Diagnosis of Breast Cancer using Averaged Proximity Measure between Samples

R.I. Andrushkiw[1], E.N. Golubeva[2], D.A. Klyushin[2], Yu.I. Petunin[2], and N.V. Boroday[3]
[1]New Jersey Institute of Technology, Newark, NJ 07102, USA
[2]Kyiv National Taras Shevchenko University, Kyiv, Ukraine
[3]R.E.Kavetsky Institute of Experimental Pathology, Oncology and Radiobiology, Kyiv, Ukraine

Abstract - *In this paper we determine the risk degree of malignancy in tumors that have been diagnosed as benign. To do this we compute the proximity measure between corresponding morphological and densitometrical indexes of digital images of interphase nuclei of buccal epithelium in patients with benign tumors, malignant tumors and individuals that are practically healthy (without tumors).*

Keywords: Breast Cancer, Diagnosis, Proximity Measure

1 Introduction

The development of a neo-plastic process in an organism is usually accompanied by changes in the functional interrelations between its organs [1]. In a series of investigations [2-4], it was proved that changes in oral mucosa are an early indicator of some pathological processes in an organism. Hence, it is possible to use buccal epithelium for the investigation of changes which are going on in epitheliocytes of oral mucosa in patients with oncological pathology. Such changes are called MAC – malignancy associated changes. Since violation in the function of organs and systems in an organism are related to changes in the functional state of a cell genome, the morphometric and densitometric parameters of epitheliocytes in buccal epithelium may be used as a criterion of MAC [5]. The use of quantitative automatic image analysis opened the possibility of estimating the content of DNA in the nuclei and compactness of the chromatin, which characterizes the functional state of cells in various pathological processes, including tumors [6].

2 Materials

We consider three groups of patients: G_1 – patients suffering from breast cancer (38 cases), G_2 – patients suffering from fibroadenomatosis (44 cases) and G_3 – group of practically healthy women (33 cases). Smears from various depths 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 DNA content stained by Feulgen was estimated using the Olympus computer analyzer, consisting of the Olympus BX microscope, Camedia C-5050 digital zoom camera and a computer. We investigated from 40 to 60 nuclei in each preparation. The DNA-fuchsine content in the nuclei of the epitheliocytes was defined as a green component of a RGB-value.

3 Methods

3.1 Proximity measure

Let $x=(x_1,...,x_n)\in G$ and $x'=(x'_1,...,x'_m)\in G'$ be samples from general populations G and G', and $x_{(1)}\le ...\le x_{(n)}$ and $x'_{(1)}\le ...\le x'_{(m)}$ be their order statistics. We test the hypothesis on the identity of absolutely continuous distribution functions $F_G(u)$ and $F_{G'}(u)$ of the general populations G and G'. Suppose that $F_G(u)=F_{G'}(u)$. Denote by $A_{ij}^{(k)}$, $k=1,2,...,m$, a random event that x'_k lies in the interval $(x_{(i)},x_{(j)})$:

$$A_{ij}^{(k)}=\left\{x'_k\in(x_{(i)},x_{(j)})\right\},\ (i<j).$$

The probability of this event is determined by the formula [7, p. 126]:

$$P(A_{ij}^{(k)})=P\left(x'_k\in(x_{(i)},x_{(j)})\right)=p_{ij}^{(n)}=\frac{j-i}{n+1}.$$

Let

$$p_{ij}^{(1)}=\frac{h_{ij}^{(n)}m+g^2/2-g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m+g^2/4}}{m+g^2},$$

$$p_{ij}^{(2)}=\frac{h_{ij}^{(n)}m+g^2/2+g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m+g^2/4}}{m+g^2},$$

where $h_{ij}^{(n)}$ is the relative frequency of the event $A_{ij}^{(k)}$ in m trials and $g=3$.

Denote by N the number of all confidence intervals $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$, $\left(N = n(n-1)/2\right)$ and by L the number of intervals $I_{ij}^{(n,m)}$ containing probabilities $p_{ij}^{(n)}$. Then we get the p-statistics:

$$h^{(n,m)} = \rho\left(x, x'\right) = \frac{L}{N}.$$

Letting $h_{ij}^{(n)} = h^{(n,m)}, m = N$, $g = 3$, we get the confidence interval for the p-statistics $h^{(n,m)}$:

$$p^{(1)} = \frac{h^{(n,m)} N + g^2/2 - g\sqrt{h^{(n,m)}(1-h^{(n,m)})N + g^2/4}}{N + g^2},$$

$$p^{(2)} = \frac{h^{(n,m)} N + g^2/2 + g\sqrt{h^{(n,m)}(1-h^{(n,m)})N + g^2/4}}{N + g^2}.$$

