Does Regression Analysis of Lung Function Data Obtained From Occupational Epidemiologic Studies Lead to Misleading Inferences Regarding the True Effect of Smoking?

Michael D. Attfield, PhD, and Thomas K. Hodous, MD

Exposure–response studies of the relationship between ventilatory function and dust exposure in workers are often quantified using linear regression methods. In coal miners, this technique has indicated that average effects of smoking and moderate dust exposure are roughly equivalent. However, the validity of direct comparison of the average effects of smoking and dust exposure has been questioned, the argument being that smoking causes severe effects in a minority, but leaves the remainder largely unaffected. This hypothesis was studied by examining distributions of FEV_1 in a group of working coal miners where mean effects associated with both smoking and dust exposure have been detected. Overall, the results suggest that comparison of average effects of smoking and dust exposure derived from linear regression analysis is valid and not misleading. © 1995 Wiley-Liss, Inc.*

Key words: lung function, coal mining, smoking, linear regression, dust exposure, health effects

INTRODUCTION

It is now almost universally accepted that smoking will lead to loss of pulmonary function [U.S. Department of Health and Human Services, 1984]. Moreover, there is ample evidence showing that loss of pulmonary function can lead to impairment, disability, and premature death [U.S. Department of Health and Human Services, 1984]. The potential for dust exposure to bring about disability and early death through causing reduced pulmonary function has long been a question of interest to occupational epidemiologists and others. To examine the relationship between dust exposure and lung function, researchers commonly use multiple regression procedures, with FEV_1 or change in FEV_1, as the response variable, and some measure of dust exposure as a predictor variable (while simultaneously controlling for age, height, smoking, and other confounders). Some recent uses of this approach, demonstrating its application to study of respiratory hazards, are gold miners [Hnizdo,

National Institute for Occupational Safety and Health, Morgantown, West Virginia 26505.
Address reprint requests to Michael D. Attfield, National Institute for Occupational Safety and Health, 1095 Willowdale Rd., Morgantown, WV 26505.
Thomas K. Hodous is now with Division of Safety Research, Morgantown.
Accepted for publication April 1, 1994.

1992], pulp and paper makers [Henneberger et al., 1989], polyurethane production workers [Wegman et al., 1982], and cotton workers [Kennedy et al., 1987].

It is among coal miners, however, where the regression approach has been applied most intensely. Studies have involved absolute values of ventilatory function against years of exposure [Hankinson et al., 1977], or against estimated dust exposure [Rogan et al., 1973; Soutar and Hurley, 1986; Attfield and Hodous, 1992; Seixas et al., 1992]. Longitudinal changes in FEV_1 in relation to dust exposure have also been examined [Love and Miller, 1982; Attfield, 1985]. These have all consistently detected effects of dust exposure on pulmonary function, both among smokers and never smokers, and without obvious synergism with smoking. More importantly, they indicate that at certain levels of dust exposure, the average effect of dust exposure can approach that of smoking. For example, in one recent study [Attfield and Hodous, 1992], an exposure of 4 mg/m^3 (a dust level commonly experienced by face workers prior to 1969 [Jacobson, 1971]) was associated with a decrement of close to 5 ml per year, similar to that associated with smoking. Hence, these findings provide prima facie evidence that dust exposure is harmful, a view which is held by some authorities [Seaton, 1983; Becklake, 1985].

Yet not all agree with this conclusion. Some consider that, in the absence of progressive massive fibrosis (PMF), coal miners suffer from little functional impairment. They basically subscribe to the view that dust exposure causes mild and reversible irritation amongst most miners, leading to what is termed "industrial bronchitis" [Morgan, 1978]. This they argue, should not be compared to smoking, with its well-known deleterious effects on health.

Morgan [1986] has explained this thesis in detail. He has postulated that the observed average effect of smoking derived from regression analysis actually reflects the combined effect of a severe decrement in ventilatory capacity in a minority of smokers together with no effect (or possibly a trivial effect) in the majority. As a result, he stated that it is invalid to compare the mean effects of the two insults.

