Thorax (1959), **14**, 300.

ON THE PROBABILITY OF CORRECT DIAGNOSIS BY PULMONARY FUNCTION TESTS

BY

R. J. SHEPHARD* AND M. E. TURNER

From the Department of Preventive Medicine, University of Cincinnati, Ohio, and the Department of Biophysics and Biometry, Medical College of Virginia, Richmond, Va., U.S.A.

(RECEIVED FOR PUBLICATION MAY 14, 1959)

For many years, chest physicians have cherished the hope that respiratory physiologists will devise an effective and simple test to distinguish normal from abnormal subjects. The limitations of well-established tests have gained increasing recognition, and this has led to a rapid multiplication in the number of test procedures. Patients are commonly subjected to a battery of 10 or more tests, but all too commonly a clear separation of normal from abnormal is still not obtained.

One difficulty is the tenuous relationship between the test procedures and the fundamental characteristics of the respiratory system. Multivariate statistical techniques such as factor analysis (Gilson and Hugh-Jones, 1955; Shephard, Turner, and Williams, unpublished) and discriminant function analysis are now beginning to explore this problem, and as it becomes better understood it may be possible to devise tests more specific to individual characteristics of the respiratory system, describing disease states in terms of alterations in these characteristics. Factor analysis is a formidable undertaking, and it has been suggested that in the immediate future more efficient use of existing data could be obtained by applying discriminant analysis. Instead of relying upon intuitive inspection of the laboratory data sheet, results of individual tests are weighted and combined in a manner that maximizes the difference between normal subjects and those suffering from the disease under study. Factor analysis also produces a weighted combination of tests; however, the criterion of weighting involves maximizing the variation of the combined value rather than maximizing discrimination. A general discussion of the objectives, similarities, and differences of factor analysis, discriminant function analysis, and other multivariate techniques is given by Kendall (1957).

Discriminant analysis has been used by Gilson and Hugh-Jones (1955) in assessing the degree of emphysema in coalworkers, and by Becklake, du Preez, and Lutz (1958) in a study of Witwatersrand gold-miners. The purpose of the present paper is to examine critically the value and limitations of the technique with reference to these examples and to a similar study carried out by us.

METHODS

TEST SITUATION.—The specific situation to which we applied this form of analysis arose in a small and rather closely knit community in an isolated part of the United States. A proportion of the employees in an industrial plant began to develop a characteristic pattern of respiratory symptoms, worsened by entering the plant. The causative agent is still under investigation, and it is sufficient to note here that in a number of cases submitted for lung biopsy the predominant histological picture was an interstitial fibrosis. The number affected was small (14 typical cases, four doubtful) and it was therefore necessary to consider all 14 cases as the "abnormal" population. Eighteen normal subjects were drawn from the same community, and a full range of physiological tests was carried out on the two groups. The effects of a small age difference (mean ages 40.4 and 43.8 years respectively) and a small sex difference (10 male, four female, and 17 male, one female) were minimized by expressing results as a percentage of expected normal values. It is still possible that other biological and sociological differences existed between the two groups, but these are thought to be minimal.

PHYSIOLOGICAL TECHNIQUES.—Portable apparatus was used throughout, enabling all tests to be carried out at the plant. Vital capacity and one second forced expiratory volume were determined by bellows spirometer (Shephard, Thomson, Carey, and Phair, 1958). Maximum inspiratory and expiratory pressures were measured by aneroid gauge. Carbon monoxide uptake and functional residual volume were estimated with a portable box bag (Carey, Phair, Shephard, and Thomson, 1957; Shephard, Carey, and Phair, 1958). Oxygen saturations were indicated by a Wood-Geraci type oximeter. Finally diaphragmatic movement was

*Present address: War Office, Chemical Defence Experimental Establishment, Porton Down, Salisbury, Wilts.

