Statistical Analyses and Artificial Neural Networks for Prognoses in Hepatitis C

Andreea Drăgulescu[1], Adriana Albu[2], Cătălin Gavriluță[3], Ştefan Filip[4], Karoly Menyhardt[5]

[1] Medicine and Pharmacology University, Timişoara, Romania, adragulescu@cardiologie.ro
[2] Department of Automation and Applied Informatics, Politehnica University, Timişoara, Romania, adriana.albu@aut.upt.ro
[3] Politehnica University, Timişoara, Romania, tr.trope@gmail.com
[4] Politehnica University, Timişoara, Romania, stefy.filip@gmail.com
[5] Politehnica University, Timişoara, Romania, karoly_m@cmpicsu.utt.ro

Abstract: This paper analyses a large database with hepatitis C virus infected patients. There are made a lot of statistical analyses on the records of this database in order to determine the evolution of biological parameters during the treatment. That's for what it was also implemented a system which offeres predictions about the same parameters, using artificial neural networks. The results of the statistical analyses and the predictions of the system indicate to the same conclusions. It encourages the use of such a system to facilitate the physicians' work.

Keywords: hepatitis C virus infection, statistical analysis, artificial neural networks

1 Introduction

Hepatitis C is a serious and frequent disease and its evolution has to be carefuly overseen during the treatment. Even the efficency of the hepatitis C treatment inproves continously, the burden of this infection will remain a major issue for the next several decades.

The patients fro this study have been kept under observation for 12 months to establish the treatment's influence on the evolution of the biological indicators. Three different treatment schemes have been instituted:

- Simple Interferon (IFN);
- Peg interferon α-2a;
- Peg interferon α-2b.

The biological parameters were determined every three months and their evolution in time was monitored, trying to establish the relations between the biological indicators values (*TGP, TGO, GGT, ARN VHC*) and time, on patient groups sampled on their answer to the treatment. There are six types of reactions to the treatment considered as *answer-code*: 0-responder, 1-no responder to IFN, 2-no responder to Peg IFN, 3-backslide in treatment with IFN, 4-backslide in treatment with Peg IFN, 5-recess any treatment.

The correlations *biological indicators – time – code of reaction* have been 3D represented as functions of *z=f(x,y)* type. The obtained results are presented in Chapter 2.

On the other hand, the article presents (in Chapter 3) an artificial neural network trained to predict the evolution of the biological indicators. By introducing the patients personal data, as well as the results of biological tests at the treatment start, the implemented system can indicate the evolution in time of the illness. The use of the neural network presents the same conclusions as the statistical analysis.

2 Statistical Analysis

The investigated population contained 193 patients registered at the Emergency Clinical Hospital in Timisoara, Gastroenterology Department between 2003 and 2005 (120 women and 73 men aged between 14 and 67 years in old).

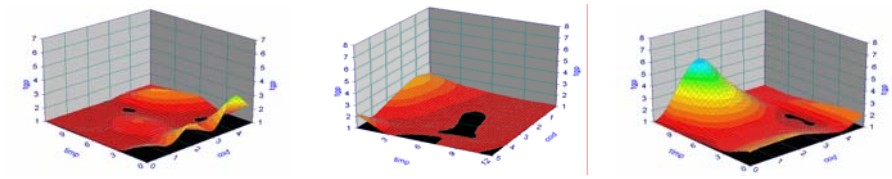
The study of the 3D correlations was made for the main biological indicators. For data processing the *TableCurve 3D* software was used. It can make the nonparametric interpolation of some 3D data multitudes by using the homogeneous grid method. Different calculating algorithms have been used (Akima [1], [2], [3], Bicubic [2], B-Spline [4], Preusser [6, 7], Renka [8], [9], [10], [11] and Watson [12]) and it has been observed that the Watson algorithm gave the biggest values of the correlation coefficient R^2. Therefore all the 3D representations and regressions established for the *z=f(x,y)* functions were represented by using this algorithm.

Even the information in the database is incomplete (there are patients which didn't follow the treatment for 12 months or were not periodically showed up for analyses) the indicated processing method can overpass it. Based on Watson algorithm the missing data have been calculated by extrapolation [12].

