Journal of Epidemiology and Community Health, 1980, **34**, 217–222

Dose-response in case-control studies

G. BERRY

From the MRC Pneumoconiosis Unit, Penarth, South Glamorgan

SUMMARY The evidence provided by a case-control study on the association between a disease and some factor is strengthened if the extent of exposure to the factor is categorised into several groups or measured on a continuous scale. Then dose-response relationships can be estimated. The methods available are illustrated by application to data on lung cancer and chrysotile asbestos exposure from Quebec in which there were three matched controls for each case. Regression-type models were fitted assuming that the relative risk of lung cancer was linearly related to an exposure measure; a covariate, smoking, was also included in the analysis. The data were first analysed ignoring the matching and secondly taking account of the matching. The methodology for the latter analysis has only recently been developed; formerly, matched studies were of necessity analysed as unmatched. Although, in this particular example, the unmatched and matched analyses gave similar results, this is not always the case and it is argued that, now that the methodology is available, matched case-control studies should be analysed taking proper account of the matching.

In the simplest form of a case-control study, cases of disease and suitable controls are each categorised into two classes indicating exposure, or no exposure, to an agent suspected of causing the disease. Such a study may provide evidence on the association between the agent and the disease, which will be strengthened, particularly if it is to be used to imply causation, if a dose-response relationship can be established.[1] The dose-response relationship can be estimated by categorising both cases and controls into more than two groups defined in terms of the extent of exposure to the agent. The controls must be chosen from the same environment as the cases.

An example of the type of material under discussion was obtained in a study of a cohort of men employed in the mining and milling of chrysotile asbestos in Quebec.[2] There were 245 deaths from lung cancer, and for each case three controls were chosen, from the population of miners and millers, matched for year of birth and still living when the case died. The results of this study, summarised in Table 1, are discussed by McDonald *et al*[2] whose report should be consulted by those interested in the epidemiological findings; in the present paper, the data are used only for illustrative purposes.

Table 1 *Dust exposure in deaths from lung cancer and in controls*

*Dust exposure**				
Range	*Mean***	*Cases*	*Controls*	*Relative risk*
Less than 6	2	49	190	1·00
6—	8	12	51	0·91
10—	18	28	92	1·18
30—	60	40	119	1·30
100—	182	33	124	1·03
300—	442	32	88	1·41
600—	771	24	39	2·39
1000—	1249	12	15	3·10
1500—	1710	6	8	2·91
2000—	2722	9	9	3·88
Total		245	735	

*Units are millions of particles per cubic foot x years evaluated up to 9 years before death of case.

**The mean dust exposures are for cases and controls combined. Data are from the study of McDonald *et al.*[2]

Although the controls were chosen in threes to match the individual cases, the data are first analysed ignoring the matching; that is, treating the controls as if they had been chosen strictly at random from the complete cohort of miners and millers. Secondly, the data are analysed using recently developed methods for matched case-control studies. The first analysis is of interest because it illustrates the methods for unmatched data, which have often been used—as in an earlier presentation of this study[3]—when the matched analysis was not yet available for the situation of multiple controls *and* multiple levels of the factor.

DOSE-RESPONSE RELATIONSHIP

Let x be a measure of exposure and let R be the relative risk of dying with lung cancer, relative to the unexposed ($x=0$). Then the simplest form of

dose-response relationships is that R is linearly related to x and, since $R = 1$ when $x = 0$,

$$R = 1 + b\,x \qquad \ldots (1)$$

where b is a parameter. This is the basic form of dose-response relationship used in this paper, although extensions to this form will be introduced as required.

