Cancer diagnostic method based on pattern recognition of DNA changes in buccal epithelium in the pathology of the thyroid and mammary glands

Dmitri A. Klyushin[(1)], Yuri I. Petunin[(1)], Roman I. Andrushkiw[(2)], Natalya V. Boroday[(3)], and Karelia P. Ganina[(3)]

[(1)] Department of Cybernetics,
Kyiv National Shevchenko University,
Kyiv, Ukraine

[(2)] Department of Mathematical Sciences and
Center for Applied Mathematics and Statistics
New Jersey Institute of Technology, Newark, NJ 07102

[(3)] R.E. Kavetsky Institute of Experimental Pathology,
Oncology and Radiology,
The National Academy of Sciences of Ukraine,
Kyiv, Ukraine

CAMS Report 0203-04, Fall 2002

Center for Applied Mathematics and Statistics

NJIT

Cancer diagnostic method based on pattern recognition of DNA changes in buccal epithelium in the pathology of the thyroid and mammary glands

D.A. Klyushin[1], Yu. I. Petunin[1] , R.I. Andrushkiw[2], N.V. Boroday[3], K.P. Ganina[3]

[1]Department of Cybernetics, Kyiv National Shevchenko University,
Kyiv, Ukraine

[2]Department of Mathematical Sciences and Center
for Applied Mathematics and Statistics,
New Jersey Institute of Technology, Newark, NJ, USA

[3]R.E. Kavetsky Institute of Experimental Pathology, Oncology and Radiology,
National Academy of Science of Ukraine, Kyiv, Ukraine

Corresponding author: R.I. Andrushkiw, Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, N.J. 07102 <Andrushkiw@njit.edu>

Keywords*: DNA, malignancy-associated changes, interphase nuclei, buccal epithelium, confidence interval, pattern recognition, breast cancer, cancer of thyroid gland.*

Abtsract. *The object of this investigation is to study, from the point of view of statistical and geometrical theory of pattern recognition, the peculiarities of the distribution of optical density of DNA in the interphase nuclei of buccal epithelium present in the pathology of the thyroid and mammary glands. Two new indices characterizing this distribution (ratio of modal class volumes and relief index) are proposed. It is shown that in malignant neoplasms of the thyroid and mammary glands the changes in the nuclei of buccal epithelium are characterized by an increase in the optical density of DNA in a range from 0.15 to 0.30 in conventional units of measure, as compared with its values in benign pathological processes. The sensitivity of the proposed criterion in the case of the diseases of the thyroid gland is equal to 76.2 % and the specificity is equal to 85.8 %. In the case of diseases of the mammary gland (excluding IDLC) we have discovered that the sensitivity of the method is equal to 94.29 % and its specificity equal to 90.91 %. In the case of diseases of the mammary gland (including IDLC) we have discovered that the sensitivity of the method is equal to 71.42% and its specificity is equal to 90.91%.*

1. Material and method. The scanograms of the interphase nuclei of buccal epithelium in patients suffering from adenocarcinoma of the thyroid gland (15 cases ATG), nodular goiter (15 cases NG), autoimmune thyroiditis (27 cases AT), and also fibroadenoma (10 cases FA), fibroadenomatosis (12 cases FAM), and cancer of the mammary gland (35 cases - 13 cases infiltrative ductal cancer (IDC), 9 cases of

infiltrative lobular cancer (ILC), 9 cases of infiltrative ductal-lobular cancer (IDLC) and 4 cases of scirrhus) are investigated. Two new texture indices characterizing the space distribution of DNA and the functional state of genome are proposed.

For the purposes of this study smears from various depth of the spinous layer were obtained (conventionally they were denoted as median and deep), after gargling and removing the superficial cell layer of the buccal mucous. The smears were dried out under room temperature and fixed for 30 min in Nikiforov's mixture. Then, a Feulgen reaction was made with cold hydrolysis in 5 n. HCl for 15 min, under the temperature t=21-22 ^{0}C. Optical density of the nuclei was registered by a cytospectrophotometer, using the scanning method wave length 575 nm and probe 0.05 mcm. We investigated from 10 to 20 nuclei in each preparation. The DNA-fuchsine content in the nuclei of the epitheliocytes was defined as a product of density and area (in terms of conventional units). The scanograms obtained as a results of the investigations of the nuclei of the cells were analyzed using statistical and geometric methods of pattern recognition[1-2].

