Thorax (1972), **27**, 547.

Dependence of the incidence of emphysema on smoking history, age, and sex

J. A. ANDERSON, M. S. DUNNILL, and R. C. RYDER

Department of Biomathematics and the Radcliffe Infirmary, Oxford, and St. Tydfil's Hospital, Merthyr Tydfil

The relationship between the incidence of emphysema and smoking history, age, sex, and percentage bronchial mucous gland volume is investigated using the series of necropsy cases reported by Ryder, Dunnill, and Anderson (1971). Using those cases originating from hospital, it is shown that the incidence of emphysema is dependent on smoking history, age, and sex. Percentage bronchial mucous gland volume is not related to emphysema when the three other variables have been allowed for. Thus sex has an effect on the incidence of emphysema over and above the difference between the sexes in the occurrence of smoking. It is also shown that it is reasonable to extrapolate these conclusions from hospital necropsy cases to the general population.

In a recent study, Ryder, Dunnill, and Anderson (1971) investigated the dependence of the incidence and severity of emphysema on age, sex, smoking habit, and percentage bronchial mucous gland volume in a series of cases at necropsy. From one and two factor analyses of the data it was concluded that age, sex, and smoking were certainly significant but the position of the percentage bronchial mucous gland volume was less certain. Since the true picture was being obscured by associations between the factors, it seemed worth while to try to clarify the situation by a multifactorial analysis. Another point worthy of investigation is the extent to which conclusions drawn from these necropsy cases can be extrapolated to the population in general. These two matters are fully examined in the present paper.

MATERIAL AND METHODS

As described in the paper by Ryder *et al.* (1971), the study is based on 352 consecutive necropsy cases excluding miners and children. Each case was classified according to its origin as:

(*a*) source I, all hospital cases excluding those referred to the coroner;

(*b*) source II, non-hospital cases of sudden death due to an unknown cause referred to the coroner for necropsy by general practitioners;

(*c*) source III, all cases of sudden traumatic death and suicide, whether hospital cases or not, referred to the coroner.

There were 226, 102, and 24 cases from sources I, II, and III respectively.

The age and sex of each case were recorded together with the smoking history (smoker or non-smoker) where available. One lung was obtained from each subject and, using the method of point counting previously described (Dunnill, 1962; Dunnill, Massarella, and Anderson, 1969; Ryder *et al.*, 1971), the percentage volume of emphysematous lung parenchyma and of normal lung parenchyma were measured, as was the percentage volume of mucous glands in the bronchial wall. The distribution of the percentage volume of emphysema is markedly skewed (Ryder *et al.*, 1971), so it was decided to classify subjects as being emphysematous only if more than 2% of the volume of their lung parenchyma was diseased. This 'definition' of emphysema differs slightly from that used by Ryder *et al.* (1971).

STATISTICAL METHOD

The relationship between the incidence of emphysema and the other observations in the study was not assessed by classical regression methods since they are best suited to situations where the dependent variable is measured on a continuous scale. Incidence is a presence or absence observation, so it was decided to use the method of logistic regression which has been advocated for such binary dependent variables by many authors, including Cox (1970). Suppose the probability of emphysema, P, given the factor $x_1, x_2, \ldots, x_p$ under consideration is given by

$$P = eS/(1+eS), \quad (1)$$

where S is a score that depends linearly on the factors. Thus

$$S = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p \quad (2)$$

Thorax: first published as 10.1136/thx.27.5.547 on 1 September 1972. Downloaded from http://thorax.bmj.com/ on April 9, 2024 by guest. Protected by copyright.

where $\alpha_0, \alpha_1, \ldots, \alpha_p$ are coefficients that have to be estimated from the data. If S is positive, the individual is more likely than not to have emphysema, the probability increasing with the score. Similarly, if S is negative, emphysema is less likely than not and for $S=0$ either possibility is equally likely.