UNMATCHED ANALYSIS

Consider data set out as in Table 2 where the cases and controls are sampled from a population. Let p_i be the probability that a member of group i in the population is a case, and let p_o be the probability that a, possibly hypothetical, non-exposed person in the population is a case. Suppose that the sampling fractions of the cases and controls are f_1 and f_2 respectively and that P_i is the consequent probability that a member of group i in the sample is a case. As is well known, for case-control studies the only functions of the p_i that can be estimated are odds ratios, that is, the approximate relative risks assuming that cases are rare. We have

Table 2 *Form of data for unmatched analysis*

Group	*Mean x*	*Cases*	*Controls*	*Total*
1	x_1	r_1	n_1-r_1	n_1
2	x_2	r_2	n_2-r_2	n_2
,	,	,	,	,
,	,	,	,	,
k	x_k	r_k	n_k-r_k	n_k
Total		$r.$	$n.-r.$	$n.$

$$\begin{aligned} P_i/(1-P_i) &= f_1/p_i/f_2\,(1-p_i) \\ &= [f_1/f_2]\;[p_o/(1-p_o)]\;[p_i/(1-p_i)]/[p_o/(1-p_o)] \\ &= \theta R_i \\ &= \theta(1+bx) \qquad \ldots (2) \end{aligned}$$

where $\theta = f_1 p_o/f_2\,(1-p_o)$ and R_i is the approximate risk in group i relative to zero exposure. Within group i, r_i is distributed binomially with probability P_i and group size n_i, and so the likelihood of the data can be written down. This enables the parameters θ and b to be estimated by maximum likelihood using iterative methods. The general method is well known and details are given in the Appendix. The formulation is a particular case of the class of generalised linear models discussed by Nelder and Wedderburn[4] and also by Nelder.[5] Test statistics are based on differences in the maximum log-likelihood obtained when fitting different models. The departure of a fitted model from a perfect fit is measured by the deviance, defined as twice the difference of the log-likelihoods of the perfect fit and the fitted model[4]; deviances are approximately distributed as chi-squared statistics.

The linear dose-response relationship (1) was fitted to the data of Table 1. The deviance between the 10 groups was 23·05 with 9 degrees of freedom (df). Fitting the linear term accounted for a reduction in deviance of 19·61 (1 df), a highly significant effect, and the maximum likelihood estimate of b was 0·00119. The residual deviance was 3·43 (8 df) providing no evidence against the linear fit. An alternative method of judging the adequacy of the linear model is to add a quadratic term in x, and see if this results in a substantial reduction in deviance, but here the reduction was less than 0·01.

The summary data in Table 1 were produced by grouping the observations according to the exposure. The number of groups and their boundaries are arbitrary and the question arises of the extent to which this might influence the conclusions. This question was investigated by dividing the data into 6, 20, and 27 groups and also by treating each of the cases and controls as individuals, equivalent to treating the data as in 980 groups each with a single member; the estimation method does not break down in this case. The results are given in the upper part of Table 3. Firstly, all estimates of b were similar, particularly when one considers that the approximate 95% confidence limits for b are from 0·0006 to 0·0024. Secondly, the significance of the contribution of the linear term is similar for all groupings. Thirdly, the adequacy of the linear model, judged either by the residual deviance or by the reduction in deviance due to including a quadratic term, is assessed similarly for all groupings (the former method is not available with no grouping since it is not valid to regard the residual deviance as approximately chi-squared when all the groups contain just one member). Hence, dividing the data into groups, even as few as 6, has little effect. This is not surprising since a similar result is known to apply for least squares regression analysis of a quantitative variable.[6] In the case of a quantal variate, grouping is necessary to give a comprehensible presentation of the data. In the Figure the data of Table 1 are plotted together with the fitted linear relationship. The relationship may be fitted on grouped or individual data as convenient. The latter increases computing time but not sufficiently to be an important consideration.

MATCHED ANALYSIS

The method of analysis appropriate to a matched case-control study with more than one control per case and where the factor has more than two levels was first given by Thomas[7] and later by Prentice and Breslow[8] and also by Breslow.[9]

Table 3 *Summary of unmatched and matched analyses*

No. of groups in unmatched analysis	Deviances (df) Between groups	Due to linear term b	Residual about linear	Due to quadratic term	Estimate of b in linear model (approximate 95% confidence limits)	
6	19·90 (5)	18·03 (1)	1·87 (4)	0·08 (1)	0·00113	(0·00056 — 0·00228)
10	23·05 (9)	19·61 (1)	3·43 (8)	0·00 (1)	0·00119	(0·00060 — 0·00238)
20	35·20 (19)	20·22 (1)	14·98 (18)	0·00 (1)	0·00122	(0·00061 — 0·00242)
27	37·86 (26)	20·27 (1)	17·59 (25)	0·05 (1)	0·00121	(0·00061 — 0·00240)
980*	1102·09 (979)	19·89 (1)	1082·20 (978)	0·02 (1)	0·00119	(0·00060 — 0·00239)
Matched analysis	—	21·37 (1)	—	0·10 (1)	0·00136	(0·00068 — 0·00274)

* Ungrouped analysis; treating observations individually.

