Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Iman M. Attia (imanattiathesis1972@gmail.com)
Institute of Statistical Studies and Research, Cairo University https://orcid.org/0000-0002-7333-9713

Research Article

Keywords: log-linear model, Multistate model, Non-Alcoholic Fatty Liver Disease, Poisson regression, Continuous time Markov chains, longitudinal studies, Maximum Likelihood estimation, Mean Sojourn Time, HOMA2-IR.

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1416828/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

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

Iman M. Attia *, ORCID , 0000-0002-7333-9713

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

Corresponding author : Iman M. Attia (imanattiathesis1972@gmail.com)

ABSTRACT *:* In the present paper, the deleterious effects of obesity, type 2diabetes and insulin resistance, systolic and diastolic hypertension on the rate of progression of fibrosis in non-alcoholic fatty liver disease (NAFLD) patients are illustrated using a new approach utilizing the Poisson regression to model the transition rate matrix. The observed counts in the transition counts matrix are used as 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 28 years recording their demographic characteristics and their timeline follow up are demonstrated. The findings revealed that insulin resistance expressed by MOMA-IR 2 had the most deleterious effects among other factors for increasing the rate of forward progression of patients from state 1 to state 2 as well as from state 2 to state 3 and from state 3 to state 4. The higher the level of HOMA-IR is, the more rapid the rate of progression is. This analysis helps the health policy makers and medical insurance managers to allocate the financial and human resources for investigating and treating high risk patients for 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.

Categories: Endocrinology, Hepatology, Epidemiology, Biostatistics, and Bioinformatics

Key Words: log-linear model, Multistate model, Non-Alcoholic Fatty Liver Disease, Poisson regression, Continuous time Markov chains, longitudinal studies, Maximum Likelihood estimation, Mean Sojourn Time, HOMA2-IR.

Introduction And Background

Continuous time Markov chains (CTMCs) are valuable mathematical and statistical tools. They are of great potentiality to be used for evaluation of disease progression over time. NAFLD is an increasingly worldwide epidemic, paralleling the rise in 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 carbohydrate diets. The pathological milestone for NAFLD is insulin resistance and hyperinsulinemia. This will eventually results in type II diabetes with all its 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].

NAFLD can be modeled using the simplest form for health, disease, and death model. It is composed of 4 states. One state for susceptible individuals with risk factors like type 2 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 is shown in Fig. (1):

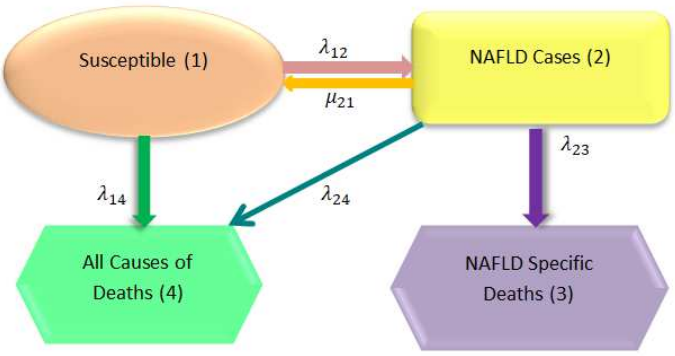

FIGURE 1: General Model Structure [2]

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 Fig.(2).