To take a specific example, for source I cases, taking age, sex, and smoking history as dependent variables, the score was estimated to be

$$S = -6{\cdot}70 + 0{\cdot}066x_1 + 1{\cdot}26x_2 + 2{\cdot}29x_3 \quad (3)$$

where x_1 is age, $x_2=1,0$ for males and females, respectively, and $x_3=1,0$ for smokers and non-smokers, respectively. This implies, for instance, that the risk of emphysema in a female is less than that in a male for any age and smoking habit group. This point is returned to in the discussion. Logistic regression is further discussed in the Appendix.

The dependence of the incidence of emphysema on age, sex, smoking history, and percentage bronchial mucous gland volume was investigated in this way using only those 173 cases from source I where the data on these variables was complete. Sources II and III were not included because there were so few cases (5) with recorded smoking histories.

RESULTS

As described in the Appendix, the logistic regression model (1) was fitted to the data from source I by maximum likelihood, taking an approach analogous to that of ordinary regression. Thus all possible selections, three at a time and four at a time, were taken from age, sex, smoking, and percentage bronchial mucous gland volume and the model (1) was estimated with each set as explanatory variables. Each of the corresponding maximized log-likelihoods is shown in Table I together with the appropriate test of significance. It was concluded that smoking history, age, and sex contributed significantly to the prediction of emphysema, even after all the other factors had been taken into account. However, percentage bronchial mucous gland volume did not satisfy this criterion, so it was decided to exclude it from further consideration. Then, the estimated form of the logistic regression S is as given in (3).

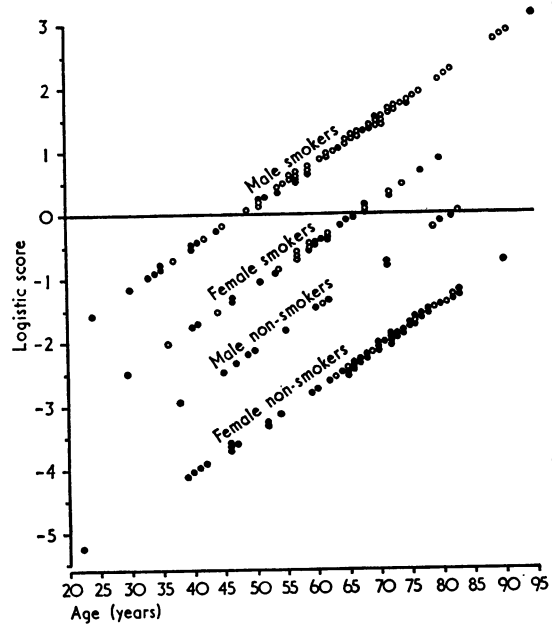

FIGURE. *The logistic score based on age, sex, and smoking history from source I cases only, plotted against age:* ○ = *emphysema*, ● = *no emphysema.*

A visual impression of the fit of the model to the data is given in the Figure by plotting for each subject the score (3) against age, distinguishing between smokers and non-smokers and males and females. On the whole, the fit is satisfactory, cases with and without emphysema tending to have positive and negative scores, respectively. However, the fit appeared to be better for males than

TABLE I

ASSESSMENT OF EFFECT OF FACTORS ON INCIDENCE OF EMPHYSEMA IN HOSPITAL SUBJECTS WITH SMOKING HISTORY RECORDED

Factors Included in Logistic Model	Minus Maximized Log Likelihood	Factor under Assessment	Twice Change in Maximized Log Likelihood	Degrees of Freedom	Significance Level
None	119·91				
BMG, age, sex, smoking	80·30 (1)		Decrease from (1) if factor omitted		
Age, sex, smoking	81·94	BMG	3·2	1	NS
BMG, sex, smoking	91·1	Age	21·6	1	0·1%
BMG, age, smoking	84·7	Sex	8·8	1	0·5%
BMG, age, sex	92·1	Smoking	23·6	1	0·1%
			Increase from (1) if factor added		
Smoking, age, sex + age×sex	80·6	Age×sex	2·6	1	NS
Smoking, age, sex + sex×smoking	81·76	Sex×smoking	0·4	1	NS
Smoking, age, sex + age×smoking	81·86	Age×smoking	0·2	1	NS