The scanogram of the DNA distribution is a rectangular matrix $R = \left\| r_{ij} \right\|_{i=\overline{1,m}}^{j=\overline{1,n}}$, where r_{ij} are values of pointwise optical density of chromatin in interphase nuclei of the cell, expressed in terms of conventional unit of measure, and n,m are the numbers of points of the scanogram along the vertical and horizontal directions, respectively. Usually a scanogram contains 8 or 9 rows and columns, hence it consists of 64 or 81 numbers. From each patients 10 to 20 cells were taken.

The first index, called the *ratio of modal class volumes*, is obtained by considering the set of all scanograms as an unarranged set of random values from some general population, and by distributing this set into 3 modal classes consisting of the random values from the predefined ranges

$$M_1 = \left\{ s_{ij} : 0 \le s_{ij} < 0.15 \right\},$$
$$M_2 = \left\{ s_{ij} : 0.15 \le s_{ij} \le 0.30 \right\},$$
$$M_3 = \left\{ s_{ij} : s_{ij} > 0.30 \right\}$$

and, finally, by calculating the ratio of volumes of the modal classes M_1 and M_2 in the *k*th scanogram:

$$V_k = \frac{cardM_1^{(k)}}{cardM_2^{(k)}},$$

where $cardM_j^{(k)}, j = 1,2$ is the number of the elements from the modal class $M_j^{(k)}$ (for example, $cardM_2^{(k)}$ is the number of points in the *kth* scanogram, where the DNA optical density varies from 0.15 to 0.30). The *ratio of modal class volumes* is characteristic for each patient and is given by the average of all scanograms:

$$V = \frac{1}{N}\sum_{k=1}^{N} V_k$$

This index is statistical in nature, since it contains the information about the distribution of the DNA optical density in the interphase nuclei of epitheliocytes in buccal epithelium.

The second index, called *relief index*, is based on the geometrical interpretation of the features of the patient's scanogram. To calculate this index we consider the patient's scanogram as a surface of the function of two arguments $s_{ij} = s(i, j)$, where (i, j) are the coordinates of the points in scanogram (see Figures 1-2) .

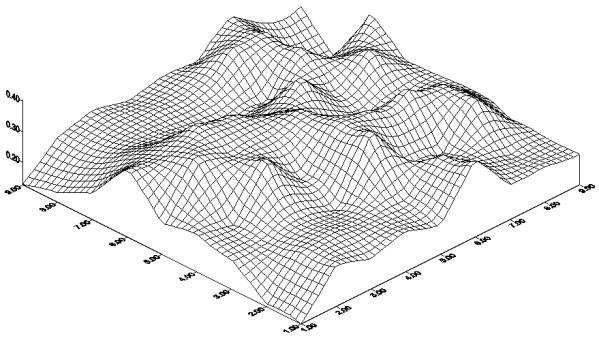

Figure 1 – The surface of DNA optical density in the scanogram of a patient suffering from nodular goiter

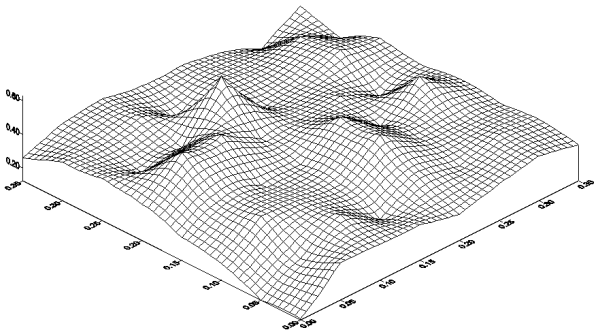

Figure 2 – The surface of DNA optical density in the scanogram of a patient suffering from infiltrative ductal cancer of the mammary gland (IDC)

To characterize the relief of the surface that corresponds to *kth* scanogram, we calculate the average slope of its slices with respect to the coordinates i and j:

$$R_k = \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n-1} \left| s_{i,j+1} - s_{i,j} \right| + \sum_{i=1}^{n-1}\sum_{j=1}^{n} \left| s_{i+1,j} - s_{i,j} \right| \right).$$

The relief index that characterizes a given patient is determined as the average of all scanograms

$$R = \frac{1}{N}\sum_{k=1}^{N} R_k ,$$

where N is the number of scanograms. By calculating the above indices for all scanograms from the training samples, we determine the corresponding confidence regions.