If the hypothesis postulated by Morgan is correct, examination of the distribution of FEV_1 for all smokers should reveal evidence of two subgroups. One small subgroup would consist of those affected smokers who are on their way to developing severe ventilatory impairment due to their habit. The much larger remainder would have a distribution of ventilatory function close to normal. Figure 1a illustrates this hypothesis in the extreme case where distributions for both subgroups are clearly discernable. (It is also possible, however, that the affected group would not be so obvious, but when subsumed into the greater mass of the majority, would cause the whole distribution to appear skewed, with a heavier and longer tail to the left as in Fig. 1b.)

Verification or refutation of this hypothesis was sought using epidemiologic data in which average effects of both smoking and dust exposure on ventilatory function have already been demonstrated [Attfield and Hodous, 1992]. Distributions of FEV_1 were examined among different smoking and age groups, the intent being to detect subgroups of affected smokers, whose presence could have unduly influenced the overall mean value for smokers.

It must be emphasized at this point, that this study has been deliberately constrained to answer a very restricted question, i.e., whether coefficients obtained by application of regression methods to occupational cohorts provide misleading summaries of the effect of smoking among the members of those cohorts, and hence make

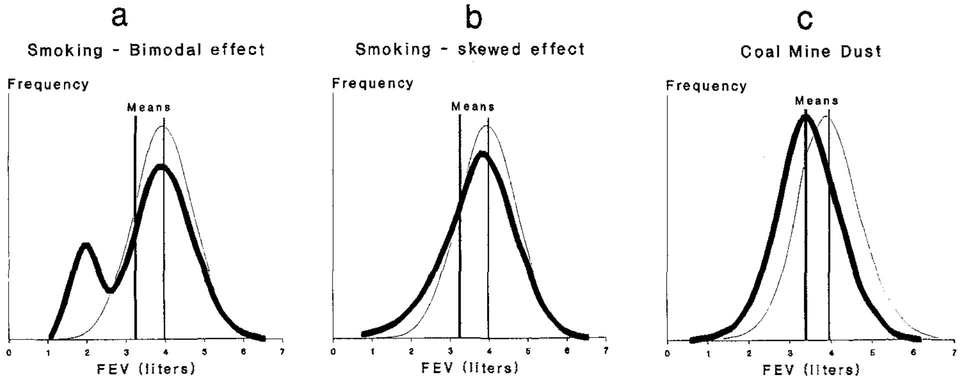


Fig. 1. Illustration of hypotheses investigated in this study. **a,b:** Possible manifestations of effect of smoking showing subgroups of affected individuals.

invalid any comparisons with observed dust exposure effects. The issue was definitely not the much bigger and more complex question of the natural history of disease in smokers and dust exposed individuals. Such a study would have required a very different approach, and involved questions of susceptibility, selection, and longitudinal changes.

MATERIALS AND METHODS

Medical Data

The medical data analyzed here were drawn from the first round of the National Study of Coal Workers' Pneumoconiosis (NSCWP). Information on ventilatory function, chest symptoms, age, height, working and smoking history, and related topics was collected on 9,078 working coal miners at mine site visits undertaken between 1969 and 1971. For further details on the methods used, see Morgan et al. [1974]. Although later data on coal miners exists in the study, these early data were preferred owing to the excellent participation achieved (91%), and to the fact that the miners, in general, had received much higher dust exposures prior to examination than have coal miners since that time (resulting from lower dust levels mandated by the 1969 Federal Coal Mine Health and Safety Act). A high participation minimizes selection effects, while a large range of dust exposures facilitates the detection of exposure–response relationships.

Data for 7,154 white males remained after exclusion of those aged <25 and >64 years, nonwhites, and those with missing information. According to self-reported smoking habits there were 1,963 never smokers, 1,348 ex-smokers, and 3,843 current smokers among the 7,154. Those classified as never smokers denied smoking more than five packs of cigarettes in their entire life.

