Biometrical Letters
Vol. 47 (2010), No. 1, 33-44

Logistic multivariate regression analysis as a tool to predict fibrosis in light-drinking chronic hepatitis C patients

M.C. Manso[1], R.M. Cerqueira[2], C. Fernandes[2], M. Correia[2]

[1]University Fernando Pessoa, Faculty of Health Sciences, Rua Carlos da Maia,
296 4200-150 Porto, Portugal / REQUIMTE-UP, e-mail: cmanso@ufp.edu.pt
[2]Gastroenterology Department, São Sebastião Hospital, Rua Dr. Cândido de Pinho, 4520-211
Santa Maria da Feira, Portugal

SUMMARY

Complications of the chronic hepatitis C virus (HCV) include the development of fibrosis, which depends on host and viral characteristics. In most studies moderate and heavy alcohol intake is a risk factor for hepatic fibrosis. However, there is a lack of data on the effects of light alcohol intake on HCV disease progression.
In this work the independent effect of significant variables ($p < 0.05$) on fibrosis was assessed using backward stepwise binary logistic multivariate regression analysis (0.05 for factor inclusion and 0.2 for exclusion). Data were collected in a retrospective cross-sectional study of 99 untreated HCV patients (79.4% male, age (mean ± SD) 35.2 ± 10.3 years) with liver biopsy who drank up to 30 g alcohol daily (8.3 ± 10.7 g). The prevalence of fibrosis (METAVIR $\geq$ 2) in this group was 40.4% (95% CI: 30.7% - 50.1%). Age, elevated GPT ($\geq$ 2 x upper normal limit), steatosis ($\geq$ 5%) and viral load ($\geq$ 800,000 UI/mL) remained in the model. The area under the curve (AUC) derived from this model was 0.776 (95%CI: 0.685 - 0.868) for fibrosis in light-drinking chronic hepatitis C patients. The sensitivity and specificity for predicting severe hepatic fibrosis were 0.64 and 0.77 respectively.

Key words : logistic model, multivariate analysis, backward stepwise method, fibrosis, chronic hepatitis C

1. Introduction

Hepatitis C virus (HCV) is a leading cause of cirrhosis and hepatocellular carcinoma in Europe and the United States. The rate of progression from chronic hepatitis C to fibrosis and cirrhosis is highly variable among patients – some individuals experience a benign clinical course for decades, while others rapidly progress to end-stage liver disease. Several host and viral factors have been linked to fibrosis progression, such as older age at infection, male gender, body mass index, alcohol intake, HCV genotype and viral load. A heavy alcohol intake of more than 50 to 60 g/day has been found in many studies to be an independent risk factor for fibrosis in HCV infection (Metwally et al., 2007). However, there is a lack of data on the threshold level of alcohol intake which negatively influences the natural course of HCV infection, as for the impact of minimal alcohol consumption on the degree of histological liver lesions. Additionally, there is increasing evidence that light drinking brings significant health benefits due to positive cardiovascular effects. However, HCV patients are generally counselled by their physicians to abstain from drinking alcohol.

Logistic regression is commonly used when the independent variables include both numerical and nominal measures and the outcome variable is binary (dichotomous), although it can also be used when the outcome has more than two values (Hosmer, Lemeshow, 1989).

One reason for the popularity of logistic regression is that many outcomes in health are nominal, actually binary, variables – they either occur or do not occur. The second reason is that the regression coefficients obtained in logistic regression can be transformed into odds ratios. So, in essence, logistic regression provides a way to obtain an odds ratio for a given risk factor that controls for, or is adjusted for, confounding variables (Dawson, Trapp, 2004).

The aim of this study was to define the role of the factors likely to explain the process of fibrosis in light-drinking HCV patients.

2. Description of the Data and Methods

Data were collected in a retrospective cross-sectional study of 99 untreated HCV patients with liver biopsy who drank up to 30 g of alcohol daily (mean age ± standard deviation: 8.3 ±10.7 g). Their age was 35.20 ± 10.23 years, ranging from 18 to 63 years (79.4% were male, 35.2 ±10.3 years). The prevalence of fibrosis (METAVIR ≥ 2) in this group was 40.4% (95% CI: 30.7%–50.1%).

