AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 120, No. 5
Printed in U.S.A.

SELECTION EFFECTS OF REPEATABILITY CRITERIA APPLIED TO LUNG SPIROMETRY

ELLEN A. EISEN,[1] JAMES M. ROBINS,[1] IAN A. GREAVES[1] AND DAVID H. WEGMAN[2]

Eisen E. A. (Harvard School of Public Health, Boston, MA 02115), J. M. Robins, I. A. Greaves and D. H. Wegman. Selection effects of repeatability criteria applied to lung spirometry. *Am J Epidemiol* 1984;120:734-42.

The potential for introducing bias in studies of pulmonary function by the exclusion of subjects with nonrepeatable measurements was examined in a cohort of Vermont granite workers followed for five years. At each annual survey, a "test failure" was defined as a test in which the two largest forced expiratory volumes in one second (FEV_1) differed by more than 200 ml. "Persistent test failure" was defined in terms of 1) the number of test failures for each worker over the six surveys and 2) the difference between the two best efforts at each survey, averaged over all surveys for each worker. The rate of FEV_1 loss was estimated for each subject based only on repeatable measurements. It is widespread practice to exclude subjects from analysis who do not perform repeatable lung function tests. The authors found that subjects with persistent test failure were losing FEV_1 at a faster rate than subjects without. The results suggest that the application of rigid repeatability criteria may bias epidemiologic findings by the exclusion of many subjects with accelerated loss of lung function.

epidemiologic methods; longitudinal studies; lung volume measurements

In 1978, the Snowbird Conference of the American Thoracic Society proposed a series of standards for spirometry (1). Later that same year, similar standards were published by the Epidemiology Standardization Project (2). The Snowbird Report included instrument specifications as well as methods for the standardized measurement of forced expiratory volumes. An acceptable curve was to be defined by the technician's observation that the subject understood the instructions and performed the test with a smooth continuous exhalation, with apparent maximal effort, with a good start, and without coughing, early termination, glottis closure, a leak, or an obstructed mouthpiece. The participants at the Snowbird workshop recommended a minimum of three acceptable forced expiratory maneuvers. In an attempt to ensure subject cooperation and maximal efforts, they further proposed that the best two of three acceptable curves should not vary by more than 100 ml or 5 per cent, whichever was greater.

Most epidemiologic studies of pulmonary function reported after 1978 have conformed with the recommended standards, although there are differences in the degree of variability tolerated between the two largest of the three acceptable curves. Common variants on the

Received for publication October 17, 1983, and in final form March 1, 1984.

[1] Department of Environmental Science and Physiology, Harvard School of Public Health, Boston, MA.

[2] Center for the Health Sciences, UCLA School of Public Health, Los Angeles, CA.

Reprint requests to Dr. Ellen A. Eisen, Occupational Health Program, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115.

This work was supported by Grant 1 R23 ES 03023-01 from the National Institute of Environmental Health Sciences.

"100-ml or 5 per cent" rule found in the literature include a simple 100-ml or 200-ml rule, a 3 per cent rule, and a 5 per cent criterion (3–10). Pulmonary function tests that fail to satisfy whatever criterion is applied are generally regarded as unreliable and are excluded from the subsequent analysis of the measurements.

The need for a standardized method to exclude submaximal values from the analysis of pulmonary function data motivated the American Thoracic Society recommendations. To the extent that satisfactory repeatability can only be achieved for true maximal efforts, the exclusion of nonrepeatable tests would certainly reduce the risk of falsely low test results. If lung disease were sometimes responsible for certain subjects being unable to perform repeatable tests, however, the widespread adoption of spirometry standards that include repeatability criteria could result in the exclusion of data from a proportion of subjects with pulmonary disease. If the application of repeatability criteria leads to the selective exclusion of diseased individuals, their application may thus be an important source of bias in epidemiologic studies that consider the possible effects of environmental agents on the lungs.

Clinical experience suggests that subjects with airflow obstruction and pulmonary disability do not have highly repeatable lung function tests. This is particularly true for asthmatics in whom repeated forced expiratory maneuvers may exacerbate airflow narrowing (11), and for subjects with other forms of airway obstruction in whom serial measurements over several minutes may show a sequential decrease in forced expiratory volumes. It is important to note that these clinical observations relate to patients who have overt disease and are usually disabled; it is presently unknown whether "early" or mild disease might similarly influence forced expiratory volumes.