TABLE I
DISCRIMINATORY CAPACITY OF RESPIRATORY TESTS

Test	Normal Subjects	Abnormal Subjects	Δ	S_Δ	Student's t Test	Distance D_1^2	Probability Weighting	Weighting by Discriminatory Analysis
Vital capacity (height standard) (% normal)	133·4	107·7	−25·7	±5·9	4·35	2·48	2,000	2·71
Vital capacity (surface area standard) (% normal)	105·8	90·2	−15·6	±5·8	2·72	—	—	—
One second forced expiratory volume (% normal)	101·3	83·9	−17·4	±6·4	2·74	—	—	—
F.E.V.$_1$/V.C. (% normal)	95·9	92·1	−3·8	±3·6	1·04	0·16	2·9	1·39
Inspiratory reserve (% V.C.)	42·8	43·1	+0·3	±3·5	0·09	—	—	—
,, ,, (corrected age and sex)	−5·66	−6·35	−0·69	±3·38	0·20	—	—	
Tidal volume (% V.C.)	20·3	24·0	+3·7	±2·7	1·38	—	—	—
,, ,, (corrected age and sex)	18·2	23·0	+4·8	±2·7	1·78	0·40	−11·2	−0·40
Expiratory reserve (% V.C.)	36·9	32·9	−4·0	±2·5	1·62	—	—	—
Expiratory reserve (corrected age and sex)	37·4	32·8	−4·6	±2·6	1·77	0·46	11·1	+0·62
Maximum inspiratory pressure (mm. Hg)	25·6	19·0	−6·6	±3·8	1·72	0·46	10·0	+0·31
Maximum expiratory pressure (mm. Hg)	35·7	24·1	−11·6	±6·0	1·95	0·46	15·1	−0·05
Carbon monoxide uptake (corrected tidal vol. and time)	29·7	22·5	−7·2	±2·8	2·84	0·83	100	+0·33
Residual volume (% expected, corrected age and sex)	100·0	105·6	+5·6	±2·8	1·97	0·46	−15·9	+0·71
Resting oxygen saturation (%)	94·4	95·2	+0·8	±1·5	—	—	—	—
Breathing oxygen 2 min.	99·1	99·9	+0·8	±0·6	—	—	—	—
Exercise 2 min.	95·0	95·9	+0·9	±2·0	—	—	—	—
Radiographic index	0·48	0·34	−0·14	±0·045	3·03	1·12	161	+0·86

estimated from full inspiratory and full expiratory films and expressed as a ratio of the radiological chest length.

MEASUREMENT OF DISCRIMINATORY CAPACITY

If the population tested were large, the degree of overlap between normal and abnormal could be determined by a simple frequency histogram, but with the present data, as with the figures of Gilson and Hugh-Jones (1955) and Becklake and others (1958), it is necessary to predict overlap on a statistical basis. For simplicity a normal distribution of the data is assumed. If the mean and standard deviation of any value for the two groups is known, then the overlap can be determined either graphically by plotting the two normal curves, or more simply from tables showing the area of different segments of the normal curve (Yule and Kendall, 1940).

The relative value of individual tests as a means of discriminating normal from abnormal is indicated in Table I. The significance of discrimination is measured by use of the standard t test for differences between two means. This test is equivalent to the test of discrimination given by Rao (1952) in the case of a single physiological test. In the present problem, vital capacity standardized on a height and age basis separates the two groups more clearly than any other test. It is adequate to establish the existence of a pathological group (chance probability of difference between normal and abnormal groups 0.05%), but fails rather badly when applied to an individual case, the probable frequency of misdiagnosis being almost 22%.

How far can the accuracy of diagnosis be improved by adding other data? Certain of the test measurements, particularly the inspiratory reserve and the oximeter readings, are so similar in the normal and abnormal groups that they may immediately be discarded. The methods of expressing vital capacity, forced expiratory volume, and tidal volume that achieve the poorer separation are also discarded. This leaves nine test measurements. To simplify subsequent analysis, values are so adjusted that the mean value for each test is unity. The "distance" between the two groups for each of the nine measurements in terms of Mahalanobis' D^2 statistic is shown in column 6 of Table I. Combination of the data may be achieved in several ways, in an attempt to increase the "distance" between the two groups of subjects. Simple addition of the adjusted scores fails to achieve any advantage over using the single vital capacity test, the frequency of misdiagnosis still being 22%. Another simple device, which is reasonable when the correlation between tests is small, is to

weight scores for each test according to the probability that the test in question will distinguish normal from abnormal. This does give a slight apparent improvement, but even using nine test measurements the probable frequency of misdiagnosis is still 19%. Other methods such as those employed in factor analysis or Pearson's "coefficient of racial likeness" could be used, but formal discriminant analysis (Rao, 1952) gives optimal weighting to the test scores. Using this technique, the probable frequency of misdiagnosis is further reduced to 15.2%. However, it must be emphasized that a formal discriminant analysis is not a light statistical undertaking when many variables are involved, and as is shown in the appendix even the improvement produced in this way is not statistically significant in the present case.

