Applications of Computer Produced Frequency Distribution Curves

II. Evaluation of the Diagnostic Significance of Test Results by Multidimensional Analysis

N. Ressler and L. S. Whitlock

A method is described for determining the diagnostic significance of test results. It is based upon the relation of a test value to a frequency distribution curve of patients with a particular disease, and to a curve of others without that disease. When the two curves overlap at the test value found, the results are evaluated in terms of the probability of the patient corresponding to either of the two curves. The evaluation of results in terms of probabilities is more descriptive than the classification of all answers as either "normal" or "abnormal" and permits the combination of probabilities obtained from a number of tests into a single, resultant probability. The most probable diagnosis of a given patient, based upon all the tests done, or the most discriminating combination of tests for the diagnosis of any disease can be determined. The method permits compensation for such factors as a patient's age and sex, or the stage of a disease. It is nonparametric, and can be performed automatically with a data-processing system.

When frequency distribution curves of normal and of abnormal individuals are compiled for various clinical tests, there is usually an overlap between the normal and the abnormal curves (1–3). If a test value of an undiagnosed patient is found to fall within the overlap region, it cannot be classified as normal or abnormal with certainty. When all test results are labeled normal or abnormal, a certain percentage of these labels must be incorrect, depending in each case upon the proportion of answers in the overlap region. The probability estimate of the patient belonging to a normal or abnormal population, how-

From the Departments of Pathology and of Biological Chemistry, The University of Michigan Medical School, and from the Department of Biostatistics, The University of Michigan School of Public Health, Ann Arbor, Mich.

Supported in part by Grants HE-10361 and GM892 from the National Institutes of Health, U. S. Public Health Service.


Received for publication Dec. 6, 1966; accepted for publication May 10, 1967.
ever, can be determined from the relative heights of the two curves at the test value found (1). A consideration of test values in terms of probability is more descriptive than labeling all answers as normal or abnormal. It also permits the individual probabilities, obtained from different tests done on an individual patient, to be combined into a single, resultant probability. The most likely diagnosis of an individual patient can thus be determined from this combined probability value. The most discriminating combination of tests among those available for the diagnosis of any particular disease can also be determined. The performance of these determinations by multidimensional analysis of the results of tests done in combination is discussed in this paper.

The relationship between the overlap of individual frequency distribution curves and the probability of a patient belonging to one of any two diagnostic classifications has been discussed in a previous communication (4). When frequency distribution curves are obtained by plotting the number of individuals against a combined, rather than an individual test result, the relationship between overlaps and probability of classification remains unchanged (i.e., the ability of a test to discriminate or to determine the diagnostic classification to which a patient belongs increases in the same manner when the overlap decreases).

The diagnostic classification of patients by test results alone is limited, since individual factors such as age, sex, or the stage of a disease are often important in the interpretation of the results. These factors can be taken into consideration and compensated for in the evaluation of results of tests done in combinations. These compensations increase the discrimination which can be achieved and are described below.

The method of analysis which is described is feasible for routine use since it can be performed automatically with a data-processing system. The determination of the probability of classification directly from the relative heights of overlapping frequency distribution curves is nonparametric (5) and independent of the shape of individual frequency distribution curves. It is free of assumptions other than that previous experience can be used as a basis for evaluation. Abraham and Caceres have previously discussed applications of nonparametric "preception" analysis to the classification of electrocardiograms (6).

**Advantages of Multidimensional Analysis**

**Compensation for Factors Pertaining to an Individual Patient**

The significance of a test result is generally affected by individual considerations such as age, sex, and race. Normal ranges for serum
cholesterol, for example, are generally considered to be about 110–250 mg./100 ml. for a 20-year-old male, but 140–320 mg./100 ml. for a 60-year-old male. In a 60-year-old female, the normal range is believed to be about 156–356 mg./100 ml. (7). Normal concentrations of serum gamma globulins in 2-month-old individuals are less than half those of a normal adult (8). Some test values depend upon factors characteristic of certain geographic locations, while others, such as the incidence of hemoglobin S, are related to race.

The effect of individual factors upon the significance of test results can be compensated by selecting only those individuals with comparable factors in the compilation of the frequency distributions. The significance of a serum cholesterol determination done on an undiagnosed patient could be evaluated, for example, by utilizing frequency distributions composed only of those individuals of the same sex, approximate age, etc. The determination of significant factors, selection of the appropriate populations for comparison, and compilation of the frequency distribution curves for these populations can be performed automatically with the use of computers. The selections can be done from all of the chart information which has been incorporated into the data-processing system.

**Compensation for the Stage of a Disease**

The particular stage of a disease at which a test is conducted, or the variation in the values of a given test at different stages, is often critical in evaluating the significance of test results. The elevation of serum enzymes after a myocardial infarction is an example. Elevations of serum lactic dehydrogenase have been found to begin about 1 day after a myocardial infarction and to last for a week or more (9). Serum concentrations of creatine phosphokinase, however, begin to rise a few hours after a myocardial infarction, and descend to normal levels in about 3 days (9). Consequently, creatine phosphokinase determinations done 4 or 5 days after a myocardial infarction, or lactic dehydrogenase determinations done within a few hours afterward, may not provide any indication of this disease.