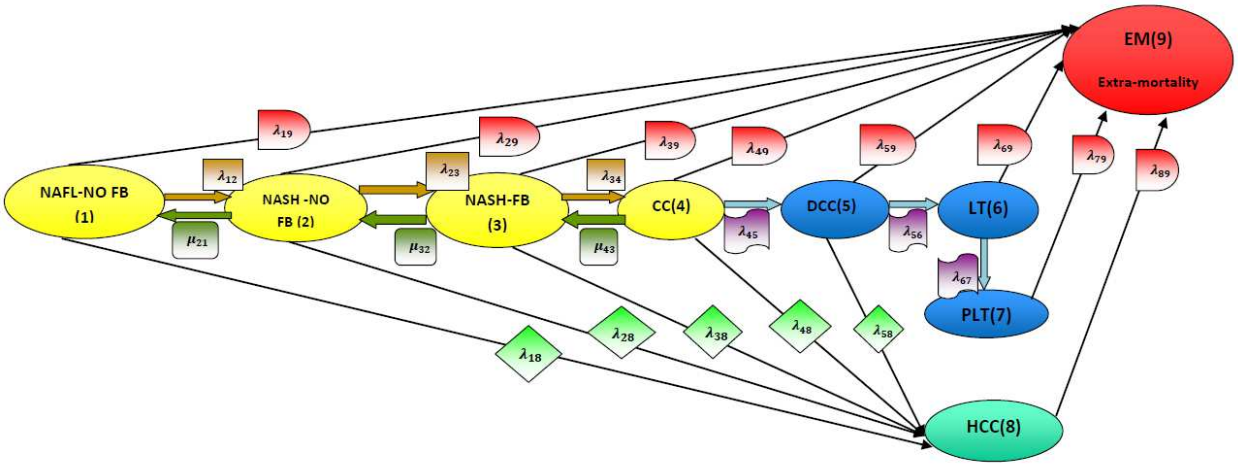

FIGURE 2: Disease Model Structure [2]

NAFL-NO FB = nonalcoholic fatty liver 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).

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. A subset of the early states, when the fibrous tissue develops, is used to relate these risk factors to the rates. Definition of each state is shown in Fig.(3)[4][5].

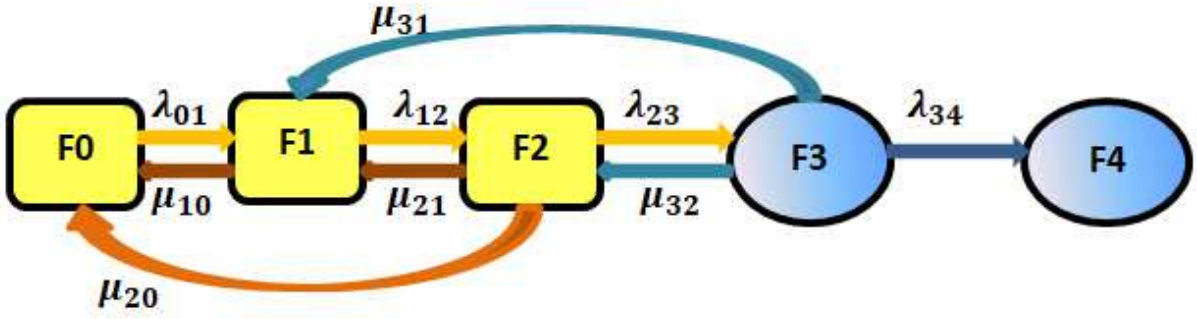

FIGURE 3: NAFLD with the evolving fibrosis stages.[3]

F0= no fibrosis (stage 0) whether hepatic steatosis is present or not. NASH-FB-1 =nonalcoholic 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.

Kalbfleisch & 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 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]. And either univariate or multivariate linear regression was used to relate the risk factors with the rate of progression. 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 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 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 , which is related to the covariates. The primary equation (1) of the model is

$$P(Y = y_i|x_i) = \frac{exp(-\lambda_i) \times \lambda_i^{y_i}}{y_i!} . \qquad (1)$$

The most common formulation (2) for the λ_i is the log-linear model:

$$\ln \lambda_i = x_i' B . \qquad (2)$$

And the expected number of events per period is given by (3) :

$$E[y_i|x_i] = var[y_i|x_i] = \lambda_i = \exp(x_i' B). \qquad (3)$$

The observed counts in the transition counts matrix are used as response variables and the covariates are the risk factors for fatty liver where the participants are subjected to same periods of follow up. Then the estimated counts obtained from running the Poisson regression are used as input to estimate the transition probability matrix

using the CTMC. To expound this procedure, a hypothetical example is used in the form of an observational prospective longitudinal study.

AI [8] used the same data in previous work , but in this article the author discusses the issue of multi-collinearity, the equi-dispersion Poisson of response variables in the presence of excess zeros, more comparisons between this work and previous works, and finally the author high lightens the benefit of such analysis to pharmacoeconomic evaluation and healthcare economics.

Review

Materials and methods

Patients

One hundred fifty participants were followed up every year for 28 years, and at each visit the characteristics of the participants were recorded like: sex (0=female, 1=male), age, BMI, LDL-chol, HOMA2_IR, systolic blood pressure as well as the diastolic pressure .For each participant the recorded value is the mean of the follow up measurements. The age is the median value.

Statistical analysis

The relationship between the response variable (counts of transitions) and the predictors were nonlinear 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. (See suppl. Mat., Appendix A).

