Estimation of the FEV_1

PD OLDHAM, TJ COLE

From the Medical Research Council Pneumoconiosis Unit, Llandough Hospital, Penarth, Glamorgan

ABSTRACT The procedure recommended by the Medical Research Council for estimating a subject's forced expiratory volume in one second (FEV_1) is to require five separate attempts, discard the first two results, and average the last three. The most popular alternatives are to use the largest of the last three or the largest of a smaller number of results. Nine different indices derived from some or all of five attempts were compared in two studies. In one 40 normal subjects were studied. In the other 335 men exposed to industrial dust, whose forced expiratory volume declined with their degree of radiological pneumoconiosis as well as with age, were studied. There were small but consistent differences between indices. The index which emerged as the best overall in both studies was the mean of the largest three results from five attempts. It was better than the recommended index for all the comparisons made, but at the same time it gave a very similar mean value for the FEV_1. Excluding the lowest two results rather than the first two from five blows is a rational procedure, and it should be formally recognised as providing the best index available.

The one second forced expiratory volume (FEV_1) is a widely used measure of ventilatory function. It has been found to be of particular value in epidemiological studies of people exposed to atmospheric pollutants because the rate of decline of FEV_1 with age is well established, and an excessive rate can be detected readily.

Successive measurements of the FEV_1 of an individual will vary, and some convention has to be adopted of how many blows should be required on one occasion, and what index constructed from the resulting FEV_1s should be used to characterise the individual on that occasion. Because the test is a measure of maximal performance, it has seemed natural to many users of it to take the largest value achieved. On the other hand general statistical laws would suggest that some form of average of the separate blows would give a more stable index, better for comparing the individual's performance with that of another or with himself on another occasion. Gilson and Hugh-Jones[1] showed that the vital capacity of an individual measured many times formed a symmetrical distribution like a normal curve, and the same would be expected for the FEV_1. Even if its distribution were highly skewed, clustered at or near the individual's true maximum with a tail of poor attempts below, the average of a small number of blows would remain a good estimate of the *average* of the real skewed distribution, whereas the largest of the small number would not estimate any particular feature of the real skewed distribution.

Several studies established that, on average, the first and second blows of a set gave smaller FEV_1s that the third and subsequent blows. The Medical Research Council (MRC)[2] therefore recommended that the best procedure would be to require five blows, the first and second being treated as practice attempts, and to report the average of the third, fourth, and fifth blows as giving the individual's FEV_1.

Although this recommendation has been widely adopted, there has always been dissatisfaction with it. It has been claimed that five blows are too many for a child or disabled person, so that the last three

Address for reprint requests: Dr PD Oldham, MRC Pneumoconiosis Unit, Llandough Hospital, Penarth, S Glam CF6 1XW.

Accepted 3 May 1983

blows do not cluster around a consistent average but show a steady decline, or are in many cases unobtainable because of the subject's fatigue. It has been claimed that a common pattern of the last three blows is for two to be alike and one widely different, usually having a lower value but sometimes a higher one. This makes the averaging of all three seem unattractive.

There has been little evidence presented to show whether or not the MRC recommendation could be improved on, or whether other indices of the FEV$_1$, which might be more convenient to use, give better or worse results in practice. There have been, however, continuing debate and uncertainty about this point. The purpose of this paper is to present the results of two investigations of rather different types, in each of which the subjects performed five blows for FEV$_1$ estimation, so that the performance of almost any index of FEV$_1$ can be compared.

Methods

STUDY 1: REPEATABILITY IN NORMAL SUBJECTS

The first study was conducted by one of us (TJC) on 40 normal subjects, 20 male and 20 female, who did five FEV$_1$ blows on each of six days (a five day week and the preceding Thursday). Over such a period nothing but random fluctuations in FEV$_1$ would be expected, so if the total variance of the 240 values of an index of FEV$_1$ is partitioned into that corresponding to differences between subjects' means and that corresponding to variation within subjects an obvious measure of the quality of a particular index is the size of its random within-subject variance as a fraction of the total, which should be as small as possible, or of the complementary fraction, which should be as large as possible.

A standard wet spirometer (Poulton) was used. The spirometer was initially calibrated for volume by means of a 1 litre syringe in steps up to 6 l. The timing was set electronically. Each day the consistency was checked by applying a standard weight with a standard orifice in place of the breathing tube. The water temperature was measured before the first blow of each subject and the reading corrected to BTPS. Each subject was asked to carry out the full forced vital capacity (FVC) procedure but only for the fourth and fifth blows was care taken to ensure that the breath was fully expired; the FVC was then recorded. The subjects took their time between blows, the operator refraining from prompting them to blow again. The usual interval was less than 20 seconds, though in a recent outpatient clinic at this unit the average interval chosen by 20 very disabled patients was 28 seconds. When there were technical faults (lack of a full inspiration, removal of mouthpiece, coughing) the attempt was repeated.