In order to distinguish patients with a given pathologic process most effectively, the data-processing system can indicate if a test should be done once or in a series, and the optional times for the tests relative to this process.

This can also be accomplished by a process of selection of the populations used for compilation of the frequency distributions. Figure 1, for example, shows frequency distribution curves for serum creatine phos-
phokinase levels. The laboratory data were obtained with normal individuals and with patients with acute myocardial infarctions at our Medical Center (10). In Fig. 1A, patients with acute myocardial infarctions, which had occurred at various times previously, are represented

![Fig. 1. Frequency distribution curves for serum creatine phosphokinase concentrations. Abnormal population consists of all patients with acute myocardial infarctions in A and of patients in whom onset of myocardial infarctions occurred 18–36 hr. previously in B. Curves have been normalized to provide value of 100 on ordinate for peak of normal distribution.](image)

in the abnormal curve. The latter is seen to overlap with the normal curve. In Fig. 1B, however, the abnormal population was restricted to patients who had a myocardial infarction between 18 and 26 hr. previously. The overlap of the curves is eliminated by this restriction, and the discrimination has been increased.

The selection process can also be carried a step further by plotting the creatine phosphokinase concentrations only of those patients with elevated levels 18–36 hr. after a possible myocardial infarction, but with normal levels 5–6 days afterward. The exclusion from the compilations of patients with elevated levels which do not return to the normal distribution range after this period of time can increase the discrimination between patients with acute myocardial infarctions and others in whom serum creatine phosphokinase levels are also elevated, but with a different time relationship.

Similar considerations apply to other diseases in which a test result is related to a changing pathologic process. In this manner, indications of the most effective times to conduct tests, or an allowance for the effect of the timing whenever a test has been done, can be provided automatically by a data-processing system.
Discrimination of Test Combinations

A combination of tests is frequently of more diagnostic specificity than any one of them alone (11, 12). Modern data-processing systems can compile frequency distribution curves based upon any combination of tests among those available which can identify a patient, most specifically, with a given disease classification (i.e., with the least overlap).

How this can be done is illustrated by the following, simplified example. Frequency distribution curves, representing data obtained by means of a hippuric acid liver function test, are plotted along the abscissa of Fig. 2 (13). Curves showing the data obtained by means of a Bromsulphalein (BSP) liver function test are plotted along the ordinate (13). Both tests had been done on the same individuals who were either normal or had liver disease. A diagonal line was then constructed, and composite frequency distribution curves were compiled along this diagonal by utilizing the same points on the graph. The point shown on the graph, for example, represents an individual with a hippuric acid value of 1.0 gm. and a BSP value of 8%. The height of each curve rep-
resents the number of points within a narrow vertical segment of values underneath it for the hippuric acid test, the number within a horizontal segment for the BSP test, and the number within a diagonal segment for the composite curve.

The per cent of the total areas of the normal and abnormal curves which overlap (area of uncertainty) was 91.8 for the BSP test, and 29.5 for the hippuric acid test. The overlap area of the diagonal frequency distribution curves, which combined the results of both tests, was reduced to 18.0%.

This procedure can be continued by representing the latter composite curves on the ordinate and curves representing values obtained from a third test such as a galactose tolerance determination, for example, on the abscissa. The resultant diagonal frequency distribution would then combine the values represented by the above, composite curve with those obtained by means of the galactose tolerance test. The same process can then be continued with a fourth test, etc. In this manner, the effect upon the discrimination of patients with liver disease, obtained by any combination of all available tests, can be determined. All tests which increase discrimination can thus be combined into a single composite frequency distribution for individuals with liver disease, and another distribution for individuals without liver disease. In practice the effect of various tests upon discrimination would probably not be determined with a data-processing system in the simplified manner utilized for this illustration (see below), but the principles of the correlations would remain the same. Combinations of tests which provide discrimination for patients with other diseases can similarly be determined by such correlations. The most discriminating combination of tests for any diagnostic classification can thus be determined.

The diagnoses of a given patient which are most probable on the basis of tests which have been done on him can also be determined by a similar process. The combined results of tests done on a patient can be related to composite frequency distribution curves based upon the tests done, for each diagnostic classification. The frequency distributions of any diagnostic classifications to which the combined test results of the patient correspond, or the relative probabilities in the case of overlaps, can then be determined. These diagnostic classifications, and the probabilities of correspondence, can be listed by the data-processing system. Any tests which would increase the discrimination for the most probable disease, which had not yet been done, could then be performed.

In order to illustrate this procedure, the composite frequency distribution curves in Fig. 2, relating BSP and hippuric acid excretion test
values to normals and to individuals with liver disease, are reproduced in Fig. 3A. If Patient a has the combined test value shown in this figure, the curves would indicate a probability of about 66% that the patient has liver disease. Any additional tests not yet done on this patient that could increase the discrimination for patients with liver disease could thus be called for by the data-processing system. The frequency distributions indicate that the probability of Patient b having liver disease is negligible. An evaluation of the most probable diagnosis for any given patient can be determined by his position on composite frequency distribution curves compiled for other additional disease classifications in the same manner.