Results

The distribution of the transition counts among the states was Poisson. The application of Lowess smoother showed the nonlinear relationship between the predictors and the response variables as shown in Fig (4). Table (1) & (2) show the transition counts for each state. (See also suppl. Mat., Appendix B).

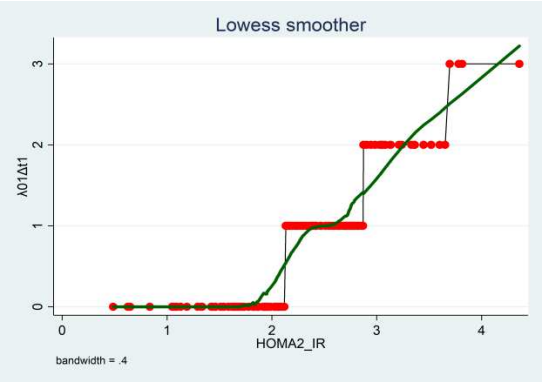

FIGURE (4): Lowess smoother showing the nonlinear relationship between the transition counts from F0 to F1 and the HOMA2-IR levels.

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 28 years

	State 0	State1	State2	State3	State4	total
State0	1909	120	15	6	0	2050
State1	36	1116	67	28	0	1247
State2	13	30	703	37	0	783
State3	11	14	23	50	22	120
State4	0	0	0	0	0	0
						4200

Initial observed rates are:

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

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

The observed counts were used as the response variables to fit the Poisson regression model. For each transition count, the model that represented the most explainable covariates with their Beta coefficients and the corresponding incidence rate ratios were illustrated in (Suppl. Mat, Appendix B). The transitions were subdivided into progressive transitions and regressive transitions. Comparison between the null model and full model was explored in supplementary materials. The main important result is that HOMA2-IR is positively correlated with all progressive transitions and is inversely related to the regressive transitions as shown in table (3) & table (4).

The continuous predictors (age, BMI, HOMA2-IR, LDL-Chol, systolic and diastolic blood pressure) were highly correlated with a correlation coefficient of 0.99, and with 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]

Progressive transitions

Table (3): Parameters for each transitions

Transition from F0 to F1						
	LDLsp2	**HOMAsp1**	**SysSP2**	**LDLsp2# HOMAsp1**	**LDLsp2# SysSP2**	**HOMAsp1# SysSP2**
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 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						
	LDLsp2	**HOMAsp1**	**SysSP2**	**LDLsp2# HOMAsp1**	**LDLsp2# SysSP2**	**HOMAsp1# SysSP2**
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 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						
	LDLsp2	**HOMAsp1**	**SysSP2**	**LDLsp2# HOMAsp1**	**LDLsp2# SysSP2**	**HOMAsp1# SysSP2**
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 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						
	LDLsp2	**HOMAsp1**	**SysSP2**	**LDLsp2# HOMAsp1**	**LDLsp2# SysSP2**	**HOMAsp1# SysSP2**
B co.(P)	0.452(0.000)	10.866 (0.000)	0.073(0.141)	-0.166(0.000)		
CI for 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)	(3357.9, 8.17e+5)	(0.976, 1.187)	(0.818 , 0.877)		

Regressive transitions

Table (4): Parameters for each transitions

Transition from F1 to F0						
	LDLsp2	**HOMAsp2**	**SysSP2**	**LDLsp2# HOMAsp2**	**LDLsp2# SysSP2**	**HOMAsp2# SysSP2**
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						
	LDLsp2	**HOMAsp2**	**SysSP2**	**LDLsp2# HOMAsp2**	**LDLsp2# SysSP2**	**HOMAsp2# SysSP2**
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						
	LDLsp2	**HOMAsp2**	**SysSP2**	**LDLsp2# HOMAsp2**	**LDLsp2# SysSP2**	**HOMAsp2# SysSP2**
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)
Transition from F2 to F0						
	LDLsp2	**HOMAsp2**	**SysSP2**	**DiasSP2**		
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						
	LDLsp2	**HOMAsp2**	**SysSP2**	**DiasSP2**		
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)		

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, $E[y_n|x_n] = \lambda_{ij} = exp(x_n' B)$, to obtain the expected counts of transition that had been accomplished by this participant during this 28 years. Then the result is rounded to the appropriate integer and summed to get all counts for this transition and then compare it to the observed counts accomplished by the all participants.

