SIMPLE METHODS OF CYBERNETIC PROCESSING
OF MEDICAL DIAGNOSIS DATA

V. S. Genes

The subject of study of medical cybernetics is the methodology and automation of medical diagnosis. A system of medical diagnosis should provide means for distinguishing between groups of diseases requiring differential medical treatment under conditions of fluctuating boundaries between various groups. The diagnostic process must not only recognize known symptoms but also provide new knowledge. In the course of diagnosis each step should be determined by the results of the preceding one (in accordance with the feedback principle), i.e., be a result of comparison of test data with the information stored in the physician's memory. It is natural that in each specific case the physician should strive for a definite diagnosis. Unfortunately, this is impossible at present with respect to a large number of diseases. For this reason, one must satisfy himself in the majority of cases by the most probable diagnosis; and, the cybernetic diagnosis system should at every stage of the examination—given an evaluation of the probability of possible diseases in the given patient—point out the next kind of tests necessary for better discrimination between the possible diseases and indicate the most simple and easy way for selecting the necessary and sufficient set of criteria for the diagnosis of each disease. It is also desirable that the system be simple and within the reach of a physician at any medical institution.

The above requirements indicate that a cybernetic system of processing medical data should consist of the following elements: 1) a subsystem for the evaluation of the diagnostic value of individual indicators of the organism's state and for selecting among them the symptoms of diseases; 2) a subsystem for evaluating indicator complexes and for selecting from them disease symptoms; 3) a subsystem for storing information about all diseases.

Subsystem for Evaluating the Diagnostic Value of Indicators. Indicators of the organism's condition which can be expressed quantitatively are analyzed by the method of characteristic intervals. For this purpose one plots on a single graph the distribution of the number of observed cases vs. the indicator value for all diseases to be differentiated (see Fig. 1). For simplicity, the figure shows the curves of only two diseases and the indicator value scale is relative (two adjacent scale divisions differ by a magnitude greater than the measuring method error). In accordance with the classical method of characteristic-interval analysis one tries to isolate on the graph zones corresponding to the observations made on patients suffering from the given disease (zone 1 for disease A and zone 5 for disease B). These are deterministic zones corresponding to absolute disease symptoms within the limits of the given group of differentiated diseases. If indicators having a value lying within one of these zones are detected in new patients, the corresponding disease may be diagnosed. The zones 2, 3, and 4 between the deterministic zones 1 and 2 are zones of overlap. These zones must be divided into subzones in accordance with the frequency of occurrence of different diseases typical of them. In this connection the following versions are possible: a) zones in which one disease occurs more frequently than other diseases, and b) zones in which the frequency of occurrence of both diseases is essentially the same. If the curves are plotted from data obtained during one and the same period of time and at the same medical institutions, the frequency of occurrence of each disease in the given zone corresponds to its probability as given by the Bayes formula. Instead of calculation one can use tables listing the significance levels of the differences between the frequencies of occurrence of different diseases in the given zone (V. S. Genes, 1967). The tables were compiled on the basis of binomial distribution for three significance levels (75-90, 95, and 99%), and for using them one has only to find the number of observations pointing to the given disease from all the observations in the given zone. The meaning of the significance levels is that in a large number of observations of similar groups of diseases the frequency of occurrence of the given disease will exceed the frequency of other diseases within the percentage of cases corresponding to the significance level. In dividing the zone of overlap into subzones one must

Academy of Medical Sciences of the USSR; Ukrainian Institute for the Advancement of Physicians.
take into account that the lower the given significance level the larger the percentage of cases falling into the zones a and the greater the probability of an erroneous diagnosis. On the other hand, if the significance level is high the probability of false diagnosis is reduced but, at the same time, the zone of indicator values in which no diagnosis at all can be stated increases. The strategy of selection of significance levels should be governed by the possibility of additional informative testing. If no such tests can be made the significance level should be lowered and vice versa. In Fig. 1 the overlap zone has been divided into three sub-zones: subzone 2 in which the significance level of disease A occurring more frequently than disease b is 99%, subzone 4 in which disease B occurs more frequently than disease A with the same significance level, and subzone 3 in which the frequency of occurrence of both diseases is the same. Practically, the zone boundaries were selected as follows. First, the number of observations of each disease separately and of both diseases together (total number of observations) was summed for scale divisions to the right of line I that separates the deterministic zone 1 from the probabilistic subzone 2 and compared with the table of significance levels: for scale divisions corresponding to indicators 17 and 18 the sum of observation of disease A is 14 of a total of 17 cases (A+B). This corresponds to the 99% significance level. The observations corresponding to the 19th scale division were then added to the respective groups of observations. Of a total number of 28 observations, 21 belonged to the disease A. This is also within the 99% significance level. The number of observations corresponding to the 20th scale division was next added to the preceding number. For disease A we got 27 observations out of a total number of 38. The significance level for such a proportion is no longer 99 but 95%. Since the diagnosis reliability changed when passing from the 19th to the 20th scale division, the subzone boundary was drawn at this point (double broken line II). A similar procedure was employed to draw the boundary between the overlap and deterministic zones for disease B. The boundary of the probabilistic subzone with a 99% significance level for disease B (double broken line III) was established in the same way. The boundaries II and III are separated by a zone in which the diseases A and B have practically the same frequency of occurrence.

After the deterministic and probabilistic zones of each disease are established, one calculates the number of cases distinguishable within these zones and uses tabulated data to determine to what percentage they amount of a total number of "specific" diseases and what is the mean error and the confidence limit of the percentage. Qualitative indicators are analyzed in the same way (the characteristic nature—whether deterministic or probabilistic—of each quality gradation for this or that disease being first analyzed with the aid of significance level tables). The quality gradations of similar informative value can then be combined (generalized). Different indicators are compared according to the rules of mathematical statistics. The most promising indicators are then selected for further processing. This information is then used for compiling cards that list the indicator denomination, the method used for its measurement, the diseases which can be distinguished to a certain measure with its aid, zone boundaries of different informative significance (symptoms in the clinical sense), the frequency of these zones in different diseases, and the probability of certain specific diseases occurring within these zones.

For each probabilistic zone it is desirable to find an auxiliary indicator which makes it possible to differentiate between diseases within the given zone. These indicators are subjected to a similar treatment but the analysis is made with only those cases which fell within the given zone. If different indicators are used for improving the diagnosis within the probabilistic zones, a notice is made of each indicator in the corresponding zone. If however the diagnosis in all or in the majority of probabilistic zones can be improved by one indicator, one plots a complex of two indicators.

Subsystem for Evaluating Indicator Complexes. The first stage in the analysis of indicator complexes involves coding of the data about patients. The code consists of the number of zones of individual indicators which have different informative values. For example, five zones have been shown in Fig. 1. Consequently, this indicator will have five code units. Each patient is assigned a code unit corresponding to the indicator zone to which his value belongs. If the value is 25 the patient is assigned the code 5, if 20 the code is 3, etc. A coding table is then compiled for all indicators taken as a single complex. The arrangement of indicators in the table determines the order of digits in the patient's code number. For example, the number of leukocytes occupies the third place in the table. Consequently, the third digit from the left in the patient's code number will reflect the state of this indicator. If the gradation of a certain indicator is more than ten, the gradation number (two digits) is separated from adjacent indicators by a semicolon. After coding the indicators of all patients whose diseases are to be differentiated, one compiles a multicoordinate table (for 2-5 indicators) or a multicoordinate card (for a practically unlimited number of indicators). Along the horizontal and vertical axes of the multicoordinate table (in two-dimensional representation) one lists in