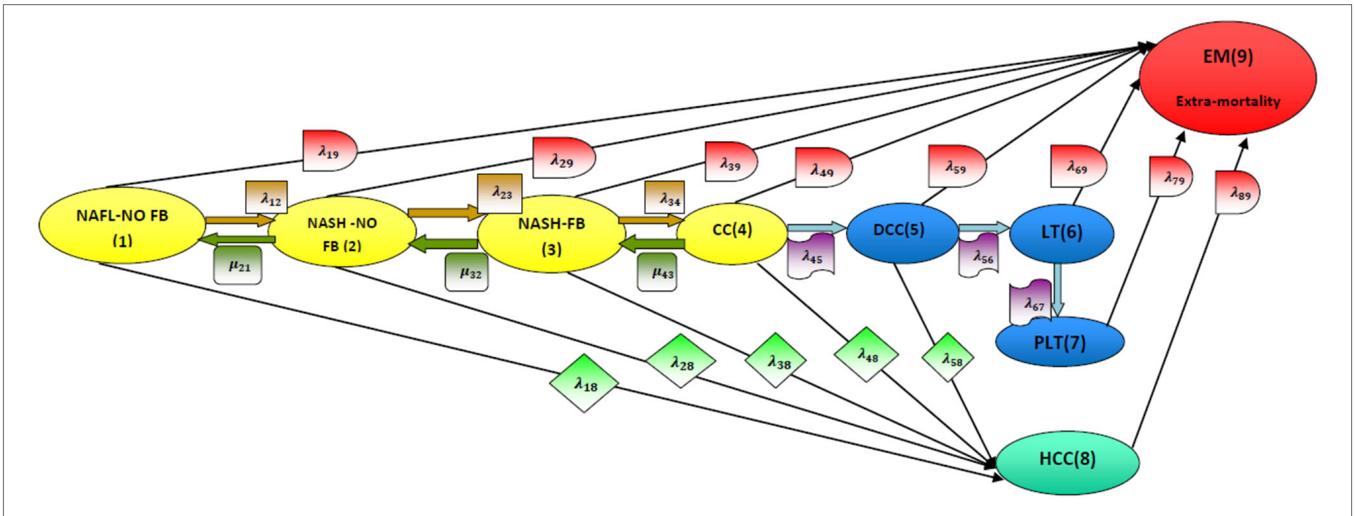

FIGURE 2 | Disease model structure [2]. NAFL-NO FB, non-alcoholic fatty liver with no fibrosis (stage 1); NASH-NO FB, non-alcoholic steatohepatitis with no fibrosis (stage 2); NASH-FB, non-alcoholic steatohepatitis with fibrosis (stage 3); CC, compensated cirrhosis (stage 4); DCC, de-compensated cirrhosis (stage 5); LT, liver transplant (stage 6); PLT, post-liver transplant (stage 7); HCC, hepatocellular carcinoma (stage 8); EM, extramortality (stage 9).

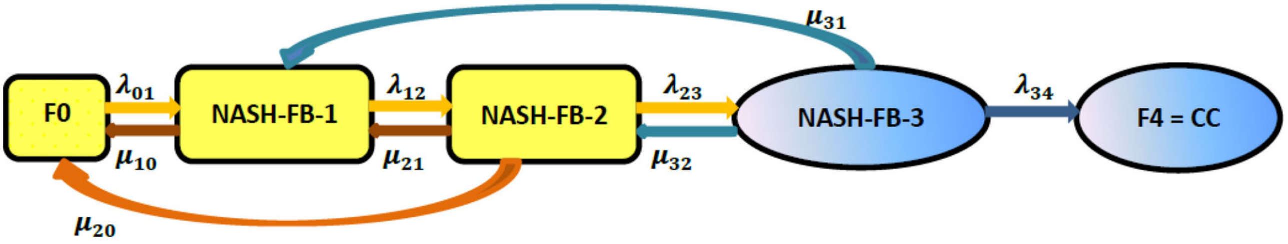

FIGURE 3 | NAFLD with the evolving fibrosis stages [4]. F0, no fibrosis (stage 0); NASH-FB-1, non-alcoholic steatohepatitis with mild fibrosis (stage 1); NASH -FB-2, NASH with moderate fibrosis (stage 2); NASH -FB-3, NASH with advanced or severe fibrosis (stage 3); CC, compensated cirrhosis (stage 4) which is the more severe or advanced form of fibrosis.

factors. Third, it recommends using continuous-time Markov chains to obtain the transition probabilities and predict the expected counts of patients in each state at a specific time point in the future. The counts of each transition can be modeled as a function of some explanatory variables reflecting the characteristics of the patients. The Poisson regression model specifies that each response y_i is drawn from a Poisson population with parameter λ_i, related to the covariates. The primary equation of the model is

$$P\left(Y = y_i | X_i\right) = \frac{\exp\left(-\lambda_i\right) \times \lambda_i^{y_i}}{y_i!}.$$

The most common formulation for the λ_i is the log-linear model:

$$\ln \lambda_i = X_i^{'}\beta = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \ldots . + b_k x_{im}.$$

where β is the k × 1 parameter vector, m is the number of predictors, and Xs are the predictors.

The expected number of events per period is given by:

$$E\left[y_i | X_i\right] = var\left[y_i | X_i\right] = \lambda_i = \exp\left(X_i^{'}\beta\right).$$

The observed counts in the transition counts matrix are used as response variables. The covariates are the risk factors for the fatty liver, where the participants are subjected to the same follow-up periods. Then, the estimated counts obtained from running the Poisson regression are used as input to estimate the transition probability matrix using the CTMC. The author clarifies this procedure using a hypothetical example in the form of an observational prospective longitudinal study.

Attia [8] used the same data in previous work. Still, in this article, the author discusses the issue of multicollinearity, the equidispersion Poisson of response variables in the presence of excess zeros, and more comparisons between this work and previous works. Finally, the author highlights the benefit of such analysis to pharmacoeconomic evaluation and healthcare economics.

MATERIALS AND METHODS

Patients

A total of one hundred fifty participants were followed up every year for 29 years, and during each visit, the characteristics

TABLE 1 | Summary of transition counts among the states.

Counts	Transition 0→ 1	Transition 1→ 2	Transition 2→ 3	Transition 3→ 4	Transition 1→ 0	Transition 2→ 1	Transition 3→ 2	Transition 2→ 0	Transition 3→ 1
0	63	96	121	128	121	127	130	138	139
1	58	43	23	22	24	17	17	11	9
2	25	9	4		3	5	3	1	2
3	4	2	2		2	1			

TABLE 2 | Observed transition counts of the patients over the 29 years.

	State 0	State 1	State 2	State 3	State 4	total
State 0	1,909	120	15	6	0	2,050
State 1	36	1,116	67	28	0	1,247
State 2	13	30	703	37	0	783
State 3	11	14	23	50	22	120
State 4	0	0	0	0	0	0
						4,200

of the participants were recorded like sex (0 = female, 1 = male), age, body mass index (BMI), low-density lipoprotein (LDL)-chol, homeostatic model assessment-insulin resistance (HOMA2-IR), and systolic blood and diastolic blood pressure. For each participant, the recorded value is the mean of the follow-up measurements. The age is the median value. The participants were followed up till the end of the study or having liver cirrhosis (F4).

Statistical Analysis

The relationship between the response variable (counts of transitions) and the predictors was non-linear as shown by Lowess smoother. Restricted cubic spline was used to obtain a suitable functional form of the predictors to fit a Poisson model using STATA 14. The CTMCs were used to obtain transition probability matrix and transition rate matrix. p-Value of <0.05 was considered statistically significant; all tests were two-sided tests (refer to **Appendix A**).