To evaluate the influence of current spirometric standards that include a repeatability criterion, we have examined the rate of decline in lung function among subjects who satisfy the repeatability criterion and among those who do not. The subjects come from a cohort of Vermont granite workers who were recruited in 1970; all were currently employed at the time of recruitment and none could be considered to have disabling lung disease at the beginning of the study.

We shall use the term *repeatability* to denote the ability of an individual at a single sitting to perform a series of expiratory maneuvers within a specified narrow range. *Test failure* will be used to describe a series of forced expirations performed on a single occasion which do not satisfy the repeatability criterion.

METHODS

Study design

An industry-wide silicosis prevention program was initiated in the Vermont granite sheds in the late 1930s. At that time, dust controls were installed in all the granite sheds which substantially reduced silica exposures, and a medical monitoring system was started. There was an initial emphasis on radiographic examinations, but in 1969, pulmonary function tests and respiratory questionnaires were added to the annual surveys. This longitudinal study consists of data collected from 1970 to 1975 in six annual medical surveys of the workers.

The population eligible for this study were the white male workers actively employed in 1970 who started work in the industry after the reduction of dust levels (after 1940). Eligibility also required that the workers were at least 25 years of age in 1970 and that they attended the medical survey in that year. After excluding those with histories of employment in other dusty trades, 713 workers were available for study.

Pulmonary function tests

Forced expiratory volume in one second (FEV_1) was measured annually for all the participants in the medical surveys. A single Stead-Wells spirometer was used for all lung function testing, and the same technician administered all tests. For each subject, a minimum of three and a maximum of five measurements were recorded. A *test failure* for a given worker was defined when the two largest FEV_1s differed by more than 200 ml during a particular test session (7). For all repeatable FEV_1s, the mean of the two best values was used. An additional measure of repeatability was defined as the difference between the two largest acceptable FEV_1s in each session (*FEV_1 difference*).

Two measures of persistent test failure were also defined: 1) the total number of test failures for each worker over the six medical surveys and 2) the average FEV_1 difference over all surveys for each worker. The latter measurement will be referred to as the *mean difference*.

Calculation of rate of FEV_1 decline

An estimate of rate of FEV_1 decline was computed for all workers who had repeatable FEV_1 measurements on at least two occasions. The rate of FEV_1 decline for each worker was estimated in a separate regression equation in which FEV_1 was regressed on follow-up time as follows:

$$\mathrm{FEV}(i,t) = a(i) - b(i)t, \qquad (1)$$

where $a(i)$ is the estimated FEV_1 level of the ith worker in 1970, $b(i)$ is the estimate of the rate of FEV_1 decline for the ith worker, and t is the duration of follow-up for each measurement (i.e., $t = 0-5$ for measurements from 1970 to 1975, respectively). $\mathrm{FEV}(i,t)$ is the mean of the two largest FEV_1s for worker i at time t. The value for $\mathrm{FEV}(i,t)$ was used only if the repeatability criterion was satisfied; otherwise, $\mathrm{FEV}(i,t)$ was treated as a missing observation.

Exposure

Lifetime silica exposure was estimated for each subject by combining his work history in the industry with extensive industrial hygiene data collected during the study period. Lifetime exposure for each individual was defined as a sum of the dust levels in each job worked, weighted by the number of years spent in that job. The derivation of exposure estimates is described in detail in a separate report (12). On average, the dust levels were approximately half the current threshold limit value for occupational exposures to silica-containing dust.

Analytic method

Neither the rate of FEV_1 decline nor the FEV_1 difference was normally distributed. Both were skewed and had a higher proportion of observations in the tails than Gaussian distributions. Furthermore, the standard error of the mean rate of FEV_1 decline varied by threefold between relevant subgroups of the cohort. In order to use the rate of FEV_1 decline as the outcome in a multivariate parametric model, a logarithmic transformation was needed to stabilize the variance and to symmetrize the distribution. A constant was added to each measurement before the transformation to avoid undefined values (13).

A matrix of Spearman correlations, rather than parametric correlations, was computed between FEV_1 differences in each survey. A simultaneous test of whether the off-diagonal elements were nonzero was performed using an approximation of the Bartlett statistic (14).

