UDC 519.23:578.087.1

STATISTIC DATA SIMULATION AT ESTIMATION OF BIOLOGICAL SYSTEM STATE

V.A. Fokin

HPE SEI Siberian state medical university, Tomsk
E-mail: fokin@ssmu.tomsk.ru

It has been shown by the method of statistic data simulation that using samples of small volume characterizing reference state of biological systems results in overstatement of difficulty degree of state integral estimation. It was stated that using asymptotic integral estimations obtained at unlimited increase of modeling reference sample volume is effective as integral estimation of biological system state.

Introduction

Analysis of biomedical data, development of methods of data extraction form them, formation of integral estimations of biosystem state represent actively developing directions of modern information technologies in medical science and practice of public health [1–4]. In mathematical formulation the problem comes to construction of algorithms and functional reflection of attribute space characterizing biosystem in one-dimensional space of estimations of this system states determined by a value of specified integral test.

Construction of integral tests of state estimation may result in obtaining efficient estimations however their use supposes accumulation of rather high volumes of reference data that is not may be always implemented in state of single experimental investigations. Therefore, statistic data simulation may be considered as efficient method of studying biosystem properties the results of which allow, on the one hand, estimating numerically statistic properties of the test itself and, on the other hand, allow determining conditions put on sampling volume required for obtaining robust generalized state estimations.

Type of test of state estimation

System S state is estimated with respect to a certain reference state S_0 of this system. The state of the system corresponding, in particular, to a healthy organism may be selected, for example, as such state. Let S_0 and S be specified reference and estimated states characterized by sets of objects $\{\vec{b}_i | i \in N_{S_0}\}$ and $\{\vec{b}_j | j \in N_S\}$ respectively. Here N_{S_0} and N_S are the sampling volumes. The value of quantitative estimation of certain object state $\vec{b}_i \in S$ may be characterized by its extent of proximity to reference state S_0 selecting which one should take into account the configuration of area in attribute space occupied by reference state, positioning of object $\vec{b}_j$ relative to this area as well as positional relationship of objects representing system reference state. Taking into account these conditions the test of integral estimation of object $\vec{b}_i$ proximity to the state S_0 may be specified in the following way [5]:

$$I_{S_0}(\vec{b}_i) = \frac{d(\vec{b}_i, S_0)}{D_{S_0}}, \qquad (*)$$

where $d(\vec{b}_i, S_0)$ is the certain extent of object $\vec{b}_i$ proximity to the set S_0; D_{S_0} is the extent of compactness of the area occupied in attribute space by objects referring to the state S_0.

Normalization to value D_{S_0} in expression (*) allows taking into account contribution of both configuration of area S_0 and positional relationship of objects in it to the obtained estimation. Let us specify the extent of compactness D_{S_0} of reference state S_0 in the following form:

$$D_{S_0} = \frac{1}{N_{S_0}} \sum_{k=1}^{N_{S_0}} \frac{1}{N_{S_0} - 1} \sum_{j=1}^{N_{S_0}-1} d(\vec{b}_k, \vec{b}_j),$$

that is as an averaged value of average distances from each object referring to the state S_0 to the rest of them. Value D_{S_0} determined in this way represents intramultiple distance [6] concrete kind of which is determined by a way of specifying distances in attribute space. The generalized distance is efficiently to be used as an extent of objects proximity in attribute space in biomedical tasks [7] as in this case interconnection of features characterizing studied bioobjects is naturally taken into account. The generalized distance between k and i objects is determined in the following way:

$$d_M(\vec{b}_k, \vec{b}_i) = (\vec{b}_k - \vec{b}_i)^T \mathbf{C}_0^{-1} (\vec{b}_k - \vec{b}_i).$$

Here $\mathbf{C}_0$ is the matrix of covariation of features characterizing state S_0. Quantitative estimation of area compactness extent characterizing state S_0 in generalized metric equals to double dimension of attribute space [5]:

$$D^*_{S_0} = D_{S_0} = 2m,$$

and expression for integral estimation of object $\vec{b}_i$ proximity to state S_0 takes a form:

$$I_{S_0}(\vec{b}_i) = \frac{1}{2m} d(\vec{b}_i, S_0),$$

where $d(\vec{b}_i, S_0)$ is calculated as an averaged generalized distance from object $\vec{b}_i$ to S_0,

$$d(\vec{b}_i, S_0) = \frac{1}{N_{S_0}} \sum_{j=1}^{N_{S_0}} d_M(\vec{b}_i, \vec{b}_j),$$

calculated using covariation matrix corresponding to reference state S_0.

