Thorax 1985;**40**:438–443

Study of lung function data by principal components analysis

H COWIE, MH LLOYD, CA SOUTAR

From the Institute of Occupational Medicine, Edinburgh

ABSTRACT As a rational approach to the many lung function tests available, we have subjected the results of a battery of six lung function measurements made in 458 coalminers to the statistical technique of principal components analysis. By this means the six test results were reduced to three principal components without important loss of information. The first component appeared to represent lung size and the second the degree of airflow obstruction, and the third detected impairment of gas transfer factor in excess of that explained by the first two components. The values of the first principal component, used to select men with abnormal lung function, identified more younger men with functional abnormalities than a method based on comparison of observed and predicted values of forced expiration volume in one second. The values of the second and third principal components were used to classify types of functional abnormality. It is concluded that this statistical technique provides a sensitive method of identifying men with unusual lung function, particularly younger men, in a population and can be used to define and quantify different aspects of lung function.

In assessing the results of lung function tests the physician may take account of several measurements representing different aspects of lung function in order to recognise patterns of functional abnormality that may aid diagnosis. By contrast, in epidemiological studies it is usual to examine each measurement singly, or sometimes as the ratio of two such measurements; and this limits the examination of patterns of abnormality.

Some previous workers[1-3] have applied multivariate techniques to the analysis of lung function variables; and we have now applied one of these techniques, that of principal component analysis,[4] to six measurements derived from simple lung function tests performed on a population of 458 coalminers, including men with abnormal lung function.

The aims of the study were to reduce the number of variables without losing information, to examine the patterns of abnormality described by the new functional components; and to study their ability to identify individuals with abnormal or unusual function.

Methods

The 458 men whose lung function has been analysed comprise a sample of miners who had worked at one colliery in South Wales at any time from 1970 to 1981. This colliery had participated for 26 years in the pneumoconiosis field research of the National Coal Board, and men were selected from these records within five 10 year age ranges from 25 years to 65 or over. Selection was weighted to include all men with higher lifetime cumulative exposures to respirable dust and a sample of men with lower exposures. Further details of this population will be published separately.

All lung function measurements were made during a five week period by the same team of trained personnel. Each man was asked to make at least three forced expirations after inspiration to total lung capacity, and flow-volume curves were recorded, an Ohio 800 electronic spirometer and a fast response XY recorder (Hewlett Packard model number 7045A) being used. The best curve was selected for analysis on the basis of technical correctness, a clearly maximum effort (peaked not rounded flow), largest forced vital capacity (FVC), and largest forced expiratory volume in one second (FEV_1) if more than one curve had the same max-

Address for reprint requests: Dr CA Soutar, Institute of Occupational Medicine, Roxburgh Place, Edinburgh EH8 9SU.

Accepted 29 January 1985

Table 1 *Criteria for clinical definition of men with unusual lung function, based on Cotes's predicted values[5] (four groups mutually exclusive)*

Group	*n*	*FEV_1 (% pred)*	*FEV_1/FVC ratio (% pred)*	*TLCO (% pred)*
A: airflow obstruction with low gas transfer	27	<80	<90	<80
B: airflow obstruction with preserved gas transfer	27	<80	<90	>80
C: restrictive defect with low gas transfer	12	<80	>90	<80
D: restrictive defect with preserved gas transfer	30	<80	>90	>80

imum FVC. From these curves measurements were made of FEV_1, FVC, and maximum expiratory flow at 50% and 25% of vital capacity ($\dot{V}max_{50}$ and $\dot{V}max_{25}$). Single breath carbon monoxide diffusing capacity (TLCO) was measured in duplicate with a transfer test B automatic spirometer system (PK Morgan Ltd, Chatham) by the method of Meade *et al.*[5] The average of the two TLCO readings was used. Other measurements derived from these manoeuvres could have been included but we wished to avoid overcomplicating this exploratory work. Measurements of each man's height and weight were made in a standard manner.