RESULTS

Summary of the transition counts among the states is shown in **Table 1**. The observed counts of the participants over the 29 years of follow-up are demonstrated in **Table 2**. The distribution of these counts was Poisson (mean = variance). The dispersion indices for the nine response variables ranged between 0.82 and 1.34. In **Appendix B**, more figures illustrate the dispersion of these response variables. They were also correlated with high statistical significance (p-value = 0.000) as shown in **Table 3**.

Initial observed rates are as follows:

$$\lambda_{01} = \frac{120}{2050} = 0.059,\ \lambda_{12} = \frac{67}{1247} = 0.0537,$$
$$\lambda_{23} = \frac{37}{783} = 0.047\ ,\ \lambda_{34} = \frac{22}{120} = 0.183$$

$$\mu_{10} = \frac{36}{1247} = 0.0288,\ \mu_{21} = \frac{30}{783} = 0.0383,$$
$$\mu_{32} = \frac{23}{120} = 0.191,\ \mu_{20} = \frac{13}{783} = 0.016,$$
$$\mu_{31} = \frac{14}{120} = 0.116$$

Although the response counts showed excess zeros, they fitted Poisson distribution and the zero inflated Poisson model. Their mean and variance were approximately equal as evident by their dispersion indices. So, Poisson regression was conducted for each transition count. Most statistical software packages conduct Poisson regression or generalized linear model utilizing log-link function with only one response variable. Thus, using STATA14, Poisson regression was conducted with one response variable. The response variable could not be used as a matrix to conduct the regression as multivariate regression with multiple response variables.

The application of Lowess smoother showed the non-linear relationship between the predictors and the response variables as shown in **Figure 4**. In **Supplementary Materials**, more figures illustrating these relationships between the different predictors and response variables are clearly shown (refer to also **Appendix B**).

The continuous predictors (age, BMI, HOMA2-IR, LDL-chol, and systolic and diastolic blood pressure) were highly correlated with a correlation coefficient of 0.99 and a condition number for data matrix (X'X) of 453.57. The condition number for the data matrix (X'X) constructed from the transformed variables used in the analysis (HOMAsp1, HOMAsp2, LDLsp2, sysPS2, diasPS2) was 54.89. These transformed variables were also highly correlated. However, the condition number did not exceed 100. Thus, this multicollinearity can be considered non-harmful, and it will not affect the analysis [9].

The observed counts were the response variables used to fit the Poisson regression model. For each transition count, the model that represented the most explainable covariates with their estimated beta coefficients and the corresponding incidence rate ratios were illustrated in **Appendix B**. The transitions were subdivided into progressive transitions and regressive transitions. The main important result is that HOMA2-IR is positively correlated with all progressive transitions and is inversely related to the regressive transitions, with control of other variables, as shown in **Tables 4**, **5**.

TABLE 3 | Correlation between the different response variables.

	F0→ F1	F1→ F2	F2→ F3	F3→ F4	F1→ F0	F2→ F1	F3→ F2	F2→ F0	F3→ F1
F0→ F1	1	0.794 (0.000)	0.794 (0.000)	0.70 (0.000)	0.798 (0.000)	0.719 (0.000)	0.693 (0.000)	0.559 (0.000)	0.548 (0.000)
F1→ F2	0.798 (0.000)	1	0.785 (0.000)	0.709 (0.000)	0.76 (0.000)	0.762 (0.000)	0.719 (0.000)	0.728 (0.000)	0.711 (0.000)
F2→ F3	0.794 (0.000)	0.785 (0.000)	1	0.82 (0.000)	0.99 (0.000)	0.928 (0.000)	0.868 (0.000)	0.768 (0.000)	0.791 (0.000)
F3→ F4	0.709 (0.000)	0.709 (0.000)	0.82 (0.000)	1	0.813 (0.000)	0.898 (0.000)	0.897 (0.000)	0.687 (0.000)	0.643 (0.000)
F1→ F0	0.798 (0.000)	0.76 (0.000)	0.99 (0.000)	0.813 (0.000)	1	0.911 (0.000)	0.867 (0.000)	0.753 (0.000)	0.778 (0.000)
F2→ F1	0.719 (0.000)	0.765 (0.000)	0.928 (0.000)	0.898 (0.000)	0.911 (0.000)	1	0.921 (0.000)	0.824 (0.000)	0.81 (0.000)
F3→ F2	0.693 (0.000)	0.719 (0.000)	0.868 (0.000)	0.897 (0.000)	0.867 (0.000)	0.921 (0.000)	1	0.798 (0.000)	0.796 (0.000)
F2→ F0	0.559 (0.000)	0.728 (0.000)	0.768 (0.000)	0.687 (0.000)	0.753 (0.000)	0.824 (0.000)	0.798 (0.000)	1	0.935 (0.000)
F3→ F1	0.548 (0.000)	0.711 (0.000)	0.791 (0.000)	0.643 (0.000)	0.778 (0.000)	0.81 (0.000)	0.796 (0.000)	0.935 (0.000)	1

In each cell, the Pearson correlation coefficient, for transitions among the different states, is shown with the significant p-value below this coefficient between the brackets.

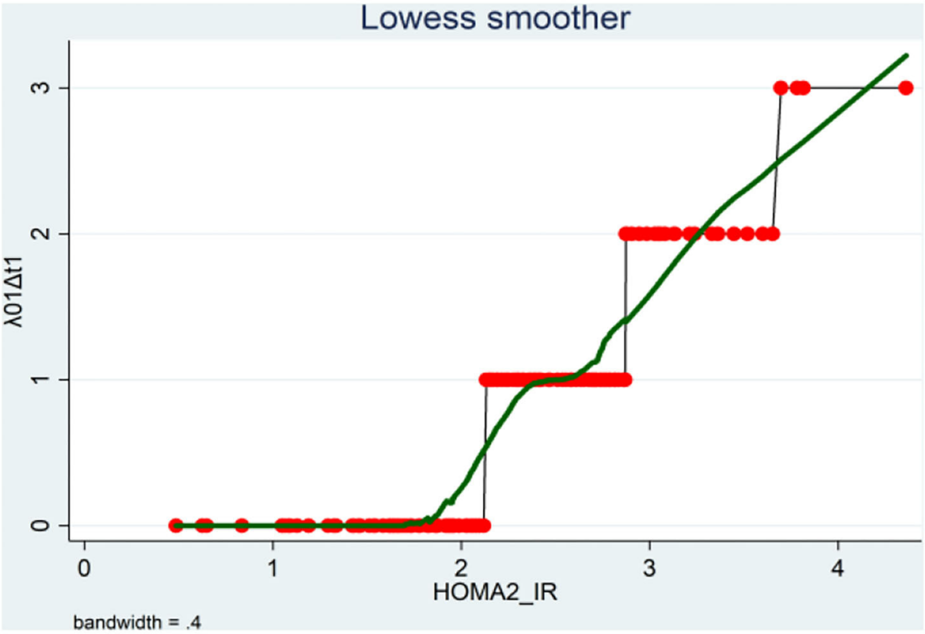

FIGURE 4 | Lowess smoother showing the non-linear relationship between the transition counts from F0 to F1 and the HOMA2-IR levels.

Progressive Transitions With Rates λ_{01}, λ_{12}, λ_{23}, λ_{34}

Persons with high insulin resistance (elevated HOMA2-IR) had 60 times the rate of transition from F0 to F1 compared to persons with normal level of HOMA2-IR (persons with normal insulin sensitivity), also the rate increased to 240 times for the rate of transition from F1 to F2, increased to 480 times for the rate of transition from F2 to F3, and increased to more than 50,000 times for the rate of transition from F3 to F4. Statistically speaking, the expected increase in log count of transition from F0 to F1 for one-unit increase in transformed HOMA is 4.096, which is highly statistically significant ($p = 0.000$). The expected increase in log count of transition from F1 to F2 for one-unit increase in transformed HOMA is 5.486, which is highly statistically significant ($p = 0.000$). The expected increase in log count of transition from F2 to F3 for one-unit increase in transformed

TABLE 4 | Parameters for each transition.