3.2 Averaging of proximity measure

Let $x = (x_1, x_2, ..., x_n)$ be a sample from the general population G, which is obtained by simple random sampling. Let $y_1 = (y_1^{(1)}, ..., y_{n_1}^{(1)})$, …, $y_K = (y_1^{(K)}, ..., y_{n_K}^{(K)})$ be similar samples, which are obtained from general populations $G_1', ..., G_K'$ accordingly. Let G^* be a group which consists of the populations $G_1', ..., G_K'$:

$$G^* = \left\{G_1', ..., G_K'\right\}.$$

Based on the obtained statistics, let us calculate the p-statistics $\rho(x, y_i)$ [7] and define the quantity

$$\overline{\rho}(x, G^*) = \frac{1}{K}\sum_{i=1}^{K} \rho(x, y_i), \qquad (1)$$

The value $\overline{\rho}(x, G^*)$ is called the averaged p-statistics between the sample *x* and the group of general population $G^* = \left\{G_1', ..., G_K'\right\}$.

3.3 Method of diagnostics

Let *P* be a patient with uknown diagnosis: breast cancer or fibroadenomatosis. Let $x_1, ..., x_n$ be the sample which consists of the areas of interphase nucleus of buccal epithelium of some patient *P*. We'll denote $x_1, ..., x_n$ as a sample *x*. The group G^* is a teaching sample, which consists of similar indexes of patients $P_1^{(1)}, ..., P_K^{(1)}$ with breast cancer, or patients $P_1^{(2)}, ..., P_m^{(2)}$ with fibroadenomatosis.

Consequently, the teaching sample with indexes of patients with cancer is

$$G_1^* = \left\{G_1^{(1)}, ..., G_K^{(1)}\right\},$$

and the teaching sample with indexes of patients with fibroadenomatosis is

$$G_2^* = \left\{G_1^{(2)}, ..., G_m^{(2)}\right\}.$$

The criterion for the diagnostics of breast cancer consists of two parts. First, patients with cancer and their averaged p-statistics in group G_1^* and their averaged p-statistics in group G_2^* are considered.

Thus, the first patient $P_1^{(1)} \in G_1^*$ is considered. After excluding $P_1^{(1)}$ from group G_1^* we compute the averaged p-statistics between $P_1^{(1)}$ and the group

$$\tilde{G}_1^{(1)} = G_1^* \setminus P_1^{(1)} = \left\{P_2^{(1)}, ..., P_K^{(1)}\right\}: \ \overline{\rho}(P_1^{(1)}, \tilde{G}_1^{(1)}).$$

Then, patient $P_2^{(1)}$ is excluded from group G_1^* and in this way the group $\tilde{G}_2^{(1)}$ is obtained:

$$\tilde{G}_2^{(1)} = G_1^* \setminus P_2^{(1)}.$$

After that, the averaged p-statistics $\overline{\rho}(P_2^{(1)}, \tilde{G}_2^{(1)})$ is computed. Then the next patient is excluded and the averaged p-statistics is computed, and so on. This method is called «one-out». The results of the computations are in table 1.

In the same way, computations of the averaged p-statistics for patients $P_i^{(1)}$, $(i = 1, ..., K)$ and group G_2^* are done. The obtained values $\overline{\rho}(P_i^{(1)}, G_2^*)$ are given in table1.

Table 1. Averaged p-statistics between patients with breast cancer and patients with fibroadenomatosis

№	Averaged *p*-statistics between patients with breast cancer and	
	patients from group with breast cancer	patients from group with fibroadenomatosis
101	0,85	0,47
130	0,86	0,48
132	0,79	0,45
135	0,88	0,49
139	0,7	0,43
154	0,78	0,46

155	0,8	0,46
156	0,84	0,48
157	0,87	0,48
159	0,7	0,42
160	0,67	0,42
161	0,83	0,48
165	0,86	0,48
170	0,89	0,47
180	0,79	0,45
183	0,83	0,48
185	0,84	0,47
191	0,86	0,48
194	0,89	0,49
196	0,8	0,46
197	0,71	0,44
198	0,78	0,45
200	0,77	0,45
201	0,79	0,47
204	0,87	0,48
208	0,77	0,46
209	0,87	0,49
210	0,86	0,48
212	0,65	0,43
34	0,79	0,45
36	0,82	0,46
37	0,69	0,43
39	0,75	0,45
41	0,86	0,48
43	0,8	0,45
46	0,79	0,46
54	0,77	0,45
87	0,74	0,44

Analysis of table 1 shows that all values of the averaged p-statistics between patients with cancer and the group of patients with fibroadenomatosis are situated between $x_{(1)} = 0,649$ and $x_{(n)} = 0,887$, and all values of the averaged p-statistics between patients with cancer and group of patients with fibroadenomatosis are situated between $x_{(1)} = 0,415$ and $x_{(n)} = 0,493$.