In view of these discouraging findings, the data of other authors were examined critically. Becklake and others (1958) studied more than 20 tests for the diagnosis of silicosis, and the best (maximal mid-expiratory flow) showed an overlap of 30.5%. By formal discriminant analysis the diagnostic error could be reduced, but still amounted to 15%. Gilson and Hugh-Jones (1955) developed a discriminant function for the diagnosis of emphysema that achieved much clearer separation (only 0.6% overlap), but it must be remembered that the cases where this degree of separation could be demonstrated had quite advanced clinical disease. Further, the advantage gained by formal discriminant analysis is not as great as their paper might suggest, since the best single test in their series (helium mixing index I_r) in itself showed an overlap of only 1.1%

Comment

The above figures suggest that pulmonary function tests will often lead to an erroneous diagnosis even with such procedures as discriminant analysis, unless the disease process is already clinically obvious. It remains to consider whether this is inevitable, or merely a consequence of our present method of approach.

Let us assume that we have succeeded in the search for a test that will measure a fundamental characteristic of the respiratory system, and the measurement is statistically perfect (that is, the correlation between test value and fundamental characteristic is unity). Let us further assume that the method of measurement has a negligible error. We are left with the physiological variation, which distributes normal individuals about the mean with a well-defined standard deviation. Suppose this to be 10%. If a part of the normal population is now affected by a disease process that produces a uniform 20% loss of pulmonary function, even assuming that our test gives a perfect measure of this loss, the probable overlap cannot be less than 15.9%. On the basis of these same assumptions, curves can be drawn relating the probable error of diagnosis to functional loss at different levels of physiological variation (Fig. 1). In practice, the minimum error of diagnosis is rather greater than these curves

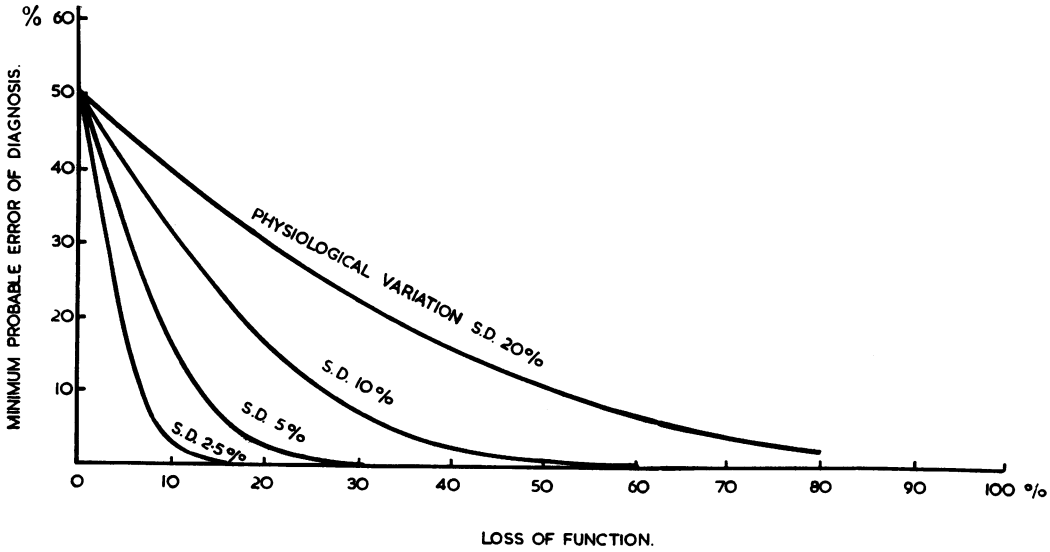

FIG. 1.—Probable error of diagnosis in relation to functional loss and physiological variation of test measurement.

would suggest, since the physiological test is rarely perfectly correlated with the fundamental characteristic, and it is unlikely that the error of measurement will be negligible. Further, the disease process cannot be expected to produce a uniform functional loss in every subject, and there is no guarantee that the test will reflect perfectly the loss of function produced by disease. However, physiological variation is more important than these ancillary factors, and it imposes an important limit on the accuracy of diagnosis by any system of function tests. This limit cannot be modified by any technique of combining tests if they do not characterize different unique aspects of physiological variation or of the disease state differential.