The n_{i+} is the total marginal transition counts out of this state , which is assumed to be constant. The estimated and obtained counts from running the Poisson 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 $\theta_0 = \frac{n_{ijr}}{n_{i+}}$ where the n_{ijr} 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).

$$\hat{Q} = \begin{bmatrix} -0.059 & 0.059 & 0 & 0 & 0 \\ 0.029 & -0.080 & 0.051 & 0 & 0 \\ 0.015 & 0.033 & -0.093 & 0.045 & 0 \\ 0 & 0.108 & 0.158 & -0.433 & 0.167 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \; where$$

$\lambda_{01} = \frac{120}{2050} = 0.059$, $\lambda_{12} = \frac{64}{1247} = 0.051$, $\lambda_{23} = \frac{35}{783} = 0.045$, $\lambda_{34} = \frac{20}{120} = 0.167$

$\mu_{10} = \frac{36}{1247} = 0.029$, $\mu_{21} = \frac{26}{783} = 0.033$, $\mu_{32} = \frac{19}{120} = 0.158$, $\mu_{20} = \frac{12}{783} = 0.015$, $\mu_{31} = \frac{13}{120} = 0.108$

Probability transition matrix is obtained from exponentiation of this Q matrix after 1 year:

$$P(\Delta t = 1) = exp(\hat{Q} \times \Delta t) = \begin{bmatrix} 0.944 & 0.055 & 0.001 & 0 & 0 \\ 0.027 & 0.925 & 0.047 & 0.001 & 0.0001 \\ 0.014 & 0.033 & 0.915 & 0.035 & 0.003 \\ 0.002 & 0.086 & 0.125 & 0.651 & 0.136 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Goodness of fit

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

Step 1: $H_0 = \; future\ state\ does\ not\ depend\ on\ the\ current\ state$
$H_1 = \; future\ state\ \ depends\ on\ the\ current\ state$

Step 2: calculate the $p_{ij}(\Delta t = 1) \; = exp(Q \times \Delta t) = \begin{bmatrix} 0.944 & 0.055 & 0.001 & 0 & 0 \\ 0.027 & 0.925 & 0.047 & 0.001 & 0.0001 \\ 0.014 & 0.033 & 0.915 & 0.035 & 0.003 \\ 0.002 & 0.086 & 0.125 & 0.651 & 0.136 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

Step3: 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 table in the same interval to get the expected counts as shown below:

	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 4: apply $\sum_{i=1}^{5}\sum_{j=1}^{5}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 1140.097 \sim \chi^2_{(5-1)(5-1)(.05)}$.

So from the above results the null hypothesis is rejected while the alternative hypothesis is accepted and the 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 be used to predict the count of patients in each state at specific time point using (4), for example if a cohort of 6000 patients with the following number in each state being $[3000 \quad 1800 \quad 1020 \quad 180 \quad 0]$ so after one year the predicted counts will be $[2895 \quad 1879 \quad 1044 \quad 154 \quad 28]$. This can be achieved by multiplying the initial count distribution of the patients with the transition probability calculated at the required specific time point , $p_{ij}(t) = exp(Qt)$

$$E\left[u_j(t)|u_j(0)\right] = \sum_{j=1,i=1}^{5} u_j(0)P_{ij}(t) \qquad i\& \; j = 1,..,5 \tag{4}$$

Let $u(0)$ be the size of patients in a specific state at specific time $t = 0$ .The initial size of patients $U(0) = u_j\,(0)$, as there are 4 transient states (F0 to F3) and 1 absorbing state (F4), where $u_j(0)$ is the initial size or the number of patients in state j at time $t = 0$ given that $u_5(0) = 0$ 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 $(0, t)$, there will be $u_j(t)$ patients in the transient states at time t, and there will also be $u_5(t)$patients in state 5 (F4= liver cirrhosis) at time t.

In addition, the state probability distribution $\pi\,(t)$, which is the probability distribution for each state at a specific time point given the initial probability distribution $\pi(0)$, can be estimated by applying the following formula: $\pi(t) = \pi(0)P(t)$. In this example, the cohort of 6000 patients has initial probability distribution of $[0.5 \quad 0.3 \quad 0.17 \quad 0.03 \quad 0]$ after one year, the state probability distribution will be $[0.4825 \quad 0.3131 \quad 0.174 \quad 0.0257 \quad 0.0046]$.