In order to obtain the maximum discrimination with any given patient, it is necessary that the frequency distribution curves be compiled individually for that patient. This is necessary not only to base the composite frequency distribution curves on the particular tests done on the patient but also to allow for a compensation of other individual factors, as described above. If Patients a and b in Fig. 3, for example,

![Graph](image)

**Fig. 3.** Diagnostic application of combined test frequency distributions. A. Frequency distributions are same as composite curves along diagonal in Fig. 2. B. Curves resulted after adjusting curves in A for effect of age on BSP values, in case of 7-year-old Patients a and b. Adjustment changes indicated probabilities of patients having liver disease (see text).

were 7-year-old children instead of adults, the composite frequency distribution curves would be altered, due to the effect of age upon the significance of the BSP values (14). Figure 3B shows the same composite frequency distributions as Fig. 3A, except that the BSP values were corrected for the effect of age in the case of a 7-year-old, in the manner described by Vink (14), before being combined with the hippuric acid values in the compilation of the composite curves. The probability of Patient a having liver disease is now close to 100% in Fig. 3B, while
the probability of Patient b having liver disease has increased from a negligible value to about 25%. (With a data-processing system set up, the frequency distributions would be compensated by a selection of the age ranges of the cases used for compilation, rather than by a mathematical correction). Other individual factors which affect the test results pertaining to a given patient could be similarly taken into consideration at the same time.

Considerations Involved in Performance of the Automated Analyses

Since it is necessary to establish composite frequency distribution curves separately for each patient, in order to determine the most probable diagnoses with the greatest accuracy, a large magnitude of compilations and analyses are required. The procedure would be practical for routine use, however, since methods have been developed which would permit these compilations and analyses to be done automatically with a data-processing system.

Methods for the direct incorporation of patient case history, chart information (15, 16), and test results (17) into a data-processing system have been described and would provide the data which would be necessary for the compilations. Data processing systems can be programmed to produce frequency distributions in a graphic form, in order to permit visual comparisons (18).

Cluster analysis and order statistics on a similarity matrix as described, for example, by Rogers and Tanimoto (19) can be adapted for determining which test combination, or which factors used for population selection in compilations of the curves, results in the greatest discrimination. Cluster analysis includes the selection, from a number of factors, of those that are most closely associated or disassociated in the distinction of different classifications. In the present instance, the factors involved are test results and all other information in the patient’s chart which have been incorporated into the data-processing system. The classifications are those which correspond to various diagnoses. Different methods of cluster analysis, performed with data-processing systems, have been discussed by Sokal and Sneath (20). The relationship between different parameters and diseases can also be investigated in this manner. A complete description of the use of cluster analysis for such purposes is beyond the scope of the present paper, but will be presented in a future communication.

The data available to the computer do not need to be restricted to the values obtained at a given hospital. Data from the literature can also be incorporated, if the methodology and results are comparable. If
there is a difference which can be corrected for mathematically, the data can still be utilized. After frequency distribution curves have been compiled for a period of time, the available data will become adequate for more demanding requirements. It would also be possible for a number of participating hospitals to share the facilities of a central data-processing system for this purpose, and to exchange data when the test methods produce comparable results.

It is not necessary to assume that changes do not occur in the population. The data-processing system can indicate a rate of change in ranges, or their mean values, beyond specified limits, as described in a previous paper (4). If it is verified that the change is not due to laboratory error, the previous data in the computer can be eliminated, or adjusted so as to compensate for the change. Variations that are related to a specific factor—such as the season of the year, pollen count, or the temperature—can be compensated for automatically.

Results indicated by any single test or group of clinical tests are not a substitute for a complete diagnosis. The present discussion has been concerned only with chemical or other quantitative clinical determinations, and has not included such factors as biopsy examinations, personality evaluations, etc., which may be important to the clinician in establishing the diagnosis.

The probability resulting from a given combination of tests is based upon previous experience. If a diagnosis which test results indicate has little probability of being correct is nevertheless confirmed, the validity of the test results can be reexamined. If no mistakes are indicated, the subsequent distribution curves and probabilities will be altered accordingly. The probability of similar, additional cases will then no longer be as low.

The discrimination achieved depends upon the validity of the diagnoses and the test values used for compilation of the frequency distributions. The validity of these values, which are used as a basis for comparison, limits the evaluation of test results, whatever method is used, whether done with a computer or mentally. Despite precautions which may be taken, a certain percentage of the diagnoses and test answers utilized for compilations may be erroneous. This limitation is taken into consideration in the method described, however, by evaluation of the overlap of the different frequency distribution curves.

Besides quantitative chemical tests, the results of other diagnostic tests which can be expressed in terms of probability can also be included in this system. The results of electronically analyzed electrocardiograms (6), for example, can be combined with the results of
serum creatine phosphokinase, lactic dehydrogenase, and other tests, for the diagnosis of myocardial infarctions. The resultant probabilities obtained from the composite frequency distribution curves will then be based upon the total information. As additional or more specific methods are developed and included in the system, the probabilities obtained will presumably become of greater value to the clinician in establishing a diagnosis.

References