BMG = percentage bronchial mucous gland

Thorax: first published as 10.1136/thx.27.5.547 on 1 September 1972. Downloaded from http://thorax.bmj.com/ on April 9, 2024 by guest. Protected by copyright.

females, suggesting that some of the coefficients in (3) ought to have different values for males and females. This was tested by estimating logistic models (1) using the 'interaction' technique to give differing coefficients, as indicated in the Appendix. Thus the effects of (i) different age coefficients for males and females, (ii) different smoking coefficients for males and females, and (iii) different age coefficients for smokers and non-smokers were investigated and are reported in Table I as the age × sex, sex × smoking, and age × smoking interactions, respectively. None of these effects was statistically significant, although the age × sex interaction gave the greatest increase in maximized log-likelihood.

It is clear that the subjects from source I, being hospital cases, are not a random sample from the population at large so some justification is required before the above conclusions can be given a greater applicability. The individuals from source III had suffered a sudden traumatic death and thus might be thought to be reasonably representative of the population. Source II subjects, not being from hospital, might also be thought to be more typical than those from source I. Unfortunately, there were only 24 cases from source III (Table II) and only 5 cases with recorded smoking histories from sources II and III combined. Thus the full dependence structure of emphysema, including smoking histories, cannot be investigated from sources II and III alone. All that can be done is to extrapolate from the conclusions on source I subjects to the population at large. The validity of this step cannot be probed directly but a close examination of the figures in Table II, which give the distribution of emphysema by age and sex for the three sources separately, suggests that it is not unreasonable. Fisher's exact test for 2 × 3 contingency tables shows that there are no significant differences between the sources for either sex in any of the age groups in Table II. This result also holds good when some of the age groups are amalgamated to give larger observed numbers. The smallest significant level obtained from any of these tests was 10% but it should be noted that, ignoring age, there is evidence of crude differences between the sources for males (χ^2 statistic is 9·37; with 2 degrees of freedom, this is significant at the 1% level). Thus, having allowed for age and sex, there is no evidence of differences between the sources in the incidence of emphysema. This suggests that the dependence of emphysema on age, sex, and smoking history is the same in patients from the three sources.

TABLE II

INCIDENCE OF EMPHYSEMA BY SOURCE, SEX, AND AGE

Age Group	Source 1		Source 2		Source 3		Total	
	E	N	E	N	E	N	E	N
				Females				
21–30	0	4	0	0	1	2	1	6
31–40	2	7	0	4	0	2	2	13
41–50	1	11	1	5	0	2	2	18
51–60	7	23	1	4	1	1	9	28
61–70	6	35	4	12	0	1	10	48
71–80	11	36	4	12	1	1	16	49
81–	2	8	2	6	0	1	4	15
Total	29	124	12	43	3	10	44	177
				Males				
21–30	0	3	0	0	0	5	0	8
31–40	3	11	2	3	0	1	5	15
41–50	6	13	2	5	1	1	9	19
51–60	11	23	8	10	2	2	21	35
61–70	16	22	17	19	3	4	36	45
71–80	16	19	16	18	0	0	32	37
81–	9	11	4	4	1	1	14	16
Total	61	102	49	59	7	14	117	175

The columns headed E and N give the numbers of subjects with emphysema and the category total, respectively.

An alternative approach to the above problem is to use the logistic regression method. Choosing source, age, and sex as explanatory variables, the question is which of them contributes significantly to the prediction of the incidence of emphysema. As before, the logistic regression model (1) was fitted to the data by maximum likelihood with different subsets of the dependent variables. Because of the different pattern of age changes in males and females observed in Table II, different age coefficients for the two sexes were allowed by introducing an age × sex interaction into the model. The results of these analyses are shown in Table III where it can be seen that the effect of source was not significant, whether or not different age coefficients for the sexes were allowed. The effects of age and sex were highly significant (0·1% level) and there was evidence that different age coefficients for males and females were required (the age × sex interaction term was significant at the 2·5% level whether or not source was allowed for).

It is concluded that there is no counter indication to extrapolation of the conclusions from source I to the population at large.