STUDY 2: SENSITIVITY OF AN EPIDEMIOLOGICAL SURVEY

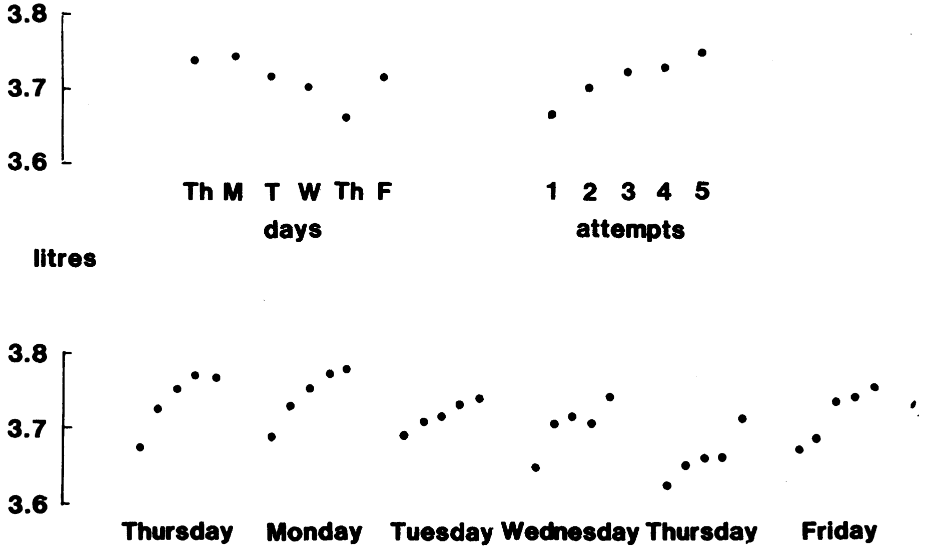
The second study was a survey of workers and exworkers in the North Wales slate industry.[3] When the index of the FEV$_1$ recommended by the MRC, standardised for height by the method of Cole,[4] was used it was found that FEV$_1$ and other measures of respiratory health depended on age, smoking, and the amount of radiological pneumoconiosis but not on any other aspect of the men's occupational history. From this survey the 355 current smokers were selected to investigate other indices of the FEV$_1$.

A bellows type digital spirometer (McDermott, Garw Electronics) was used. The temperature of the metal plate forming the top of the bellows was recorded before the first blow of each subject, and the readings were then corrected to BTPS. The procedure was otherwise the same as for the first study.

Since the amount of pneumoconiosis had been converted to a numerical score by the method of Oldham[5] any index of FEV$_1$ could be related, by linear regression, to the ages and amounts of pneumoconiosis of the subjects; and the quality of a particular index of FEV$_1$ may be measured by the size of the deviations from regression as a fraction of the total, which should be as small as possible, or the complementary fraction, which in this case is the square of the multiple correlation coefficient, R^2, which should be as large as possible. Further, in this case the size of each regression coefficient as a multiple of its standard error can be examined, to see whether some indices are relatively more efficient at detecting the loss of FEV$_1$ associated with the amount of pneumoconiosis than the loss, of no relevance to occupational disease, associated with age.

POSSIBLE INDICES OF FEV$_1$

A very large number of possible indices of FEV$_1$ can be invented, and we have chosen those which have been recommended or used by others and those which might have to be adopted when less than five successful blows have been recorded. The following indices have been examined: (1) the MRC recommendation, the average for the last three blows of five; (2) the average for all five blows; (3) the largest result from five blows; (4) the largest of the last three results—recommended by Peto in Fletcher *et al*[6]; (5) the average of the largest three results from five blows; (6) the average for the first three blows; (7) the largest of the first three results; (8) the average of the first two results; (9) the larger of the first two results. Indices 6 to 9 are evidently of



The results for the 40 normal subjects, averaged over subjects and attempts, subjects and days, and subjects alone.

interest in cases where the investigator was unable or unwilling to obtain five blows.

In addition, an optimal index was obtained for each set of data, by finding that set of weights which was such that the weighted average of the five results (in original order or in order of magnitude) had the smallest residual variation as a fraction of the total. Use of the absolute value of the residual variation for this purpose[6 7] is unsatisfactory since the most stable index found in this way may not respond to the real sources of variation, whose detection is the basic purpose of any study. We need maximal variation between persons, subject to maximal consistency of measurements of the same person.