Statistic simulation

The main problem when using tests based on multidimensional methods of data analysis stipulated by low volumes of samplings characterizing reference state that results in significant variability of estimations obtained at their use. Specific character of biomedical data is reflected in it and first of all their wide intra- and interin-

dividual variability the consequence of which is the fact that repeat measurement at one and the same sampling may result in different quantitative values of estimated characteristics. Estimation of statistic features of integral test (*) suggested before represents nontrivial problem the solution of which using only analytical approaches is stipulated by significant difficulties and in some practical cases is impossible.

In this case investigation of statistic features of integral test may be efficiently implemented by the methods of statistic simulation the results of which allow, on the one hand, estimating numerically statistic features of the test and, on the other hand, allow determining conditions put on conditions of forming reference samplings required for obtaining robust estimators.

Statistic features of integral test were estimated in two stages. At the first stage M of sample set X_k ($k=\overline{1,M}$) of specified volume corresponding to statistic features of reference state S_0, represented by a certain sample set of objects $X:\{\vec{b}_i | i \in \overline{1,N_{S_0}}\}$ was simulated. The obtained sequences of values imitate taking samplings from one and the same set and, therefore, are free of errors stipulated by the influence of intra- and interindividual variabilities of biological data. Then for each set X_k values of estimations $I_{S_0,k}(\vec{b})$ the distribution of which was further used for studying statistic features of integral test were computed. Here vector $\vec{b}$ characterizes object for which the estimation is carried out. In particular, vector corresponding to sample representative of state S, for example, vector corresponding to the centre of class may be examined as it.