	LDLsp2	HOMAsp1	SysSP2	LDLsp2# HOMAsp1	LDLsp2# SysSP2	HOMAsp1# SysSP2
Transition from F0 to F1						
$\hat{b}$ co.(P)	0.523 (0.032)	4.096 (0.000)	−0.628 (0.070)	−0.179 (0.011)	0.003 (0.000)	0.151 (0.122)
CI for $\hat{b}$ co	(0.046, 1.000)	(3.452, 4.740)	(−1.308, 0.052)	(−0.317, −0.041)	(0.002, 0.003)	(−0.040, 0.342)
IRR	1.687	60.097	0.534	0.836	1.003	1.163
CI for IRR	(1.047, 2.718)	(31.569, 114.4)	(0.270, 1.054)	(0.728, 0.960)	(1.002, 1.003)	(0.960, 1.408)
Transition from F1 to F2						
$\hat{b}$ co.(P)	0.311 (0.432)	5.486 (0.000)	−0.314 (0.564)	−0.105 (0.367)		0.079 (0.616)
CI for $\hat{b}$ co	(−0.465, 1.086)	(4.366, 6.606)	(−1.383, 0.754)	(−0.332, 0.123)		(−0.231, 0.389)
IRR	1.364	241.179	0.730	0.901		1.083
CI for IRR	(0.628, 2.962)	(78.690, 739.192)	(0.251, 2.126)	(0.717, 1.131)		(0.794, 1.476)
Transition from F2 to F3						
$\hat{b}$ co.(P)	−1.480 (0.031)	6.174 (0.046)	2.497 (0.010)	0.390 (0.042)	−0.001 (0.687)	−0.655 (0.017)
CI for $\hat{b}$ co	(−2.823, −0.137)	(0.112, 12.237)	(0.602, 4.391)	(0.014, 0.766)	(−0.005, 0.004)	(−1.191, −0.118)
IRR	0.228	480.318	12.143	1.477	0.999	0.520
CI for IRR	(0.059, 0.872)	(1.118, 2.06e+5)	(1.826, 80.754)	(1.014, 2.151)	(0.995, 1.004)	(0.304, 0.889)
Transition from F3 to F4						
$\hat{b}$ co.(P)	0.452 (0.000)	10.866 (0.000)	0.073 (0.141)	−0.166 (0.000)		
CI for $\hat{b}$ co	(0.345, 0.559)	(8.119, 13.613)	(−0.024, 0.171)	(−0.201, −0.131)		
IRR	1.571	52375.984	1.076	0.847		
CI for IRR	(1.412, 1.748)	(3,357.9, 8.17e+5)	(0.976, 1.187)	(0.818, 0.877)		

B co. (p), the estimated B coefficient with p-value in the brackets; CI for B co., confidence interval for the estimated B coefficient; IRR, incidence rate ratio; CI for IRR stands for confidence interval for IRR; LDLsp2, the transformed LDL variable using restricted cubic spline method; HOMAsp1, the transformed HOMA2-IR variable using restricted cubic spline method; sysSp2, the transformed systolic blood pressure variable using restricted cubic spline method; LDLsp2 # HOMAsp1, interaction between the 2 variables.

HOMA is 6.174, which is not highly statistically significant ($p = 0.046$). The expected increase in log count of transition from F3 to F4 for one-unit increase in transformed HOMA is 10.866, which is highly statistically significant ($p = 0.000$).

Regressive Transitions With Rates μ_{10}, μ_{21}, μ_{32}, μ_{20}, μ_{31}

Persons with high insulin resistance (elevated HOMA2-IR) had 0.011 times the rate of transition from F1 to F0 compared to persons with normal level of HOMA2-IR (persons with normal insulin sensitivity), also the rate decreased to 0.037 times for the rate of transition from F2 to F1, decreased to 0.005 times for the rate of transition from F3 to F2, decreased to 0.066 times for the rate of transition from F2 to F0, and decreased to 0.084 times for the rate of transition from F3 to F1. Statistically speaking, the expected decrease in log count of transition from F1 to F0 for one-unit increase in transformed HOMA is 4.489, which is not statistically significant ($p = 0.13$). The expected decrease in log count of transition from F2 to F1 for one-unit increase in transformed HOMA is 3.288, which is not statistically significant ($p = 0.242$). The expected decrease in log count of transition from F3 to F2 for one-unit increase in transformed HOMA is 5.214, which is not statistically significant ($p = 0.103$). The expected decrease in log count of transition from F2 to F0 for one-unit increase in transformed HOMA is 2.713, which is highly statistically significant ($p = 0.000$). The expected decrease in log count of transition from F3 to F1 for one-unit increase in transformed HOMA is 2.476, which is highly statistically significant ($p = 0.000$).

Validation and Residual Analysis

Poisson model fitted the data. When comparing the full model to the null model, there was a marked decrease in the deviance goodness of fit. Also, the akaike information criteria (AIC) and bayesian information criteria (BIC) were less than their values in the null model, indicating the full model improvement. In addition, there was an increase in the pseudo-R^2, indicating the ability of the model to predict the outcome better than the null model. The output results of the null model for each of the transition counts are shown in **Tables 6**, **7**.

The observed rates were approximately equal to the estimated rates after running the Poisson model as shown in **Table 8**.

Analysis of residuals especially Pearson residuals, for all transitions, revealed that they were not normally distributed. The Q-Q plot for these residuals did not exhibit normality. The Pearson dispersion statistics for each count was less than one supporting no evidence of overdispersion of the fitted model despite the apparent excess zeros (**Appendix B**, Table 21). Generalized Poisson regression did not fit the data. In **Appendix C**, more figures of these residuals are presented [10, 11].

This observational study aims to obtain preliminary and explanatory ideas about the effects of each risk factor on the different transition counts among the states. This

TABLE 5 | Parameters for each transition.

	LDLsp2	HOMAsp2	SysSP2	LDLsp2# HOMAsp2	LDLsp2# SysSP2	HOMAsp2# SysSP2
Transition from F1 to F0						
B co.(P)	−0.454 (0.063)	−4.489 (0.130)	1.340 (0.000)	0.290 (0.002)	−0.010 (0.005)	−0.286 (0.048)
CI for B co	(−0.932, 0.024)	(−10.294, 1.316)	(0.729, 1.951)	(0.102, 0.478)	(−0.017, − 0.003)	(−0.571, −0.002)
IRR	0.635	0.011	3.820	1.337	0.990	0.751
CI for IRR	(0.394, 1.024)	(0.000, 3.730)	(2.074, 7.034)	(1.108, 1.612)	(0.983, 0.997)	(0.565, 0.998)
Transition from F2 to F1						
B co.(P)	−0.128 (0.499)	−3.288 (0.242)	0.913 (0.000)	0.152 (0.022)	−0.010 (0.003)	−0.114 (0.317)
CI for B co	(−0.499, 0.243)	(−8.800, 2.224)	(0.519, 1.307)	(0.022, 0.282)	(−0.017, −0.003)	(−0.338, 0.109)
IRR	0.880	0.037	2.492	1.164	0.990	0.892
CI for IRR	(0.607, 1.275)	(0.000, 9.244)	(1.681, 3.694)	(1.022, 1.326)	(0.983, 0.997)	(0.713, 1.116)
Transition from F3 to F2						
B co.(P)	0.302 (0.154)	−5.214 (0.103)	0.422 (0.142)	0.002 (0.984)	−0.012 (0.006)	0.132 (0.375)
CI for B co	(−0.113, 0.716)	(−11.478, 1.05)	(−0.142, 0.987)	(−0.198, 0.202)	(−0.02, −0.003)	(−0.16, 0.425)
IRR	1.352	0.005	1.526	1.002	0.998	1.142
CI for IRR	(0.893, 2.047)	(0.000, 2.859)	(0.868, 2.683)	(0.821, 1.223)	(0.98, 0.997)	(0.852, 1.529)