2.1 The TGP Biological Indicator Analysis

The *TGP=f(time, answer-code)* indicator analysis, as a answer- to the *IFN* treatment (Figure 1a), shows, for answer-code 0 patients, an emphasized decrease

of the *TGP* indicator in the first three months of treatment from the initial value *TGP_0=2.190* to the value *TGP_3=1.109,* followed by a milder decrease until the end of investigation where *TGP_12=1*. For the other answer-codes, the representation shows some differencies. For example, the maximum value of the *TGP* indicator *TGP_0= 3.525* for answer-code 4 and a smaller value *TGP_0=2.590* for answer-code 2. The represented surface allows the predicted evolution of all the answer-codes to be followed. The missing values can be obtained from the graphic representation. Therefore, for answer-code 5 for example, where there were values only for the initial phase and for the 3 month period, it can be observed that the representation gives calculated information even for the following months. For the 6 months phase the *IFN* treatment leads to a value *TGP_6=1.5,* then it rises to *TGP_9=1.8* and decreases again in the next period of time. Such values are readable on the graphic for all answer- codes in all temporal phases of the study.



a) patients treated with *IFN* b) patients treated with *Peg INF α-2a* c) patients treated with *Peg IFN α-2b*

Figure 1
The variation as function of time and answer code of the *TGP* indicator

The analysis of the same indicator for *Peg IFN α-2a* treatment (Figure 1b) shows a practically continuous decrease of the indicator's value from *TGP_0=2.018* to *TGP_12=1.093* for patients with answer-code 0.

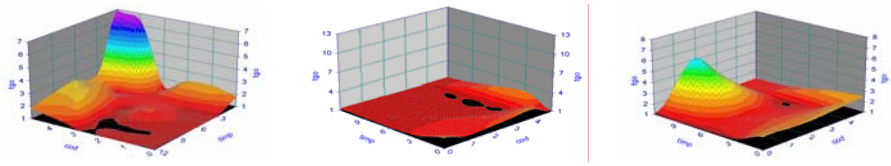
The representation verifies correctly enough the data for answer- code 4, where *TGP_0=1.190* and once with 6th month of the study it should stabilize at the indicator's value of 1. However, it can be observed from the representation that the technique of data extrapolation changes the *TGP* value in the period of 9-12 month. Therefore, at the end it is not obtained *TGP* measured value 1, but an estimated value of *1.008*, which represents an error of 0.8%. In conclusion, the verification of the existent data in the representation leads to the certitude that the estimated data are valid, having an acceptable error coefficient.

The same observations are valid for the surface *TGP=f(time, answer-code)* for the treatment with *Peg IFN α-2b* (Figure 1c). It can be observed that patients with answer- code 2 do not react to this treatment and that the indicator's value rises from *TGP_0=2.184* to *TGP_12=5.258*. For patients with answer- codes 1 and 3, for whom there is no information in the database, the representation in Figure 1b allows a relatively correct prediction, its correlation coefficient being $R^2=0.75$.

This way of approaching the problem makes the estimation of the *TGP* indicator's values a 70% true. It is calculated through the summing up of the probabilities induced by the correlation coefficients values and the statistical belief used.

2.2 The TGO Biological Indicator Analysis

The *TGO=f(time, answer code)* indicator analysis has been done in the same manner. It is represented in Figure 2 (a, b and c), for the three types of treatments.



a) patients treated with *IFN* b) patients treated with *Peg INF α-2a* c) patients treated with *Peg IFN α-2b*

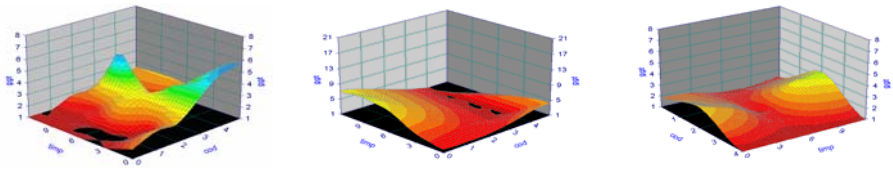
Figure 2
The variation as function of time and answer code of the *TGO* indicator

By comparing Figures 2b and 2c it can be observed that the *Peg IFN α-2a* treatment is more efficient than the *Peg IFN α-2b* treatment because the decrease of the indicator's value is continuous for the former and it is predicted a lower final value than for the latter. The representations allow the *TGO* indicator's value to be assessed for patients with answer-codes 3 and 5 which are not present in the database and which have been estimated through the Watson algorithm. These data are very important, especially for the answer-code 5 patients that have stopped the treatment because of collateral effects.