Thus, the process of making a diagnosis involves the evaluation of the above indices based on patients' scanograms, and checking whether the point (V, R) belongs to the corresponding confidence region (see Figures 3-5).

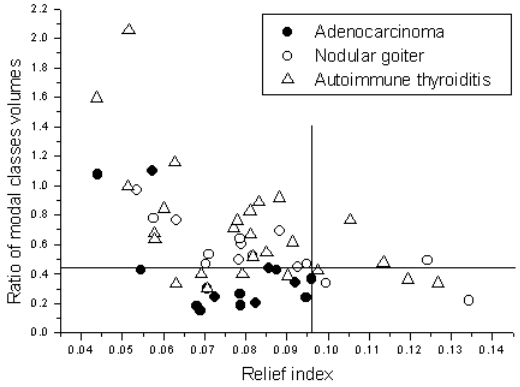

Figure 3 – Confidence regions for the relief indices and ratios of modal classes volumes for patients suffering from diseases of the thyroid gland

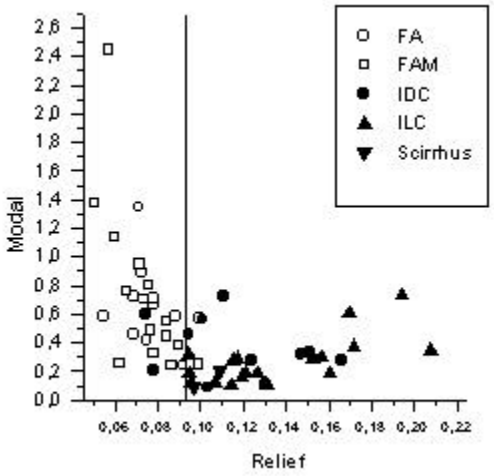

Figure 4 – Confidence regions for the relief indices and ratios of modal classes volumes for patients suffering from diseases of the mammary gland (excluding IDLC)

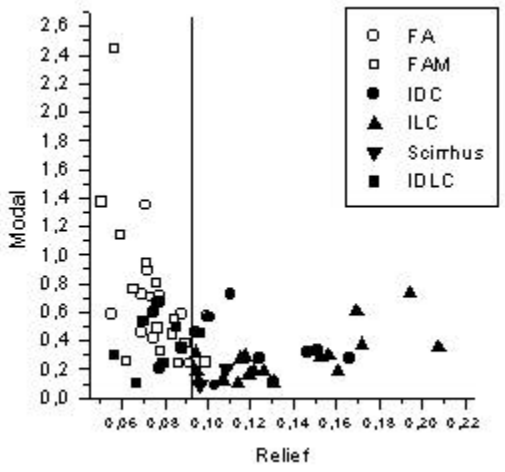

Figure 5 – Confidence regions for the relief indices and ratios of modal class volumes for patients suffering from diseases of the mammary gland (including IDLC)

2. Mathematical means of computer diagnosis: confidence intervals. Many problems in natural sciences and technology reduce to the following: stipulate an interval $I=(a,b)$ that contains the value x of general population G with a given probability β (for example, $\beta = 0.95$), $p(x\in(a,b))=\beta$. This interval $I=(a,b)$ is called a confidence interval for the population G (or simply a confidence interval), the numbers a and b are called the lower and upper confidence limits respectively, the number β is the confidence level, and the value $\alpha = 1-\beta$ is the significance level. The confidence interval (a,b) can be constructed with the help of the Chebyshev inequality[3]

$$p(|x-m(x)|\ge\lambda\sigma(x))\le 1/\lambda^2$$

$$(then\ \alpha\le 1/\lambda^2),$$

however, the confidence level is very crudely estimated. In this connection a large number of Chebyshev-type inequalities have been derived that refine the Chebyshev inequality under certain restriction on the distribution of x (see[3]). Nevertheless, these modifications of the Chebyshev inequality do not allow us to justify the empirical "3σ *rule*", which asserts that for distributions occurring in practice

$$p(|x-m(x)|\ge 3\sigma(x))\le 0.05,$$

where $m(x),\sigma^2(x)$ are the expectation and variance of x, respectively.