DISCUSSION

The first analysis showed that in source I cases, only smoking history, age, and sex were significantly related to the incidence of emphysema. This suggested that the significance of the percentage bronchial mucous gland volume and the age-sex interaction found in the first analysis was simply a product of the omission of smoking history. Thus although percentage bronchial

Thorax: first published as 10.1136/thx.27.5.547 on 1 September 1972. Downloaded from http://thorax.bmj.com/ on April 9, 2024 by guest. Protected by copyright.

TABLE III

ASSESSMENT OF FACTORS AFFECTING INCIDENCE OF EMPHYSEMA IN PATIENTS FROM ALL SOURCES

Factors Included in Logistic Model	Minus Maximized Log Likelihood	Factor under Assessment	Twice Change in Maximized Log Likelihood	Degrees of Freedom	Significance Level
None	242·70				
Source, age, sex	191·92	Source, allowing for age × sex	5·04	2	NS
Age, sex	194·59	Source, not allowing for age × sex	5·34	2	NS
Source, age	227·38	Sex	70·92	1	0·1%
Source, sex	206·35	Age	28·86	1	0·1%
Source, age, sex, age × sex	187·71	Age × sex, allowing for source	8·42	1	0·5%
Age, sex, age × sex	190·23	Age × sex, not allowing for source	8·72	1	0·5%

mucous gland volume and the incidence of emphysema appear to be related when they are considered in isolation (Ryder *et al.*, 1971), this is no longer true when the effects of age, sex, and smoking history are allowed for first. In addition, Ryder *et al.* (1971) found that the percentage bronchial mucous gland volume was associated with smoking history but not with age and sex. This suggests that there is not a causal connection between the percentage bronchial mucous gland volume and the incidence of emphysema, but that both are increased by the smoking habit.

Ryder *et al.* (1971) found there was no significant difference between male and female smokers nor between male and female non-smokers in the incidence of emphysema. This would lead one to expect that the effect of sex would not be significant having accounted for smoking habit. However, the more sensitive analysis here indicates that sex has an effect on the incidence of emphysema over and above the difference between the sexes in smoking history. This could be due to differences in smoking technique, amount smoked or length of smoking history between the sexes.

Although there are clear differences between the age trends in the incidence of emphysema of males and females (Table II), the age × sex interaction was not significant, having allowed for smoking history. This suggests that this apparent difference between the sexes is a by-product of the smoking habit and, in fact, smoking is much more common in males of the decades 60 and over than in females.

Although strictly these conclusions apply only to a hospital necropsy population (source I), the second analysis showed that there was no evidence of differences in the incidence of emphysema between the sources in the various age and sex groups. This was confirmed by the logistic regression analysis which showed that having allowed for age, sex, and an age × sex interaction, the effect of source did not contribute significantly to the incidence of emphysema. This was a somewhat conservative test since, as explained, it was not possible to take account of the smoking history. Thus there are few crude differences between the sources and even these disappear when age and sex are allowed for. It is a reasonable conjecture, then, that these differences were caused by the varying age and sex distributions from source to source and, hence, that the dependence pattern of the incidence of emphysema is the same in subjects from all three sources.

Since the effect of smoking is believed to be cumulative, it is perhaps surprising that there is no evidence of a significant interaction between age and smoking history in the incidence of emphysema. However, it must be remembered that there were few subjects in the study with a recorded history of smoking younger than 40 years. Ryder *et al.* (1971) also noted that the incidence of emphysema remained fairly constant between 50 and 90 years. It thus seems likely that a good sample of cases from the younger age groups would be required to make a more definite pronouncement on this subject.

REFERENCES

Cox, D. R. (1970). *The Analysis of Binary Data*, p. 61. Methuen, London.

Dunnill, M. S. (1962). Quantitative methods in the study of pulmonary pathology. *Thorax*, **17**, 320.

—— Massarella, G. R., and Anderson, J. A. (1969). A comparison of the quantitative anatomy of the bronchi in normal subjects, in status asthmaticus, in chronic bronchitis, and in emphysema. *Thorax*, **24**, 176.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Ann. hum. Genet.*, **20**, 309.