Using the above information each year in addition to the costs of investigations and treatments and the quality adjusted life years for the patients, pharmaco-economic evaluation can be assessed in three different categories: The cost-benefit analysis, the cost-effectiveness analysis, and the cost utility analysis.[10] [11]

This approach differs from the one used by Rustgi *et al.* [12] 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 time, any cohort of patients can then be followed up utilizing the information gained from sampling the general population.

Discussion:

To elucidate the agreements and comparisons as regard the findings in this study with the previous studies, a brief summary of some of the previously conducted studies highlighting the effects of various factors on the rate of progression of fibrosis in NAFLD patients, is described in the following discussion.

Hui *et al*. [13] 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). Nearly, more than half of them developed progressive fibrosis as compared to the initial biopsy; this was due to the fact that these patients were initially suffering from steatohepatitis, although there was no significant correlation between the degree of steatohepatitis and the degree of fibrosis between the 2 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 biopsy. The changes in these parameters also showed no significant correlation with changes in the scores of steatosis, necroinflammation or fibrosis. Nevertheless, there was

a negative correlation, although non-significant, between the change in the score of fibrosis and each of the changes: in the BMI, in plasma total cholesterol levels, in plasma triglyceride levels and in glycosylated hemoglobin. During the course of the follow up, 2 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 one 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.* [14] conducted a study on 22 patients who had liver biopsy with evidence of NASH and found that 31.8% (7 patients=P group) had progression of liver fibrosis over a median follow-up of 4.7 years. The other group was 68.2% (15 patients=NP group) and did not show any progression 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), while; the rate of progression in the P group was 1.85/6.59=0.28. There were no statistical difference as regards the clinical, biochemical, grade of steatosis, and grade of inflammation between the 2 groups except for presence of obesity and higher BMI (progressor were more obese with higher BMI than the nonprogressor) whether this was performed during the initial liver biopsy or the final liver biopsy. 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 2 groups including the BMI.

Adam *et al.* [7] conducted a study on 103 patients who had performed 2 liver biopsies with mean follow up period of 3.2 ± 3 years (range = 0.7-21) between the first and the second biopsy. A total of 38 patients were progressors, 35 patients were stable, and 30 patients were regressors. No clinical or biochemaical variables were statistically different between 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, 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 that there was no significant correlation between rate of progression and HOMA.

There are many studies performed by Ekstedt *et al.* [15], Teli *et al.* [16], Pais *et al.*[17], Argo *et al.* [18], Evans *et al.* [19], Hamaguchi *et al.* [20], and Wong *et al.* [21]. They reported the fibrosis progression rates utilizing 2 paired biopsies and the reader can refer to them.

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 key players for progression are the HOMA2-IR, LDL-chol and systolic blood pressure explaining about 35% to 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 rate of remission from F2 to F0 and from F3 to F1. Poisson regression model explained that the same factors and the interactions between them were responsible for about 60% to 70% of variability in the rates of remission among the states. Each of the high HOMA2-IR level and high LDL level significantly decreases the effect of the other on the rate of progression from F0 to F1 and from F3 to F4, and this can be a protective mechanism in attempt to slow down the rate of progression of fibrosis. It is also notable that each of the low HOMA2-IR level and low LDL level significantly increases the effect of the other on the rate of remission from F1 to F0 and from F2 to F1, and this can be a protective mechanism in attempt to accelerate the rate of regression of fibrosis. This can also be achieved with the help of rigorous control of the blood level of insulin, glucose, cholesterol and blood pressure. It is also notable that each of the high level of systolic blood pressure and low LDL level significantly decreases the effect of each other on the rate of remission from F1 to F0, from F2 to F1, and from F3 to F2. This point to the possibility that controlling the most deleterious factors like hyperinsulinemia and hypercholesterolemia, even in absence of strict control of hypertension, can still has a beneficial effect on repressing

the fibrogenesis. This in association with life style modification, in the form of physical exercise and low caloric diet, has a great impact on arresting the process of fibrogenesis.