STATISTICAL METHODS

Principal components analysis[4] was applied to results from the six lung function measurements in this population. The application of this technique results in a set of linear combinations of the original variables, each of which explains a percentage of the total variation in the data. These linear combinations are called components. The first principal component describes by the values and signs of its coefficients the line of best fit to the observations and explains more of the variation than any other such line. The second and subsequent components are derived similarly and sequentially, each explaining as much as possible of the variation unexplained by the preceding components. Each component is independent of those preceding it (at right angles in a graphical analogy), and sequential components explain progressively less of the total variation.

The components are not independent of the scales in which the original variables are measured. If any one variable is of greater magnitude, and hence has a larger variance than the others (TLCO in this example), then it will dominate the first principal component. The component will then not fairly represent the relative importance of each variable's contribution. This difficulty was overcome by standardisation of each variable to have unit variance, by subtracting the mean from each observed value and dividing by the standard deviation, before the calculation of the components.

After each component had been estimated, a component score was calculated for each man by entering his lung function measurements into the equation. These scores may be interpreted as numerical quantities on a continuous scale that represent an aspect of lung function characterised by the principal component concerned. In this paper the component scores have been used to identify men with poor lung function. Standard linear regression techniques were applied to the component scores to allow for the influence of age, height, and weight.

The selection of men with unusual function by the above method has been compared with the selection of such men by comparison of observed lung function data with published predicted values. Cotes's predicted values[6] for FEV_1, FEV_1/FVC ratio, and TLCO were used to classify men with different types of abnormal function. Men were selected if their FEV_1 was less than 80% of the predicted value, and the criteria for four mutually exclusive types are shown in table 1. Functional types A and B were chosen to represent airflow obstruction, types C and D to represent restriction. Men of type A are distinguished from type B and men of type C from type D by their lower TLCO, indicating a more severe degree or type of the obstructive or restrictive defects.

Table 2 *Weighting coefficients for the first four principal components with six lung function variables*

Component	*FEV_1*	*FVC*	*FEV_1/FVC ratio*	*$\dot{V}max_{50}$*	*$\dot{V}max_{25}$*	*TLCO*	*% variance*
1st	−0.46	−0.38	−0.38	−0.44	−0.43	−0.35	73.5
2nd	0.18	0.58	−0.57	−0.32	−0.21	0.41	14.5
3rd	0.23	0.40	−0.24	0.12	0.17	−0.83	7.2
4th	−0.26	−0.22	−0.52	0.05	0.77	0.14	3.2