	LDLsp2	HOMAsp2	SysSP2	DiasSP2
Transition from F2 to F0				
B co.(P)	0.076 (0.335)	−2.713 (0.000)	−0.123 (0.010)	0.358 (0.001)
CI for B co	(−0.079, 0.231)	(−4.102, −1.324)	(−0.216, −0.030)	(0.143, 0.573)
IRR	1.079	0.066	0.884	1.430
CI for IRR	(0.924, 1.260)	(0.017, 0.266)	(0.806, 0.970)	(1.154, 1.773)
Transition from F3 to F1				
B co.(P)	0.145 (0.038)	−2.476 (0.000)	−0.129 (0.004)	0.276 (0.003)
CI for B co	(0.008, 0.282)	(−3.769, −1.183)	(−0.216, −0.042)	(0.093, 0.459)
IRR	1.156	0.084	0.879	1.318
CI for IRR	(1.008, 1.326)	(0.023, 0.306)	(0.805, 0.959)	(1.098, 1.582)

B co. (p), the estimated B coefficient with p-value in the brackets; CI for B co., confidence interval for the estimated B coefficient; IRR, incidence rate ratio; CI for IRR stands for confidence interval for IRR; LDLsp2, the transformed LDL variable using restricted cubic spline method; HOMAsp2, the transformed HOMA2-IR variable using restricted cubic spline method; sysSp2, the transformed systolic blood pressure variable using restricted cubic spline method; DiasSp2, the transformed diastolic blood pressure variable using restricted cubic spline method; LDLsp2 # HOMAsp1, the interaction between the 2 variables.

Poisson regression is not aiming for future prediction of counts. Although the residuals are not normally distributed, such analysis can give fair provisional ideas about the effects of the risk factors. The Poisson model gives unbiased estimates for the regression coefficients, but these coefficients' statistical significance should be cautiously taken.

CTMCs Utilize the Estimated Counts From Log-Linear Model to Obtain the Transition Probability Matrix

For each of the transitions from state (i) to state (j), where λ_{ij} denotes the counts of transition from state (i) to state (j), and after running the Poisson model, the linear predictor $\ln \lambda_{ij} = X_n^{'}B$ for each participant (n) is exponentiated, $E\left[y_n \middle| X_n\right] = \lambda_{ij} = \exp\left(X_n^{'}B\right)$, to obtain the expected counts of transition that this participant had accomplished during this 29 years. Then, the result is rounded to the appropriate integer and summed to get all counts for this transition and then compared to the observed counts accomplished by all participants.

The n_{i+} is the total marginal transition counts out of this state, which is assumed to be constant. The estimated counts from running the Poisson model will be substituted in the transition count table. Because the marginal counts are assumed to be the same and when using the initial rates calculated as $\theta_0 = \frac{n_{ij}}{n_{i+}}$ where the n_{ij} is the transition counts from state i to state j, the Q matrix can be estimated. (Hint: the numerators below are the estimated counts obtained from running the Poisson regression).

$$\hat{Q} = \begin{bmatrix} -\lambda_{01} & \lambda_{01} & 0 & 0 & 0 \\ \mu_{10} & -(\lambda_{12}+\mu_{10}) & \lambda_{12} & 0 & 0 \\ \mu_{20} & \mu_{21} & -(\lambda_{23}+\mu_{21}+\mu_{20}) & \lambda_{23} & 0 \\ 0 & \mu_{31} & \mu_{32} & -(\lambda_{34}+\mu_{32}+\mu_{31}) & \lambda_{34} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

TABLE 6 | Comparison between null and full model as regards the progressive transitions.

	Cons. Co.	C.I. of CO.	Log pseu. like	Pseudo R^2	Deviance GOF	Pearson GOF	AIC	BIC
Transition F0 to F1								
Null model	−0.223 $P = 0.007$	(−0.385, −0.061)	−171.273	0.000	149.236 $P = 0.4792$	122.5 $P = 0.944$	344.55	347.56
Full model	−9.510 $P = 0.000$	(−10.930, −8.089)	−110.43	0.355	27.55 $P = 1$	24.458 $P = 1$	234.86	255.94
Transition F1 to F2								
Null model	−0.806 $P = 0.00$	(−1.046, −0.566)	−130.82	0.000	146.133 $P = 0.551$	150.16 $P = 0.458$	263.64	266.65
Full model	−14.884 $P = 0.000$	(−17.555,−12.213)	−67.887	0.481	20.27 $P = 1$	18.12 $P = 1$	147.77	165.84
Transition F2 to F3								
Null model	−1.4 $P = 0.000$	(−1.767,−1.032)	−95.146	0.000	127.853 $P = 0.894$	194.08 $P = 0.007$	192.29	195.3
Full model	−20.866 $P = 0.004$	(−35.160, −6.572)	−37.87	0.6020	13.29 $P = 1$	12.42 $P = 1$	89.73	110.81
Transition F3 to F4								
Null model	−1.92 $P = 0.00$	(−2.307,−1.532)	−64.23	0.000	84.46 $P = 1$	128 $P = 0.89$	130.46	133.47
Full model	−34.034 $P = 0.000$	(−41.608, −26.459)	−26.97	0.58	9.94 $P = 1$	8.96 $P = 1$	63.94	78.99

TABLE 7 | Comparison between null and full model as regards the regressive transitions.

	Cons. Co.	C.I. of CO.	Log pseu. like.	Pseudo R^2	Deviance GOF	Pearson GOF	AIC	BIC
Transition F1 to F0								
Null model	−1.427 $P = 0.000$	(−1.795, −1.059)	−93.039	0.000	124.25 $P = 0.931$	189 $P = 0.015$	188.08	191.08
Full model	−5.916 $P = 0.000$	(−6.912, −4.921)	−38.14	0.59	14.46 $P = 1$	13.55 $P = 1$	90.29	111.36
Transition F2 to F1								
Null model	−1.609 $P = 0.000$	(−2.024, −1.195)	−83.54	0.000	117.021 $P = 0.975$	200 $P = 0.003$	169.08	172.09
Full model	−7.666 $P = 0.000$	(−8.875, −6.457)	−29.96	0.64	9.86 $P = 1$	8.97 $P = 1$	73.92	94.99
Transition F3 to F2								
Null model	−1.875 $P = 0.000$	(−2.307, −1.444)	−68.208	0.000	94.574 $P = 0.999$	166.13 $P = 0.16$	138.42	141.43
Full model	−7.363 $P = 0.000$	(−8.855, −5.871)	−26.37	0.61	10.89 $P = 1$	9.77 $P = 1$	66.74	87.81
Transition F2 to F0								
Null model	−2.446 $P = 0.000$	(−3.009, −1.882)	−45.487	0.000	66.3 $P = 1$	160. $P = 0.253$	92.97	95.98
Full model	−7.034 $P = 0.000$	(−8.015, −6.053)	−15.63	0.656	6.65 $P = 1$	7.36 $P = 1$	41.26	56.31
Transition F3 to F1								
Null model	−2.446 $P = 0.000$	(−3.048, −1.843)	−46.18	0.000	69.133 $P = 1$	183.15 $P = 0.029$	94.36	97.37
Full model	−7.584 $P = 0.000$	(−8.934, −6.235)	−14.18	0.693	5.14 $P = 1$	6.09 $P = 1$	38.36	53.42

Cons.Co, constant coefficient; C.I. of CO., confidence interval of constant; Log pseu.like., Log pseudolikelihood.

