# Statistical Model to Determine the Factors that Affect Liver Cancer

by

Merna Atef Shafeek Aziz

# Statistical Model to Determine the Factors that Affect Liver Cancer

**Dr. Medhat Mohamed Ahmed Abdel Aal**
*Professor of Mathematics and Statistics Department*
*Faculty of Commerce, Ain Shams University*

**Merna Atef Shafeek Aziz**
Faculty of Commerce, Ain Shams University

## Abstract:

This study aims to identify the risk factors that affect the survival of liver cancer (hepatocellular carcinoma) patients, using the statistical models to estimate the probability of survival of hepatocellular carcinoma (HCC) patients, especially in the light of spread of hepatitis C virus infection, and its impact on the probability of a decline in survival of these patients. Some of the risk factors such as hepatitis C virus, patient's age group, patient's gender, alpha fetoprotein (AFP), Performance Status (PS), and Child Score has been studied in this study using discriminant analysis model, logistic regression mode, artificial neural networks model, and classification and regression tree model for prediction of survival of hepatocellular carcinoma patients, and concluded that the discriminant analysis model, logistic regression model, artificial neural network analysis focusing on multilayer perceptron network algorithms, and classification and regression tree model, showed that hepatitis C virus variable, age groups variable, Child Score variable, Alpha fetoprotein variable are risk factors that affect the decrease in survival of HCC patients.

## Keywords

Primary Liver Cancer, Hepatocellular Carcinoma, Hepatitis C virus, Discriminant Analysis, Logistic Regression Analysis, Artificial Neural networks, Classification and Regression Tree.

# 1. Introduction:

## 1.1 Hepatocellular Carcinoma (HCC) Overview

Hepatocellular carcinoma (HCC), a major health problem worldwide, is one of the most common primary neoplasms of the liver and one of the most common solid tumors in the world.. Hepatocellular carcinoma is one of the 130 major causes of morbidity and mortality in the world. HCC is the commonest primary cancer of the liver, it accounts for between 85% and 90% of primary liver cancer, thus representing the major histological subtype of primary liver malignancies. HCC is one of the few cancers with clearly defined major risk factors. Hepatocellular carcinoma is a major health problem in Egypt and its incidence is increasing. It is the second most common cancer in men and the sixth most common cancers in women. HCC has a rising incidence in Egypt mostly due to high prevalence of viral hepatitis and its complications.

Hepatocellular carcinoma in Egypt is currently undergoing a shift in the relative importance of HBV and HCV as primary risk factors. Egypt is among the endemic countries for hepatitis C. Viral hepatitis C and HCC are major health problems and the burden of HCC has been increasing in Egypt, with a two-fold increase in incidence in the past 10 years.

## 1.2 The increasing importance of HCV infection in the etiology of liver cancer (HCC)

Hepatocarcinogenesis is a multistep process and involves multiple cellular signaling pathways. HCV as the major risk factors leading to the development of HCC, have been implicated in disrupting several cellular transformation. Current advances in gene expression profile have improved approach to the pathogenesis of HCC. The heterogeneity of genetic events observed in HCV-related HCCs has suggested that complex mechanisms underlie malignant transformation induced by HCV infection.

### 1.3 Overview on Hepatitis C Virus

**"Hepatitis"** means inflammation of the liver. The liver is a vital organ that processes nutrients, filters the blood, and fights infections. When the liver is inflamed or damaged, its function can be affected. Hepatitis can be caused by a variety of different viruses such as hepatitis A, B, C, D and E. Heavy alcohol use, toxins, some medications, and certain medical conditions can also cause hepatitis.

Hepatitis C is a contagious liver disease that results from infection with the Hepatitis C virus. When first infected, a person can develop an "acute" infection, which can range in severity from a very mild illness with few or no symptoms to a serious condition requiring hospitalization.

**Acute** Hepatitis C is a short-term illness that occurs within the first 6 months after someone is exposed to the Hepatitis C virus. For reasons that are not known, 15%–25% of people "clear" the virus without treatment. Approximately 75%–85% of people who become infected with the Hepatitis C virus develop "chronic," or lifelong, infection.

**Chronic** Hepatitis C is a long-term illness that occurs when the Hepatitis C virus remains in a person's body. Over time, it can lead to serious liver problems, including liver damage, cirrhosis, liver failure, or liver cancer.