So far as simple measurements like height and weight are concerned, the range of physiological variation is readily assessed since measurement errors are small. In one large survey (Morant and Gilson, 1945), the height of fit young men showed a standard deviation of about 3.6%, while the standard deviation for weight was 11.5%. Respiratory measurements do not carry the same confidence, but if attention is paid to technique the physiological variation probably still exceeds other sources of error. Assuming this view to be correct, then from published normal figures (Table II) it would seem that, even after standardizing for age, sex, and size, the physiological variation is never less than 5%, and usually

TABLE II

ESTIMATED STANDARD DEVIATION OF SOME COMMON RESPIRATORY TESTS STANDARDIZED FOR AGE AND SEX AS PERCENTAGE OF THE MEAN (COEFFICIENTS OF VARIATION)

	Gilson and Hugh-Jones (1955)	Becklake and others (1958)	Present Series
Vital capacity (%) ..	10·5	15·6	—
,, ,, standardized for size (%) ..	—	14·1	11·7
M.V.V. (%)	14·3	22·7	—
F.E.V./V.C. (%) ..	—	—	8·7
R.V./T.L.C. (%) ..	18·7	64·4	14·6
CO uptake (%) ..	7·7	—	16·9
Helium mixing index I_0 (%)	30·0	16·7	—

exceeds 10%. Thus if a 5% error in diagnosis is considered acceptable, this cannot be attained until there is a functional loss of at least 30%. Considered in anatomical terms, this is quite extensive disease, and it is not surprising that at this stage it is often very obvious to the clinician.

In the present group of abnormal subjects, vital capacity, forced expiratory volume, and carbon monoxide uptake agreed quite closely, each indicating a mean functional loss of some 20%, and the discrimination achieved is approaching closely to the limit imposed by physiological variation. The emphysema subjects selected by Gilson and Hugh-Jones (1955) had more severe disease, some aspects of lung function being reduced by as much as 65%. Reference to Fig. 1 shows that 0.6% overlap is again close to the limit of discrimination possible under these conditions. The figures of Becklake and others (1958) occupy an intermediate position, with some 30% loss of function. The observed overlap of 15% is perhaps a little disappointing with this degree of functional loss, and, since the standard deviation of some of their test measurements is rather large, it is probable that discrimination has been limited as much by varying co-operation by patients as by physiological variation.

In order to make pulmonary function tests more useful, it will be necessary to study further the causes and control of physiological variation. One obvious factor affecting many tests is body size, and by standardizing for this variable the variation of some test measurements can be decreased by 1–2%. Diet and time of testing also produce some differences in respiratory test values (Dodds, 1921 ; Shephard, 1955). However, a large part of the physiological scatter remains uncontrolled for the foreseeable future.

Over a short period, the problem can be overcome by using the patient as his own control. In the present experiments five patients showing symptoms were exposed to the supposed toxic environment for a further four days, and all showed a decrease in test score over the period of exposure. It is possible that tests such as the vital capacity would have an increased usefulness if repeated at regular intervals by the patient's physician ; a sudden departure from the previously established rate of ageing should give valuable indication of the onset of disease. An even stronger case could be made for the pre-employment testing of respiratory function in those who are to be exposed to industrial hazards.

Very good results can sometimes be obtained by careful assessment of the patient's symptomatic assessment of his condition, the reason being once again that the reference line is the previous state of the patient and not some arbitrary " normal " value. In the present experiments data obtained by a modified Cornell questionnaire showed a large change of score in the abnormal group (Table III). However, in this case the standard deviation of the scores was also large, and the discrimination did not equal that provided by the respiratory tests. Despite careful framing of the questions, the score seems to have reflected the

TABLE III
POSITIVE RESPONSES TO MODIFIED CORNELL QUESTIONNAIRE

	Normal Group Mean ± S.D.	Abnormal Group Mean ± S.D.	Abnormal / Normal (%)
"Respiratory" questions	$\frac{2\cdot44\pm2\cdot50}{17}$	$\frac{6\cdot00\pm3\cdot33}{17}$	246
"Alimentary" questions	$\frac{2\cdot06\pm2\cdot09}{18}$	$\frac{5\cdot43\pm2\cdot73}{18}$	264
General questions ..	$\frac{4\cdot72\pm4\cdot11}{46}$	$\frac{6\cdot29\pm4\cdot11}{46}$	153

general mood of the patient more than the specific respiratory condition since positive responses to questions concerning the alimentary tract equalled positive responses to those concerning the respiratory tract.

Despite the rather pessimistic nature of these findings, there is a place for respiratory function tests in some situations where the expected loss of function is small and no previous control values are possible. As in the present example, they can be used to give objective evidence of the presence of disease in an industrial environment. They can be used to compare different forms of therapy, to assess prognosis, and to investigate the pathological changes produced by disease. However, in every case, the range of physiological variation must be remembered, and the number of subjects in normal and abnormal groups adjusted to achieve the desired level of confidence.