RESULTS

Study population

The rate of FEV_1 decline could not be calculated for 56 workers who had attended only one survey nor for 39 who failed to perform repeatable FEV_1s on at least two occasions. The mean character-

istics of the remaining 618 subjects are shown in table 1. The workers had an average baseline FEV_1 which was 95 per cent of that predicted for their age and height, using a standard equation based on healthy, asymptomatic white males (15). Their mean rate of FEV_1 decline was 51 ml/year.

The cohort was divided into "dropouts" (those who left during the study) and "survivors." Those who left the industry were slightly older and had a lower baseline FEV_1 (per cent predicted). The mean difference was approximately equal in the two groups, although its variance was greater among the dropouts. The dropouts also had a greater rate of FEV_1 decline than the survivors: 69 ml/year versus 48 ml/year, respectively. The proportion of current smokers was similar in the two groups.

Because of the restriction imposed on the date of hire (1940), the cohort was relatively young, and only 11 of the 103 dropouts left because of retirement. The 11 retirees were found to be losing 42 ml/year; this rate of decline was less than that of workers who quit before retirement age. Additional data on the dropouts have been published in a separate report (16). The remainder of this analysis concentrates on those 515 workers who remained actively employed throughout the entire study period.

Does test failure occur randomly?

The overall test failure rate throughout the study was 11 per cent using the 200-ml repeatability criterion. This rate did not vary significantly between surveys (table 2). We first examined whether test failure was equally likely for all subjects at each testing session. A goodness of fit test showed that the variability in test failure rates was greater than expected under the binomial assumption ($p = 0.001$). Thus, a greater proportion of workers was found to have multiple test failures than would have been expected if test failures had occurred randomly, indicating that some individuals had persistent test failure (table 3).

The associations between degrees of repeatability among surveys were also illustrated by the Spearman correlation matrix of FEV_1 differences in each survey for each worker. Although small, the strength of the correlations increased (from $r = 0.03$ to $r = 0.14$) as the number of years between surveys decreased (from five to one). All the correlations between FEV_1 differences from survey to survey were positive and simultaneously greater than zero ($p < 0.001$). This suggests that workers who performed a poorly repeatable FEV_1 at one survey were more likely to perform poorly repeatable tests at subsequent surveys.

Having demonstrated that persistent

TABLE 1

Characteristics of the cohort of Vermont granite workers: dropouts and survivors

	Dropouts	Survivors	Total
No. of workers	103	515	618
Age (years)*	42.7 ± 1.0	39.7 ± 0.4	40.2 ± 0.4
FEV_1† level (% predicted)*	92% ± 1.6%	96% ± 0.7%	95% ± 0.6%
FEV_1 mean difference (ml)*	93.5 ± 6.6	90.0 ± 2.5	90.4 ± 2.3
FEV_1 decrement (ml/year)*	69.4 ± 12.6	47.8 ± 3.9	51.4 ± 3.9
Silica exposure ($mg/m^3 \times$ years)*	0.773 ± 0.060	0.703 ± 0.024	0.714 ± 0.023
% never smoked	18	17	17
% current smokers	58	60	60

* Values are means ± standard errors.

† Forced expiratory volume in one second.

TABLE 2

Survey effects; study of Vermont granite workers

	Survey no.					
	1	2	3	4	5	6
No. of participants	542	477	452	438	429	446
FEV_1* difference (ml)†	92 ± 5	94 ± 5	94 ± 5	94 ± 5	96 ± 6	94 ± 6
FEV_1 (% predicted)	96	95	94	94	94	94
Test failure rate	0.10	0.12	0.11	0.11	0.11	0.11

* Forced expiratory volume in one second.

† Values are means ± standard errors.

test failure does occur in some individuals, we proceeded to examine the more critical question of whether the subjects with persistent test failure were a random subset of the cohort with respect to the rate of FEV_1 decline.