It seems plausible to think that for any random variables $x\in S_n$, such that

$$S_n=\left\{x: x=\sum_{i=1}^{n} y_i, y_i\ independent, F_{y_i}\equiv F(u)\right\}$$

the "3σ *rule*" is fulfilled since the limit theorems are valid, but this is not correct. Indeed, the following assertion holds true[4] : for all $\lambda \ge 1$ and every natural number n the following inequality holds

$$\sup_{x\in S_n} p(|x-m(x)|\ge\lambda\sigma(x))\ge\left(1-1/\lambda^2\right)/\lambda^2.$$

Thus, for every natural number n

$$\sup_{x\in S_n} p(|x-m(x)|\ge 3\sigma(x))\ge 8/81\approx 0.0987.$$

The problem of the justification of the "3σ *rule*" has been successfully solved for unimodal distribution in the way suggested by C.F.Gauss[5].

Recall that the random variable x is said to be unimodal if its probability density $f(u)$ has only one local maximum.

More precisely, the variable x is unimodal if there exists a point a such that the distribution function of x is convex in the domain $(-\infty,a)$ and concave in the domain (a,∞).

It turns out that the classical Gauss inequality

$$p(|x-m(x)|\ge\lambda\sigma(x))\le 4/9\lambda^2,$$

which is valid for all symmetric unimodal random variables (i.e. $f(a-u)=f(a+u)\quad\forall u\in R^1$) will be correct for all unimodal (not necessarily symmetric) random variables x, for all $\lambda>\sqrt{8/3}$ (see[4,6]). In particular for $\lambda = 3$

$$p(|x-m(x)|\ge 3\sigma(x))\le 4/81\approx 0.049<0.05.$$

This refinement of the Gauss inequality is called the Vysochanskij-Petunin inequality[7,8]; at present there exist many generalizations and extensions of this inequality[7-9].

Another idea for solving the problem of construction of the confidence limits containing the bulk of the general population is based on order statistics. Suppose G is some general

population with unknown distribution function $F(u), x_1, x_2, ..., x_n$ is a sample obtained from *G* as the result of simple random sampling. If we rearrange the sample values in increasing order $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$, then we obtain the order statistics $x_{(i)}$, $i = 1, 2, ..., n$. It is shown[6,10] that for the population with a continuous distribution

$$p\left(x_{n+1} \in \left(x_{(i)}, x_{(j)}\right)\right) = \frac{j-i}{n+1} \, (i < j), \quad (1)$$

where x_{n+1} is a sample value that does not depend on the sample $x_1, x_2, ..., x_n$, obtained by means of simple random sampling.

Consider two measurable (Borel) functions of *n* variables $f(u_1,...,u_n)$, $g(u_1,...,u_n)$, $g(u_1,...,u_n)$ satisfying the inequality

$$f(u_1,...,u_n) \le g(u_1,...,u_n) \quad \left(\forall (u_1,...,u_n) \in R^n\right);$$

With the help of the function $f(u_1,...,u_n)$, $g(u_1,...,u_n)$ and the sample $x_1, x_2, ..., x_n$ we can construct the random confidence interval

$$I = \left(f(x_1,...,x_n), g(x_1,...,x_n)\right)$$

for the bulk of the general population *G*; this interval *I* is said to be an invariant confidence interval if

$$p(x_{n+1} \in I) =$$

$$= p\left(x \in \left(f(x_1,...,x_n), g(x_1,...,x_n)\right)\right) = const$$

for any general population *G* with the continuous distribution function $F(u)$.

It is shown[11] that the following statement is true: let $f(u_1,...,u_n)$ and $g(u_1,...,u_n)$ be continuous symmetric functions satisfying the inequality

$$f(u_1,...,u_n) \le g(u_1,...,u_n) \quad \left(\forall (u_1,...,u_n) \in R^n\right)$$

which coincide on the set from R^n with zero Lebesque measure. In order that the confidence interval

$$\left(f(x_1,...,x_n), g(x_1,...,x_n)\right)$$

be invariant it is necessary and sufficient that $f(x_1,...,x_n) = x_{(i)}, g(x_1,...,x_n) = x_{(j)} (i < j)$ where $x_{(i)}, x_{(j)}$ are some order statistics constructed with the help of sample $x_1, x_2, ..., x_n$. Therefore, a set B_n of all confidence intervals consists only of rational numbers $0, \frac{1}{n+1}, \frac{2}{n+1}, ..., 1$, namely

$$B_n = \left\{0, \frac{1}{n+1}, \frac{2}{n+1}, ..., 1\right\}.$$

The results mentioned above allow us to construct the confidence intervals for the bulk of the distribution corresponding to the given significance level with the help of ordered statistics or the 3σ-rule. In this work we use the confidence intervals constructed with the help of ordered statistics. These intervals are indicated on Figures 3-5 at the axis of coordinates.