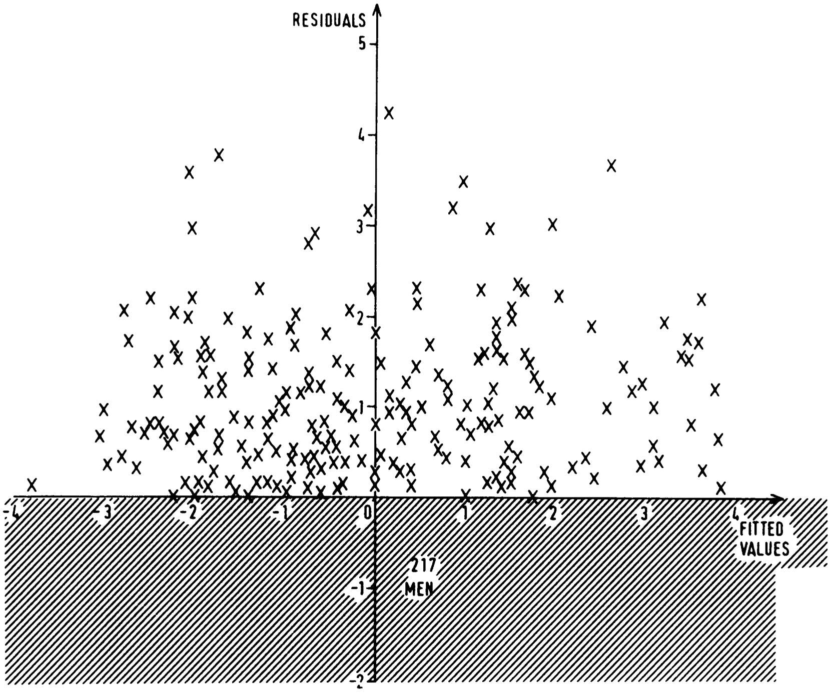

Fig 1 *Residuals (observed values minus values predicted by age, height, and weight) for first principal component scores. Comparison with fitted values (the value predicted) indicates a satisfactory fit of the statistical model. The results for 217 men with negative residuals (the hatched area and below—indicating good lung function) are not shown.*

Results

SELECTION OF INDIVIDUALS WITH POOR LUNG FUNCTION

The principal components obtained by use of the lung function variables FEV_1, FVC, TLCO, $\dot{V}max_{50}$, $\dot{V}max_{25}$, and FEV_1/FVC ratio are shown in table 2, with the proportion of the total variance in the lung function measurements explained by each component. The size and algebraic sign of the coefficients indicate respectively the variables' relative importance and the direction of the contribution to the component considered.

The first three components together accounted for 95% of the variance. The first component can be seen to be a general measure of lung function, in which each variable is about equally weighted. While interpretation of these components is subjective this component could be thought to represent lung size, and it alone explains over 70% of the variation of the measurements. Individuals with relatively small lungs (or poor lung function) would tend to have large positive scores. Since any measure of lung size is likely to be related to physique and age, the effects of these factors were taken into account by regressing the scores for the first principal component on age, height, and weight. The residuals from this analysis, calculated by substitution of age, height, and weight in the regression equations and subtraction of these values from the original component scores, are shown in figure 1. These residuals provide a way of identifying men with unusually poor lung function, by selection of all men with residuals greater than a chosen value. Taking an arbitrary residual value of greater than or equal to 1.0 defines a set of 98 men with the poorest lung function.

CLASSIFICATION OF LUNG FUNCTION ABNORMALITY

The lung function of these 98 men was classified further by the values of the second and third principal components. The second component, with large positive coefficients for FVC and TLCO and large negative coefficients for FEV_1/FVC ratio and $\dot{V}max_{50}$ (table 2), provides a contrast between these pairs of measurements. This therefore would probably distinguish between those with airflow obstruc-

Table 3 *Comparison of the classifications of men*

	Classification by Cotes's predicted values[6]					
	Unusual function			*Normal function*		
Residual	*>1.0*	*70 MEN*			*28 MEN*	
		Mean	*SD*		*Mean*	*SD*
	Age	52.9	12.26	Age	36.2	8.63
	FEV_1 % pred	59.9	13.93	FEV_1 % pred	86.3	5.28
Classification by principal components analysis	TLCO % pred	79.7	20.51	TLCO % pred	95.5	15.68
	$\frac{FEV_1}{FVC}$ % pred	8.18	14.08	$\frac{FEV}{FVC}$ % pred	90.25	6.57
Residual	*<1.0*	*26 MEN*			*334 MEN*	
		Mean	*SD*		*Mean*	*SD*
	Age	62.3	7.07	Age	45.6	11.81
	FEV_1 % pred	72.0	5.82	FEV_1 % pred	97.9	13.45
	TLCO % pred	88.1	19.40	TLCO % pred	100.9	17.35
	$\frac{FEV_1}{FVC}$ % pred	94.6	13.00	$\frac{FEV}{FVC}$ % pred	101.0	9.62

tion and those with restrictive disease. This follows since individuals with airways obstruction would be expected to have a relatively low FEV_1 compared with FVC and so a low FEV_1/FVC ratio, while those with restrictive disease would be expected to have low FVC as well as low FEV_1 and hence a high FEV_1/FVC ratio.