Dust Exposure Estimates

Cumulative respirable dust exposure estimates from starting work until date of medical examination were generated for each miner [Attfield and Morring, 1992]. These were based on occupation-specific dust concentration estimates from a large industrial hygiene survey undertaken between 1968 and 1969 at 29 underground coal mines, together with dust sample data collected by coal mine operators after 1969 as mandated by the 1969 Federal Coal Mine Health and Safety Act. Cumulative dust exposures were generated from the products of dust concentration and years worked in jobs as reported by each miner. The exposure estimates were converted to units of gram-hr per cubic meter (gh/m^3) through use of a factor of 1.74 gh to one mg-year assuming a miner worked 1,740 hr per year on average.

Regression Analysis of Smoking and Dust Effects on FEV$_1$

Linear regression analysis was used to relate FEV$_1$ to smoking and dust exposure, allowing for age and height. Smoking was examined using a categorical variable (never smokers, ex-smokers, current smokers) together with a term for pack-years. Estimated cumulative dust exposure was used to measure occupational exposure. See Attfield and Hodous [1992] for further details of this analysis.

Analysis of Nature of Smoking Effect

The hypothesis that a subgroup of the data had an undue effect on the findings concerning smoking was first studied by examination of influential observations using the methods described by Belsley et al. [1980]. Specifically, the influence diagnostic DFBETA was employed to identify individuals whose data caused the observed coefficient for pack-years to change markedly from that obtained through omission of their data from the analysis. DFBETA is calculated by taking the difference between the two estimates of the coefficient (with and without the specific observation) and dividing by the standard error of the coefficient. Belsley et al. [1980] recommend a cut-off point of $2/\sqrt{n}$, where n is the number of observations.

Another influence statistic examined was the studentized residual, RSTUDENT. This statistic is useful for identifying observations that unduly affect the regression findings in general, and therefore could help, in this case, to isolate the hypothesized influential subgroup. It is obtained by first taking the difference between the residual (observed − predicted FEV$_1$, in this case) derived from the model estimated from all of the data and that calculated from the data omitting the observation in question, and then dividing by the standard error of the residual. A cut-off limit of 2 is recommended for this statistic.

The methods advocated by Belsley et al. [1980] probably work best for small datasets and when there are few major outliers. Conversely, they may fail to detect problems when applied to very large datasets with substantial numbers of more moderate outliers. For this reason, a further, graphical, analysis was undertaken of the data. In this, distributions of FEV$_1$ were examined in eight subsets defined on the basis of four age ranges (25–34, 35–44, 45–54, 55–64) and two smoking status groups (never smokers, current smokers). Within each subgroup, deviations of the raw FEV$_1$ values from the mean (FEV$_1$ − mean) were calculated, grouped into 0.25 l intervals, charted, and the resulting distributions compared across smoking groups within each age range both visually and using a chi-square test of homogeneity. A similar analysis was undertaken using medians rather than means, using residuals from model fitting on age and height within each group, and using percent predicted FEV$_1$ values obtained using the prediction equation of Knudson et al. [1976]. In the latter, the values were centered on zero by subtracting the mean percent predicted

TABLE I. Mean FEV_1, Mean FEV_1 Percent Predicted, Age, Height, and Numbers of Observations by Smoking Group and Age Group*

	Smoking	Age groups				
Variable	status	25–34	35–44	45–54	55–64	All
Mean FEV_1 (1)	NS	4.25	3.92	3.59	3.21	3.69
	ES	4.19	3.83	3.42	3.02	3.47
	CS	4.10	3.60	3.20	2.80	3.43
Mean FEV_1 (% predicted)	NS	100.9	100.9	100.1	98.4	100.0
	ES	98.6	98.2	95.0	92.7	95.3
	CS	96.9	92.5	89.6	85.2	91.2
Mean age (year)	NS	29	40	50	58	46
	ES	30	41	50	59	48
	CS	29	40	50	58	44
Mean height (in)	NS	70	69	69	68	69
	ES	70	70	69	69	69
	CS	70	69	69	69	69
Mean cumulative exposure (gh/m^3)	NS	25	84	142	183	117
	ES	27	89	142	181	130
	CS	24	86	139	184	108
Number	NS	277	267	466	338	1963
	ES	218	377	793	575	1348
	CS	868	927	1368	680	3843

*NS = never smokers; ES = ex-smokers; CS = current smokers.

value for each smoking–age combination, then grouped into ranges of 5%, charted, and compared as described above.