## 2. Nature of the problem

Hepatocellular carcinoma (HCC) has a rising incidence in Egypt mostly due to high prevalence of viral hepatitis C and its complications. The problem of increasing incidence of hepatocellular carcinoma rate accompanying the increase in death rate of this disease, has needed, to identify the risk factors and causes, and to develop assumptions that may affect the decrease in survival of hepatocellular carcinoma (HCC) patients.

Egypt has the highest hepatitis C virus (HCV) prevalence worldwide, which has accompanying great dangers to the liver. Due to the increasing incidence of chronic viral hepatitis C, there is a state of uncertainty in answering the question: Is hepatitis C virus a dangerous and major factor that decreases the survival of hepatocellular carcinoma (HCC) patients or accompanied by other factors.

## 3. Objectives of the Study

The early detection and evaluation of risk factors which might affect the decrease of survival rate of hepatocellular carcinoma (HCC) patients is very important. The prediction of risk factors is an important pivot in saving the lives of these patients and may help doctors to focus on these factors and inform patients to avoid it. The usage of statistical methods to identify risk factors would help to improve the survival of hepatocellular carcinoma (HCC) patients.

**This study proposes to:**

1. Identify the independent variables that affect the survival of hepatocellular carcinoma (HCC) patients' group membership, and propose a statistical model to explain the relationship between the studied covariates and survival of hepatocellular carcinoma (HCC) patients.
2. Establishing a classification system using artificial neural networks model to determine group membership.

## 4. Source of data and Variables of the Study

From (1150) registered patients at Ain Shams university hospitals, radiation oncology & nuclear medicine department, gastroenterology unit, Cairo, Egypt. The data contains different types of gastrointestinal cancer patients covering a period of 5 years from year 2011 to year 2015. Only (212) patients meet the study assumptions as follows:

a) Alive patients who have hepatocellular carcinoma (HCC).
b) Dead patients who have hepatocellular carcinoma (HCC).

## 5. The Used Statistical Techniques

1. Discriminant Analysis.
2. Logistic Regression.
3. Artificial Neural Network.
4. Classification of Regression Tree.

# 6. Statistical Application Techniques

## 1.6 Discriminant Analysis:

Table (1) shows that the three steps were taken, each step includes another variable and therefore these three were included in the variables in the analysis table, because each predictor was adding some predictive power to the discriminant function. At step 1, the hepatitis C virus is the first variable to enter the model, then at step 2, the child score variable entered the model, and the age groups variable entered the model at the last step (step3) with F- values of 122.251, 24.948, and 15.819 respectively. As shown in table below.

**Table (1): Variables in the analysis, Stepwise method**

| Step | | Tolerance | F to Remove | Wilks' Lambda |
|---|---|---|---|---|
| 1 | Hepatitis C Virus | 1.000 | 97.714 | |
| 2 | Hepatitis C Virus | .894 | 124.717 | .967 |
| | Child Score | .894 | 26.607 | .682 |
| 3 | Hepatitis C Virus | .888 | 122.251 | .893 |
| | Child Score | .893 | 24.948 | .630 |
| | age groups | .994 | 15.819 | .605 |

The significance of the discriminant function is provide by Wilks' lambda value of 0.563, and a chi-square value of 119.928 corresponding to p-value of 0.000, which indicates the statistical significance of the discriminatory power of the discriminant function, and the corresponding function explains the group membership well, as shown in table (2) below, the result is statistically significant and acceptable, and all variables in the equation together improve the prediction.

**Table (2): Significance of the discriminant function, Stepwise method**

| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
|---|---|---|---|---|
| 1 | .563 | 119.928 | 3 | .000 |

Classification function coefficients table below provides Fisher's linear discriminant function coefficients, the coefficients estimates are used to estimate the probability of survival rate of HCC patients. Table (3) shows that the final model variables are hepatitis C virus, child score, and age with fisher coefficients. groups

**Table (3): Classification function Coefficients, Stepwise method**

| | Survival | |
|---|---|---|
| | dead | alive |
| age groups | 6.309 | 5.249 |
| Hepatitis C Virus | -2.078 | 3.459 |
| Child Score | 3.928 | 2.301 |
| (Constant) | -18.287 | -14.092 |

Fisher's linear discriminant functions

At the 0.05 level of significance, table (4) shows that the model is a good fit for the data with just three predictors as all are significant with p value < .05. The table below also shows the decrease in the wilks' lambda values from 0.682 at step 1 to 0.605 at step 2 tilling reaching 0.563 at the last step; this decrease in the wilks' lambda values indicates improvement in the model performance.