Results

STUDY 1

The figure shows the results of averaging the results of the individual blows of the normal subjects in various ways. The upward trend over the five attempts is clear. There is also a cyclic trend from day to day over the period of the study.

Table 1 shows performance as expressed by the nine indices. The table shows the means, the standard deviations within subjects (which correspond to the repeatability of each index) in units of FEV_1 and as percentages of the mean, and the square of the intraclass correlation coefficient. The larger this last is the higher the proportion of the total variation which corresponds to the real differences between subjects. The rankings of these measures of quality are also given.

The mean of all five observations is best by each criterion. The mean of the largest three readings is, however, almost as good. Surprisingly, the mean of the first three readings comes next, by each criterion, and the largest of the five readings comes fourth. The MRC recommendation (mean of the last three readings) is slightly better than Peto's recom-

Table 1 *Nine indices of FEV_1 in 40 normal subjects (numbers in parentheses indicate rank order)*

Index	*Mean (l)*	*SD within subjects*	*Coefficient of variation %*	*R^2*
1 MRC index (mean of last three results from five blows)	3·733	0·089 (4 =)	2·38 (5)	0·990 (4 =)
2 Mean of all five	3·713	0·082 (1)	2·20 (1 =)	0·992 (1 =)
3 Maximum of all five	3·795	0·089 (4 =)	2·34 (4)	0·990 (4 =)
4 Maximum of last three (Peto)	3·783	0·092 (6 =)	2·43 (6)	0·990 (4 =)
5 Mean of largest three	3·755	0·083 (2)	2·20 (1 =)	0·992 (1 =)
6 Maximum of first three	3·760	0·096 (9)	2·56 (9)	0·988 (9)
7 Mean of first three	3·696	0·086 (3)	2·33 (3)	0·991 (3)
8 Mean of first two	3·683	0·092 (6 =)	2·50 (8)	0·989 (7 =)
9 Larger of first two	3·729	0·093 (8)	2·49 (7)	0·989 (7 =)

mendation[6] (maximum of the last three).

None of the nine indices is conspicuously bad; even the worst, the largest of the first three blows, has a coefficient of variation of only 2·56%, compared with 2·20% for the best indices. The means vary considerably.

The best attainable index is a weighted average of the five blows in the order in which they were obtained. It has a standard deviation of 0·081 and an R^2 of 0·992.

STUDY 2

Table 2 shows the performance of the nine indices in study 2. In this case the table shows the regression coefficients of the index on age and amount of pneumoconiosis; their standard errors; the corresponding values of Student's *t*; and, as before, the "error" standard deviation and the square of the multiple correlation coefficient. The differences between the coefficients on age are negligible; rounded to 3 decimal places they are virtually indistinguishable, although the values of *t*, derived from the unrounded coefficients and standard errors, show some variation. The regression coefficients on amount of pneumoconiosis vary more, as do the corresponding values of *t*; but even the worst index, the mean of the first two FEV_1s, still shows a significant relationship to amount of pneumoconiosis.

The multiple correlation coefficients indicate that the mean of the largest three results from all five blows is formally the best index, but it is clearly not materially better than the MRC index, the mean of the last three blows, and several of the other indices. The largest attainable R^2 is 0·583, given by a weighted average of the results of the five blows in order of magnitude (the weights being not obviously interpretable); so that indices 1–5 and 7 are all almost optimal. The most consistently successful index by any criterion is the mean of the largest three results.

DISTRIBUTIONS OF DEVIATIONS

Examination of the actual distributions of deviations (from the subject means in the first set of data, from the regression lines in the second) shows no conspicuous differences. In the first set of data the distributions of the deviations from the mean of all five results and from the mean of the last three were slightly skew and the latter slightly too peaked. The mean of the largest three results and the largest of the last three give reasonably normal curves. Deviations from the regression lines in the second set of data show lesser differences from index to index.

Discussion

There have been few previous examinations of the merits of different indices of FEV_1. Published reports are reviewed by Fletcher *et al*[6] (appendix B, section B1) and may be summarised as follows: (1) In inexperienced subjects the first three blows show an increase, which then levels off; but in experienced subjects the first blow is as large as any.[8 9] (2) The mean of three readings usually has a lower variance than the maximum.[10] (3) The higher of the first two readings has a smaller variance than their

Table 2 *Performance of nine indices of FEV_1: data from study 2, on 335 smokers from a slate works (numbers in parentheses indicate rank order)*