Among the explanatory variables, 2 types were considered: demographics and laboratory data. Table 1 summarizes demographic and laboratory data for the total patient population and univariate comparison according to degree of hepatic fibrosis, considering no or light Fibrosis as METAVIR F0 and F1 respectively, and severe Fibrosis as METAVIR F2, F3 and F4. Demographic variables included gender, age and body mass index (BMI). Laboratory data included glutamic-pyruvate transaminase (GPT), gamma glutamil transferase (GGT), Ferritin, HCV load (assessed by qualitative PCR assay with lower detection limit 50 UI/mL), HCV genotyping and hepatic steatosis. Daily alcohol intake was also considered. Several variables that were of a continuous nature were dichotomized based on values that are usually present in scientific literature.

In this work the independent effect of significant variables on fibrosis was assessed using backward stepwise binary logistic multivariate regression analysis (0.05 for factor inclusion and 0.2 for exclusion), using SPSSv15.0 for Windows.

Associations between severe fibrosis and covariates were assessed by unadjusted odds ratio (OR), as estimated using the Mantel-Haenszel statistic analysis (Table 2). The odds ratio is one of a range of statistics used to assess the risk of a particular outcome (or disease) if a certain factor (or exposure) is present. The odds ratio is a relative measure of risk, telling us how much more likely it is that someone who is exposed to the factor under study will develop the outcome as compared to someone who is not exposed (Bland, Altman, 2000). The odds of an event happening is the probability that the event will happen divided by the probability that the event will not happen. The observed odds ratio, lets say OR = 4.00 for age (Table 2) (meaning that someone who has HCV and has an age ≥ 40 years is 4 times more likely to develop the outcome – hepatic severe fibrosis – as compared to someone who has HCV and is under the age of 40 years) is not in the centre of the confidence interval, because of the asymmetrical nature of the odds ratio scale. The odds ratio is 1 when there is no relationship, and the null hypothesis that the odds ratio is 1 can be tested by the usual χ^2 test for a two-by-two table (Bland, Altman, 2000).

Table 1. Demographic and laboratory data for the total patient population and univariate comparison according to degree of hepatic fibrosis. Continuous variables are expressed as mean (± standard deviation), median [inter quartile range] as well as minimum and maximum values, and categorical data as a frequency (n) and percentage (%).

Variables		All	Non severe Fibrosis (METAVIR F0/1)	Severe Fibrosis (METAVIR F2/3/4)	p-value
Gender	n	99	59	40	0.001*
Female; Male	%	20.6; 79.4	27.1; 72.9	12.5; 87.5	
Age (years)	n	99	59	40	0.017**
	Mean(±SD)	35.20 (±10.23)	33.08 ^{b} (±8.87)	38.32 ^{a} (±11.36)	
	Min–max	18–63	18–63	18–62	
BMI (kg/m^2)	n	95	59	36	0.035***
	Mean(±SD)	23.89 (±3.08)	23.30 (±2.93)	24.73 (±3.13)	
	Me [IQR]	23 [4]	23 ^{b} [3.5]	24 ^{a} [3.7]	
	Min–max	17–35	17–30	20–35	
BMI (kg/m^2)	n	99	59	40	0.387*
<25; 25-29.9; ≥30	%	62.6; 29.3; 8.1	66.1; 27.1; 6.8	57.5; 32.5; 10.0	NS
GPT (×N****)	n	98	59	39	0.001***
	Mean(±SD)	2.69 (±2.30)	2.32 (±2.37)	3.25 (±2.09)	
	Me [IQR]	2 [2]	2 ^{b} [1]	3 ^{a} [3]	
	Min–max	1–15	1–15	1–10	
GGT (×N****)	n	99	59	40	0.007***
	Mean(±SD)	1.89 (±1.83)	1.60 (±1.67)	2.32 (±1.99)	
	Me [IQR]	1 [1]	1 ^{b} [0.6]	1.25 ^{a} [2]	
	Min–max	1–10	1–10	1–10	
Ferritin (ng/L)	n	90	51	39	
	Mean(±SD)	201.80 (±164.91)	187.96 (±180.99)	219.9 (±141.4)	0.098***
	Me [IQR]	169 [162,5]	166 [153]	172 [203]	NS
	Min–max	0–1188	0–1188	44–657	
Ferritin (ng/L)	n	90	51	39	0.105*
<300; ≥300	%	80.0; 20.0	84.3; 15.7	74.4; 25.6	NS
	n	95	57	38	
Viral load	Mean(±SD)	963110(±1592375)	650364(±673912)	1432230(±2319023)	0.028***
(UI/mL)	Me [IQR]	664000 [713000]	501187 ^{b} [776000]	845500 ^{a} [591750]	
	Min–max	0–12700000	0–4510000	2740–12700000	
Viral load (UI/mL)	n	98	59	39	0.002*
<8x10^5; ≥8x10^5	%	58.2; 41.8	66.1; 33.9	46.2; 53.8	
HCV genotyping	n	96	57	39	<0.001*
g1; g3; g4	%	60.4; 27.1; 12.5	57.9; 24.6; 17.5	64.1; 30.8; 5.1	
Steatosis (%)	n	99	59	40	<0.001*
0-4; 5-29; ≥30	%	60.6; 21.2; 18.2	72.9; 18.6; 8.5	42.5; 25.0; 32.5	
	n	99	59	40	
Alcohol intake	Mean(±SD)	8.33 (±10.71)	9.41 (±10.51)	6.75 (±10.95)	0.131***
(g/day)	Me [IQR]	0 [20]	10 [20]	0 [20]	NS
	Min–max	0–30	0–30	0–30	