Characteristics of subjects with persistent test failure

Because participation in the medical surveys was voluntary, some workers did not attend all the surveys. There was no correlation, however, between the mean rates of FEV_1 decline and the number of surveys attended ($r = 0.04$, $p = 0.31$). Therefore, in spite of the unequal number of opportunities for failure, the relationship between persistent test failure and the rate of FEV_1 decline was unconfounded by the degree of participation. As shown in table 3, 62 per cent of the cohort had no test failures, 25 per cent had only one, and 12 per cent had more than one test failure; the corresponding mean rates of FEV_1 decline for the three groups were 46, 37, and 81 ml/year, respectively.

The subjects with persistent test failure, as defined by at least two test failures, had greater variability in their rates of FEV_1 decline than subjects without persistent failure. This was probably because of the heterogeneity of the population of individuals with persistent failure and the fewer number of test successes per individual available for the estimation of their rates of decline. When the transformed rate was treated as the outcome in a regression analysis, the association between persistent test failure and rate of FEV_1 decline was significant ($p = 0.05$) after controlling for age, height, silica exposure, and current smoking status.

As a more sensitive indicator of persis-

TABLE 3

Frequency of test failure among survivors in the six annual surveys of Vermont granite workers

	No. of test failures		
	0	1	2
No. of workers	323	131	61
Age (years)*	40.0 ± 0.5	39.4 ± 0.7	38.5 ± 1.0
Height (cm)*	173.8 ± 0.3	174.2 ± 0.6	173.6 ± 0.8
FEV_1† level (% predicted)*	96% ± 0.8%	96% ± 1.4%	96% ± 2.2%
FEV_1 decrement (ml/year)*	45.9 ± 3.4	37.2 ± 8.0	81.1 ± 21.2
Silica exposure ($mg/m^3 \times$ years)*	0.637 ± 0.031	0.716 ± 0.051	0.757 ± 0.062
% never smoked	15	17	25
% current smokers	58	66	57

* Values are means ± standard errors.

† Forced expiratory volume in one second.

tent test failure, the differences between the two largest FEV_1s at each session were averaged over all the surveys for each worker. As the mean difference increased, so did the rate of FEV_1 decline. This relationship is illustrated in figure 1.

FEV_1 differences were found to be correlated between successive surveys; that is, an individual's FEV_1 differences at surveys adjacent to test failures were likely to be larger than those for the surveys that were preceded and followed by test successes. Thus, inclusion of FEV_1 measurements from surveys adjacent to test failures could influence estimates of the rates of FEV_1 decline. If test successes occurred in the early surveys and were followed by test failures, the last valid FEV_1 would more likely underestimate the "true" FEV_1 than the earlier measurements, thereby causing the rate of FEV_1 decline to be overestimated. Conversely, test failures in the early surveys followed by successes could result in underestimation of the rate of FEV_1 decline.

To assess the impact of these possible biases on our results, the juxtaposition of test failures and successes was examined among subjects with a mean difference of at least 200 ml. Without knowledge of the calculated rates of FEV_1 decline, two of the authors (E.A.E. and J.M.R.) independently divided subjects into three groups according to the relative ordering of their failures and successes: 1) those with test patterns likely to cause overestimation of FEV_1 decline (successes followed by failures); 2) those likely to be unbiased (successes and failures interspersed); and 3) those likely to cause underestimation of FEV_1 decline (failures followed by successes). The means of the observed rates of FEV_1 decline for these three groups were in the anticipated order: 171 ml/year, 99 ml/year, and 67 ml/year, respectively. It is important to note that even subjects in groups 2 and 3 (with unbiased or underestimated rates of FEV_1 decline) had, on average, greater rates of FEV_1 decline than the members of the cohort who had a mean difference