RESULTS

Basic Statistics

Table I gives the number of observations for each smoking and age group, as well as the mean FEV_1 values, the mean percent predicted FEV_1, and the mean age, height, and cumulative dust exposure. Median FEV_1 values (not shown) were almost identical to the means. The well-known age-related and smoking-related declines in FEV_1 are immediately apparent.

Regression Estimates of the Smoking Effect

Linear regression analysis of FEV_1 allowing for age, height, and geographical region revealed clear effects ($p < .0001$) of smoking and dust exposure. The regression coefficients for the latter two factors are shown in Table II, and reveal a 5 ml loss in FEV_1 for each pack-year, and 0.7 ml per gh/m^3 (or equivalently, a 1.2 ml loss for each mg-year/m^3). (An identical effect of pack-years was found when current smokers were analyzed separately). These regression coefficients represent the estimated average effects associated with each variable in its relationship with FEV_1 in the studied individual.

Analysis of Influential Observations

Using the suggested criterion of $2/\sqrt{n}$ for DFBETA led to detection of 169 observations that were influencing the pack-years coefficient towards a greater effect

TABLE II. Regression Coefficients and *t* Statistics Abstracted From Full Regression Model of FEV_1 on Age, Height, Region (Coefficients not Shown), Smoking, and Cumulative Dust Exposure

Variable	Coefficient	*t*
Constant (ml)	−1,702	
Smoking status relative to never smokers		
Current smokers	−208	−10.0[a]
Ex-smokers	−36	−1.6
Pack-years	−4.7	−9.9[a]
Estimated cumulative dust exposure (gh/m^3)	−0.69	−5.5[a]
R^2	0.47	
Residual d.f.	7,126	

[a] $p < .0001$.

on FEV_1 (consistent with the stated hypothesis). Individually, however, the effect of each on the coefficients was trivial, being about 4% of the standard error on average (i.e., about 0.2 of −4.7). These 169 were more than counterbalanced by 205 observations that, using the same criterion, were deemed to be influential in the opposite direction. Curiously, both sets of individuals were older than the remainder of the cohort (52 and 53 years, compared to 45), had greater dust exposure, and greater pack-years. To assess the overall effect of these groups on the findings, the regression model on FEV_1 was refitted after their deletion from the dataset. The results showed an increase in the magnitude of the smoking effect, whereas a decrease was expected on the basis of the hypothesis.