* Chi-square test. NS – non significant differences; $^{a, b}$ – different letters indicate significant differences of the expected value of the variable between severe fibrosis and no severe fibrosis groups, according to the ** t test or the *** Mann-Whitney U test. **** N – upper normal limit

Table 2. Variables associated with severe fibrosis in the univariate analysis.

Covariates		N	Odds Ratio[a)] [95% C.I.]	p-value
Gender	Female	21	1.00	
	Male	78	2.60 [0.87; 7.82]	0.088
Age (years)	< 40	71	1.00	
	≥ 40	28	4.00 [1.59; 10.08]	0.003
BMI (kg/m^2)	< 25	63	1.00	
	≥ 25	33	1.64 [0.70; 3.84]	0.258
	< 30	92	1.00	
	≥ 30	7	2.07 [0.44; 9.82]	0.358
GPT (×N*)	< 2	37	1.00	
	≥ 2	62	3.11 [1.26; 7.66]	0.014
GGT (×N*)	< 2	68	1.00	
	≥ 2	31	2.38 [0.99; 5.66]	0.051
Ferritin (ng/L)	< 300	72	1.00	
	≥ 300	18	1.85 [0.65; 5.26]	0.246
Viral load (UI/mL)	< 600000	50	1.00	
	≥ 600000	48	1.64 [0.73; 3.71]	0.233
	< 800000	57	1.00	
	≥ 800000	41	2.27 [0.99; 5.21]	0.052
HCV	genotyping 1 or 4	70	1.00	
	genotyping 3	26	1.36 [0.55; 3.39]	0.502
Steatosis	< 5%	60	1.00	
	≥ 5%	39	3.64 [1.55; 8.51]	0.003
	< 30%	81	1.00	
	≥ 30%	18	5.20 [1.68; 16.10]	0.004

a) Mantel-Haenszel Common Odds Ratio Estimate; * N – upper normal limit

As the literature contains different cut points for some of the variables under study, different dichotomizations were produced: BMI with a cut point of 25 (separating underweight and normal weight patients from overweight/obese) and of 30 (separating overweight or less from obese patients), viral load with a cut point of 600,000 UI/mL and another at 800,000 UI/mL, and hepatic steatosis with a cut point of 5% (presence/absence) and another at 30 % (severe cases). It is possible to see (Table 2) that as the severity of the covariate increases, so does the OR, showing a risk increase for all cases, although BMI and viral load were not shown to be significantly associated risk factors for severe fibrosis (p>0.05).

3. Logistic Regression Modeling

Logistic regression is commonly used when the independent variables include both numerical and nominal measures and the outcome variable is binary (dichotomous), although it can also be used when the outcome has more than two values (Hosmer, Lemeshow, 1989).