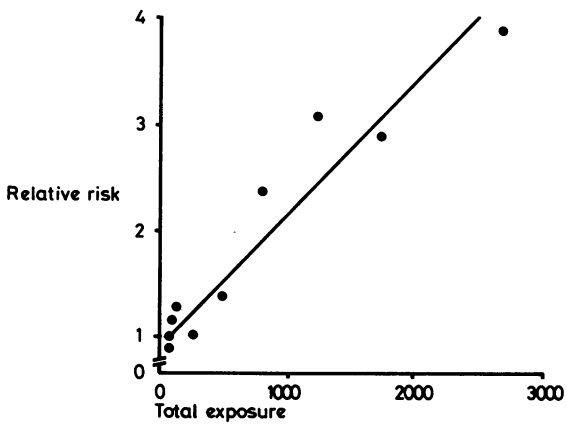

Figure *Relationship between relative risk of lung cancer and exposure to asbestos (data of Table 1).*

Suppose there are n cases each with m controls. Let x_{io} be the exposure for the i-th case and x_{ij} for its j-th control, and R_{io} and R_{ij} the corresponding relative risks as defined by the dose-response relationship (1). Then consider the set consisting of the i-th case and its controls. Given that this set consists of individuals with relative risks R_{ij} and that just one member of the set is a case, then the probability of a particular member of the set being the case is proportional to its relative risk. Therefore the probability P_{io} that the one case is the observed case is given by

$$P_{io} = R_{io} / \sum_{j=0}^{m} R_{ij}$$

$$= (1+bx_{io}) / \sum_{j=0}^{m} (1+bx_{ij})$$

The likelihood of the observations is the product of the above probabilities over the n cases and the method of maximum likelihood can be applied (Appendix).

Applying the method to the data gave the results in the lower part of Table 3. The maximum likelihood estimate for b was 0·00136, with approximate 95% confidence limits of 0·00068 and 0·00274, and the test statistic, twice the gain in log-likelihood, was 21·37 (1 df). The test statistic for the inclusion of a quadratic term was only 0·10 (1 df). Comparing these results with the ungrouped unmatched analysis (Table 3), the matched analysis gave a slightly higher estimate of b and the test statistic for its significance was a little higher, 21·37 compared with 19·89.

The smoking habits were available for about 60% of the men in the study; 140 of the cases had known smoking habits and at least one control with known smoking habits. The effect of smoking and the combination of asbestos and smoking have been analysed in this subgroup of the data. Men were classified as non-smokers, $s = 0$, or smokers, $s = 1$. The data are given in relation to smoking in Table 4. The cases have either 1, 2, or 3 matched controls and the method copes without difficulty with differing numbers of controls per case (Appendix).

Various models for relative risk were fitted (Table 5). Fitting asbestos exposure alone (B) gave an estimate of the linear effect of 0·00146 with a test statistic of 11·92 (1 df). Fitting smoking alone (C)

Table 4 *Smoking habits of cases and matched controls*

NON-SMOKING CASES			SMOKING CASES			NO. OF CONTROLS	
Case	Control(s)	No. of sets	Case	Control(s)	No. of sets	N	S
N	N	1	S	N	16	17	
	S	2		S	19		21
N	NN	4	S	NN	14	36	
	NS	5		NS	24	28	29
	SS	4		SS	21		50
N	NNN	2	S	NNN	4	18	
	NNS	0		NNS	5	10	5
	NSS	2		NSS	7	9	18
	SSS	0		SSS	10		30
		20			120	119	153