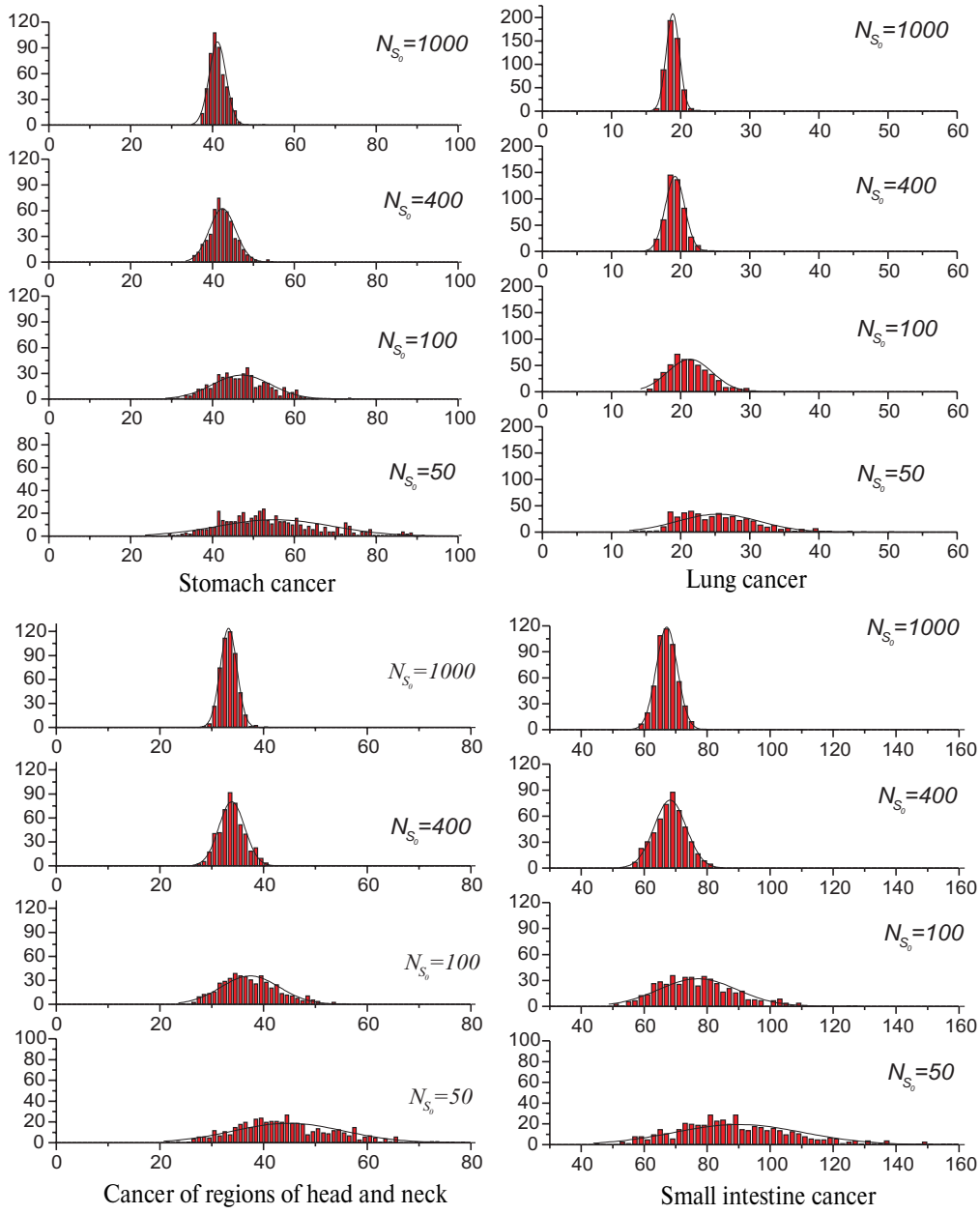

Fig. 1. *Frequency histograms I_{S_0} for different volumes N_{S0} of reference state. On abscissa – the value I_{S_0} on ordinate axis – the frequency value. The curve – approximation by normal distribution*

At the second stage the influence of such factors as sampling volume, ratio between sampling volume and quantity of jointly analyzed indices on the value of estimations was studied. Depending on the fact whether the law of multidimensional data distribution for statistic simulation of sets of their values is known or not different techniques may be applied [8–10]. Statistic features of integral test were estimated by calculating of average value of integral index

$$\hat{I}_{S_0}(\vec{b}) = \frac{1}{M}\sum_{k=1}^{M} I_{S_0,k}(\vec{b}),$$

of average square deviation

$$\hat{\sigma}_I(\vec{b}) = \frac{1}{M}\sum_{k=1}^{M} (I_{S_0,k}(\vec{b}) - \hat{I}_{S_0}(\vec{b}))^2.$$

To estimate integral test variability the variation coefficient was calculated

$$V = \frac{\hat{\sigma}_I(\vec{b})}{\hat{I}_{S_0}(\vec{b})} \cdot 100\,\%$$

and $(1-p)\cdot 100$ % confidence interval as the interval containing values I_{S_0} being between $p/2\cdot 100$ % and $(1-p/2)\cdot 100$ % number of all values of integral test in a ranged estimation row. Here p is the proper level of statistical significance. Such nonparametric way of estimation of confidence interval allows estimating it without any suppositions relative to the form of distribution law and statistic features of integral test.

Results of simulation

Data of scanning electron microscopy (SEM) of surface architectonics of red blood cells obtained by composite authors [11–13] by the results of examination of patients at some cancer localizations of II–III stages as well as healthy people served as the initial data for forming model samplings and statistic estimations of suggested integral test. As a form of red blood cells and their capacity to deformation is a sequence of a complex of violations of features, organization and metabolism of single components of red blood cells stipulated by presence of proper pathologic process then SEM data may be used for integral estimation of degree of changes occurring in red blood system by selected complex of features.