Index	*Mean (l)*	*SD about regression on age and amount of radiological pneumoconiosis and coefficients of variation (%)*		*Regression on age*	*Regression on amount of pneumoconiosis*	*R^2*
1 MRC index (mean of last three results from five blows)	3·112	0·892 (5)	28·7 (5)	− 0·041 ± 0·002 (t = 18·4) (2)	− 0·164 ± 0·061 (t = 2·71) (3)	0·578 (2)
2 Mean of all five	3·029	0·886 (1)	29·3 (6)	− 0·040 ± 0·002 (t = 18·38) (4)	− 0·157 ± 0·060 (t = 2·61) (6)	0·576 (4 =)
3 Maximum of all five	3·200	0·887 (2 =)	27·7 (1)	− 0·040 ± 0·002 (t = 18·39) (3)	− 0·162 ± 0·060 (t = 2·70) (4)	0·577 (3)
4 Maximum of last three (Peto)	3·178	0·895 (7)	28·2 (3 =)	− 0·041 ± 0·002 (t = 18·25) (6)	− 0·169 ± 0·061 (t = 2·78) (1)	0·575 (6)
5 Mean of largest three	3·142	0·887 (2 =)	28·2 (3 =)	− 0·040 ± 0·002 (t = 18·47) (1)	− 0·164 ± 0·060 (t = 2·74) (2)	0·579 (1)
6 Maximum of first three	3·159	0·888 (4)	28·1 (2)	− 0·040 ± 0·002 (t = 18·36) (5)	− 0·161 ± 0·060 (t = 2·66) (5)	0·576 (4 =)
7 Mean of first three	2·974	0·894 (6)	30·1 (8)	− 0·040 ± 0·002 (t = 17·93) (7)	− 0·152 ± 0·062 (t = 2·45) (8)	0·562 (7)
8 Mean of first two	2·905	0·913 (8)	31·4 (9)	− 0·040 ± 0·002 (t = 16·77) (9)	− 0·147 ± 0·066 (t = 2·24) (9)	0·528 (9)
9 Larger of first two	3·053	0·914 (9)	30·0 (7)	− 0·040 ± 0·002 (t = 17·19) (8)	− 0·164 ± 0·065 (t = 2·54) (7)	0·544 (8)

mean.[11] (4) In follow up studies of the same subjects the largest of three readings is more reproducible than the mean from one occasion to the next.[12]

Fletcher and colleagues'[6] own data were a sequence of seven annual surveys of the same men. Five readings of FEV_1 were obtained on each occasion, but only the last three were available for analysis. The criterion used was the linearity of the change of FEV_1 for each man, and they found that the maximum of the three blows showed a smaller deviation from linearity than the mean, and that the optimal combination of the three blows in order of magnitude was virtually the same as choosing the largest and discarding the others. They therefore recommend very strongly the use of the largest of the last three blows.

A later study by Ferris,[7] based on three sets of data, suggested that the mean of the largest three of five measurements was best, though their maximum was slightly better if the criterion used was that fewer subjects should show an increase in FEV_1 over six years.

A recent study by Ullah *et al*[13] concluded that choosing the one largest measurement was insensitive, and strongly recommended averaging as many attempts as could conveniently be obtained. This conclusion was reached from the absence of trends and skewness of distribution in runs of 10 or 20 blows at one or two minute intervals. These authors also draw attention to the possibility that useful information may be contained in the actual pattern of individual blows, which is lost when only a single index is used. Their study extends the work of Gilson and Hugh-Jones[1] by showing that the FEV_1 is normally distributed in patients as well as in normal subjects. It also shows that the first blow is either the highest or the lowest (out of 10) more often than expected, and significantly so for patients. This emphasises that the first blow (and to a lesser extent the second as well) is more variable than later blows, which may mean that it is potentially informative.

Our results (fig) show a consistent rise from blow 1 to blow 5 on each of the days. Ullah *et al*[13] by contrast found no overall trend in FEV_1 with time. This may be related to their extended sampling regimen, attempts being made every one or two minutes rather than at the more usual interval of 30 seconds or less. Our findings indicate that a mean FEV_1 based on five attempts should be appreciably higher than a mean based on only three. Thus their recommendation to use "as many observations as can be conveniently obtained" is inherently unsatisfactory, as the result is likely to depend on how many attempts are made. In any case it is unwise to leave the number of attempts unspecified, as this could lead to even less standardisation than exists now.

In our results indices based on means had smaller variances than indices using maxima, whether measured as day to day scatter or as unexplained variation about regression lines on related variables. In particular, the index preferred by Fletcher *et al*,[6] the largest of the last three blows, was far from optimal in both our studies. Although Fletcher *et al* regarded this index as optimal, its advantage over the mean of the last three blows (the MRC index) was measured by a reduction of standard error of only 3·6%. Moreover, the criterion used, least scatter about a linear decline of FEV_1 with age, may not be appropriate in that the authors conclude that FEV_1 is lost with age at an accelerating rate, albeit with a very small acceleration. Fletcher *et al* suggested that, had all five blows been available to them, the maximum of the five would have been found to be best of all. Our studies show the maximum of five to be a very good index but inferior to the mean of all five.