2.3 The GGT Biological Indicator Analysis

The *GGT=f(time, answer-code)* indicator was similarly analyzed. The obtained graphical representations are presented in Figure 3 (a, b and c) for the three types of treatment.

It is proved that the *Peg IFN α-2b* treatment is the most efficient, as the *GGT* indicator's value decreases from *GGT_0=4.974* to *GGT_12=1.314*. The same treatment has a fluctuant influence on the code 2 patients. In the first 6 months the treatment leads to a decrease value of the indicator from *GGT_0=2.843* to *GGT_6=1.158,* followed by a come back to *GGT_12=3.472*.



a) patients treated with *IFN* b) patients treated with *Peg INF α-2a* c) patients treated with *Peg IFN α-2b*

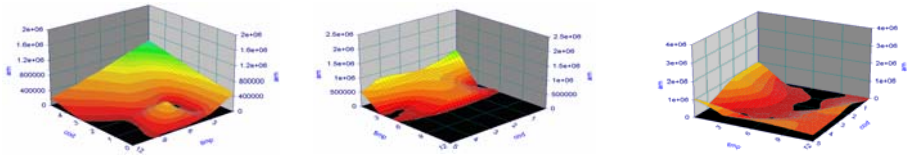
Figure 3
The variation as function of time and answer code of the *GGT* indicator

The data in Figure 3b definitely show the inefficiency of the *Peg IFN α-2a* treatment over the value of the *GGT* indicator for answer-code 0 patients. The representation presents an estimation of the reaction of the answer-code 2 patients, for whom there was no data starting with the 6th month of the study. For the same patients, it is estimated the value *GGT_12=5*, the correlation coefficient of the representation being $R^2=0.74$. Answer-code 2 patients are no-responders to the *PegIFN* treatment, therefore it is important to know what is the predicted value of the indicator on the areas in which the database is incomplete.

2.4 The ARN VHC Biological Indicator Analysis

The variation of the *ARN VHC=f(time, answer-code)* indicator is represented in Figure 4 (a, b and c) for the three types of treatment. Regarding the evolution of this indicator, all treatments are efficient.

It is observed that the patients which have been no-responders to the *IFN* treatment (answer-code 1) have a positive evolution in the first three months of treatment, followed by a decrease in the next period, but at a lower value of the indicator than the initial. From now the missing data can be completed by predicting that an eventual continuation of the treatment could lead to a significant decrease of the *ARN VHC* value. As a result, the patient would become a responder.



a) patients treated with *IFN* b) patients treated with *Peg INF α-2a* c) patients treated with *Peg IFN α-2b*

Figure 4
The variation as function of time and answer code of the *ARN VHC* indicator

In addition to this, the figure gives probable information of the indicator's values for answer-code 3 patients, information that does not exist in the database. For these, with a probability of 65%, the initial value of the indicator could be *ARN VHC_0≅ 750000*. So, this information could be taken into consideration together with another indicators, in order to attach a code, as realistic as possible, to the patients considered in a future study. Figures 4b and 4c are useful for the prediction of the values of the *ARN VHC* indicator for patients with answer-codes 4 and 5, for whom the database is also incomplete.

3 Predictions using Artificial Neural Networks

The same conclusions as before regarding the evolution of the biological indicators during the hepatitis C treatment have been obtained throw the implementation of a system based on artificial neural networks. This system will predict the values of the *TGP*, *TGO*, *GGT* and *ARN VHC* after 3, 6, 9, and 12 months of treatment.

Artificial neural networks are a branch of the artificial intelligence and they have been developed to reproduce human reasoning and intelligence. The initial idea was that, in order to reproduce intelligence, it would be necessary to build systems with architecture similar to the brain one. Therefore, artificial neural networks are built by the interconnection of certain primary elements, whose structure is similar to the biological neuron. Like the human brain, these artificial neural networks are able to recognize patterns, manage data and, most important, they have the ability of learning [5] (ability to adapt to the informational environment specific to the problem they are solving).