Ryder, R. C., Dunnill, M. S., and Anderson, J. A. (1971). A quantitative study of bronchial mucous gland volume, emphysema and smoking in a necropsy population. *J. Path.*, **104**, 59.

Thorax: first published as 10.1136/thx.27.5.547 on 1 September 1972. Downloaded from http://thorax.bmj.com/ on April 9, 2024 by guest. Protected by copyright.

APPENDIX

The logistic form for the probability of emphysema (E) given the factors $x_1, x_2, \ldots, x_p$ in equation (1) is

$$\text{pr}(E/x) = e^S/(1+e^S) \quad (1)$$

where

$$S = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p \quad (2)$$

and $\alpha_0, \alpha_1, \ldots, \alpha_p$ (α) are constants that are estimated from the data and $x_1, x_2, \ldots, x_p$ (x) is the set of observations made on an individual. The factors included in the score S may be varied at will leading to a regression type of situation (Cox, 1970). For each choice of the factors, α was estimated by maximum likelihood. Other methods have been suggested for this problem (Haldane, 1955) but in this context, maximum likelihood seems most appropriate. Just as in ordinary regression, the model (2) for the score can be modified to allow for 'interactions'. For example, suppose that the effect of age and sex on the incidence of emphysema is being investigated and, suppose further, that it is suspected that the form of the logistic score, S, is different for males and females. The simplest approach would be to fit two models:

$$\left.\begin{aligned} \text{for males: } S_M &= \alpha_0 + \alpha_1 x \\ \text{for females: } S_F &= \beta_0 + \beta_1 x \end{aligned}\right\} \quad (3)$$

where x denotes age. However, an entirely equivalent approach is to fit one model to both males and females. Thus

$$S = \gamma_0 + \gamma_1 x + \gamma_2 y + \gamma_3 xy,$$

where x as before is age and $y = 0,1$ for females and males, respectively. The term $\gamma_3 xy$ is an interaction term. This gives:

$$\left.\begin{aligned} \text{for females: } S &= \gamma_0 + \gamma_1 x \\ \text{for males: } S &= \gamma_0 + \gamma_2 + (\gamma_1 + \gamma_3)x \end{aligned}\right\} \quad (4)$$

Thus by equating coefficients in (3) and (4) in the obvious way the two approaches of separate logistic scores and combined scores with an interaction term are entirely equivalent.

The significance of one of the coefficients in the score (1) is assessed by calculating twice the change in the maximized log-likelihood with and without the corresponding term in the score. This statistic is taken to be approximately distributed as a chi-square variate with one degree of freedom under the null hypothesis that the true value of the coefficient is zero. If it is desired to test the significance of k coefficients considered together, the corresponding change in twice the maximized log-likelihood has approximately a chi-square distribution with k degrees of freedom.

The results of the computations on the cases from source I to assess the effects of age, sex, smoking history, and percentage bronchial mucous gland volume on emphysema are given in Table I. In the first half of the table 'main' effects are evaluated. The likelihood is first maximized with all the factors in and then, in turn, each factor is dropped and the decrease in maximized log-likelihood is noted. The contribution of a factor was said to be statistically significant if twice this decrease exceeded the appropriate chi-square percentage point. In the second half of Table I, interactions are investigated in a similar way except that now the terms are added to the logistic model (1), one at a time and their effects on the maximized log-likelihood are noted.

The results of the computations on subjects from all sources are shown in Table III. The procedure here was very slightly different from the above. For example, the effect of differences between the sources has been tested with and without the age × sex interaction being taken into account.

In all these analyses, the analogy of ordinary regression is very helpful in appreciating the motivation behind the various steps of the analysis.

This work was assisted by a personal M.R.C. research grant to one of us (M.S.D.).

Thorax: first published as 10.1136/thx.27.5.547 on 1 September 1972. Downloaded from http://thorax.bmj.com/ on April 9, 2024 by guest. Protected by copyright.