Statistic estimation of I_{S_0} was carried out using the developed computer program [14] simulating samplings corresponding to volumes N_{S_0} equal 50, 100, 200, 400, 600, 800 and 1000 of observations with the assumption that reference sampling data satisfy multidimensional normal distribution law. Each sampling was simulated from 100 to 1000 times with a pitch 100 according to which statistic estimations of test value variability were calculated. The results of simulation of statistic characteristics of integral test I_{S_0} and its variability for estimating the state of red blood system by SEM data at different cancer localizations for certain values of N_{S_0} and M are given in Fig. 1 and 2.

It follows form the analysis of the obtained results that the value of reference sampling volume N_{S_0} influences significantly the obtaining of robust estimators. In particular, at low volumes of samplings the wide variability of value I_{S_0} is observed. The variation coefficient at N_{S_0}=50 amounts at the average to 20...25 % for all examined states decreasing to 4...8 % at sampling volumes N_{S_0}=1000. The dependence of computed average magnitudes of integral test values I_{S_0} on model sampling volume for oncological disease of different localization corresponding to a number of model samplings M=500 is given in Fig. 2.

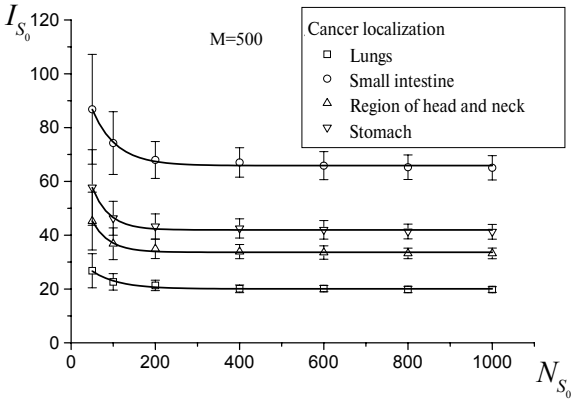

Fig. 2. *Dependence of integral test value I_{S_0} on model sampling volume N_{S_0} of reference state*

For other volumes of model samplings dependences have the similar form. Confidence intervals corresponding to 95 % are indicated by sections. Interesting result of statistic simulation is the fact that the value of estimation depends on the volume of reference sampling and small volumes of samplings result in high estimations of integral test average value. Therefore, examination of asymptotic estimations obtained at unlimited increase of the volume of simulated reference sampling may be of interest.

Conclusion

1. Statistic simulation of the data of biological system reference state is the efficient method of forming samplings of multidimensional data allowing decreasing the variability of integral estimation of the examined system state in conditions of low volumes of initial experimental data.
2. Volume of reference sampling is a significant parameter for quantitative estimation of biological system state and low volumes of reference samplings along with great variability of the obtained integral estimation also result in overrating of average value of the obtained integral estimation.
3. It is efficient to use the asymptotic integral estimation obtained at unlimited increase of the volume of model reference sampling as integral estimation of biological system state.