3. Statistical tests. The classical theory for the test of hypothesis using statistical criteria was created in the first half of XX century and was based on the Neumann-Pearson fundamental lemma[12]. This lemma allows one to obtain a powerful test in the case of two simple alternative hypotheses H and H'. We can construct this test and calculate its probability of errors of the first and second kind if we know the distribution functions $F_H(u_1,...,u_n)$, $F_{H'}(u_1,...,u_n)$ corresponding to these hypotheses exactly. Unfortunately, in practice we never know these distribution functions. That is why it is necessary to construct statistical criteria for the test of hypotheses that are based on training samples and not on the distribution functions $F_H(u_1,...,u_n)$, $F_{H'}(u_1,...,u_n)$.

Suppose G and G' are two general populations and let $x = (x_1,...,x_n)$ and $y = (y_1,...,y_m)$ be samples from G and G', respectively. We assume for simplicity that these samples are obtained by means of simple sampling. Let $z = (z_1,...,z_k)$ be a sample from the general populations G or G' and let's

assume that we do not know to which population the sample z belongs. Let $H=\{z\in G\}$ and $H'=\{z\in G'\}$, and suppose that (a,b) is a confidence interval for the bulk of *G*, and (a',b') is a confidence interval for G'. We can construct these intervals with the help of estimates of the mathematical expectation and variance in accordance with the "3σ *rule*" (or, more precisely, the "3s *rule*"):

$$(a,b)=(\overline{x}-3s,\overline{x}+3s),\ (a',b')=$$
$$=(\overline{x}'-3s',\overline{x}'+3s')\ ,$$

where

$$\overline{x}=\frac{1}{n}\sum_{k=1}^{n}x_k,\ s^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)^2,$$
$$\overline{x}'=\frac{1}{m}\sum_{k=1}^{m}x_k',\ s'^2=\frac{1}{m-1}\sum_{k=1}^{m}\left(x_k'-\overline{x}'\right)^2$$

or with the help of the order statistics

$$a=x_{(i)},\ b=x_{(j)}\ (i<j),\ a'=x'_{(r)},$$
$$b'=x'_{(t)}\ (r<t)\ .$$

Assume for simplicity $z=z_1$ (i.e. *k* = 1). Then the statistical criterion based on the training samples[13] has the form

1) if $z\in(a,a')$ then *H* is accepted;

2) if $z\in(b,b')$ then H' is accepted;

3) if $z\in(a',b)$ then decision is not accepted.

The probability of the error of the first kind is

$$\alpha^*=p\left(H'\middle|H\right)=p\left(z\in(b,b')\middle|H\right)\le$$
$$\le p\left(z\notin(a,b)\middle|H\right)=\alpha$$

(≈ 0.05 *for example*)

where α is a significance level of the confidence interval (a,b). Similarly the probability of the error of the second kind is

$$\beta^*=p\left(H\middle|H'\right)=p\left(z\in(a,a')\middle|H'\right)\le$$
$$\le p\left(z\notin(a',b')\middle|H'\right)=\alpha'$$

(≈ 0.05 *for example*).

Now, let us consider a general case where *k* is any natural number $z=(z_1,...,z_k)$. If the order statistics $x_{(1)},x_{(n)},y_{(1)},y_{(m)}$ satisfy the inequality

$$x_{(1)}\le y_{(1)}\le x_{(n)}\le y_{(m)}\ ,$$

then $a=x_{(1)},b=x_{(n)},a'=y_{(1)},b'=y_{(m)}$.

Let ϑ be any number from $[0,1)$, (i.e. $0\le\vartheta<1$), and let $k\ge2$. Now, let us define a number $r=r(\vartheta,k)$ by the formula $r=r(\vartheta,k)=\left[k/(1+\vartheta)\right]$, where [*a*] is an integer part of the number *a*. A statistical criterion T_ϑ for the test of hypothesis H and H' on the basis of the sample $z=(z_1,...,z_k)$ is defined in the following way[14]:

1) H is accepted if at least *r* sample values of the sample being investigated $z=(z_1,...,z_k)$ belong to the interval $(-\infty,y_{(1)})$;

2) if at least *r* value from the sample $z=(z_1,...,z_k)$ belong to the interval $(x_{(n)},\infty)$ then the hypothesis H' is accepted;

3) in any other cases the decision in not accepted.