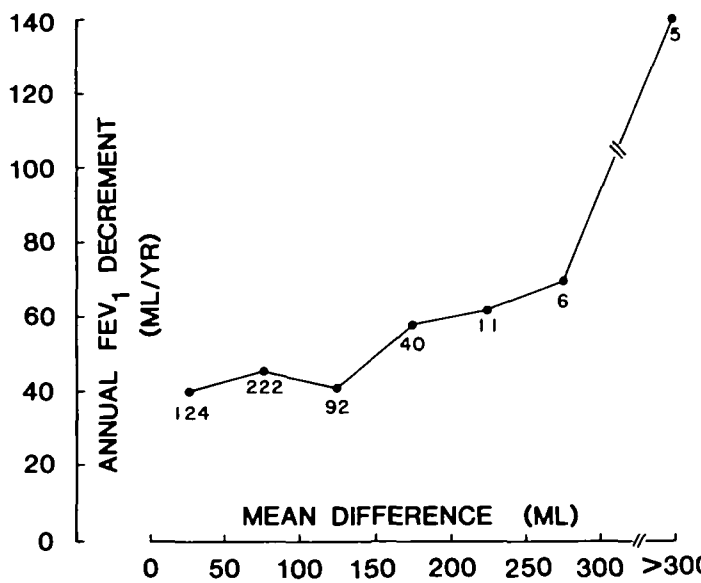

FIGURE 1. The average rates of FEV_1 loss are shown for individuals grouped by the following intervals of average FEV_1 difference: <50 ml, 50–100 ml, 100–150 ml, 150–200 ml, 200–250 ml, 250–300 ml, >300 ml. The number of subjects in each interval is indicated.

of less than 200 ml (44 ml/year). Therefore, the increased rate of FEV_1 decline observed among subjects with persistent test failure cannot be explained by a systematic pattern of test failures and successes causing an overestimation of their rates of FEV_1 declines.

Discussion

Whenever an individual performs an FEV_1 maneuver, there exists a hypothetic, "true" maximal value which could be attained at that particular time. How closely the largest FEV_1 measurement approaches that true value cannot be determined; that is, there is an unobservable measurement error inherent in any such test. However, there also exists for any particular test an observed variability associated with each forced expiratory effort, which we have indexed by the difference between the two largest FEV_1 measurements at each survey for each subject (FEV_1 difference). Underlying the use of a repeatability criterion is the assumption that the observed variation between blows reflects the unobservable error of the measurement; thus, a maximum allowable FEV_1 difference (e.g., 200 ml) can be used to identify, and therefore exclude, measurements which are most likely to underestimate substantially the true FEV_1. We have found, however, that a nonrepeatable test may reflect not only measurement error but also pathology in the tested subject. The findings reported here suggest that the exclusion of study subjects without repeatable pulmonary function tests may result in selection bias.

We have reviewed some of the epidemiologic studies of pulmonary function reported since the recommendations of the Snowbird Workshop. In cross-sectional studies, the proportion of subjects excluded from analysis because of the unsatisfactory repeatability of their tests was usually less than 10 per cent, and often less than 5 per cent (3, 5, 7). Evidence from our longitudinal study suggests that some of those subjects with a test failure would have been likely to fail again if retested and, moreover, that these individuals would on average lose FEV_1 at a faster rate than those who had repeatable tests. The effects of excluding a small percentage of subjects in a cross-sectional study of an environmental pulmonary hazard would depend on the prevalence of pulmonary disease in the study population. If the prevalence of disease were high in relationship to a particular exposure, and if the study population were sufficiently large, the exclusion of a small percentage of the group because of test failures would be unlikely to influence greatly the overall exposure-response relationship, even if all those excluded had lung disease. On the other hand, if the prevalence of disease were low (less than 5 per cent), the exclusion of even a few subjects with disease related to the environmental exposure might result in no significant association between exposure and disease being observed. This bias to the null may be particularly important for occupational settings in which pulmonary disease may be an idiosyncratic response, such as a specific allergy causing occupational asthma. In such a setting, self-selection of affected workers away from hazardous exposures frequently occurs. If we then further exclude affected individuals because of poor repeatability, the likelihood of finding a pulmonary effect may be reduced substantially.

In recently published longitudinal studies (4, 6, 8–10) (table 4), the proportions of subjects excluded for test failures were higher than in cross-sectional studies because inclusion was often made conditional upon having repeatable pulmonary function tests in each of several surveys. An extreme example is a recent study of air pollution (9), in which the authors excluded 41 per cent of their study population by requiring a "completely ac-

TABLE 4

Selected studies of pulmonary function reported after standardization report of the Snowbird Workshop

Study (reference no.)	No. of surveys	Repeatability rule: two largest efforts must be	% excluded by the rule
Bosse et al. (4)	2	Such that observer felt maximal effort had been made in both surveys	25*
Diem et al. (6)	9	Within 3% in at least three of the nine surveys	20
Jones et al. (8)	2	Within 3% in both surveys	10†
Van der Lende et al. (9)	4	Within 100 ml for FEV_1 *and* within 150 ml for forced vital capacity in all four surveys	41
Wegman et al. (10)	3	Within 5% or 100 ml in all surveys	8

* No differences were found between those excluded and the rest of the cohort with respect to age or percentage of current smokers.