TABLE 8 | The comparison between observed and estimated response rate after fitting Poisson model.

		Comparison between observed and estimated progressive counts				Comparison between observed and estimated regressive counts	
		Observed response count	**Estimated mean response count**			**Observed response count**	**Estimated mean response count**
0→ 1	Mean	0.8	0.8	1→ 0	Mean	0.24	0.24
	Variance	0.658	0.619		Variance	0.305	0.314
1→ 2	Mean	0.45	0.45	2→ 1	Mean	0.2	0.2
	Variance	0.45	0.45		Variance	0.268	0.284
2→ 3	Mean	0.25	0.25	3→ 2	Mean	0.15	0.15
	Variance	0.32	0.318		Variance	0.171	0.173
3→ 4	Mean	0.15	0.15	2→ 0	Mean	0.09	0.09
	Variance	0.126	0.126		Variance	0.093	0.101
				3→ 1	Mean	0.09	0.09
					Variance	0.106	0.11

where

$$\begin{aligned}
\lambda_{01} &= \frac{120}{2050} = 0.059,\ \lambda_{12} = \frac{64}{1247} = 0.051,\\
\lambda_{23} &= \frac{35}{783} = 0.045\ ,\ \lambda_{34} = \frac{20}{120} = 0.167\\
\mu_{10} &= \frac{36}{1247} = 0.029,\ \mu_{21} = \frac{26}{783} = 0.033,\\
\mu_{32} &= \frac{19}{120} = 0.158,\ \mu_{20} = \frac{12}{783} = 0.015,\\
\mu_{31} &= \frac{13}{120} = 0.108
\end{aligned}$$

The probability matrix at any specific time point in the future can be obtained by exponentiation of this matrix because the chain is homogenous continuous-time Markov chains with constant rates over time. This result can also be obtained by solving the forward Kolmogorov differential equations, which will yield the same result as the exponentiation of the estimated Q matrix (refer to **Appendix D**).

The transition probability matrix is obtained by exponentiation of this estimated $\hat{Q}$ matrix after 1 year:

$$P\left(t = 1\right) = \exp\left(\hat{Q}t\right) = \begin{bmatrix} P_{00} & P_{01} & P_{02} & P_{03} & P_{04} \\ P_{10} & P_{11} & P_{12} & P_{13} & P_{14} \\ P_{20} & P_{21} & P_{22} & P_{23} & P_{24} \\ P_{30} & P_{31} & P_{32} & P_{33} & P_{34} \\ 0 & 0 & 0 & 0 & P_{44} \end{bmatrix}$$

$$= \begin{bmatrix} 0.944 & 0.055 & 0.001 & 0 & 0 \\ 0.027 & 0.925 & 0.047 & 0.001 & 0.0001 \\ 0.014 & 0.033 & 0.915 & 0.035 & 0.003 \\ 0.002 & 0.086 & 0.125 & 0.651 & 0.136 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Goodness of Fit for the Multistate Markov Model

To calculate goodness of fit for multistate model used in this example, it is like the procedure used in contingency table:

TABLE 9 | The expected transition counts after one year of the follow-up.

	State 0	State 1	State 2	State 3	State 4
State 0	1934.175	112.955	2.87	0	0
State 1	34.168	1153.101	58.484	1.122	0.125
State 2	11.275	25.604	716.367	27.248	2.506
State 3	0.276	10.356	14.94	78.144	16.284
State 4	0	0	0	0	0

Step 1: $H_0 =$ *future state does not depend on the current state*
$H_1 =$ *future state depends on the current state*
Step 2: After obtaining the estimated Q matrix, the probability matrix is calculated in time interval equals one because the participants' follow-up period was done every year.

$$p_{ij}\left(\triangle t = 1\right) = \exp\left(\hat{Q} \times \triangle t\right) = \begin{bmatrix} P_{00} & P_{01} & P_{02} & P_{03} & P_{04} \\ P_{10} & P_{11} & P_{12} & P_{13} & P_{14} \\ P_{20} & P_{21} & P_{22} & P_{23} & P_{24} \\ P_{30} & P_{31} & P_{32} & P_{33} & P_{34} \\ 0 & 0 & 0 & 0 & P_{44} \end{bmatrix}$$

$$= \begin{bmatrix} 0.9435 & 0.0551 & 0.0014 & 0 & 0 \\ 0.0274 & 0.9247 & 0.0469 & 0.0009 & 0.0001 \\ 0.0144 & 0.0327 & 0.9149 & 0.0348 & 0.0032 \\ 0.0023 & 0.0863 & 0.1245 & 0.6512 & 0.1357 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Step 3: Calculate the expected counts in this interval.

$$E_{ij} = n_{i+}\, P_{ij}\left(t\right).$$

$$n_{1+} = 2050,\ n_{2+} = 1247,\ n_{3+} = 783,\ n_{4+} = 120$$

Multiplying each row in the probability matrix with the corresponding total marginal counts in the observed transition counts table in the same interval yields the expected counts as shown in **Table 9**.

Step 4: The observed counts, O_{ij}, are shown in **Table 2**. The expected counts, E_{ij}, are obtained from the previous step and are shown in **Table 9**. Then, apply the Pearson statistic formula which yields a value of 1,140.097 with high statistical significance ($p = 0.000$). Apply $\sum_{i=1}^{5}\sum_{j=1}^{5}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = 1140.097 \sim \chi^2_{(5-1)(5-1)(.05)}$.

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

Health Economics

This transition probability matrix can predict the count of patients in each state at specific time point, for example, if a cohort of 6,000 patients with the following number in each state is $\left[3000\ 1800\ 1020\ \ 180\ 0\right]$, after 1 year the predicted counts will be $\left[2895\ 1879\ 1044\ \ 154\ 28\right]$. This count can be achieved by multiplying the initial count distribution of the patients with the transition probability calculated at the required specific time point, $p_{ij}\left(t\right) = \exp\left(\hat{Q}t\right)$

$$E\left[u_j\left(t\right)|u_j\left(0\right)\right] = \sum_{j=1,i=1}^{5} u_j\left(0\right)P_{ij}\left(t\right) \quad i,\ j = 1,..,5$$

Let $u\left(0\right)$ be the size of patients in a specific state at specific time $t = 0$. The initial size of patients is $U\left(0\right) = u_j\ \left(0\right)$, as there are 4 transient states (F0 to F3) and 1 absorbing state (F4), where $u_j\left(0\right)$ is the initial size or the number of patients in state j at time $t = 0$ given that $u_5\left(0\right) = 0$, i.e., initial size of patients in state 5 (absorbing state) is zero at initial time point = 0. As the transition or the movement of the patients among states is independent, at the end of the whole time interval $\left(0, t\right)$, there will be $u_j\left(t\right)$ patients in the transient states at time t, and there will also be $u_5\left(t\right)$patients in state 5 (F4 = liver cirrhosis) at time t.

In addition, the state probability distribution $\pi\ \ \left(t\right)$, which is the probability distribution for each state at a specific time point given the initial probability distribution $\pi\left(0\right)$, can be estimated by applying the following formula:

$$\pi\left(t\right) = \pi\left(0\right)P\left(t\right).$$

In this example, the cohort of 6,000 patients has initial probability distribution of $\left[0.5\ 0.3\ 0.17\ \ 0.03\ 0\right]$, after 1 year, the state probability distribution will be $\left[0.4825\ 0.3131\ 0.174\ \ 0.0257\ 0.0046\right]$.

Pharmaco-economic evaluation can be assessed in three categories: the cost-benefit analysis, the cost-effectiveness analysis, and the cost-utility analysis. The evaluation utilizes the predicted number of patients in each state estimated every year, the state probability distribution predicted every year, the costs of investigations and treatments, and the quality adjusted life years for the patients [12, 13].