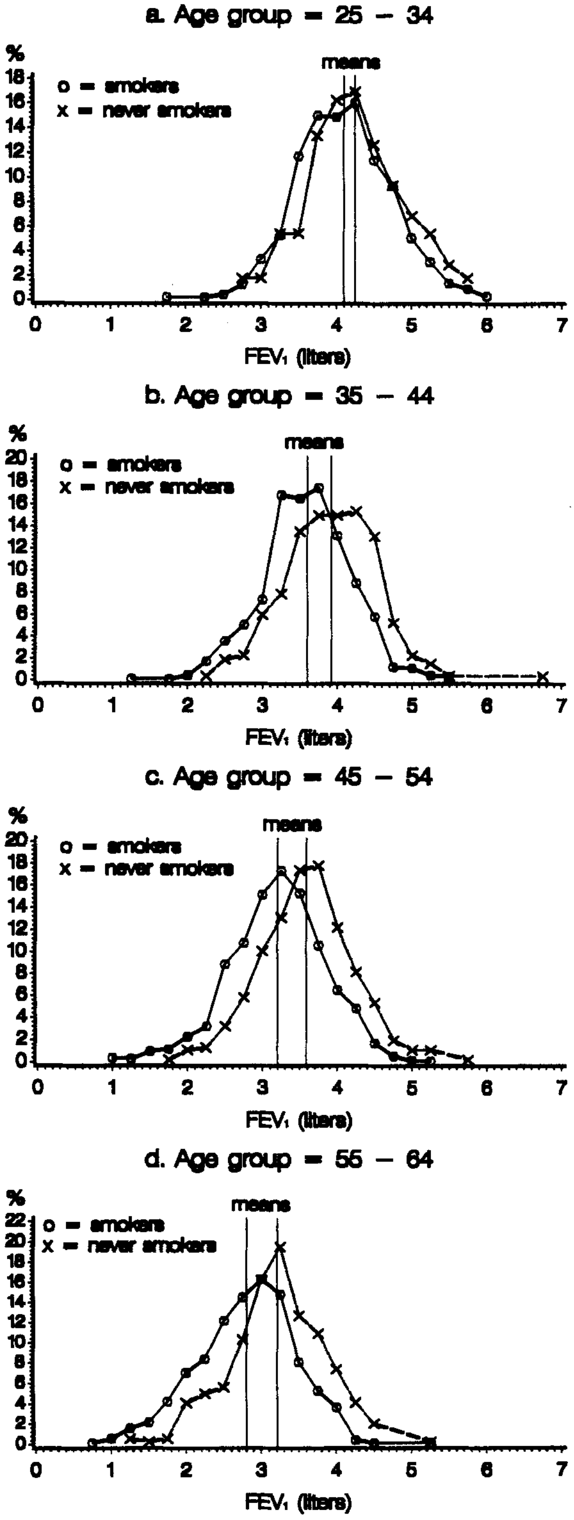
An alternative approach, using the RSTUDENT statistics with the suggested criterion of 2, gave rise to 355 possibly influential observations. Of these, 105 were consistent with lower than predicted FEV_1 observations. The model fitted to the data without the influential cases was little different to that for all of the data, the coefficient for pack-years being −4.2 compared to −4.7 for the full set, while the corresponding estimate for dust exposure were −0.62 compared to −0.69.

Graphical Analysis

Overall, the analysis of influential observations has failed to identify any obvious subgroup of individuals whose results had unduly affected the estimation of the effect of smoking. Hence, no evidence has been found from this investigation to support the postulated hypothesis. However, it is likely that this method is not very suitable for identification of subgroups of less obvious outlying points, particularly in large datasets. For this reason, a graphical analysis was undertaken.

When the distributions of FEV_1 for all four age groupings of smoking and never smoking miners are viewed as a whole (Fig. 2), it is clear that the main feature of these data is an age-related progressive deterioration apparently affecting the FEV_1 of all smokers, and that any effect due to a minority of severely affected individuals seems to be very minor and secondary.

In order to investigate further, distributions of FEV_1 deviations for smokers and never smokers were compared by age group. Figure 3 shows distributions of FEV_1 − mean FEV_1 for current smokers superimposed on those for never smokers. No obvious sign of skewness in the left-hand tail of the smokers distribution for any age group is seen. There is a suggestion of an effect in the oldest age group, but a



a. Age group = 25 – 34

b. Age group = 35 – 44

c. Age group = 45 – 54

d. Age group = 55 – 64

Fig. 2. Frequency distribution (%) of observed FEV_1 values by age group and smoking status.