† This 10% includes subjects excluded because of a poor start or failure to maintain effort, and indicates poor repeatability.

ceptable lung function reading all four times (e.g., maximum effort according to technician, shape curve satisfactory, difference among three VC readings no more than 150 ml, among three FEV_1 readings no more than 100 ml)." The impact of repeatability restrictions is therefore potentially greater in longitudinal studies.

The exclusion of such a large number of subjects is avoidable. In our study, for example, by requiring that repeatable FEV_1s be present in just two of the six surveys, only 3 per cent of the survivor cohort were excluded. Although one may choose to exclude nonrepeatable *measurements* from the estimation of the rates of decline in lung function, *subjects* should not be excluded from the study as long as they have two or more repeatable tests that permit their rate of decline to be estimated. Investigators should also consider the order of occurrence of test failures and successes within individuals and examine whether estimates of lung function decline could be systematically affected by the sequence of failures and successes.

Editor's note. The findings in the preceding paper by Eisen et al. have somewhat broader application than may be apparent at first glance. In addition to other measures of ventilatory function, they probably apply to periodic tests of any characteristic which require maximal physical or mental effort, such as grip strength; Master's two-step test and its modern successor, the treadmill; or tests of memory and calculating ability.

REFERENCES

1. American Thoracic Society. ATS statement—Snowbird workshop on standardization of spirometry. Am Rev Respir Dis 1979;119:831-8.
2. Ferris BG (Principal Investigator). Epidemiology standardization project. Am Rev Respir Dis 1978;118(part 2):55-88.
3. Banks DE, Morring KL, Boehlecke BA, et al. Silicosis in silica flour workers. Am Rev Respir Dis 1981;124:445-50.
4. Bosse R, Sparrow D, Rose CL, et al. Longitudinal effect of age and smoking cessation on pulmonary function. Am Rev Respir Dis 1981; 123:378-81.
5. Broder I, Mintz S, Hutcheon M, et al. Comparison of respiratory variables in grain elevator workers and civic outside workers of Thunder Bay, Canada. Am Rev Respir Dis 1979;119: 193-200.
6. Diem JE, Jones RN, Hendrick DJ, et al. Five year longitudinal study of workers employed in a new toluene diisocyanate manufacturing plant. Am Rev Respir Dis 1982;126:420-8.
7. Ferris BG, Speizer FE, Bishop YMN, et al. Spirometry for an epidemiologic study: deriving optimum summary statistics for each subject. Bull Eur Physiopathol Respir 1978;14:145-55.
8. Jones RN, Diem JE, Glindmeyer H, et al. Mill

effect and dose-response relationships in byssinosis. Br J Ind Med 1979;36:305-13.
9. Van der Lende R, Kok TJ, Reig RP, et al. Decrease in VC and FEV-1 with time indicators for effects of smoking and air pollution. Bull Eur Physiopathol Respir 1981;17:775-92.
10. Wegman DH, Musk W, Main DM, et al. Accelerated loss of FEV-1 in polyurethane production workers: a four year prospective study. Am J Ind Med 1982;3:209-15.
11. Gayrard P, Orehek J, Grimaud C, et al. Bronchoconstrictor effects of a deep inspiration in patients with asthma. Am Rev Respir Dis 1975;3:433-9.
12. Eisen EA, Smith TJ, Wegman DH, et al. Estimation of long-term dust exposures in the Vermont granite sheds. Am Ind Hyg Assoc J 1984;45:89-94.
13. Mosteller F, Tukey JW. Data analysis and regression. Reading, MA: Addison-Wesley, 1977.
14. Morrison DF. Multivariable statistical methods. New York: McGraw-Hill, 1967.
15. Knudson RJ, Slatkin RC, Lebowitz D, et al. Maximal expiratory flow-volume curve. Am Rev Respir Dis 1976;113:587-600.
16. Eisen EA, Wegman DH, Louis TA. Effects of selection in a prospective study of forced expiration in Vermont granite workers. Am Rev Respir Dis 1983;128:587-91.