The probability of the error of the first kind is

$$\alpha^*=p\left(H'\middle|H\right)=$$
$$=\frac{B(n+k-r+1,r)}{B(k-r+1,r)}=$$
$$=\frac{(k-r+1)(k-r+2)...k}{(n+k-r+1)(n+k-r+2)...(n+k)}=$$
$$=o\left(\frac{1}{n^{r-1}}\right)$$

and the probability of the error of the second kind is

$$\beta^*=p\left(H\middle|H'\right)=$$
$$=\frac{(k-r+1)(k-r+2)...k}{(m+k-r+1)(m+k-r+2)...(m+k)}$$
$$=o\left(\frac{1}{m^{r-1}}\right)$$

where *B*(*a*,*b*) is a beta-function.

Denote by CD a procedure of the non-acceptance of decision. Assume that the simultaneous distribution functions $F_G(u_1,...,u_n)$

and $F_{G'}(u_1,...,u_n)$ of the samples $z=(z_1,...,z_n)$, where $z\in G$ and $z'\in G'$, are known and construct the statistical criterion T for the test of hypothesis $H=\{z\in G\}$ and $H'=\{z\in G'\}$ satisfying the conditions

$$p(H'|H)\le\alpha^*, p(H|H')\le\beta^*$$

and

$$p(CD|H)+p(CD|H')\to\min,$$

where α^* and β^* are fixed number. This criterion is called an optimal condition.

Denote by $W,V,S\subset R^n : W\bigcup V\bigcup S=R^n$ the regions of decision acceptance of the criterion T: $z\in W\Rightarrow H$ is accepted, $z\in V\Rightarrow H'$ is accepted , $z\in S\Rightarrow$ decision is not accepted. It is shown[15] that these regions are defined by the likelihood ratio

$$h(u)=\frac{f_{H'}(u_1,...,u_n)}{f_H(u_1,...,u_n)},$$

where f_H ($f_{H'}$) are the probability densities.

4. Estimation of method sensitivity and specificity. We consider the probabilities of the errors of the first and second kinds for the proposed method: suppose the main hypothesis H is that the patient is suffering from adenocarcinoma of thyroid gland, and the alternative competing hypothesis H' is that the patients is suffering from nodular goiter or autoimmune thyroiditis. As is well-known[12], the probability of error of the first kind is defined as the probability of rejection of the hypothesis H when it holds true. In our case it is equal to $p(H'|H)$. To estimate the sensitivity $p(H|H)$, we use the formula shown below.

In estimating the sensitivity and specificity of the method, we assume that the indices R and V are independent random variables. This assumption is acceptable, since the correlation coefficient between R and V is -0.493, which indicates the absence of a strong dependence between these two indices. In this connection the probability of the errors defined above is only an approximation of the true probability. Therefore,

$$p(H|H)=p(R\le 0.0960, 0\le V\le 0.439)\approx$$

$$\approx p(R\le 0.0960)\,p(0\le V\le 0.439).$$

Since the number of ATG-patients $n=15$,

$$p(R\le 0.0960)=\frac{15}{16}=0.9375;$$

$$p(0\le V\le 0.439)=p(V_{(0)}\le V\le V_{(13)})=$$

$$=\frac{13}{16}=0.8125.$$

Hence, the sensitivity of the method is

$$p(H|H)\approx 0.7617.$$

Respectively, the probability of the error of the first kind is 0.2383.

The probabilty of the second kind is defined as the probability of acceptance of hypothesis H in the case when the alternative competing hypothesis H' holds true. This probability can be estimated approximately in terms of the frequency of the random events that the NG- and AT-patients are recognized as ATG-patients. In our investigation the total number of the NG- and AT-patients was 42, and 6 AT-patients were erroneously recognized as ATG-patients. Note that all NG-patients were recognized correctly. Therefore,

$$p(H|H')\approx\frac{6}{42}\approx 0.142.$$

Thus, the specificity of the proposed test is equal to 85.8 %.