The third component is dominated by TLCO, which has a much larger coefficient than any other variable, and provides a measure by comparison with FVC and FEV_1. Thus a high value of the score for the third component identifies men whose TLCO is lower than average, after the abnormalities described by the first two components have been taken into account. If the residuals of both the second and the third components are taken into account, after being regressed on age, height, and weight as for the first component, then men may be classified into four mutually exclusive groups, consisting of the four possible combinations of positive and negative values of the second and third principal components. Possibly these groupings could

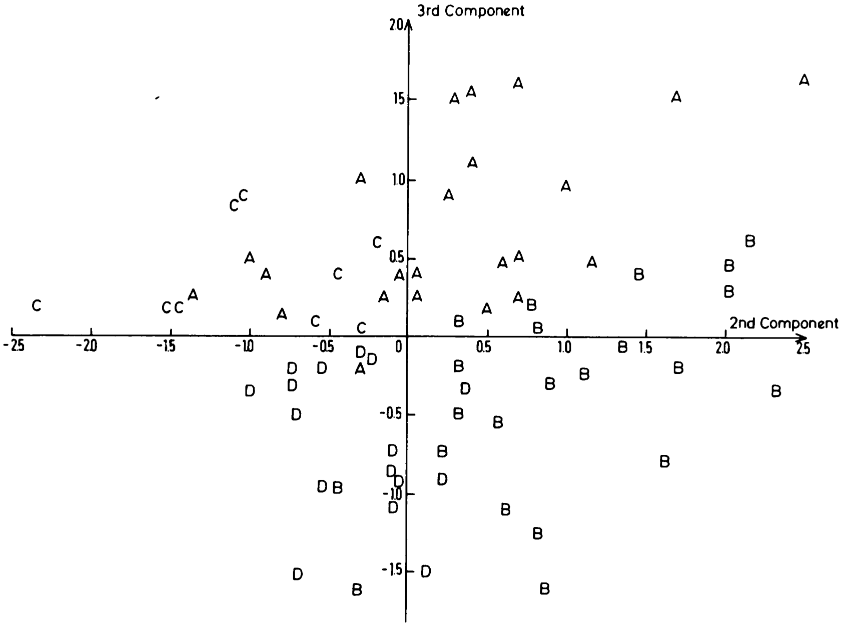

Fig 2 *Residuals (observed value minus value predicted by age, height, and weight) for second and third principal component scores. Only the results from 70 men whose residual first component score was >1·0 and men who also were included in groups A–D by the clinical criteria described in table 1 are shown here.*

describe men with airflow obstruction (with and without reduction of TLCO) and men with restrictive defects (with and without reduction of TLCO).

The fourth principal component accounted only for 3% of the variance and interpretation of its coefficients in clinical terms is not obvious. This component was not therefore included in further analysis of these men.

COMPARISON WITH SELECTION OF ABNORMAL MEN BY ANOTHER METHOD

The potential value of this method of analysis in the identification of men with unusual lung function was examined by selecting groups of men by comparison of their lung function with published predicted values, and comparing these men with the 98 selected by using the residuals from the first principal component. Selection based on Cotes's predicted values, by the criteria described in table 1, yielded 96 cases. The degree of compatibility of the two groups of cases is summarised in table 3. Seventy men were defined as cases by both criteria, leaving 26 who were included only by using Cotes's predicted values and 28 who were included only by using principal components. A useful comparison of the two methods is to look at these two sets of 26 and 28 men. Table 3 shows that the major difference between them was in age, the principal components method tending to select more younger men and fewer older men. The difference in percentage predicted FEV_1 is a consequence of the selection criteria; the men selected by using Cotes's predicted value were all constrained to have a percentage predicted FEV_1 of less than 80%, whereas no such constraint was imposed in the principal components analysis.

A comparison of the classification of types of lung function abnormality by the two methods is shown for the 70 men classified as abnormal or unusual by both methods in figure 2. The second principal component, as the coefficients suggested, distinguishes between men with obstructive and with restrictive types of defect (groups A and B as distinct from C and D); and the third component identifies men with low TLCO (groups A and C as distinct from B and D). The graph shows, however, that there are some differences in classification of individuals by the two methods.