From the above it follows that the diagnosis of patients with breast cancer was made without error. Let H denote the hypothesis that a patient has cancer and let $\bar{H}$ be the alternative hypothesis that a patient has fibroadenomatosis. Then the probability of type I error is equal to zero: $P(\bar{H}/H) = 0$. So for all patients with breast cancer the diagnosis is correct.

For the diagnostics of patients with fibroadenomatosis we used averaged p-statistics between patients with fibroadenomatosis and group G_1^*, as well as the averaged p-statistics between patients with fibroadenomatosis and group G_2^*. The results of the computation are given in table 2.

Table 2. Averaged p-statistics between patients with fibroadenomatosis and patients from the group with breast cancer and the group with fibroadenomatosis

	Averaged *p*-statistics between patients with fibroadenomatosis and	
	patients from group with breast cancer	patients from group with fibroadenomatosis
158	0,81	0,46
162	0,82	0,47
17	0,83	0,48
1	0,84	0,47
203	0,82	0,47
33	0,76	0,45
401	0,32	0,34
402	0,33	0,34
403	0,32	0,34
406	0,32	0,34
407	0,33	0,34
418	0,33	0,34
419	0,32	0,33
422	0,33	0,34
423	0,33	0,34
424	0,33	0,34
434	0,32	0,34
435	0,33	0,34
440	0,32	0,34
443	0,33	0,34
459	0,33	0,34
460	0,33	0,34
464	0,33	0,34
472	0,33	0,34
473	0,33	0,34
478	0,32	0,34
47	0,82	0,47
486	0,32	0,34
490	0,33	0,34
491	0,32	0,34
494	0,32	0,33
496	0,32	0,34

498	0,32	0,33
499	0,32	0,34
500	0,32	0,34
501	0,33	0,34
506	0,33	0,35
507	0,33	0,34
509	0,33	0,34
510	0,33	0,35
57	0,8	0,47
59	0,69	0,43
61	0,87	0,49
63	0,84	0,48

Analysis of the data in the first column of table 2, using confidence interval $I=(0,649;0,887)$ constructed by order statistics $x_{(1)}=0,649$ and $x_{(n)}=0,887$, shows that 11 patients with fibroadenomatosis were diagnosed as having cancer. Indexes of patients numbered 58, 162, 17, 1, 203, 33, 47, 57, 59, 61, 63 belong to the confidence interval $I=(0,649;0,887)$. The rest of the patients (33 persons) were diagnosed correctly.

Hence, the error is equal to 25%. The same results were obtained using the second cofidence interval $I=(0,415;0,493)$ and the second column from table 2 with averaged p-statistics $\overline{\rho}(P_i^{(2)},\tilde{G}_i^{(2)})$. In this case the same patients were diagnosed incorrectly.

In order to decrease this error we apply the second part of the diagnostics. Thus, to increase the accuracy of the criterion we consider the data of the group of practically healthy women.

We compute the averaged p-statistics α_2 between patients with fibroadenomatosis and group G_2^*, as well as averaged p-statistics β_2 between patients with fibroadenomatosis and group G_3^*. Then we calculate the ratio of the obtained averaged p-statistics α_2 and β_2. This ratio is denoted as γ_2:

$$\gamma_2=\frac{\alpha_2}{\beta_2}.$$

The results of the computations are given in table 3.

Table 3. Ratio of averaged p-statistics $\gamma_2=\alpha_2/\beta_2$ for patients with fibroadenomatosis

	Ratio of averaged *p*-statistics γ_2
158	0,755
162	1,18
17	1,074
1	1,015
203	1,263
33	0,706
401	1,17
402	1,173
403	1,171
406	1,203
407	1,173
418	1,171
419	1,169
422	1,179
423	1,175
424	1,16
434	1,16
435	1,176
440	1,178
443	1,203
459	1,149
460	1,179
464	1,158
472	1,175
473	1,171
478	1,17
47	0,882
486	1,206
490	1,169
491	1,181
494	1,167
496	1,209
498	1,158
499	1,181
500	1,172
501	1,204
506	1,175
507	1,162
509	1,167
510	1,166
57	1,366
59	0,567
61	1,066
63	1,2

Similarly, we obtain the ratio between α_1 and β_1, where α_1 is the averaged p-statistics between all breast cancer patients and the group of women patients with

breast cancer, and β_1^* is averaged p-statistics between all breast cancer patients and the group of practically healthy women:

$$\gamma_1 = \frac{\alpha_1}{\beta_1}.$$

The results of the computations is given in table 4.

