UDC 004.67

ROC ANALYSIS AS A BINARY CLASSIFICATION TOOL

V. DASHKO, P. IVANOVA, V. LIVINSKAYA
Belarusian-Russian University, Mogilev, Belarus

The article presents a study that demonstrates the possibility of applying the method of space in the analysis of the results of clinical research.

The main purpose of this work was to identify statistically significant differences in the indicators of the assessment of the hemodynamics of the liver using reohepatography (RGG) of patients who have various pathologies and undergo treatment in the regional hospital of Mogilev [1]. This technique is based on fixing changes in the resistance of live tissue in an alternating electric field of high frequency. With the use of the Rheo-Spectrum (Neurosoft) software, certain physical characteristics of the rheographic curve (Figure 1) describing the dynamics of resistance of live tissue for a specific time interval (Table 1) were obtained.

Table 1. – Parameters of the rheographic curve

The RRG parameter	Designation
amplitude of the arterial part of the wave	A_{art}
time of pulse wave propagation from the heart	Q_x
Systolic maximum. Rheograms	A_{syst}
Time of the rising part of the wave	α
rapid blood filling time	α_1
slow blood filling time	α_2
total systole time	T_{total}
duration of catacrota	T_{cat}

As factors, we considered various parameters recorded by the device, as well as, calculated by the authors, the area under the rheographic curve (RC) as a single integral indicator. Using the R language, a dataset of 9 columns was formed from separate files in text format containing information about each patient, eight of which were physical parameters of the curve, and the ninth was responsible for the patient's belonging to one of 2 groups: patients with one of 3 pathological conditions: multiple organ dysfunction syndrome (SPOD), patients with liver cirrhosis, patients with chronic diseases of the stomach and pancreas (gastro), and the control group - practically healthy. The assignment of a patient to one of the two groups is described by a binary variable that takes two values: 1-the patient is sick, 0-the patient is healthy.

Table 2 shows the AUC values obtained by analyzing the results of measurements of the resistance of living tissue in a high-frequency alternating electric field.

Table 2. –ROC analysis results

Test Variable	Area	Std. error	Type 1 Error	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Area of RC	0,763	0,053	0,000	0,660	0,867
α_1	0,598	0,061	0,138	0,478	0,719
α_2	0,632	0,059	0,047	0,517	0,746
T_{cat}	0,663	0,058	0,014	0,550	0,776
RI	0,690	0,061	0,004	0,570	0,810
DIA	0,584	0,061	0,204	0,465	0,703
V_{max}	0,801	0,046	0,000	0,710	0,892

Analyzing the results obtained, it is concluded that the area under the rheographic curve can serve as the best predictor of pathology.

According to the available parameters, using the VBA macro, the areas under the curve were calculated for each patient using the area method:

Significant differences in the groups were identified using the nonparametric Mann-Whitney statistical test. This test is intended to compare two independent groups of a small amount of quantitative data.

The method is based on determining whether the zone of overlap of values between two variation series (a ranked series of parameter values in the first sample and the same in the second sample) is small enough. The U-test is suitable for comparing small samples: each of the samples must have at least 3 attribute values. The condition for applying the Mann-Whitney U-test is the absence of matching attribute values in the compared groups (all numbers are different) or a very small number of such matches.

The lower the value of the criterion, the more likely it is that the differences between the parameter values in the samples are reliable [2].

The result of testing the hypothesis that there are no differences in the selected groups for all parameters of the rheographic curve is presented in table 3.

Table 3. – No-Difference Hypothesis Tests

Rheographic curve parameters	U	Z	p-level
Q_x, sec	785,5000	-0,499723	0,617271
α_1, sec	644,0000	1,773791	0,076099
α_2, sec	613,0000	2,052916	0,040082
T_{total}, sec	585,5000	2,300526	0,021419
T_{cat}, sec	579,5000	2,354550	0,018546
RI, c.u.	391,0000	4,051808	0,000051
DIA, %	693,5000	1,328092	0,184149
V_{max}. Ohms/s	339,5000	4,448602	0,000009
V_{avr}, Ohms/s	323,0000	4,599324	0,000004
Area under curve	314,0000	4,745117	0,000002

Comparing the values of the error of the first kind of p-level with the value of 0.05, we conclude that the best factor separating the groups can be the integral indicator of the area under the curve.