Table 5 *Smoking and exposure relationships fitted to matched data*

Relative risk		2 x log-likelihood	Maximum likelihood estimates b_1	b_2	b_3
(A)	$R = 1$	−294·06	—	—	—
(B)	$R = 1+b_1x$	−282·14	0·00146	—	—
(C)	$R = 1+b_2s$	−253·47	—	4·53	—
(D)	$R = 1+b_1x+b_2s$	−240·65	0·00540	8·84	—
(E)	$R = 1+b_1x+b_2s+b_3xs$	−240·50	0·00487	8·32	0·00244
(F)	$R = (1+b_1x)(1+b_2s)$	−243·03	0·00170	4·73	—

gave an estimate of effect of 4·5, that is, relative risk due to smoking of 5·5, and a test statistic of 40·59 (1 df). Next, model (D) includes both smoking and asbestos exposure with the two effects combined additively. The inclusion of exposure as well as smoking gave a test statistic of 12·82 (1 df) as evidence of an effect due to asbestos after allowing for the effect of smoking. Model (E) includes an extra term, the product of smoking and asbestos, allowing the effect of asbestos exposure to be dependent on smoking, but this led to a negligible improvement. Finally, model (F) includes both smoking and asbestos exposure but, in contrast to model (D), the two effects are combined multiplicatively. For these data the additive model fitted better although there was not much discrimination between the two possibilities; the relative likelihood of the data with the additive and multiplicative models was 3·3 to 1 (exponential of 1·19, the difference in log-likelihood).

The estimates of the exposure and smoking effects, b_1 and b_2, were both considerably increased when the additive model (D) was used compared with fitting each effect ignoring the other (B and C). In contrast, when the multiplicative model (F) was used the estimates were similar to those obtained when each effect was estimated ignoring the other. Of course any correlation between exposure and smoking in the population under study would result in a modification of the estimates of effect when both effects are fitted together. In the absence of such correlation the estimates when both are fitted together will be similar to the separate estimates only when the combined model is multiplicative. If the additive model were appropriate then ignoring a factor having a positive effect, as would have to be done if the factor had not been measured or its existence was unknown, would result in underestimates of the remaining effects.

At the time when the data were supplied smoking habits were known for only 60% of the men but the possible influence of this is not discussed here because the purpose of this paper is to illustrate methods of analysis, not to present epidemiological results. Readers interested in these are referred to McDonald *et al.*[2] All the evidence on the combination of asbestos and smoking in producing lung cancer was reviewed by Saracci.[10]

Discussion

The aim of this paper is to illustrate the powerful methods which are now available for the analysis of case-control studies. These methods allow the analysis of data using regression type models. The actual type of dose-response relationship used in this paper is one which is appropriate to agents which increase the incidence of a tumour, but the methods are more general. Other relationships may be more appropriate in other situations. The matched analysis has usually been considered in terms of an exponential dose-response relationship and there are theoretical advantages in this form but, as observed by Thomas,[11] not all biological models can be so expressed. He discussed the type of relationship considered in this paper and observed that in some circumstances there may be difficulties in fitting the relationship. Difficulties would certainly arise with a factor which had a protective effect since b would then be negative and the relationship given in equation (1) would give impossible negative relative risks above some value of x. The relationship would be inappropriate in such a case and analysis using an exponential model would be indicated.

The matched and unmatched analysis gave similar results, and in this case allowing for the matching only slightly increased the sensitivity of the analysis. This would be expected if exposure was unrelated to the matching variables[12] but otherwise ignoring the matching would result in conservative estimates as did occur to a slight extent. Breslow *et al*[13] also discussed the conditions under which the unmatched analysis would be expected to give similar results to the matched, and gave an example in which ignoring the matching led to substantial bias.

When information is missing then the matched analysis ignores some subjects with complete information, since all the controls of cases with missing data and those cases whose controls all had missing information would not contribute. In the example discussed the matched analysis including smoking was on 140 cases and 272 controls but the corresponding unmatched analysis was on 145 cases and 447 controls. The results of the latter analysis are not given in this paper but the parameter estimates were similar to those of the matched analysis. However, because of the larger number of subjects, the test statistics were higher; that is, the unmatched analysis was more sensitive. However, the inclusion of controls without the corresponding cases negates the matching included in the study design and, if the

matched analysis had not been carried out, one could not be confident that the unmatched analysis was not substantially biased.