This approach differs from the one used by Rustgi et al. [14] who depends on calculating the cost-effectiveness analysis by following a cohort of patients, all starting at the same initial state till death. While in the approach proposed in this article, sampling the population and estimating the transition probability matrix to predict the counts in the future, any cohort of patients can be followed up utilizing the information gained from sampling the high-risk population.

DISCUSSION

The following discussion elucidates the agreements and comparisons between the findings in this study with the findings in the previous one high-lightening the effects of various factors on progression rate of fibrosis in NAFLD patients.

Hui et al. [15] conducted a study on 17 patients who had previous liver biopsy showing evidence of steatosis with or without the presence of necroinflammation and fibrosis. Those patients underwent second liver biopsies with a median of 6 years apart (range: 3.8–8 years). More than half of them developed progressive fibrosis compared to the initial biopsy; because that these patients suffered from steatohepatitis, although there was no significant correlation between the degree of steatohepatitis and the degree of fibrosis between the two biopsies. However, the correlation was significant between the initial stage of fibrosis and the fibrosis grade in the second biopsy. Also, the clinical and laboratory parameters were not statistically significant between the recorded values during the first and the second biopsies. The changes in these parameters also showed no significant correlation with changes in the scores of steatosis, necroinflammation, or fibrosis. There was a negative correlation, although non-significant, between the change in the score of fibrosis and each of the changes: in the BMI, plasma total cholesterol levels, plasma triglyceride levels, and glycosylated hemoglobin. During the follow-up, two patients developed type II diabetes and one developed hypertension but without progression of fibrosis, their initial biopsy revealed F0, and the second one was also F0. Another patient developed type II diabetes with evolution of the fibrosis from F0 to F2, and another 2 patients developed hypertension with advancement of fibrosis from F0 to F1.

Fassio et al. [16] conducted a study on 22 patients who had liver biopsy with evidence of NASH and found that 31.8% (7 patients = P group, progressors) had progression of liver fibrosis over a median follow-up of 4.7 years. The other group was 68.2% (15 patients = NP group, non-progressors) and did not progress over a median follow-up of 4.3 years. The rate of progression in the entire population was estimated as 0.059 fibrosis units per year (mean difference in fibrosis score divided by mean interval in years between the first and second biopsies = 0.32/5.34 = 0.059), the rate of progression in the P group was 1.85/6.59 = 0.28. There was no statistical difference as regards the clinical, biochemical, grade of steatosis, and grade of inflammation between the two groups except for the presence of obesity and higher BMI (progressor was more obese with higher BMI than the non-progressor) whether this was performed during the initial liver biopsy or the final liver biopsy. Within each group, the gradients between the final and basal results were not statistically significant as regards the clinical, biochemical, grade of steatosis, and grade of inflammation between the two groups including the BMI.

Adams et al. [7] conducted a study on 103 patients who had performed two liver biopsies with mean follow-up period of 3.2 ± 3 years (range = 0.7–21) between the first and the second biopsies. A total of 38 patients were progressors, 35 patients were stable, and 30 patients were regressors. No clinical or biochemical variables were statistically different among the progressors, stable, and regressors. The rate of fibrosis change varied from −2.05 to 1.7 stages/year and calculated as stated in the introduction. Using univariate regression model, the presence of diabetes, AST/ALT ratio, steatosis grades, and fibrosis stage were the only significant variables. By multivariate linear regression analysis and adjusting for age and BMI, only the presence of diabetes and earlier fibrosis stage were significantly associated with a higher rate of fibrosis progression. He also found no significant correlation between rate of progression and HOMA.

There are many studies performed by Ekstedt et al. [17], Teli et al. [18], Pais et al. [19], Argo et al. [20], Evans et al. [21], Hamaguchi et al. [22], and Wong et al. [23]. The reader can refer to them (refer to **Appendix E**).

The findings of the present study demonstrate that HOMA2-IR has a positive and a statistically significant effect on progression of fibrosis among the different states. Running multivariate Poisson regression reveals that the main players for progression are the HOMA2-IR, LDL-chol, and systolic blood pressure explaining about 35–60% of variability in the rates of progression. However, HOMA2-IR has a negative effect that is not statistically significant on the rate of remission or regression from F1 to F0, from F2 to F1, and from F3 to F2, but it is statistically significant on the rate of remission from F2 to F0 and from F3 to F1. Poisson regression model explained that the same factors and their interactions were responsible for about 60–70% of variability in the rates of remission among the states. The high HOMA2-IR levels significantly decrease the effects of high LDL levels on the progression rate from F0 to F1 and from F3 to F4. Thus, this interaction can be a protective mechanism to slow down the progression rate of fibrosis. The low HOMA2-IR levels significantly increase the effect of low LDL levels on the remission rate from F1 to F0 and from F2 to F1. Thus, this interaction can be a protective mechanism to accelerate the remission rate of fibrosis. The rate of fibrosis decreases with the help of rigorous control of the blood level of insulin, glucose, cholesterol, and blood pressure. The high levels of systolic blood pressure significantly decrease the effect of low LDL levels on the remission rate of fibrosis from F1 to F0, from F2 to F1, and from F3 to F2. Thus, controlling the most harmful factors like hyperinsulinemia and hypercholestrolemia, even in the absence of strict control of hypertension, can still benefit repressing the fibrogenesis. Lifestyle modification, in the form of physical exercise and a low caloric diet, and controlling the risk factors greatly impact arresting the process of fibrogenensis.

The newly emerging anti-fibrotic drugs will also help physicians treat fibrogenesis. In the FLINT study conducted on 283 non-cirrhotic patients taking obeticholic acid (OCA), 25 mg daily; the improvement in the histology detected by NAFLD activity score (NAS) was two points or more with no deterioration of fibrosis, and 35% of patients taking OCA had a decrement in fibrosis score by at least one stage in comparison with 19%in the placebo arm. REGENERATE study (still in progress, with the estimated primary completion date is on September 2025 as shown on clinicaltrials.gov official site) will evaluate safety and efficacy of obeticholic acid (OCA) in NASH patients with fibrosis who are randomized to a daily dose of 25 mg, 10 mg, and placebo, with endpoints like amelioration of fibrosis by at least one stage and decaying of NASH with no deterioration of fibrosis. At 18 month of randomization, liver biopsy revealed statistically significant histological amelioration of fibrosis and decaying of NASH with no deterioration in fibrosis for both 10 and 25 mg doses. In the GOLDEN study, conducted on 274 NASH patients, 120 mg elafibranor taken daily for 52 weeks induced decaying of moderate to severe NASH in a meaningfully higher percentage of patients than placebo; these patients also showed lowering in fibrosis stage compared to non-resolving NASH patients. The RESOLVE-IT trial (last update was on 30 November 2020, as shown on clinicaltrials.gov official site, but the study is still in progress according to Guirguis et al. [24]) emerged in May 2020 had shown that 19.2% of patients, on 120 mg daily elafibranor, had NASH decay without deterioration of fibrosis compared to 14.7% in the placebo group, which was not statistically significant. Furthermore, 24.5% of patients had shown fibrosis amelioration of more than one stage compared to 22.4% in the placebo group, which was also not statistically significant. In CENTAUR trial, conducted over 289 patients taking cenicriviroc (CVC), 150 mg daily and placebo for 52 weeks, no comparative betterment in NAS between NASH group and placebo was seen; however, there was one stage or more amelioration of fibrosis with no deterioration of NASH in the group taking the CVC compared to placebo group. The AURORA trial (primary completion dates were October 2021 according to clinicaltrials.gov site while the completion date will be October 2028 according to Guirguis et al. [24]) will evaluate long-term safety and efficacy of 150 mg daily CVC for the treatment of fibrosis in NASH adult at 2 phases: the first has endpoint of at least one stage amelioration of fibrosis without deterioration of NASH at month 12, and phase 2 has endpoint that is cirrhosis, liver-related outcome as HCC, and all causes of mortality. In a small, open-label, randomized phase II trial including 72 biopsy-proven NASH patients (NAS $\geq$ 5 and stage 2–3 liver fibrosis) receiving 18 mg daily selonsertib for 24 weeks, there was a significant improvement in liver disease activity, fibrosis, stiffness, liver fat content, and progression to cirrhosis [25].