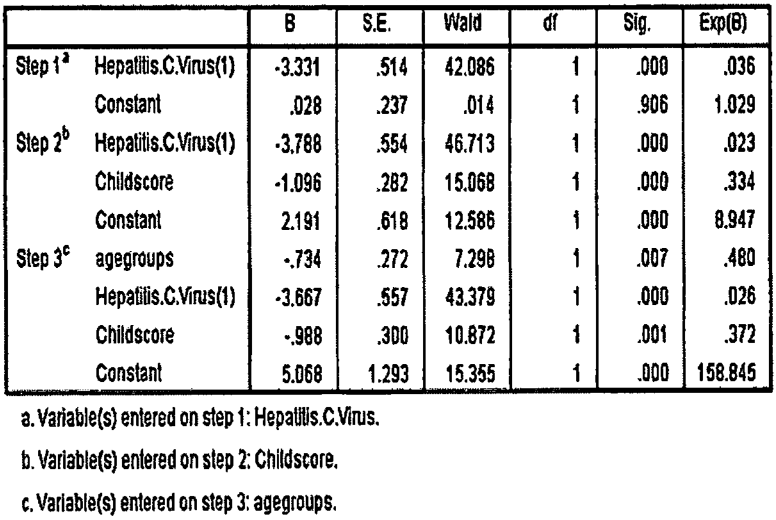
**Table (4): Wilks' Lambda, Stepwise method**

| Step | Number of Variables | Lambda | df1 | df2 | df3 | Exact F | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Statistic | df1 | df2 | Sig. |
| 1 | 1 | .682 | 1 | 1 | 210 | 97.714 | 1 | 210.000 | .000 |
| 2 | 2 | .605 | 2 | 1 | 210 | 68.118 | 2 | 209.000 | .000 |
| 3 | 3 | .563 | 3 | 1 | 210 | 53.905 | 3 | 208.000 | .000 |

## 6.2 Binary Logistic Regression Analysis

Variables in the equation table show the coefficients estimation, and Wald test that tests the significance of the coefficients of the stepwise logistic model, as shown in table (5).

Table (5): Variables in the Equation; Stepwise Method



| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Hepatitis.C.Virus(1) | -3.331 | .514 | 42.086 | 1 | .000 | .036 |
| | Constant | .028 | .237 | .014 | 1 | .906 | 1.029 |
| Step 2[b] | Hepatitis.C.Virus(1) | -3.788 | .554 | 46.713 | 1 | .000 | .023 |
| | Childscore | -1.096 | .282 | 15.068 | 1 | .000 | .334 |
| | Constant | 2.191 | .618 | 12.586 | 1 | .000 | 8.947 |
| Step 3[c] | agegroups | -.734 | .272 | 7.298 | 1 | .007 | .480 |
| | Hepatitis.C.Virus(1) | -3.667 | .557 | 43.379 | 1 | .000 | .026 |
| | Childscore | -.988 | .300 | 10.872 | 1 | .001 | .372 |
| | Constant | 5.068 | 1.293 | 15.355 | 1 | .000 | 158.845 |

a. Variable(s) entered on step 1: Hepatitis.C.Virus.

b. Variable(s) entered on step 2: Childscore.

c. Variable(s) entered on step 3: agegroups.

From the above table, the estimated coefficients are used to estimate the probability of survival rate of HCC patients, as follows:

$$\begin{aligned}\text{logit (P)} = 5.068 &- 0.734\,(\text{age groups})\\ &- 3.667(\text{Hepatits C Virus})\\ &- 0.988(\text{Child Score})\end{aligned}$$

The logit (P) above indicates that, HCV decreases the survival rate of HCC patients. Also, the less the child score, the more increase in survival rate of HCC patients. Regarding the age groups the survival rate of HCC patients decreases.

The significance of the estimated coefficients is indicated by the significance of the corresponding Wald statistics. The more the value of Wald test, the more the importance of the variable. As shown in table (5), at step 1, the hepatitis C virus is the first variable to enter the model with Wald statistic of 43.379. Then at step 2, the child score variable is the second variable to enter the model with Wald statistic of 10.872. Then at the last step, the age groups variable is the last variable to enter the model with Wald statistic of 7.298. At level of significant 0.05 for Wald test, the three covariates are statistically significant.

The results show that the hepatitis C virus is the most important independent variable affecting the survival rate of HCC patients, followed by child score, and finally age groups.