According to accepted recommendations[16], the frequency of a random event may be considered to be practically coincident with its probability if the number of trials exceeds 30. Hence, in our case the probability of the error of the second kind may be considered to be computed exactly for all practical purposes.

Analogously, we can estimate the sensitivity and specificity of the method of diagnosis of breast cancer:

1) in the case when IDLC is ignored

$$p(H|H)=p(R\ge 0.0925)\approx\frac{33}{35}=0.9429,$$

$$p(H\,|\,H') \approx \frac{2}{22} \approx 0.0909\,.$$

2) in the case when IDLC is taken into account

$$p(H\,|\,H) = p(R \geq 0.0925\,) \approx \frac{25}{35} = 0.7142,$$

$$p(H\,|\,H') \approx \frac{2}{22} \approx 0.0909\,.$$

Thus, we see that the IDLC may be recognized among other hystological forms of breast cancer on the basis of the analysis of the relief index of the scanograms.

5. Results. It is shown that in malignant neoplasms of the thyroid and mammary glands the changes in the nuclei of buccal epithelium are characterized by an increase in the optical density of DNA in a range from 0.15 to 0.30 in conventional units of measure, as compared with its values in benign pathological processes. The sensitivity of the proposed criterion in the case of the diseases of the thyroid gland is equal to 76.2 % and the specificity is equal to 85.8 %. In the case of diseases of the mammary gland (excluding IDLC) we have discovered that the sensitivity of the method is equal to 94.29 % and its specificity equals to 90.91 %. In the case of diseases of the mammary gland (including IDLC) we have discovered that the sensitivity of the method is equal to 71.42 % and its specificity is equal to 90.91 %.

For comparison, in western Europe all patients suspected of breast cancer are recommended for clinical examinations, mammography and/or ultrasound investigation and fine-needle aspiration biopsy (FNAB). The complex of these diagnostic tests is called Triple Assessment. The accuracy of each individual test is as follows: clinical investigation, 84%; mammography, 78% (depending on the age of patient and the size of tumor this value may decrease); FNAB, 91%. The sensitivity of Triple Assessment, when at least one test is positive, varies from 95 to 100% , and the accuracy is 89%. If all three tests are positive, then the sensitivity increases up to 99%. In this connection it was noted[17-20] that some of the tests have significant disadvantages, or risks, associated with them. First, the tests depend strongly on the experience and skill of the diagnostician, and repetition of certain tests is not desirable, or is limited. Second, the accuracy of mammography depends significantly on the type and size of the tumor. Finally, FNAB is an invasive procedure, which can provoke insemination of the puncture channel by malignant cells, causing a recurrence of the tumor and the need for a more radical surgical procedure. Thus, repetition of FNAB is absolutely undesirable.

Concerning the diagnostics of cancer of the thyroid gland, it is well known[21] that "the fine-needle aspiration biopsy is the main method in the evaluation of solitary thyroid nodules and dominant nodules within multinodular goiters"[22]. As noted[21], "if the procedure is done properly, it should have a false-negative rate of less than 5% and a false-positive rate of approximately 1%"[23]. But this method has one drawback: "even in skilled hands, however, approximately 10% of biopsy findings are nondiagnostic"[21]. So, quite often diagnosticians have to repeat FNAB several times, risking the possibility of insemination of the puncture channel by malignant cells. In contrast, our method is non-invasive, has high accuracy and sensitivity, and does not have any contraindications to repeated tests[24].

REFERENCES

.

1. Petunin Yu.I, Klyushin D.A., Andrushkiw R.I., Ganina, K.P.& Boroday N.V. 2001 Computer-aided differential diagnosis of breast cancer and fibroadenomatosis based on malignancy associated changes in buccal epithelium. Automedica. **19**(3/4):135–164.
2. PetuninYu.I.,Klyushin,D.A.& Andrushkiw R.I. 1997. Nonlinear algorithm of pattern recognition for computer-aided diagnosis of breast cancer. Nonlinear Analysis. **30**(8): 5431–5436.
3. Kendall M.G.& A. Stuart. 1958. The advanced theory of statistics. v.1. Distribution theory. 2nd ed.: Hafner Publishing Company .
4. Vysochanskij D.F. & Petunin Yu.I. 1980. Justification of the 3σ *rule* for unimodal distribution. Theor. Probability and Mathem. Statistics. **21**: 25–36 (In Russian).