Logistic regression assumes an underlying linear relationship between a dichotomous dependent variable and one or more independent variables. The plot of such data always results in two parallel lines, each corresponding to a value of the dichotomous dependent variable. Because the two parallel lines are difficult to model, one can alternatively create categories for the independent variable and compute the mean of the dependent variable value for the respective categories. The resultant plot based on these categories' means will, therefore, appear linear in the middle, much like what one would expect to see on an ordinary scatter plot, but curvilinear at both ends. Such a shape is often referred to as sigmoidal or S-shaped (Peng et al., 2001).

One problem in working with S-shaped data is that the extremes are difficult to model. Another problem is that the errors are typically neither normally distributed nor constant across the entire range of data. Therefore, the ordinary least squares solution cannot help a researcher to derive a regression equation from the data. The logistic regression approach solves these problems by applying the logit transformation to the dependent variable (Kutner et al., 2004).

Without the logit transformation, a dichotomous dependent variable is typically recorded as the percentage probability of a particular outcome, say experiencing severe fibrosis after hepatitis C infection. By definition, this measure must fall between 0 and 1. Hence its odds, defined as the ratio of the probability of the outcome divided by one minus that probability (p/(1-p)), varies from 0 to infinity (the abovementioned odds ratio). The natural log transformation of the odds, also referred to as logit, varies from negative infinity to positive infinity, and it is possible to relate it to any independent variable in a similar fashion as in a simple linear regression. For example, the log of the percentage of patients who experience severe fibrosis versus the percentage who did not experience severe fibrosis can be hypothesized to be linearly related to HCV. A mathematical formulation of such a relationship looks like this:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1(\mathrm{HCV}) \tag{1}$$

To derive the estimates of both beta coefficients for a given data set (Hosmer, Lemeshow, 1989), a researcher may employ the maximum likelihood method, which is readily available in statistical software such as SPSS or SAS, among others. When the beta coefficients are estimated, it becomes possible to evaluate the log of the odds of severe fibrosis versus no severe fibrosis for future patients.

Using the same logic underlying the simple logistic regression equation, it is possible to construct a more complex model that incorporates several explanatory variables.

This complex model is in the form of multiple regression equations, that is a multivariate regression model. The construction of such a model, using several nominal scaled variables and/or at least interval scaled variables, is very well described in Hosmer, Lemeshow (1989).

For the severe fibrosis patient data set, it was hypothesized that the following linear relationship might exist:

$$\begin{aligned}\ln\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_1 X_1 + +\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \\ &+ \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9\end{aligned} \tag{2}$$

where β_0 is a constant, β_1 to β_9 are the variables' coefficients, and X represents the respective variables: X_1 = gender (0=female; 1=male), X_2 = age (one year increase), X_3 = body mass index (BMI: $0 < 25$; $1 \geq 25$), X_4 = glutamic-pyruvate transaminase (GPT: $0 < 2$; $1 \geq 2$), X_5 = gamma glutamil transferase (GGT: $0 <2$; $1 \geq 2$), X_6 = ferritin ($0 < 300$; $1 \geq 300$), X_7 = viral load (0 < 800,000 UI/mL;

$1 \geq 800{,}000$ UI/mL), X_8 = hepatitis C virus genotyping (HCV: 0 = genotyping 1 and 4; 1 = genotyping 3) and X_9 = hepatic steatosis ($0 < 5\%$; $1 \geq 5\%$).

Alternatively, one can express the same functional relationship by taking the antilog function of Equation (2) on both sides and obtain a direct estimate of the probability of severe fibrosis:

$$P(\text{Severe Fibrosis} = 1) = \\ = \frac{e^{\beta_0 + \beta_1 X_1 + + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9}}{1 + e^{\beta_0 + \beta_1 X_1 + + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9}} \tag{3}$$

All the variables (age, gender, body mass index, GPT, GGT, ferritin, HCV genotyping, viral load and steatosis) were included in a stepwise logistic regression analysis, using Equation 3, in SPSS (vs. 15.0). The backward stepwise selection algorithm used a p-value of 0.05 for factor inclusion and 0.2 for exclusion. It was verified that the predictors gender (X_1), BMI (X_3), GGT (X_5), ferritin (X_6), viral load (X_7) and HCV (X_8) were insignificant at p = 0.05. Variables or predictors that were significantly associated with severe hepatic fibrosis in the univariate analysis (Table 2), age (X_2), elevated GPT (X_4) and steatosis (X_9), remained in the model (Table 3).