The newly emerging anti-fibrotic drugs will also help physicians treating the fibrogenesis. In FLINT study, conducted on 283 non-cirrhotic NASH patients taking obeticholic acid (OCA), 25 mg daily; the improvement in the histology detected by NAFLD activity score (NAS) was 2 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 in clinicaltrials.gov official site) will evaluate safety and efficacy of Obeticholic Acid (OCA) in NASH patients with fibrosis whom are randomized to a daily dose of 25 mg, 10 mg, and placebo, with end points 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 mg and 25 mg doses. In 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 in comparison to placebo, moreover; these patients also showed lowering in fibrosis stage compared to non-resolving NASH patients. RESOLVE-IT trial (last update was on November 30, 2020, as shown in clinicalrials.gov official site, but the study is still in progress according to [22]) 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 with placebo group. The AURORA trial (primary completion dates were October 2021 according to clinicaltrails.gov site and October 2028 according to[22]) will evaluate long-term safety and efficacy of 150 mg daily CVC for the treatment of fibrosis in NASH adult patients 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 end point 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 significant improvement in liver disease activity, fibrosis, stiffness, liver fat content, and progression to cirrhosis [23].

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

The strength of this study is the conduction of multiple frequent repeated observations over a long time period of follow up on a large number of at high risk participants for developing NAFLD and performing a liver biopsy in each visit. Although this may be realistically infeasible to be done in each visit, using non-invasive techniques[24][25] can substitute the invasive liver biopsy and this may be consider a weakness of the study. But the advantage of techniques like MRI and machine learning [26], to assess the liver texture and correlate these finding with the histological findings in liver biopsy, can overcome this weakness. Liver biopsy can also be reserved in situations where the results of non-invasive tests are inconclusive and these non-invasive tests decrease the number of liver biopsies each patient may encounter to do. The proposed follow up period is too long to wait for the obtained results and this can be overcome by using adaptive clinical trials.

Conclusion:

The distribution of the counts was Poisson (mean=variance); that is to mean, these counts were equidispersed. However, all the counts showed excess zeros except the transition from F0 to F1 where the zeros constituted 44% of the total count of this transition. Tlhaloganyang and Sakia [27] found that the equidispersed counts data with excessive zeros can be modeled with Poisson regression which is 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. 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. Nevertheless, the raw predictors as well as the transformed predictors were highly correlated. But the condition number is below 100, which is not harmful for the analysis as shown in the results. Vatcheva *et al.* [28] 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 mullticollinaerity like: principle component analysis, partial least squares regression, and ridge regression analysis. Akram *et al.* [29] used principle component ridge type estimator for the inverse Gaussian regression model. Many investigators like: Liu [30], Kabria and Lukman [31], Lukman *et al.* [32], and 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 Mansson and Shukur [33], Mansson *et al.*[34], Lukman *et al.*[35], [36], and Qasim *et al.* [37]. 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. It was not also used for prediction and the condition number was less than 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 can obtain their state probability distribution at different time points.

To sum up, 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 in different time points. This helps health policy makers and insurance managers for allocating the human and financial resources to investigate and treat the at high risk patients for developing NAFLD. The Poisson regression model is utilized to relate these high risk covariates to the rate of transmission 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 , DOI=10.24433/CO.8778229.V3

The code to estimate the Q transition rate matrix for the observed transition counts using continuous time Markov chains are published in the code Ocean site with following URL:
https://codeocean.com/capsule/6377472/tree/v2 , DOI=10.24433/CO.2144346.V2

The dataset is present on IEEE*D*ata*P*ort site with the following URL:
https://ieee-dataport.org/documents/fibrosis-nfld#files , with the following DOI: 10.21227/dr5j-gs46

Abbreviations:
CC: compensated cirrhosis (stage 4), CTMC: continuous time Markov chains, DCC: de-compensated cirrhosis (stage 5), EM: extra-mortality (stage 9), HCC: hepato-cellular carcinoma (stage 8), LT: liver transplant(stage 6), NAFLD: non-alcoholic fatty liver disease, NAFL-NO FB: nonalcoholic fatty liver 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 (stage 7), 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 no competing interests of any type.

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
AI carried 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, carried the methodology by creating the model, software programming and implementation, supervision, writing, drafting, editing, preparation, and creation of the presenting work.

Acknowledgement
Not applicable

Disclosures
Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships**: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships**: All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

[1] Z. M. Younossi, T. Gramlich, C. A. Matteoni, N. Boparai, and A. J. McCullough, "Nonalcoholic fatty liver disease in patients with type 2 diabetes," *Clinical Gastroenterology and Hepatology*, vol. 2, no. 3, pp. 262–265, 2004.

[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.