The improvement in the model performance is witnessed by three measures, as shown in the model summary table below. The three measures are the -2log likelihood, the Cox-Snell R Square, and the Nagelkerke R Square.

The decrease in value -2log likelihood statistic from 141.627 at step 1 to 124.280 at step 2 till reaching 116.046 at the last step, indicates improvement in the model performance.

The Cox-Snell R square increase in value from 0.270 at step 1 to 0.327 at step 2 till reaching 0.353 at step 3 indicates improvement in the model performance. This improvement is also shown by the increase in the Nagelkerke R square value from 0.431 to 0.564, it can be concluded that the coefficients in the model can explain 56.4% of the total variance in the dependent variable. As shown in table (6).

**Table (6): Model Summary; Stepwise Method**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 141.627 | .270 | .431 |
| 2 | 124.280 | .327 | .523 |
| 3 | 116.046 | .353 | .564 |

The value of the Hosmer and Lemeshow goodness-of-fit static computed for the stepwise model is 10.267 and the corresponding p-value computed from chi-square distribution with 8 degrees of freedom is 0.174. This indicates that the insignificance of the difference between the observed and predicted classification and the model seems to fit quite well. As shown in table (7).

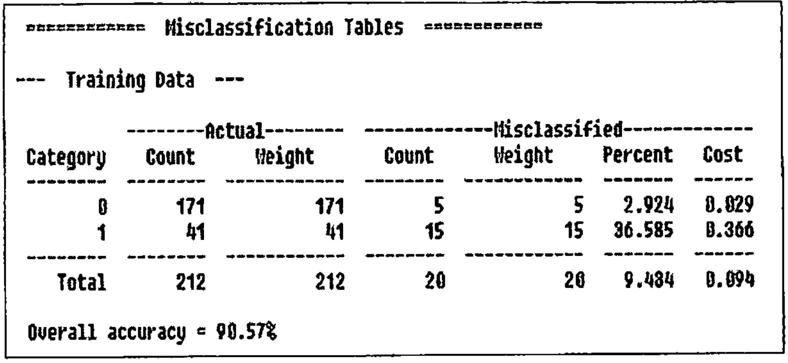
**Table (7): Hosmer and Lemeshow Test**

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 3 | 10.267 | 7 | .174 |

## 6.3 Artificial Neural Network

Table (8) shows the percentage of the misclassification for the training dataset is 9.43%, and the overall accuracy of the model is 90.57%. This indicates that the model seems to fit quite well. This indicates that the model seems to fit quite well and it can be depended on to classify a new HCC patient with a susceptibility of survival rate with 90.57% correct classification.

**Table (8): Misclassification summary table**



```
============ Misclassification Tables ============

--- Training Data ---

          --------Actual--------  ------------Misclassified------------
Category   Count       Weight      Count      Weight    Percent   Cost
--------  --------  ------------  --------  ----------  -------  ------
       0      171          171         5          5       2.924   0.029
       1       41           41        15         15      36.585   0.366
--------  --------  ------------  --------  ----------  -------  ------
   Total      212          212        20         20       9.484   0.094

Overall accuracy = 90.57%
```

A Receiver Operating Characteristic (ROC) chart displays the true positive rate (TPR) for predictions of a specific category on the vertical (Y) axis and the false positive rate (FPR) on the horizontal (X) axis. As shown in figure (1), the area under the ROC curve for the model is 0.923 which is considered an excellent classification.