Table 3. Independent predictors of severe (Metavir F2/3/4) hepatic fibrosis in patients with chronic hepatitis C: multiple regression analysis [a)].

Covariates		Variable estimate	S.E.	Odds Ratio [95% C.I.]	p-value
Age (years)	-			1.00	
	+1year	0.064	0.027	1.07* [1.01; 1.12]	0.018
GPT ≥ 2 (N**)	No			1.00	
	Yes	1.412	0.591	4.10 [1.29; 13.07]	0.017
Viral load ≥ 800000 (UI/mL)	No			1.00	
	Yes	0.772	0.515	2.16 [0.79; 5.94]	0.134
Steatosis ≥ 5%	No			1.00	
	Yes	1.408	0.550	4.09 [1.39; 12.02]	0.010
Constant		-4.459	1.190	0.012	0.000

*OR [95% C.I.] calculated for a 10-year increase = 1.90 [1.12; 3.22]; ** N – upper normal limit

[a)] Backward stepwise binary logistic multivariate regression analysis (0.05 for factor inclusion and 0.2 for exclusion).

Although viral load (X_7) was not shown to be a significant predictor, it was maintained in the final model due to the known importance of the large volume of the viral load relation with a more aggressive course of the disease. A good quality model was obtained (Table 4), and can be presented as Equation 4.

$$P(\text{Severe Fibrosis} = 1) = \\ = \frac{e^{-4.459+0.064\cdot \mathit{Age}+1.412\cdot \mathit{GPT}+0.772\cdot \mathit{Viral\ load}+1.408\cdot \mathit{Steatosis}}}{1+e^{-4.459+0.064\cdot \mathit{Age}+1.412\cdot \mathit{GPT}+0.772\cdot \mathit{Viral\ load}+1.408\cdot \mathit{Steatosis}}} \qquad (4)$$

According to this model, the log of the odds of experiencing severe fibrosis versus no severe fibrosis was positively related to the patient's age, GPT level, viral load and steatosis.

Table 4. Model assessment quality.

Model Summary	-2 Log likelihood	91.102
	Cox & Snell R^2	0.245
	Nagelkerke R^2	0.329
Hosmer and Lemeshow Test	χ^2 (df=8)	7.140
	p-value	0.522
Overall Correct model Predicted Percentage		71.4%

The resulting multivariate model goodness-of-fit (Table 4) was assessed using the –2loglikelihood test, the Cox, Snell R^2 and the Nagelkerke's (or pseudo) R^2. The Hosmer-Lemeshow test revealed good fit with a value of 7.14 (p = 0.522). This logistic regression model produced a receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.776 (95% CI: 0.685–0.868) for fibrosis in light-drinking chronic hepatitis C patients (Figure 1), which means that the model can be considered useful for predicting severe liver fibrosis (an AUC of 0.8 or higher would indicate an excellent diagnostic accuracy) (Hosmer, Lemeshow, 1989).

ROC curves measure the amount of separation between the distribution of a model's results in the disease population from the distribution of the model's results in the non-disease population. If the distribution of the model's results for the disease and non-disease completely overlap, then the ROC curve is a line from (0,0) to (1,1). The more separated the distributions, the closer the ROC curve is to the upper left-hand corner, being perfect when the curve reaches the upper left-hand corner.