[3] I. M. Attia, "Novel Approach of Multistate Markov Chains to Evaluate Progression in the Expanded Model of Non-alcoholic Fatty Liver Disease," *Frontiers in Applied Mathematics and Statistics*, vol. 7, 2022, Accessed: Mar. 03, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fams.2021.766085

[4] Z. M. Younossi *et al.*, "Economic and clinical burden of nonalcoholic steatohepatitis in patients with type 2 diabetes in the US," *Diabetes Care*, vol. 43, no. 2, pp. 283–289, 2020.

[5] S. Singh, A. M. Allen, Z. Wang, L. J. Prokop, M. H. Murad, and R. Loomba, "Fibrosis progression in nonalcoholic fatty liver vs nonalcoholic steatohepatitis: a systematic review and meta-analysis of paired-biopsy studies," *Clinical gastroenterology and hepatology*, vol. 13, no. 4, pp. 643–654, 2015.

[6] J. D. Kalbfleisch and J. F. Lawless, "The analysis of panel data under a Markov assumption," *Journal of the American Statistical Association*, vol. 80, no. 392, pp. 863–871, 1985.

[7] L. A. Adams, P. Angulo, and K. D. Lindor, "Nonalcoholic fatty liver disease," *Cmaj*, vol. 172, no. 7, pp. 899–905, 2005.
[8] M. A. A.-E. A. Elreesh, "Prognostic Factors for Evolution of Non Alcoholic Fatty Liver Disease Patients Utilizing Poisson Regression and Continuous Time Markov Chains," 2021.
[9] A. Lazaridis, "A note regarding the condition number: the case of spurious and latent multicollinearity," *Quality & Quantity*, vol. 41, no. 1, pp. 123–135, 2007.
[10] J. O'Hara *et al.*, "Cost of non-alcoholic steatohepatitis in Europe and the USA: the GAIN study," *JHEP Reports*, vol. 2, no. 5, p. 100142, 2020.
[11] M. Noureddin *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*, vol. 159, no. 5, pp. 1985–1987, 2020.
[12] V. K. Rustgi, S. B. Duff, and M. I. Elsaid, "Cost-effectiveness and potential value of pharmaceutical treatment of nonalcoholic fatty liver disease," *Journal of medical economics*, no. just-accepted, pp. 1–29, 2022.
[13] A. Y. Hui *et al.*, "Histological progression of non-alcoholic fatty liver disease in Chinese patients," *Alimentary pharmacology & therapeutics*, vol. 21, no. 4, pp. 407–413, 2005.
[14] E. Fassio, E. Álvarez, N. Domínguez, G. Landeira, and C. Longo, "Natural history of nonalcoholic steathepatitis: a longitudinal study of repeat liver biopsies," *Hepatology*, vol. 40, no. 4, pp. 820–826, 2004.
[15] M. Ekstedt *et al.*, "Long-term follow-up of patients with NAFLD and elevated liver enzymes," *Hepatology*, vol. 44, no. 4, pp. 865–873, 2006.
[16] M. R. Teli, O. F. W. James, A. D. Burt, M. K. Bennett, and C. P. Day, "The natural history of nonalcoholic fatty liver: a follow-up study," *Hepatology*, vol. 22, no. 6, pp. 1714–1719, 1995.
[17] R. Pais *et al.*, "A systematic review of follow-up biopsies reveals disease progression in patients with non-alcoholic fatty liver," *Journal of hepatology*, vol. 59, no. 3, pp. 550–556, 2013.
[18] C. K. Argo, P. G. Northup, A. M. S. Al-Osaimi, and S. H. Caldwell, "Systematic review of risk factors for fibrosis progression in non-alcoholic steatohepatitis," *Journal of hepatology*, vol. 51, no. 2, pp. 371–379, 2009.
[19] C. D. J. Evans, K. A. Oien, R. N. M. MacSween, and P. R. Mills, "Non-alcoholic steatohepatitis: a common cause of progressive chronic liver injury?," *Journal of clinical pathology*, vol. 55, no. 9, pp. 689–692, 2002.
[20] E. Hamaguchi *et al.*, "Histological course of nonalcoholic fatty liver disease in Japanese patients: tight glycemic control, rather than weight reduction, ameliorates liver fibrosis," *Diabetes care*, vol. 33, no. 2, pp. 284–286, 2010.
[21] V. W.-S. Wong *et al.*, "Disease progression of non-alcoholic fatty liver disease: a prospective study with paired liver biopsies at 3 years," *Gut*, vol. 59, no. 7, pp. 969–974, 2010.
[22] E. Guirguis, Y. Grace, A. Bolson, M. J. DellaVecchia, and M. Ruble, "Emerging therapies for the treatment of nonalcoholic steatohepatitis: A systematic review," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 41, no. 3, pp. 315–328, 2021, doi: 10.1002/phar.2489.
[23] N. Alkhouri, F. Poordad, and E. Lawitz, "Management of nonalcoholic fatty liver disease: Lessons learned from type 2 diabetes," *Hepatology Communications*, vol. 2, no. 7, pp. 778–785, 2018, doi: 10.1002/hep4.1195.
[24] L. Petitclerc, G. Sebastiani, G. Gilbert, G. Cloutier, and A. Tang, "Liver fibrosis: Review of current imaging and MRI quantification techniques," *Journal of Magnetic Resonance Imaging*, vol. 45, no. 5, pp. 1276–1295, 2017.
[25] 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," *Annals of medicine*, vol. 43, no. 8, pp. 617–649, 2011.
[26] K. Schawkat *et al.*, "Diagnostic accuracy of texture analysis and machine learning for quantification of liver fibrosis in MRI: correlation with MR elastography and histopathology," *European radiology*, vol. 30, no. 8, pp. 4675–4685, 2020.
[27] B.P. Tlhaloganyang* and R.M. Sakia, "Zero inflated Poisson distribution in equidispersed data with excessive zeros," *Research Journal of Mathematical and Statestical Sciences*, vol. 8, no. 1, pp. 31–34, 2020.
[28] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar, "Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies," *Epidemiology (Sunnyvale)*, vol. 6, no. 2, p. 227, Apr. 2016, doi: 10.4172/2161-1165.1000227.
[29] M. Akram, M. Amin, A. Lukman, and S. Afzal, "Principal component ridge type estimator for the inverse Gaussian regression model," *Journal of Statistical Computation and Simulation*, pp. 1–30, Jan. 2022, doi: 10.1080/00949655.2021.2020274.