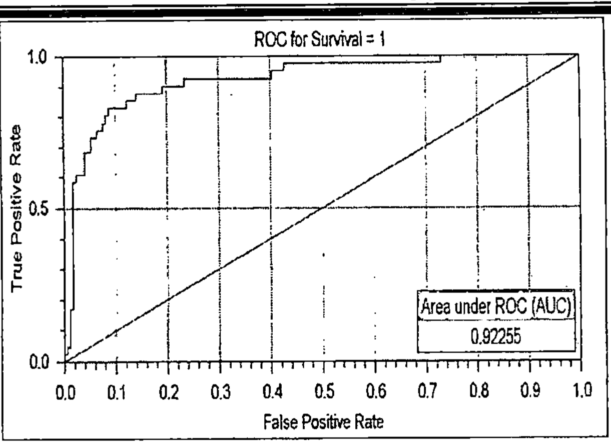


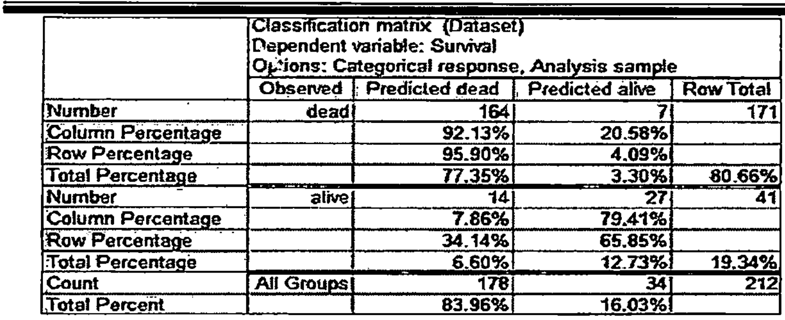
Figure (1): ROC curve, MLPNN Model

## 6.4 Classification and Regression Tree

The classification matrix below shows the accuracy of the model, the output result summarizes the observed group and the predicted group. The classification results (Table 9) reveal that the overall correctly specified group percentage “hit ratio” is 90.09% of all HCC and are correctly classified into ‘alive’ or ‘dead’ groups. It reflects the overall predictive accuracy of the model. “Dead” were classified with slight better accuracy 95.90% than “alive” 65.85% where 95.90% of all HCC dead patients are correctly classified (specificity), and 4.09% are incorrectly classified (1- specificity). So, 65.85% of all HCC alive patients are correctly classified (sensitivity), and 34.14% are incorrectly classified.

This indicates that the model seems to fit quite well and it can be depended on to classify a new HCC patient with a susceptibility of survival rate with 90.09% correct classification.

Table (9): The classification matrix



| | Classification matrix (Dataset)<br>Dependent variable: Survival<br>Options: Categorical response, Analysis sample | | | |
|---|---|---|---|---|
| | Observed | Predicted dead | Predicted alive | Row Total |
| Number | dead | 164 | 7 | 171 |
| Column Percentage | | 92.13% | 20.58% | |
| Row Percentage | | 95.90% | 4.09% | |
| Total Percentage | | 77.35% | 3.30% | 80.66% |
| Number | alive | 14 | 27 | 41 |
| Column Percentage | | 7.86% | 79.41% | |
| Row Percentage | | 34.14% | 65.85% | |
| Total Percentage | | 6.60% | 12.73% | 19.34% |
| Count | All Groups | 178 | 34 | 212 |
| Total Percent | | 83.96% | 16.03% | |

The ROC curve is created by plotting the true-positive rate (sensitivity) over the false-positive rate (1-specificity). As shown in figure (2), the area under the ROC curve is 0.905.

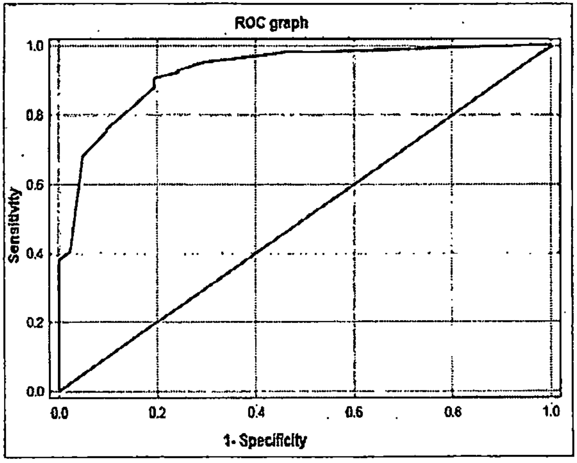


Figure (2): ROC Curve, CART Model

## 7. The Results

The risk factors of decreasing a survival of HCC patients used in this study are gender, age groups, hepatitis C virus, child score, alpha fetoprotein, and performance status. When applying discriminant analysis, the significant testing for this model using wilk's lambda test showed that hepatitis C virus, child score, and age groups are the significant factors. Therefore, the final model variables for the discriminant analysis model are hepatitis C virus as the first factor, then child score, and age groups. Also, when applying binary logistic regression analysis, the significant testing for this model using Wald test and likelihood ratio showed that hepatitis C virus, child score, and age groups are the significant factors. Therefore, the final model variables for the logistic regression model are hepatitis C virus as the first factors, then child score, and age groups.