A basic statistical principle is that the design determines the analysis. For a quantitative variable with matching there would be no argument that the analysis would have to take account of the matching, for example, using a paired *t*-test. Now that the general methodology for the analysis of matched categorical data is available there seems little reason to continue to ignore the matching in the analysis.

I thank Professor J. C. McDonald for supplying me with data from the Quebec asbestos study, Professor D. C. Thomas for his comments on an earlier draft, and Dr P. D. Oldham and Professor F. D. K. Liddell for their helpful comments and encouragement throughout.

Appendix

Unmatched analysis

The log-likelihood, L, of the data of Table 2 is given by:

$$L = \sum_{i=1}^{k} \left\{ r_i \ln P_i + (n_i - r_i)\ln(1 - P_i) \right\}$$

and

$$P_i = \theta R_i/(1 + \theta R_i)$$

The form of dose-response relationship considered is

$$R_i = 1 + b_1 x_{1i} + b_2 x_{2i}$$

where there are two variables, x_{1i} and x_{2i}, available for each group i (in the example given in the text the two variables were x and x^2).

Therefore

$$L = R \ln\theta + \sum_{i=1}^{k} \left\{ r_i \ln R_i - n_i \ln(1 + \theta R_i) \right\}$$

The dose-response relationship can be substituted for R_i in the above equation for L, which is a function of the data and the parameters θ, b_1 and b_2. The first and second derivatives of L with respect to θ, b_1 and b_2 can be written down, enabling the maximum-likelihood estimates of these parameters to be obtained iteratively using the Newton-Raphson technique. A suitable procedure is to start with $\theta = r./(n. - r.)$ and $b_1 = 0$ and to omit the b_2 term. The maximum-likelihood estimates of θ and b_1, in the simpler model with $b_2 = 0$, are then obtained. If it is required to include b_2 the maximum-likelihood estimates of θ and b_1 can be used as starting values together with $b_2 = 0$.

The best fitting dose-response relationship is that obtained when the relative risks are as actually observed. The log-likelihood of this, L', is obtained by substituting r_i/n_i for P_i in the expression for L (omitting from summation any groups with $r_i = 0$ or $r_i = n_i$). For any other relationship the deviance of the fit is given by

$$D = 2(L' - L)$$

with degrees of freedom equal to the number of groups minus the number of estimated parameters. Deviances, and reductions in deviance on adding terms to a model, are approximately distributed as χ^2. Following the steps of the above estimation procedures gives firstly, at the starting values, the deviance between groups, that is, about the null relationship $R = 1$; secondly, the deviance about $R = 1 + b_1 x_1$ and, thirdly, the deviance about $R = 1 + b_1 x_1 + b_2 x_2$. Therefore the reductions in deviance due to x_1, and due to x_2 after allowing for x_1, are available.

The above calculations can be carried out using the computer programme for fitting generalised linear models, GLIM-3.[14] The necessary link function is not standard and the method of coping with this was given by Thompson and Baker.[15]

Matched analysis

Let the sets of a case and its matched controls be represented by suffix $i = 1,2,\ldots,n$. Let suffix j identify the members of each set; $j = 0$ for the case and $j = 1,2,\ldots,m$ for the m controls. Let the dose-response relationship be as in the unmatched analysis where x_{1ij} and x_{2ij} are the variable values.

Then the log-likelihood, L, is given by:

$$L = \sum_{i=1}^{n} \ln P_{io}$$

where

$$P_{io} = R_{io} / \sum_{j=0}^{m} R_{ij}$$

Therefore

$$L = \sum_{i=1}^{n} \left[\ln R_{io} - \ln \left\{ \sum_{j=0}^{m} R_{ij} \right\} \right]$$

The estimation procedure is similar to that of the unmatched analysis, except that there is no equivalent expression for the log-likelihood of the best fitting relationship (L') and therefore no test of the adequacy of the fitted relationship. A test statistic for the inclusion of a term in a model is twice the corresponding increase in L.