To further test the possibility of using the area under the rheogram as a predictor of the patient's condition, the ROC analysis was performed. During the ROC-analysis for the factor that acts as a classifier (in our case, the area under the rheographic curve) special metrics are calculated, which are the characteristics of the quality of the partition - sensitivity (the relative index of correctly classified positive cases) and specificity (the proportion of incorrectly classified negative cases), according to which the ROC-curve is constructed.

The ROC-analysis method assumes that for each of the ordered values of the factor that determines the division into two groups (for example, whether the patient is ill or not), the proportion of truly positive cases (Se or sensitivity) and the proportion of truly negative cases that were correctly identified by the model (specificity or Sp) are calculated. Next, for all the calculated values, a ROC-curve is constructed in the coordinate system, in which the values of 1-Sp are plotted along the abscissa axis, and the values of Se are plotted along the ordinate axis. To quantify the classification results, the area (AUC) under this curve and the 95% confidence interval are calculated for each factor. The split is considered qualitative if the value 0.5 does not fall into the confidence interval, and the AUC itself is>0.5. Figure 1 shows the ROC-curve for the predictor of area under the rheographic curve.

The ideal diagnostic test should have an r-shaped characteristic curve passing through the upper-left corner, in which the proportion of true positive cases is 1, and the proportion of false positive cases is 0. The closer the characteristic curve passes to the value (0; 1), the higher the effectiveness of the test. A test in which the characteristic curve passes closer to the diagonal of the graph and does not resemble an r-shape is useless. In our case, the diagnostic test is effective, because it resembles the letter "r", and is also located above the diagonal of the graph.
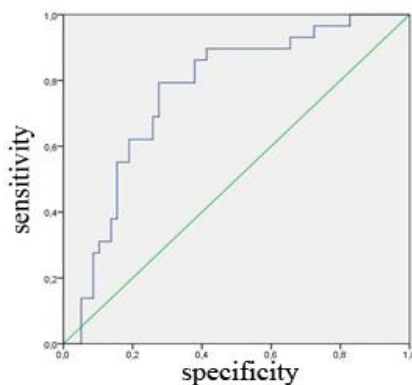

Fig. 1. – Roc-curve for the indicator of area under the rheographic curve as a predictor of disease

Quantification of the Roc-curve (AUC and 95 % interval calculated for the area by formula (1) is presented in Table 4.

$$SE(AUC)=\sqrt{\frac{AUC(1-AUC)+(n-1)(Q_1-AUC^2)+(m-1)(Q_2-AUC^2)}{n*m}},$$

(1)

For multiplicity it is indicated:

$$Q_1=\frac{AUC}{2-AUC};$$

$$Q_2=\frac{2*AUC^2}{1+AUC}.$$

Table 4. – Quantification of the Roc-curve

Area	Std. error	Asymptotic 95% Confidence Interval	
		Lower Bound	Upper Bound
0,765	0,053	0,662	0,869

REFERENCES

1. Точило С.А. Интегративный показатель состояния артериального печеночного кровотока у пациентов при критических состояниях / С.А.Точило, А.Л. Липницкий, А.В. Марочков, А.А. Антипенко, О.Л.Борисов, В.А. Ли-винская //Вестник Витебского государственного медицинского университета. 2019. Т. 18. № 3. С. 52-60.
2. Библиотека постов MEDSTATISTIC об анализе медицинских данных [Электронный ресурс]. - Режим доступа: https://medstatistic.ru/methods/methods2.html. - Дата доступа: 28.03.2021.