FLINT, GOLDEN, and CENTAUR are phase IIb placebo-controlled randomized control trials (RCTs), whereas REGENERATE, RESOLVE-IT, and AURORA are randomized, placebo-controlled, double-blinded, multicenter phase III trials.

The distribution of the counts was Poisson distribution (mean = variance); that is to mean, these counts were equidispersed. However, all the counts showed excess zeros except for the transition from F0 to F1 where the zeros constituted 42% of the total count of this transition. Tlhaloganyang and Sakia found that the equidispersed counts data with excessive zeros can be modeled with Poisson regression, the best model to represent the data. Also, the AIC scores obtained by them after running Poisson regression, on their tested data whether simulated or real, were less than the AIC scores after running ZIP on the same datasets [26]. In this article, the predictors were normally distributed, and applying the restricted cubic spline transformation was used to better specify the functional form of these predictors. The raw predictors and the transformed predictors were highly correlated. But the condition number obtained from the transformed predictors is below 100, which is not harmful for the analysis as shown in the Results. Vatcheva et al. [27] highlighted the fact that the majority of researchers do not mention the multicollinearity diagnostics when running the regression models, discussed the causes and effects of this lack, and proposed some remedies to treat multicollinearity such as: principal component analysis, partial least squares regression, and ridge regression analysis. Akram et al. [28] used principal component ridge type estimator for the inverse Gaussian regression model. Many investigators such as, Liu [29], Kibria and Lukman [30], and Lukman et al. [31] had proposed different techniques to manage the multicollinearity problem between the predictors when running regression models. Some of them, who developed methods for Poisson regression, are Månsson and Shukur [32], Månsson et al. [33], Lukman et al. [34], Lukman et al. [35], and Qasim et al. [36]. In this paper, none of these methods were used as the Poisson model was mainly used to give preliminary vision about the effects of the high-risk factors on the transition counts. Also, it was not used for prediction, and the condition number was <100. Once the estimated counts were obtained, they were fed to the CTMC to estimate the transition rate matrix and transition probability matrix at any specified time point. Thus, physicians can follow a cohort of any patients in various states and obtain their state probability distribution at different time points.

The strength of this study is the conduction of multiple frequent repeated observations over a long period of follow-up on a large number of high-risk participants for developing NAFLD and performing a liver biopsy during each visit. Although this may be realistically infeasible during each visit, non-invasive techniques [37, 38] can substitute the invasive liver biopsy. The advantage of techniques like MRI and machine learning [39], to assess the liver texture and correlate these findings with the histological findings in liver biopsy, can overcome this weakness. Liver biopsy can also be reserved in situations where non-invasive tests are inconclusive. These non-invasive tests decrease the number of liver biopsies each patient may encounter. The proposed follow-up period is too long to wait for the obtained results, which can be overcome by using adaptive clinical trials. The weakness of the study is the presence of dependency among the response variables which was not treated by the statistical analysis used in this study. A copula modeling discrete random vectors like the counts in this study can be used in future analysis. However, a copula of discrete vectors is not fully identifiable and thus causes serious inconsistencies [40], especially when modeling nine variables like the variables used in this study.

CONCLUSION

In the present study, running Poisson regression model is used to obtain the expected counts of transition among states. These counts are used as input into the homogenous CTMC. Using this CTMC, the transition rate matrix is estimated, and thus, the probability of progression of participants from specific state to another one at specific time point can be estimated by exponentiation of this rate matrix. This probability matrix at any specific time point multiplied by the initial probability distribution of a cohort of patients can be used to predict the number of the participants in each state later on at different time points. This predicted number of participants helps health policymakers and insurance managers allocate the human and financial resources to investigate and treat the high-risk patients for developing NAFLD. The Poisson regression model relates these high-risk covariates to the transition rates among states. Also, this approach can be used in the clinical trials to assess the effectiveness of the newly emerging anti-fibrotic drugs. The epidemiologists can utilize this methodology to estimate the effect of risk factors on the incidence rates of progression and remission among the different states of liver fibrosis due to NAFLD.

This hypothetical study is coded by stata-14 and is published in code ocean site with the following URL: https://codeocean.com/capsule/4752445/tree/v3.

The code to estimate the Q transition rate matrix for the observed transition counts using continuous-time Markov chains is published in the code Ocean site with following URL: https://codeocean.com/capsule/6377472/tree/v2.

The code for solving the forward Kolmogorov equations using the estimated Q rate matrix is published in the code Ocean site with following URL: https://codeocean.com/capsule/7258626/tree/v1.

The dataset is present on IEEE Data Port site with the following URL: https://ieee-dataport.org/documents/fibrosis-nfld#files, with the following doi: 10.21227/dr5j-gs46.

REGENERATE study URL: https://clinicaltrials.gov/ct2/show/NCT02548351.

RESOLVET-IT study URL: https://clinicaltrials.gov/ct2/show/NCT02704403.

A medical appendix briefly clarifies the stages of fibrosis due to NAFLD. See also the presentation (in the **Supplementary Materials**).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

IA carried out the conceptualization by formulating the goals, aims of the research article, formal analysis by applying the statistical, mathematical and computational techniques to synthesize and analyze the hypothetical data, the methodology by creating the model, software programming and implementation, supervision, writing, drafting, editing, preparation, and creation of the presenting work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2022.899247/full#supplementary-material