A sensitivity analysis was performed to assess the relative significance of input parameters in the system model and to rank the importance of the variables. The global sensitivity of the input variables against the output variable was expressed as the ratio of the network error with a given input omitted to the network error with the input included. A ratio of 1 or lower indicates that the variable diminishes network performance and should be removed. The training data set was also used to calculate the variable sensitivity ratio for the artificial neural network model. The variable sensitivity ratio values for the outcome variable (survival of HCC patients) in relation to gender, age groups, hepatitis C virus, child score, alpha fetoprotein, and performance status. In the ANNs model, hepatitis C virus was the most influential (sensitive) parameter affecting survival of HCC patients followed by child score followed by age groups and alpha fetoprotein were variables associated with survival of HCC patients. All variable sensitivity ratios values exceeded 1, which indicated that the network performed better when all variables were considered.

The results of the classification and regression tree model provided the first important splitting with hepatitis C virus variable which was selected as the best predictor of survival rate of HCC patients. To a lesser extent, age groups, child score, and alpha fetoprotein were variables associated with survival of HCC patients.

This indicated that the most important effective risk factors which predict survival of HCC patients are: hepatitis C virus, child score, age groups, and alpha fetoprotein. The other variables in the dataset did not show any interaction with survival rate.

The performance predictive accuracy of the classification techniques and effectiveness of each technique used to predict of survival HCC patients shown in table (10).

Table (10): The Performance of Classification Techniques

| Classification Technique Accuracy | Discriminant Analysis | Logistic Regression | Neural Networks | Classification and Regression Trees |
|---|---|---|---|---|
| Sensitivity | 80.5% | 73.2% | 63.4% | 65.85% |
| Specificity | 88.3% | 92.9% | 97.1% | 95.90% |
| Hit ratio | 87 % | 89.2% | 91% | 90.1% |
| Area under Roc curve | 0.831 | 0.844 | 0.923 | 0.905 |

As shown in table (10):

- The discriminant analysis and logistic regression model have the highest sensitivity when artificial neural network model and the classification and regression tree model have the highest specificity.
- Artificial neural network model and classification and regression tree model appear to be the most effective as they have the highest percentage of correct predictions (hit ratio) for survival rate of HCC patients 91%, 90.1% respectively, followed by logistic regression model 89.2% and discriminant analysis 87%.
- The full models have large areas under the ROC curve. It was found that artificial neural network model have the largest area under the ROC curve, and the largest hit ratio. As a result, this model will be reliable in this study.

## 8. Conclusions:

It was concluded by all the classification techniques used in this study that hepatitis C virus variable is the first risk factor of the decrease in survival of hepatocellular carcinoma patients. Because of the high prevalence of HCV infection and the high rate of HCC occurrence in patients with hepatitis C virus, HCC is currently the main cause of death in patients having HCV. As proven by artificial neural networks model, the child score variable is the second risk factor causing the decrease in survival of hepatocellular carcinoma patients when it increases. Also, age groups variable was concluded to be the third major risk factor; the elder HCC patients have less survival rate than the younger age groups. Finally, the last variable and the least to decrease the survival rate is alpha fetoprotein. It was concluded that HCC patients with higher AFP levels show a higher mortality rate, which appears to be attributable to the growth promoting properties of AFP.

# References

[1] Breiman JH, L.O., RA Friedman, Charles J. Stone, ***"Classification and Regression Trees"***. 1984: Chapman & Hall.

[2] David Kriesel, ***"A Brief Introduction to Neural Networks"***. 2005: dkriesel.com.

[3] David W. Hosmer ,Stanley Lmeshow, ***"Applied Logistic Regression"***. 2000: Join Wiley & Sons.

[4] Elizabeth M. Lehman, A.S.S., Kadry Ismail, et al, ***"Patterns of hepatocellular carcinoma incidence in Egypt from a population-based cancer registry"***. Hepatology Research, May 2008. **38**(5): p. 465–473.

[5] Joseph F. Hair, J., William C. Black., Barry J. Babin, Rolph, E. Anderson, ***"Multivariate Data Analysis"***. 7th ed. 2010: Prentice Hall.

[6] ***Logistic Regression Analysis.com homepage.*** **Available from:** http://logisticregressionanalysis.com/

[7] Mark L. Wilson, C., Susan K. Murray, Zhenhua Yang, R.Palmer Beasley, ***"Dynamics of Liver Disease in Egypt: Shifting Paradigms of A Complex Etiology"***. 2008: Elizabeth M.Lehman.

[8] Staff of the Comprehensive Cancer Center's Multidisciplinary Liver Tumor Clinic, ***"Liver Cancer Patient Handbook"***. 2011: University of Michigan Comprehensive Cancer Center.