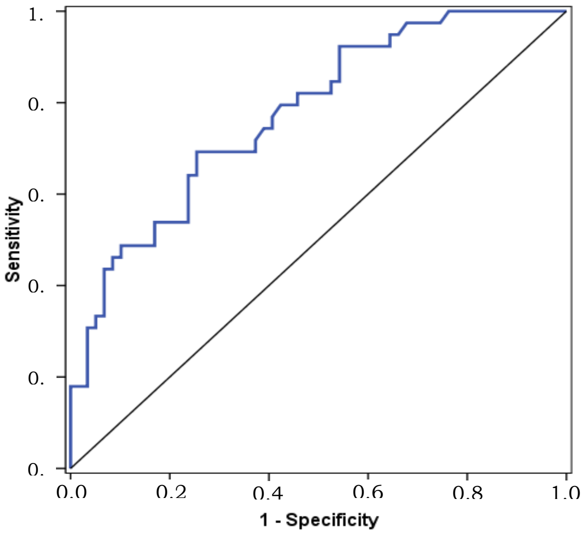

Figure 1. Receiver operating characteristic curve for the model for predicting severe liver fibrosis.

Furthermore, the AUC can be interpreted as the probability that the result for a randomly chosen subject exceeds that for a randomly chosen non-disease subject (Bamber, 1975).

For an outpoint of probability of 0.31, the model sensitivity and specificity for predicting severe hepatic fibrosis were 0.64 and 0.77 respectively.

The predictive accuracy of logistic regression models has its own limitations and can be compromised by converting continuous values into dichotomous variables. Moreover biologic measurements fluctuate, interact with other variables, and are potentially modifiable by medication, all contributing to the limitation of the precision of risk estimation. To account for each and every variable and interaction would produce a model that would be too cumbersome for clinical use. However, the versatility of logistic regression prediction can be utilized to determine levels of risk and direct clinical decisions.

The logistic equation, namely the model stated in Equation 4, can be used to find the probability of an outcome (e.g. the presence of severe fibrosis) for any given individual. For instance, let us calculate the probability that a 48-year-old patient with elevated GPT, with a viral load higher than 800,000 UI/mL and 10% steatosis, has severe fibrosis (Equation 5).

$$P(\text{Severe Fibrosis} = 1) = \frac{e^{-4.459+0.064\times 48+1.412\times 1+0.772\times 1+1.408\times 1}}{1+e^{-4.459+0.064\times 48+1.412\times 1+0.772\times 1+1.408\times 1}} = 0.901 \tag{5}$$

On the other hand, the probability that a 48-year-old patient with low GPT, with a viral load lower than 800,000 UI/mL and zero steatosis has severe fibrosis is 0.200 (Equation 6).

$$P(\text{Severe Fibrosis} = 1) = \frac{e^{-4.459+0.064\times 48+1.412\times 0+0.772\times 0+1.408\times 0}}{1+e^{-4.459+0.064\times 48+1.412\times 0+0.772\times 0+1.408\times 0}} = 0.200 \tag{6}$$

4. Conclusion

In this study, three variables were found to be significantly associated with severe hepatic fibrosis in univariate analysis and remained independently associated with fibrosis under the multiple regression analysis: age, GPT levels ≥ 2 x upper normal limit and steatosis ≥ 5%. Although viral load (≥ 800,000 UI/mL) was not significantly associated with fibrosis, it was however maintained in the model due to the known importance of the large volume of the viral load relation with a more aggressive course of the disease. This logistic regression model for fibrosis in light-drinking chronic hepatitis C patients can be considered useful for predicting severe liver fibrosis.

REFERENCES

Bamber D. (1975): The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 12(4): 387–415.

Bland J.M., Altman D.G. (2000): Statistics Notes – The odds ratio. British Medical Journal 320: 1468.

Dawson B., Trapp R.G. (2004): Basic and Clinical Biostatistics. 4th Ed., McGraw-Hill, New York.

Hanley J.A., McNeil B.J. (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36.

Hosmer D.W., Lemeshow S. (1989): Applied Logistic Regression. John Wiley & Sons, New York.

Kutner M., Nachtsheim C., Neter J. (2004): Applied Linear Regression Models. 4th edition. McGraw-Hill.

Metwally M.A., Zein C.O., Zein N.N. (2007): Predictors and Noninvasive Identification of Severe Liver Fibrosis in Patients with Chronic Hepatitis C. Digestive Diseases Sciences 52: 582–588.

Peng C.J., Manz B.D., Keck J. (2001): Modeling categorical variables by logistic regression. American Journal of Health Behavior 25(3): 278–284.