5. Gauss C.F. Theoria Combinationis Observationum Erroribus Minimis Obnoxiae. Pars Prior. Pars Posterior. Supplementum. Theory of the Combination of Observations Least Subject to Errors. Part One. Part Two. Supplement. 1995. Translated by G.W. Stewart. Classics in Applied Mathematics Series, Society for Industrial and Applied Mathematics, Philadelphia.
6. Petunin Yu.I. 1981. Use of the Theory of Random Processes in Biology and Medicine. Naukowa Dumka Publishing Company (in Russian).
7. Pukelsheim F. 1994. The three sigma rule. Amer. Statist. **48**: 88–91.
8. Sellke T. Generalized Gauss-Chebyshev inequalities for unimodal random variables. Technical Report #94/17: 1–11, Department of Statistics, Purdue University.
9. Dharmadhikari S. & Joag dev K. 1988. Unimodality, Convexity, and Applications. Academic Press, 1988.
10. Madreimov I. & Yu.I. Petunin. 1982. Characterization of the uniform distribution with the help of the order statistics. Theory of Probab. and Math. Statist. **27**: 96–102 (In Russian).
11. Bairamov I. & Yu.I. Petunin. 1990. Structure of the invariant confidence intervals containing the underlying main distributed mass. Theor. Probability and its Applications. **35**(1): 25–36 (In Russian).
12. Van der Waerden B.L. 1969. Mathematische Statistic (English transl. of 2nd (1965) ed.): Springer-Verlag.
13. Petunin Yu.I., Klyushin D.A., Savkina M.Yu & R.I.Andrushkiw. 1998. Mathematical background comp-aided cytogenetic method of diagnosis of breast cancer. In Proc. XIII Sum. School on Biometrics: 203–215.
14. Bairamov I. & Petunin Yu.I. 1991. Statistical tests based on training samples. Cybernetics. **3**: 74–77. (In Russian).
15. Petunin Yu.I. & S.A. Matveychuk. 1994. Test of hypothesis with the help of the statistical criteria using a procedure of the non-acceptance decision. Dokl. Acad. Sci. **336**(3): 301–303 (In Russian).
16. Cramer H. 1946. Mathematical methods of statistics: Princeton University Press.
17. Ciatto S., Cariaggi P., Bulgaresi P., Confortini M., Bonardi R. Fine needle aspiration cytology of the breast: review of 9533 consecutive cases // The Breast. - 1993. - 2. - P. 87-90.
18. Ciatto S., Catarzi S., Morrone D., Rosselli Del Turco M. Fine needle aspiration cytology of nonpalpable breast lesions: US versus stereotaxic guidance // Radiology. - 1993. - 188. - P.195-199.
19. Dixon J.M., Anderson T.J., Lamb J., Nixon S.J., Forrest A.P.M. Fine needle aspiration cytology in relationship to clinical examination and mammography in the diagnosis of solid breast mass // Br. J. Surg. - 1984. - 71. - P. 593-596.
20. Layfield M., Glasgow B.J., Cramer H. Fine-needle aspiration in the management of breast masses // Pathol-Annu. - 1989. - 24 - P.23-62.
21. Singer P.A., Cooper D.S., Daniels G.H., Ladenson P.W., Greenspan F.S., Levy E.G., Braverman L.E., Clark O.H.. McDougall I.R, Ain K.V., Dorfman S.G. 1996. Treatment Guidelines for Patients With Thyroid Nodules and Well-Differentiated Thyroid Cancer. Arch Intern Med. **156**:2165-2172
22. Mazzaferri EL. 1993. Management of solitary thyroid nodule. *N Engl J Med.* **328**:553-559.
23. Grant CS, Hay ID, Gough IR, McCarthy PM, Goellner JR. 1989. Long-term follow up of patients with benign thyroid fine-needle aspiration cytologic diagnoses. *Surgery*.**106**:980-986.
24. Klyushin D.A., Petunin Yu.I., Andrushkiw R.I., Boroday N.V. & Ganina K.P. 2002. Analysis of Malignancy-Associated DNA Changes in the Nuclei of Buccal Epithelium in the Pathology of the Thyroid and Mammary Glands. *Ann. N.Y.Acad. Sci.* **980**: 1-12 (2002) (to appear)