REFERENCES

1. Younossi ZM, Gramlich T, Matteoni CA, Boparai N, McCullough AJ. Nonalcoholic fatty liver disease in patients with type 2 diabetes. *Clin Gastroenterol Hepatol.* (2004) 2:262–5. doi: 10.1016/S1542-3565(04)00014-X
2. Younossi ZM, Blissett D, Blissett R, Henry L, Stepanova M, Younossi Y, et al. The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. *Hepatology.* (2016) 64:1577–86. doi: 10.1002/hep.28785
3. Attia IM. Novel approach of multistate markov chains to evaluate progression in the expanded model of non-alcoholic fatty liver disease. *Front Appl Math Stat.* (2022) 7:766085. doi: 10.3389/fams.2021.766085
4. Younossi ZM, Tampi RP, Racila A, Qiu Y, Burns L, Younossi I, et al. Economic and clinical burden of nonalcoholic steatohepatitis in patients with type 2 diabetes in the US. *Diabetes Care.* (2020) 43:283–9. doi: 10.2337/dc19-1113
5. Singh S, Allen AM, Wang Z, Prokop LJ, Murad MH, Loomba R. Fibrosis progression in nonalcoholic fatty liver vs nonalcoholic steatohepatitis: a systematic review and meta-analysis of paired-biopsy studies. *Clin Gastroenterol Hepatol.* (2015) 13:643–54. doi: 10.1016/j.cgh.2014.04.014
6. Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. *J Am Stat Assoc.* (1985) 80:863–71. doi: 10.1080/01621459.1985.10478195
7. Adams LA, Sanderson S, Lindor KD, Angulo P. The histological course of nonalcoholic fatty liver disease: a longitudinal study of 103 patients with sequential liver biopsies. *J Hepatol.* (2005) 42:132–8. doi: 10.1016/j.jhep.2004.09.012
8. Attia IM. Prognostic factors for evolution of non alcoholic fatty liver disease patients utilizing poisson regression and continuous time markov chains. *Int J Res Eng Sci.* (2021) 9:61–71.
9. Lazaridis A. A note regarding the condition number: the case of spurious and latent multicollinearity. *Qual Quant.* (2007) 41:123–35. doi: 10.1007/s11135-005-6225-5
10. Hilbe JM. Modeling count data. In: Lovric, editor. *International Encyclopedia of Statistical Science.* Springer (2011). 836–9.
11. Cameron AC, Trivedi PK. *Regression Analysis of Count Data.* 2nd ed. Cambridge University Press (2013).
12. O'Hara J, Finnegan A, Dhillon H, Ruiz-Casas L, Pedra G, Franks B, et al. Cost of non-alcoholic steatohepatitis in Europe and the USA: the GAIN study. *JHEP Rep.* (2020) 2:100142. doi: 10.1016/j.jhepr.2020.100142
13. Noureddin M, Jones C, Alkhouri N, Gomez EV, Dieterich DT, Rinella ME, et al. Screening for nonalcoholic fatty liver disease in persons with type 2 diabetes in the United States is cost-effective: A comprehensive cost-utility analysis. *Gastroenterology.* (2020) 159:1985–7. doi: 10.1053/j.gastro.2020.07.050
14. Rustgi VK, Duff SB, Elsaid MI. Cost-effectiveness and potential value of pharmaceutical treatment of nonalcoholic fatty liver disease. *J. Med. Econ.* (2022) 25:347–355. doi: 10.1080/13696998.2022.2026702
15. Hui AY, Wong V, Chan H, Liew C, Chan J, Chan F, et al. Histological progression of non-alcoholic fatty liver disease in Chinese patients. *Aliment Pharmacol Ther.* (2005) 21:407–13. doi: 10.1111/j.1365-2036.2005.02334.x
16. Fassio E, Álvarez E, Domínguez N, Landeira G, Longo C. Natural history of nonalcoholic steathepatitis: a longitudinal study of repeat liver biopsies. *Hepatology.* (2004) 40:820–6. doi: 10.1002/hep.1840400411
17. Ekstedt M, Franzén LE, Mathiesen UL, Thorelius L, Holmqvist M, Bodemar G, et al. Long-term follow-up of patients with NAFLD and elevated liver enzymes. *Hepatology.* (2006) 44:865–73. doi: 10.1002/hep.21327
18. Teli MR, James OFW, Burt AD, Bennett MK, Day CP. The natural history of nonalcoholic fatty liver: a follow-up study. *Hepatology.* (1995) 22:1714–9. doi: 10.1002/hep.1840220616
19. Pais R, Charlotte F, Fedchuk L, Bedossa P, Lebray P, Poynard T, et al. A systematic review of follow-up biopsies reveals disease progression in patients with non-alcoholic fatty liver. *J Hepatol.* (2013) 59:550–6. doi: 10.1016/j.jhep.2013.04.027
20. Argo CK, Northup PG, Al-Osaimi AMS, Caldwell SH. Systematic review of risk factors for fibrosis progression in non-alcoholic steatohepatitis. *J Hepatol.* (2009) 51:371–9. doi: 10.1016/j.jhep.2009.03.019
21. Evans CDJ, Oien KA, MacSween RNM, Mills PR. Non-alcoholic steatohepatitis: A common cause of progressive chronic liver injury? *J Clin Pathol.* (2002) 55:689–92. doi: 10.1136/jcp.55.9.689
22. Hamaguchi E, Takamura T, Sakurai M, Mizukoshi E, Zen Y, Takeshita Y, et al. Histological course of nonalcoholic fatty liver disease in Japanese patients: Tight glycemic control, rather than weight reduction, ameliorates liver fibrosis. *Diabetes Care.* (2010) 33:284–6. doi: 10.2337/dc09-0148
23. Wong VW-S, Wong GL-H, Choi PC-L, Chan AW-H, Li MK-P, Chan H-Y, et al. Disease progression of non-alcoholic fatty liver disease: a prospective study with paired liver biopsies at 3 years. *Gut.* (2010) 59:969–74. doi: 10.1136/gut.2009.205088
24. Guirguis E, Grace Y, Bolson A, DellaVecchia MJ, Ruble M. Emerging therapies for the treatment of nonalcoholic steatohepatitis: a systematic review. *Pharmacotherapy.* (2021) 41:315–28. doi: 10.1002/phar.2489
25. Alkhouri N, Poordad F, Lawitz E. Management of nonalcoholic fatty liver disease: lessons learned from type 2 diabetes. *Hepatol Commun.* (2018) 2:778–85. doi: 10.1002/hep4.1195
26. Tlhaloganyang BP, Sakia RM. Zero inflated Poisson distribution in equidispersed data with excessive zeros. *Res J Math Stat Sci.* (2020) 8:31–4.
27. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology.* (2016) 6:227. doi: 10.4172/2161-1165.1000227
28. Akram M, Amin M, Lukman A, Afzal S. Principal component ridge type estimator for the inverse Gaussian regression model. *J Stat Comput Simul.* (2022) 1–30. doi: 10.1080/00949655.2021.2020274
29. Liu K. Using liu-type estimator to combat collinearity. *Commun Stat Theory Methods.* (2003) 32:1009–20. doi: 10.1081/STA-120019959
30. Kibria BMG, Lukman AF. A new ridge-type estimator for the linear regression model: simulations and applications. *Scientifica.* (2020) 2020:e9758378. doi: 10.1155/2020/9758378
31. Lukman AF, Ayinde K, Binuomote S, Clement OA. Modified ridge-type estimator to combat multicollinearity: application to chemical data. *J Chemom.* (2019) 33:e3125. doi: 10.1002/cem.3125
32. Månsson K, Shukur G. A Poisson ridge regression estimator. *Econ Model.* (2011) 28:1475–81.
33. Månsson K, Kibria BMG, Sjölander P, Shukur G. Improved liu estimators for the poisson regression model. *Int J Stat Prob.* (2012) 1:2–6. doi: 10.5539/ijsp.v1n1p2
34. Lukman AF, Adewuyi E, Månsson K, Kibria BMG. A new estimator for the multicollinear Poisson regression model: simulation and application. *Sci Rep.* (2021) 11:3732. doi: 10.1038/s41598-021-82582-w
35. Lukman AF, Aladeitan B, Ayinde K, Abonazel MR. Modified ridge-type for the Poisson regression model: simulation and application. *J Appl Stat.* (2021) 1–13. doi: 10.1080/02664763.2021.1889998
36. Qasim M, Kibria BMG, Månsson K, Sjölander P. A new Poisson Liu Regression Estimator: method and application. *J Appl Stat.* (2020) 47:2258–71. doi: 10.1080/02664763.2019.1707485
37. Petitclerc L, Sebastiani G, Gilbert G, Cloutier G, Tang A. Liver fibrosis: review of current imaging and MRI quantification techniques. *J Magn Reson Imaging.* (2017) 45:1276–95. doi: 10.1002/jmri.25550
38. Musso G, Gambino R, Cassader M, Pagano G. Meta-analysis: Natural history of non-alcoholic fatty liver disease (NAFLD) and diagnostic accuracy

of non-invasive tests for liver disease severity. *Ann Med.* (2011) 43:617–49. doi: 10.3109/07853890.2010.518623

39. Schawkat K, Ciritsis A, von Ulmenstein S, Honcharova-Biletska H, Jüngst C, Weber A, et al. Diagnostic accuracy of texture analysis and machine learning for quantification of liver fibrosis in MRI: correlation with MR elastography and histopathology. *Eur Radiol.* (2020) 30:4675–85. doi: 10.1007/s00330-020-06831-8
40. Geenens G. Copula modeling for discrete random vectors. *Depend Model.* (2020) 8:417–40. doi: 10.1515/demo-2020-0022

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.