The system presented here was developed using feed-forward neural network with back-propagation learning algorithm. Such a network receives a series of inputs and its outputs are the results of the problem. Between the two levels (input and output) there can be a number of hidden levels. The elements on each level (neurons) are interconnected throw links called synapses. These have a weight, which can be modified along the training of the network.

In this case, the system has been designed as a network of neural networks. Each neural network has a layer of 10 hidden neurons, a single output unit and a variable number of inputs. For each of the four biological indicators that have been studied, there are four layers of neural networks. The networks on the first layer receive as inputs: patient's age, sex, location (rural/urban), treatment scheme, Knodell score, hepatic fibrosis score and value of the parameter for which the prediction is made, at the initial moment (before the treatment start). These networks have as output the value of the biological parameter at 3 months. On the following layers the networks have the same structure as the first layer ones, but

they have in addition, as inputs, the outputs of the networks on the former layers; therefore, the networks on the last layer will have not 7 inputs (as the networks on the first layer) but 10 (the initial inputs and the values of biological indicators at 3, 6, and 9 months). Figure 5 describes the architecture of this neural network.

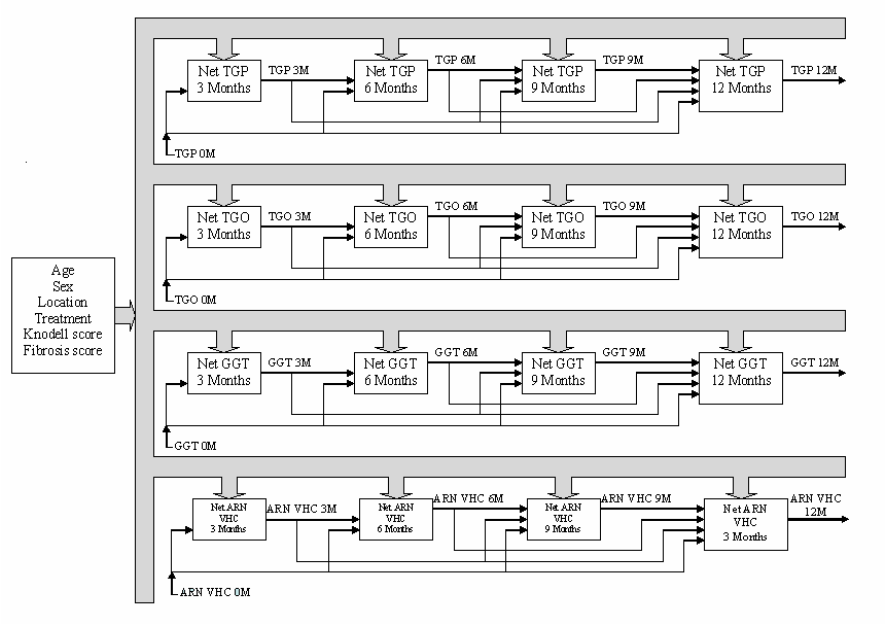


Figure 5
The architecture of the artificial neural network

The advantage of this architecture is that the input data are processed separate for each biological indicator. The disadvantage is that the errors are propagated through the system because the result of the networks from the first level (together with their errors) are used in the following levels. But this disadvantage can be minimised by lerning process.

The application has been projected in the Matlab 7.0 environment, which has a toolbox totally dedicated to the neuronal calculus. The system offers for each evaluated biological indicator predictions regarding the next 12 months evolution, indicating its growing tendency, its stabilizing or decreasing tendency (Figure 6).

The user has to choose a range regarding the age of the patient, the sex, the location where the patient lives (rural/urban), the treatment (IFN, Peg interferon α-2a or Peg interferon α-2b) and has to introduce the values of the Knodell score and of the fibrosis score. It is also necessary to introduce the values of the biological indicators before the treatment.

Looking at the predicted tendency of the biological indicators during the treatment, a physician can estimate if the patient will respond to a treatment or not.