REFERENCES

1. Bogomolov A.V., Gridin L.A., Kukushkin Yu.A., Ushakov I.B. Diagnostics of man state: mathematical approaches. – Moscow: Meditsina, 2003. – 464 p.
2. Genkin A.A. New information technology of medical data analysis (ОМИС software). – St.-Petersburg: Politekhnika, 1999. – 191 p.
3. Dyuk V., Emannuel V. Information technologies in biomedical investigations. – St.-Petersburg: Piter, 2003. – 528 p.
4. Armitage P., Berry G. Statistical Methods in Medical Research. – 3rd ed. – Oxford: Blackwell Scientific Publication, 1994. – 620 p.
5. Fokin V.A. Estimation criterion of complex biosystem state // Bulletin of the Tomsk Polytechnic University. – 2004. – V. 307. – № 5. – P. 136–138.
6. Tu J., Gonsales R. Principles of pattern recognition. – Moscow: Mir, 1978. – 416 p.
7. Konradov A.A. Statistic approaches to the analysis of multidimensional heterogeneous biological systems // Radiatsionnaya Biologiya, Radioecologiya. – 1994. – V. 34. –Issue 6. – P. 877–886.
8. Ermakov S.M., Mikhailov G.A. Statistic simulation. 2d issue. – Moscow: Nauka, 1982. – 296 p.
9. Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. // CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph. 38. – Philadelphia: SIAM, 1982. – 92 p.
10. Manly B.F.J. Randomization, Bootstrap and Monte Carlo Methods in Biology. – London: Chapman and Hall/CRC, 1997. – 424 p.
11. Novitskiy V.V., Ryazantseva N.V., Stepovaya E.A., Bystritskiy L.D., Tkachenko S.B. Atlas. Clinical pathomorphism of red blood cells. – Moscow: GEOTAR-MED, 2003. – 208 p.
12. Novitskiy V.V., Stepovaya E.A., Goldberg V.E., Kolosova M.V., Koreshkova K.G., Sokolova I.B., Bulavina Ya.V. Reversible aggregation and surface architectonics of red blood cells of peripheral blood of patients with lung cancer before and during antineoplastic polychemotherapy // Experimental and clinical pharmacology. – 1999. – V. 62. – № 5. – P. 28–30.
13. Novitskiy V.V., Stepovaya E.A., Goldberg V.E., Kolosova M.V., Ryazantseva N.V., Korchin V.I. Red blood cells and malignant growths. – Tomsk: STT, 2000. – 288 p.
14. Certificate № 2006614010 the RF. Computer program «StatSys» / V.A. Fokin, I.S. Khakimov, O.Yu. Nikiforova; Application № 2006613281; Stated 29.09.2006; Published 22.11.2006.

Received on 04.10.2007

UDC 681.3.06:518.5.001.57:612.821.7

TECHNIQUE OF AUTOMATED HYPNOGRAM CONSTRUCTION

E.S. Zakharov, P.P. Kravchenko*, A.A. Skomorokhov

«Medikom MTD», Taganrog
E-mail: office@medicom-mtd.com
*Technological institute of Southern federal university, Taganrog
E-mail: kravch@tsure.ru

The technique of automated sleep stage recognition and hypnogram construction has been considered. For partition of initial polysomnogram by segments obtained as a result of patient sleep monitoring the signal energy is analyzed using nonlinear energy controller. Frequency weighted energy is calculated for all registered signals then averaging and segmentation occur according to monitored signals behavior. Secondary index vector which is used at transition from segments to fixed duration periods is formed for segments. One or another sleep stage is finally assigned to the period by correlation analysis. Accuracy of the developed algorithm is connected with quantity of considered secondary indices, maximally detailed description of sleep stage characteristics and realization of training by manually prepared examples.

Rapid development of computer science and microtechnology during last decades made it possible to apply diagnostic equipment allowing recording continuously and processing wide range of physiological indices. It gave, in its turn, a possibility to specialists to investigate in detail life activity of human body in different states including sleep [1–3]. At the present the majority of scientists and doctors pay much attention to sleep quality and man state in this period of his life. It is known that in sleep many chronic and pathologic diseases which can be poorly diagnosed or often can not be diagnosed at all at wakefulness appear. Man health and efficiency depends on sleep quality. The importance of sleep for healthy living is a conventional fact. The first attempts of recording physiological indices by primitive equipment were made more than 100 years ago. Since that time a number of standards and agreements was developed in this field. In particular, in 1968 a manual Rechtschaffen & Kales (*R&K*) was published and was accepted as a standard [4]. Sleep stages and typical features of each of them were described in it in detail. It was suggested to divide sleep into five stages: stages with slow eye movement *I, II, III, IV* and stage of rapid eye movement (REM). When studying sleep a wide set of physiological indices is recorded. They are different in their nature:

- electroencephalogram (EEG) – brain transistance;
- electrocoulogram (ECG) – eye-bulb movement in two derivations relative to counterlateral referents;
- electromyogram (EMG) – muscle tonus;
- electrocardiogram (ECG);