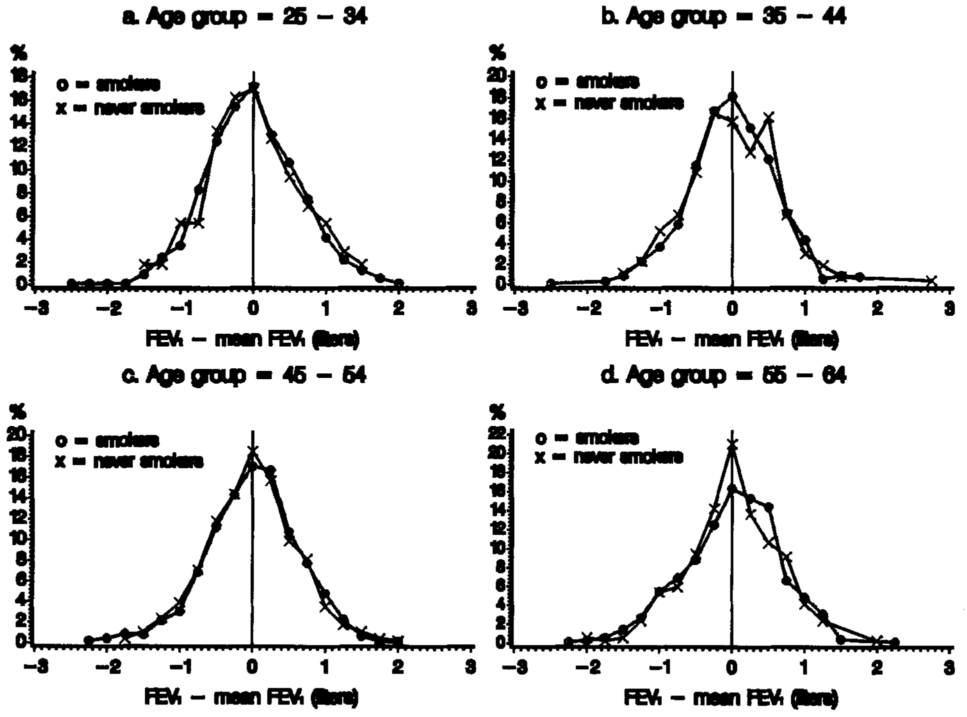


Fig. 3. Frequency distribution (%) of deviations of FEV_1 from mean by age group. Current and never smokers compared.

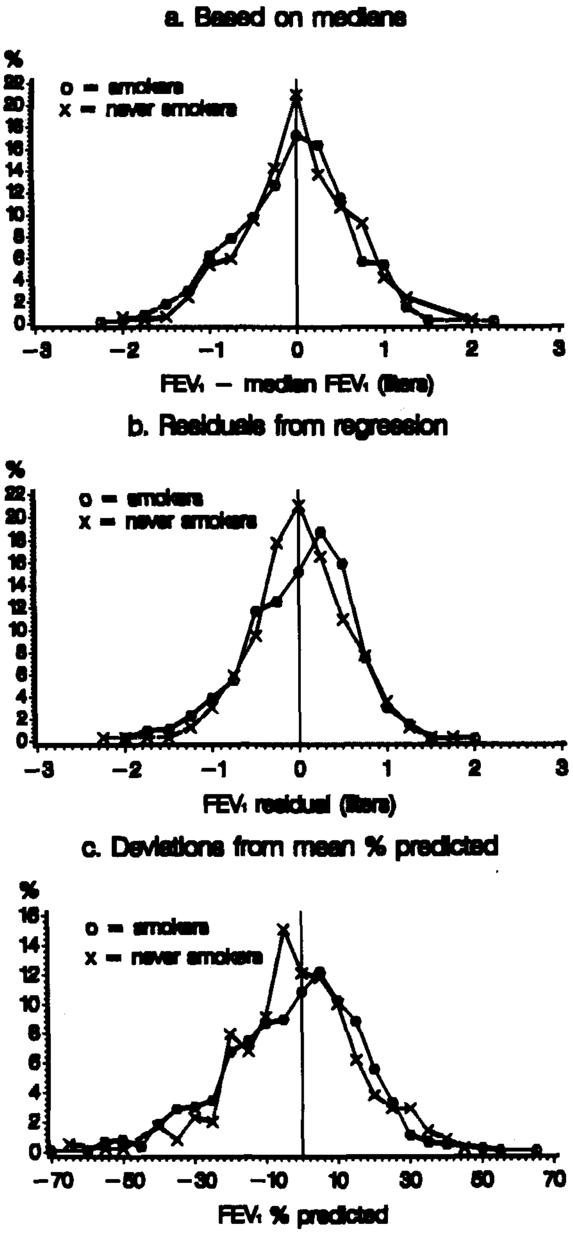
chi-square test undertaken to compare the distributions failed to reveal a significant difference.