[30] K. Liu, "Using Liu-Type Estimator to Combat Collinearity," *Communications in Statistics - Theory and Methods*, vol. 32, no. 5, pp. 1009–1020, Jan. 2003, doi: 10.1081/STA-120019959.
[31] B. M. G. Kibria and A. F. Lukman, "A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications," *Scientifica*, vol. 2020, p. e9758378, Apr. 2020, doi: 10.1155/2020/9758378.
[32] A. F. Lukman, K. Ayinde, S. Binuomote, and O. A. Clement, "Modified ridge-type estimator to combat multicollinearity: Application to chemical data," *Journal of Chemometrics*, vol. 33, no. 5, p. e3125, 2019, doi: 10.1002/cem.3125.
[33] K. Månsson and G. Shukur, "A Poisson ridge regression estimator," *Economic Modelling*, vol. 28, no. 4, pp. 1475–1481, 2011.
[34] K. Månsson, B. M. G. Kibria, P. Sjölander, and G. Shukur, "Improved Liu Estimators for the Poisson Regression Model," *International Journal of Statistics and Probability*, vol. 1, Apr. 2012, doi: 10.5539/ijsp.v1n1p2.
[35] A. F. Lukman, E. Adewuyi, K. Månsson, and B. M. G. Kibria, "A new estimator for the multicollinear Poisson regression model: simulation and application," *Sci Rep*, vol. 11, no. 1, p. 3732, Feb. 2021, doi: 10.1038/s41598-021-82582-w.
[36] A. F. Lukman, B. Aladeitan, K. Ayinde, and M. R. Abonazel, "Modified ridge-type for the Poisson regression model: simulation and application," *Journal of Applied Statistics*, pp. 1–13, Feb. 2021, doi: 10.1080/02664763.2021.1889998.
[37] M. Qasim, B. M. G. Kibria, K. Månsson, and P. Sjölander, "A new Poisson Liu Regression Estimator: method and application," *Journal of Applied Statistics*, vol. 47, no. 12, pp. 2258–2271, Sep. 2020, doi: 10.1080/02664763.2019.1707485.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarymaterialsAppendixA.docx
- SupplementarymaterialsAppendixB.docx
- SupplementarymaterialsAppendixC.docx