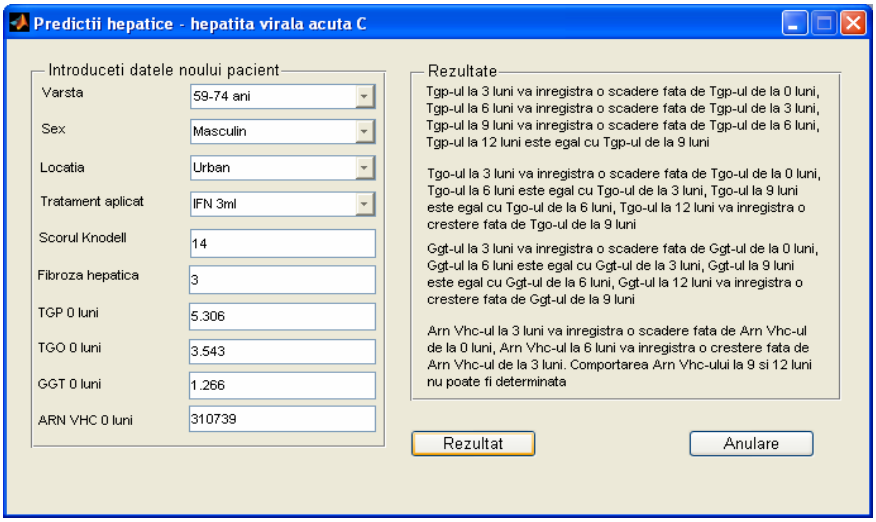


Figure 6
The prediction of biological indicators evolution

Conclusions

The statistical analysis, as well as, the implemented system offers the possibility to predict the evolution of patients in time. The hepatitis C treatment is very expensive and severe side effects can appear very often. Therefore, it is important to identify those patients who most probably can react to the treatment, so that the others can be protected from a treatment with no benefits. That's for what the use of such a system can support the physician decision concerning the treatment.

References

[1] Akima H.: A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points, ACM Transactions on Mathematical Software, Vol. 4, 1978, pp. 144-159

[2] Akima H.: Algorithm 760: Rectangular-Grid-Data Surface Fitting that has the Accuracy of a Bicubic Polynomial, ACM Transactions on Mathematical Software, Vol. 22, No. 3, Sept. 1996, pp. 357-361

[3] Akima H.: Algorithm 761: Scattered-Data Surface Fitting that has the Accuracy of a Cubic Polynomial, ACM Transactions on Mathematical Software, Vol. 22, No. 3, Sept. 1996, pp. 362-371

[4] De Boor C: A Practical Guide to Splines, Springer-Verlag, 1978, pp. 332-346

[5] Maiellaro P. A., Cozzolongo R., Marino P.: Artificial Neural Networks for the Prediction of Response to Interferon Plus Ribavirin Treatment in Patients with Chronic Hepatitis C, Current Pharmaceutical Design, 2004, Vol. 3, pp. 2101-2109

[6] Preusser A.: Efficient Formulation of a Bivariate Nonic C2-Hermite Polynomial on Triangles, ACM Transactions on Mathematical Software, Vol. 16, No. 3, Sept. 1990, pp. 246-252

[7] Preusser A.: Algorithm 684: C1 and C2-Interpolation on Triangles with Quintic and Nonic Bivariate Polynomials, ACM Transactions on Mathematical Software, Vol. 16, No. 3, Sept. 1990, pp. 253-257

[8] Renka R: Algorithm 660: QSHEP2D: Quadratic Shepard Method for Bivariate Interpolation of Scattered Data, ACM Transactions on Mathematical Software, Vol. 14, No. 2, June 1988, pp. 149-150

[9] Renka R: Multivariate Interpolation of Large Sets of Scattered Data, ACM Transactions on Mathematical Software, Vol. 14, No. 2, June 1988, pp. 139-148

[10] Renka R: Algorithm 751: TRIPACK, A Constrained Two-Dimensional Delaunay Triangulation Package, ACM Transactions on Mathematical Software, Vol. 22, No. 1, March 1996, pp. 1-8

[11] Renka R: Algorithm 752: SRFPACK, Software for Scattered Data Fitting with a Constrained Surface under Tension, ACM Transactions on Mathematical Software, Vol. 22, No. 1, March 1996, pp. 9-17

[12] Watson D.: Nngride, An Implementation of Natural Neighbor Interpolation, David Watson, P.O. Box 734, Claremont, WA 6010, Australia, 1994