In terms of repeatability, there is remarkably little to choose between the indices we have examined. Even those based on the first two or three blows are not dramatically worse than those based on five. In terms of the actual level of FEV_1, choice of index is far more important. Reference values for FEV_1 have usually been based on the MRC index, and indicate a loss of FEV_1 of 0·031 l per year of age in symptomless men.[14] In study 2 differences between the indices expressed in terms of apparent differences in years of age cover 9·5 years, from the mean of the first two blows, 2·905 l, to the maximum of all five blows, 3·200 l, a difference of 0·295 or 9·5 times 0·031 l. Results of different studies using these different indices could not safely be compared. Only the mean of the largest three blows lies within one year of age of the MRC index in both our studies, and so could safely be adopted without a major change of standard. Indeed, we think that this index should replace the MRC index in general use. Its outstanding advantage is that it removes the element of subjective judgment that often arises when one of the last three of five FEV_1s is noticeably low; should it be classed as an unsuccessful effort and be discarded or must it be retained and included in the average?* Equally, in cases where the usual pattern of successive FEV_1s is not seen, and the first or second attempt is larger than some of the last three attempts, the new index does not demand what seems to be the abandonment of common sense in

*It may be of interest to point out that in samples of three from a normal distribution the middle observation is more than four times nearer one of the extremes than the other on more than a third of occasions; only if the distance to one extreme exceeds 32·57 times the distance to the other is the sample significantly ($p = 0·05$) suggestive of non-normality.[15]

pursuit of an arbitrary rule. Provided that the basic principle of requiring five attempts is adhered to, the mean of the largest three of these will produce no bias and no increase of random error and will accord more with natural instinct for what constitutes a reasonable index than does the MRC index.

References

[1] Gilson JC, Hugh-Jones P. The measurement of the total lung volume and breathing capacity. *Clin Sci* 1949;**7**:185–216.

[2] Medical Research Council. Definition and classification of chronic bronchitis for clinical and epidemiological purposes. *Lancet* 1965;i:775–9.

[3] Glover JR, Bevan C, Cotes JE, *et al*. Effects of exposure to slate dust in North Wales. *Br J Ind Med* 1980;**37**:152–62.

[4] Cole TJ. Linear and proportional regression models in the prediction of ventilatory function. *J R Stat Soc Series A* 1975;**138**:297–338.

[5] Oldham PD. Numerical scoring of radiological pneumoconiosis. In: Walton WH, ed. *Inhaled particles III*. Old Woking, Surrey: Unwin, 1971:621–30.

[6] Fletcher C, Peto R, Tinker C, Speizer FE. *The natural history of chronic bronchitis and emphysema*. Oxford: Oxford University Press, 1976.

[7] Ferris BG. Epidemiology standardisation project. *Am Rev Respir Dis* 1978;**118**:suppl.

[8] Ashford JR, Forwell GD, Routledge R. A study of the repeatability of ventilatory tests, anthropometric measurements, and answers to a respiratory symptoms questionnaire in working coal-miners. *Br J Ind Med* 1960;**17**:114–21.

[9] Freedman S, Prowse K. How many blows make an FEV? *Lancet* 1966;ii:618–9.

[10] Ferris BG, Anderson DO, Zickmantel R. Prediction values for screening tests of pulmonary function. *Am Rev Respir Dis* 1965;**91**:252–61.

[11] Lowe CR, Pelmear PL, Campbell H, Hitchens RAN, Khosla T, King TC. Bronchitis in two integrated steel works. I. Ventilatory capacity, age and physique of non-bronchitic men. *Br J Prev Soc Med* 1968;**22**:1–11.

[12] Stebbings JH jun. Chronic respiratory disease among non-smokers in Hagerstown, Maryland. II. Problems in the estimation of pulmonary function values in epidemiological surveys. *Environ Res* 1971;**4**:163–92.

[13] Ullah MI, Cuddihy V, Saunders KB, Addis GJ. How many blows really make an FEV_1, FVC, or PEFR? *Thorax* 1983;**38**:113–8.

[14] Cotes JE. *Lung function: assessment and application in medicine*. 4th ed. Oxford: Blackwell Scientific Publications; 1979.

[15] Youden WJ. Sets of three measurements. *Scientific Monthly* 1953;**77**:143–7.