The method can be adapted to cope with different numbers of controls per case. If the i-th case has m_i controls then the upper limit of the summations over j become m_i instead of m.

Confidence intervals

The square roots of the diagonal terms in the inverse matrix of negative second derivatives at the maximum likelihood solution are approximate standard errors of the parameter estimates. In both the unmatched and matched analyses the log-likelihoods were asymmetric in the parameter values about the maximum likelihood estimates so that the confidence intervals should be calculated to be similarly asymmetric. The confidence intervals given in the text and in Table 3 were calculated assuming that the estimate of the logarithm of the parameter was distributed approximately normally (s.e.$(\ln\hat{b})$ = s.e.$(\hat{b})/\hat{b}$). A better, but more tedious, method would be to calculate the log-likelihood surface for a range of parameter values and to take as 95% confidence interval those values of the parameter which gave a log-likelihood within 1·92 (half of the 5% significant value of χ^2 with 1 df) of the maximum log-likelihood. This method would not involve the assumption that the estimate of a particular transform of the parameter was distributed normally. Table 6 gives the 95% confidence intervals of the coefficient on exposure for both the unmatched and matched analyses, calculated in three different ways. For both analyses the intervals calculated assuming $\ln(\hat{b})$ normal were closer to the intervals obtained from the likelihood surface than those obtained assuming $\hat{b}$ normal, although the logarithmic transformation over-adjusted for asymmetry.

Table 6 *Confidence intervals of exposure coefficients*

	Unmatched	*Matched*
Estimate, $\hat{b}$	0·00119	0·00136
Standard error of $\hat{b}$	0·00042	0·00049
95% CONFIDENCE INTERVAL		
Symmetric in $\hat{b}$	0·00037 — 0·00202	0·00041 — 0·00232
Symmetric in $\ln(\hat{b})$	0·00060 — 0·00239	0·00068 — 0·00274
From likelihood surface	0·00052 — 0·00222	0·00060 — 0·00256

References

[1] Hill AB. The environment and disease: association or causation. *Proc R Soc Med* 1965; **58:** 295–300.

[2] McDonald JC, Liddell FDK, Gibbs GW, Eyssen GE, McDonald AD. Dust exposure and mortality in chrysotile mining, 1910–75. *Br J Ind Med* 1980; **37:** 11–24.

[3] Liddell FDK, McDonald JC, Thomas DC. Methods of cohort analysis: appraised by application to asbestos mining. *J R Stat Soc A* 1977: **140:** 469–91.

[4] Nelder JA, Wedderburn RWM. Generalized linear models. *J R Stat Soc A* 1972; **135:** 370–84.

[5] Nelder JA. Log linear models for contingency tables: a generalization of classical least squares. *Appl Stat* 1974; **23:** 323–9.

[6] Haitovsky Y. *Regression Estimation from Grouped Observations.* London: Griffin, 1973.

[7] Thomas DC. Addendum to paper by Liddell *et al. J R Stat Soc A* 1977; **140:** 483–5.

[8] Prentice RL, Breslow NE. Retrospective studies and failure time models. *Biometrika* 1978; **65:** 153–8.

[9] Breslow NE. The proportional hazards model: applications in epidemiology. *Communications in Statistics—Theory and Methods* 1978; **A7**(4): 315–32.

[10] Saracci R. Asbestos and lung cancer: an analysis of the epidemiological evidence on the asbestos-smoking interaction. *Int J Cancer* 1977; **20:** 323–31.

[11] Thomas DC. General relative risk models for survival time and matched case-control analysis. *Biometrics* 1980 (in press).

[12] Seigel DG, Greenhouse SW. Validity in estimating relative risk in case-control studies. *J Chronic Dis* 1973; **26:** 219–25.

[13] Breslow NE, Day NE, Halvorsen KT, Prentice RL, Sabai C. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epidemiol* 1978; **108:** 299–307.

[14] Baker RJ, Nelder JA. *The GLIM system, release 3.* Oxford: Numerical Algorithms Group, 1978.

[15] Thompson R, Baker RJ. Composite link functions in generalised linear models. *Appl Stat* 1980 (in press).