Results similar to these were obtained when the lung function data were standardised for age, height, and weight before transformation and subsequent analysis by the principal components method. We preferred to present in this paper the results obtained when the standardisation was applied after the analysis, since in future work we shall be comparing the unstandardised residuals from the principal components analysis with explanatory variables such as smoking habit and amount of exposure to respirable dust, while simultaneously allowing for age, height, and weight by multiple regression methods. Similar results were also obtained when differences of observed values from Cotes's predicted values of lung function were used in the analysis instead of percentages of predicted values.

Discussion

This work was intended as a rational approach to the many lung function tests available—extracting the maximum information about lung function from these tests in combination, while summarising the information contained in them in relatively few new variables. Such an approach unifies the analysis of the results of multiple tests, and may also provide insight into associations between the results of one test and others and into the various functional patterns of abnormality which the lung may express.

While principal components analysis may often lead to a substantial reduction in the number of variables, it is well recognised that the components found may not always suggest a sensible physical or biological interpretation.[7] In this case not only was a reduction in variables from six to three possible, but each of these three principal components could be interpreted in familiar clinical terms, for the first component appeared to represent lung size, the second to distinguish between obstructive and restrictive types of defects, and the third to identify men with greater impairment of gas transfer factor than could be explained by the defects described by the first two components.

These three interpretations bear a remarkable resemblance to the three fundamental attributes of the lungs described by Gilson and Hugh Jones[1] on the basis of another form of multivariate analysis and a largely dissimilar set of lung function tests—namely, static size, ventilation, and gas distribution and transfer. Similarly, Macdonald and Cole,[3] after yet another form of multivariate analysis on flow-volume loop data from normal subjects, found the first component (discriminant, in their case) to be dominated by FVC and the second by peak expiratory flow rate. On the other hand, Cotes,[2] applying the principal components technique, interpreted the first component as describing airway obstruction, though perhaps this difference was a reflection of the selection of his study group, which consisted of patients with airway obstruction.

Thus a conventional empirical classification and three out of four based on unprejudiced statistical analyses each broadly support the ability of the others to classify presence and type of functional abnormality. Possibly principal components analysis (or other forms of multivariate analysis) would permit a more precise assessment of functional abnormalities than may be achieved by using lung function tests singly, and indeed the components method identified more younger men with unusual lung function than the conventional method. Further work will be needed to establish the practical importance of these observations, in particular whether the younger men so identified can be shown to have other unusual features.

The components define and quantify different aspects of lung function. They may be used not only to identify individuals with unusual lung function but also to provide new composite continuous lung function variables for epidemiological studies. Possibly they will eventually be useful in assessing the presence and degree of different types of lung pathology, by distinguishing not only restrictive from obstructive disease but perhaps emphysema from other causes of airflow obstruction.

Further work on these lines is planned, incorporating other tests of lung function. This may identify some that are effectively redundant in that they may contribute little additional information to the main principal components, and so enable the epidemiologist and the clinician to concentrate on those tests which give the most useful information.

References

1 Gilson JC, Hugh-Jones P. *Lung function in coalworkers' pneumoconiosis.* London: Medical Research Council, 1955:202–14. (Special report series No 290.)

2 Cotes JE. Lung volume indices of airway obstruction: a suggestion for a new combined index. *Proc R Soc Med* 1971;**64**:1232–5.

3 MacDonald JB, Cole TJ. The flow volume loop: reproducibility of air and helium-based tests in normal subjects. *Thorax* 1980;**35**:64–9.

4 Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933;**24**:417–41, 498–520.

5 Mead F, Saunders MJ, Hyett F, Reynolds JA, Pearl N, Cotes JE. Automatic measurement of lung function. *Lancet* 1965;ii:573–5.

6 Cotes JE. *Lung function: principles and application in medicine 4th ed.* Oxford: Blackwell Scientific Publications, 1979:369–85.

7 Chatfield C, Collins AJ. *An introduction to multivariate analysis.* London: Chapman and Hall, 1980.