Table 4. Ratio of averaged p-statistics $\gamma_1 = \alpha_1 / \beta_1$ for the patients with breast cancer

	Ratio of averaged *p*-statistics γ_1
101	1,719
130	1,469
132	1,473
135	1,833
139	0,945
154	1,644
155	1,473
156	1,628
157	1,651
159	0,96
160	0,794
161	1,721
165	2,182
170	1,686
180	1,2
183	2,26
185	1,565
191	1,677
194	2,002
196	2,291
197	2,167
198	1,115
200	2,251
201	1,906
204	1,97
208	2,104
209	1,927
210	2,09
212	2,204
34	1,273
36	2,089
37	2,098
39	1,99
41	1,097
43	2,025
46	1,1
54	1,1
87	0,963

Analysis of table 4 shows that the ratio γ_1 is situated between minimal $x_{(1)} = 0,794$ and maximal 2,291 order statistics. So, the confidence interval $I = (0,794; 2,291)$ covers the main distributed mass of the general population for γ_1.

On the other hand, the data from table 3 shows that the ratio of the averaged p-statistics γ_2, of patients with indexes 158, 33 and 59, does not belong to the confidence interval I. So, these patients are diagnosed as patients with fibroadenomatosis. Hence, only 8 patient are diagnosed incorrectly. After applying the second part of the criterion, the type II error is equal to 18,18%.

Let $H_0 = H$ denote the hypothesis that a patient has breast cancer and let $H_1 = \bar{H}$ be the hypothesis that a patient has fibroadenomatosis. Then $P(\bar{H}/H) = 0$, $P(H/\bar{H}) = 0,1818$.

4 Conclusions

Let us formulate the main conclusions based on the obtained results:

1) If after applying the criterion we diagnose fibroadenomatosis, then the probability of such event is close to one. The probability of the event that these patients have breast cancer is practically zero. The results are unexpected, since according to medical statistics the error in the diagnosis of fibroadenomatosis is approximately 20%.

2) If after applying the criterion we diagnose breast cancer, then the probability of such event is equal to 81,8%. The probability of not detecting a patient suffering from breast cancer is equal to zero.

In order to the increase the accuracy of detecting cancer, one can use another method [8, 9] in conjunction with the method discussed above. In that case all patients with fibroadenomatosis are diagnosed correctly, and patients with breast cancer are diagnosed with an error of 7,9%.

The application of the two methods together gives an accuracy of 92% and sensitivity of 100% .

5 References

[1] Shabalkin I.P. et al. (2000) Violation of systemic relations in an organism under cancerogenesis. Proc. Russian Academy of Science. 375 (3):404-409 (in Russian).

[2] Nieburgs H.F. et al. (1962) Buccal all changes in patients with malignant tumors. Lab. Invest. 11(1):80-88.

[3] Ogden G.R. at al. (1990) The effect of distant malignancy upon quantitative cytologic assessment of normal oral mucosa. Cancer. 65(3):477-480.

[4] Mayansky A.N. et al. (2004) Reactivity of buccal epithelicytes: indication of local and general homeostasis (survey). Clinical laboratory diagnostics, 8:31-34 (in Russian).

[5] Avtandilov T.T. et al. (2004). Ploidometrical diagnosis of precancerous processes and cancer of cervix on cytological preparations. Clinical laboratory diagnostics . 11:45-47 (in Russian).

[6] Linder J. (1994) Imaging cytometry: applications in diagnostic pathology. / Anal. And Quant. Cytology and Histology. 16(1):53-57.

[7] Klyushin D.A., Petunin Yu.I. (2008) Evidence medicine. Application of statistical methods – Williams , Moscow, (in Russian).

[8] R. Andrushkiw, D. Klyushin, M. Boroday, Yu. Petunin, K. Golubeva (2009) Determination of risk degree and making diagnosis of cancer by averaging of p-statistics – BIOCOMP'09, Las Vegas, Nevada, USA, 2009, pp 892-895.

[9] Klyushin D.A., Petunin Yu.I., Golubeva K.M. (2010) Computer citogenetic diagnostics of breast cancer and fibroadenomatosis based on areas of nucleus of buccal epithelium – Bulletin of University of Kiev, Series: Physics and Mathematics, №1, Kiev, 2010, P. 138-143 (in Russian).