The same picture emerged when the analysis was repeated using medians, residuals from linear regression, and percent predicted FEV_1 values. Only for the oldest age group was there again a hint of a subgroup of affected smokers (Figure 4).

DISCUSSION

The analysis of influential observations revealed no obvious trace of subsets of affected individuals that could have affected the model fitting in such a way as to lead to misleading interpretation of the effect of smoking. Similarly, the graphical analysis showed that age-specific distributions of FEV_1 for current smokers, although lower on average, were very similar in shape to those for never smokers. Even in the oldest age group, where there was a suggestion of skewness, the difference between the distributions for the smokers and never smokers was not statistically significant.

Despite the lack of statistical significance, does the slightly heavier left-hand tail in the distribution for the smokers in the oldest age group explain the overall decrement between the means for the smokers and never smokers in that age group? To explore this question, the mean FEV_1 for the smokers in this group was recalculated after removal of observations in the left-hand tail to make the percentage frequencies equal those for never smokers. It led to an increase in the mean for



a. Based on medians

b. Residuals from regression

c. Deviations from mean % predicted

Fig. 4. Distribution (%) of deviations of FEV_1 from median, of residuals, and of percent predicted FEV_1 from mean. 55–64-year-old current and never smokers compared.

smokers of 0.06 l, far short of the 0.34 l change needed to bring the mean in line with that for the never smokers.

In fact, a quick calculation demonstrates that a minority of severely affected smokers could not have given rise to the actual observed mean difference in FEV_1 between the smokers and never smokers observed in the oldest age group studied. If the mean for the smokers is assumed to depend on the combination of the means of two subgroups, one of which is severely affected and the other having the same mean as the never smokers, the expression for their overall mean is given by

$$\bar{x}_{all} = \bar{x}_{sev} \cdot p + \bar{x}_{nor} \cdot (1 - p)$$

where $\bar{x}_{all}$ is the mean for all smokers, $\bar{x}_{sev}$ is the mean for the minority of those severely affected, $\bar{x}_{nor}$ is the mean of those unaffected, and p is the proportion of all smokers affected by their habit. Now, if we assume, as did Morgan [1986], that $p = .13$, and also that the mean FEV_1 values for the smokers and never smokers are as reported (i.e., $\bar{x}_{all} = 2.80$ and $\bar{x}_{nor} = 3.21$ l, respectively), then

$$\bar{x}_{sev} = \frac{2.80 - 3.21 \times 0.87}{.13} = 0.06 \text{ l}.$$

In other words, the average FEV_1 for the severely affected smokers would have to be 0.06 l, a level too low to support life.

It is important not to misunderstand the purpose of this study. It is not about smoking effects and their manifestation in the general population. Nor do its results contradict those from general population studies, e.g., Burrows et al. [1977], which show progressive distributional changes in FEV_1 with amount smoked. Active miners must be fit enough to go underground and work in difficult conditions. Hence, those who develop disability (from whatever cause) will remove themselves, or be removed, from the active work force, and therefore, from a study group of active workers. This censoring of affected individuals may well lead to lack of skewness in the left tail of the FEV_1 distributions, as observed in the present study. However, those less affected will remain in the workforce, and it is the average effect of smoking in that group that the regression coefficient on smoking is measuring.

Nor is the study about the nature of the dust exposure effect. Instead, it is simply an examination of a postulated hypothesis: i.e., that regression analysis fails to provide a valid summary of the smoking effect in a working cohort. Taking all of the findings presented here into account, it is clear that the evidence does not support this hypothesis. Instead, the average effect does appear to largely and validly reflect a general divergence in mean FEV_1 of the smokers from the never smokers with age in the studied cohort. In conclusion, these results suggest that comparison of smoking and dust exposure coefficients estimated through application of regression methods to lung function data from working cohorts is valid and not misleading with respect to the effects actually experienced by those workers.