SUMMARY

The diagnostic limitations of pulmonary function tests are considered. In a disease syndrome characterized by interstitial fibrosis and giving a mean functional loss of some 20%, the best single test (vital capacity) shows an overlap of 22% between normal and abnormal subjects. A formal discriminant analysis, including the results of eight other tests, reduces the overlap to 15.2%, but the improvement is not statistically significant. It seems probable that the main factors limiting the discriminatory capacity of laboratory tests are the underlying physiological variation and redundancy in added tests.

REFERENCES

Becklake, M. R., du Preez, L., and Lutz, W. (1958). *Amer. Rev. Tuberc.*, **77**, 400.
Carey, G. C. R., Phair, J. J., Shephard, R. J., and Thomson, M. L. (1957). *A.M.A. Arch. industr. Hlth*, **16**, 225.
Dodds, E. C. (1921). *J. Physiol. (Lond.)*, **54**, 342.
Gilson, J. C., and Hugh-Jones, P. (1955). *Spec. Rep. Ser. med. Res. Coun. (Lond.)*, No. 290.
Kendall, M. G. (1957). *A Course in Multivariate Analysis*. Charles Griffin, London.
Morant, G. M., and Gilson, J. C. (1945). R.A.F. Flying Personnel Research Committee Report, F.P.R.C. No. 933, p. 1.
Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. John Wiley, New York.
Shephard, R. J. (1955). *J. Physiol. (Lond.)*, **127**, 498.
—— Carey, G. C. R., and Phair, J. J. (1958). *J. appl. Physiol.*, **12**, 79.
—— Thomson, M. L., Carey, G. C. R., and Phair, J. J. (1958). Ibid., **13**, 189.
Yule, G. U., and Kendall, M. G. (1940). *An Introduction to the Theory of Statistics*, 12th ed. Griffin, London.

APPENDIX

RESULTS OF FORMAL DISCRIMINANT ANALYSIS

The nine variables selected for formal discriminant analysis have been listed in Table I, and will be referred to subsequently as $x_1, \ldots x_9$. Table I shows that three variables x_1, x_9, and x_7, are able to discriminate significantly by themselves. Combination of these three discriminators increases the separation of normal from abnormal, but the improvement is not statistically significant:

Variables	Estimate of Distance*
x_1	$\hat{D}_1^2 = 2\cdot48$
x_1, x_9	$\hat{D}_2^2 = 2\cdot96$
x_1, x_9, x_7	$\hat{D}_3^2 = 3\cdot17$

Is $\hat{D}_3^2 > \hat{D}_1^2$?, F (2,28 d.f.) = 1·56 (not significant).

It is also of some interest to compare the information that can be obtained by spirometry alone (variables $x_1 \ldots, x_6$) with that given by spirometry and the box-bag (x_1–x_8) and by spirometry, box-bag, and radiographic measurement (x_1–x_9):

Variables	Distance
x_1	$\hat{D}_1^2 = 2\cdot48$
$x_1 \ldots, x_6$	$\hat{D}_6^2 = 3\cdot36$
$x_1 \ldots, x_8$	$\hat{D}_8^2 = 3\cdot48$
$x_1 \ldots, x_9$	$\hat{D}_9^2 = 3\cdot86$

(*a*) Is $\hat{D}_6^2 > \hat{D}_1^2$?, F (5,25 d.f.) = 0·71 (not significant).
(*b*) Is $D_8^2 > \hat{D}_6^2$?, F (2,23 d.f.) = 0·19 ,,
(*c*) Is $\hat{D}_9^2 > \hat{D}_8^2$?, F (1,22 d.f.) = 1·12 ,,

It is evident that little improvement is effected in this case by expanding the number of measurements made with one piece of apparatus, or by adding to the pieces of apparatus used. The probable explanation is that the first test is already approaching the limits of discrimination imposed by physiological variation and the fact that the different tests are correlated.

* The estimate of Mahalanobis' D^2 measure of group distance is given by

$$\hat{D}_p^2 = l_1 d_1 + l_2 d_2 + \ldots l_p d_p$$

where $l_1, l_2, \ldots l_p$ are the P weightings determined by discriminant analysis, and $d_1, d_2, \ldots d_p$ represent the mean differences between normal and abnormal on each of the P tests. See Rao (1952) for justification of this measure.