REFERENCES

Attfield MD (1985): Longitudinal decline in FEV_1 in United States coalminers. Thorax 40:132–137.

Attfield MD, Hodus TK (1992): Pulmonary function of U.S. coal miners related to dust exposure estimates. Am Rev Respir Dis 14:605–609.

Attfield MD, Morring K (1992): The derivation of estimated dust exposures for U.S. coal miners working before 1970. Am Ind Hyg Assoc J 53:248–255.

Becklake MR (1985): Chronic airflow limitation: Its relationship to work in dusty occupations. Chest 88:608–617.

Belsley DA, Kuh E, Welsch RE (1980): "Regression Diagnostics." New York: John Wiley & Sons, pp 11–30.

Burrows B, Knudson RJ, Cline MG, Lebowitz MD (1977): Quantitative relationships between cigarette smoking and ventilatory function. Am Rev Respir Dis 115:195–205.

Hankinson JL, Reger RB, Fairman RP, Lapp NL, Morgan WKC (1977): Factors influencing expiratory flow rates in coal miners. In Walton WH (ed): "Inhaled Particles IV." Oxford: Pergamon Press, pp 737–755.

Henneberger PK, Eisen EA, Ferris BG Jr (1989): Pulmonary function among pulp and paper workers in Berlin, New Hampshire. Br J Ind Med 46:765–772.

Hnizdo E (1992): Loss of lung function associated with exposure to silica dust and with smoking and its relation to disability and mortality in South African gold miners. Br J Ind Med 49:472–479.

Jacobson M (1971): Respirable dust in bituminous coal mines in the U.S. In Walton WH (ed): "Inhaled Particles III." Old Woking: England Unwin Brothers, pp 903–917.

Kennedy SM, Christiani DC, Eisen EA, Wegman DH, Greaves IA, Olenchock SA, Ye T, Lu P (1987): Cotton dust and endotoxin exposure-response relationship in cotton textile workers. Am Rev Respir Dis 135:194–200.

Knudson RJ, Slatin RC, Lebowitz MD, Burrows B (1976): The maximal expiratory flow-volume curve. Normal standards, variability, and effects of age. Am Rev Respir Dis 115:587–600.

Love RG, Miller BG (1982): Longitudinal study of lung function in coal miners. Thorax 37:193–197.

Morgan WKC (1978): Industrial bronchitis. Br J Ind Med 35:285–291.

Morgan WKC (1986): On dust, disability, and death. Am Rev Respir Dis 134:639–641.

Morgan WKC, Handelsman L, Kibelstis JS, Lapp NL, Reger R (1974): Ventilatory capacity and lung volumes of US coal miners. Arch Environ Health 28:182–189.

Rogan JM, Attfield MD, Jacobsen M, Rae S, Walker DD, Walton WH (1973): Role of dust in the working environment in development of chronic bronchitis in British coal miners. Br J Ind Med 30:217–226.

Seaton A (1983): Coal and the lung. Thorax 38:241–243.

Seixas NS, Robins TG, Attfield MD, Moulton LH (1992): Exposure-response relationships for coal mine dust and obstructive lung disease following enactment of the Federal Coal Mine Health and Safety Act of 1969. Am J Ind Med 21:715–734.

Soutar CA, Hurley JF (1986): Relation between dust exposure and lung function in miners and ex-miners. Br J Ind Med 43:307–320.

U.S. Department of Health and Human Services (1984): "The Health Consequences of Smoking: Chronic Obstructive Lung Disease. A Report of the Surgeon General." Washington, DC: U.S. Printing Office, pp 92.

Wegman DH, Musk AW, Main DM, Pagnotto LD (1982): Accelerated loss of FEV-1 in polyurethane production workers: A four-year study. Am J Ind Med 3:209–215.