Relative risk estimation from vital statistical data: validation, a pitfall and an alternative method

P B S SILCOCKS AND M MURPHY

From the Department of Clinical Epidemiology & Social Medicine,¹ St George's Hospital Medical School, London; and the Division of Medical Statistics,² Office of Population Censuses & Surveys and Department of Community Medicine & Medical Statistics, University of Southampton Medical School

SUMMARY A previously described method of obtaining an estimate of relative risk from routinely available data was applied to data on cigarette consumption and lung cancer mortality to test its validity. Some shortcomings of the method were noted and an alternative approach using weighted logistic regression gave results closer to those predicted on the basis of other studies, without the disadvantages of the original technique.

In 1979 Beral, Chilvers, and Fraser¹ described a method of estimating relative risk using vital statistical data. The advantage of their technique is its cheapness and ability to provide at least an order-of-magnitude estimate as a preliminary to specifically designed studies of a problem. Although Beral and her co-workers showed that relative risks obtained by this and more conventional methods were similar in a number of instances, they pointed out that further applications to different sets of data would be valuable in further validating the approach. We decided to do this for lung cancer, a disease for which the main causative factor is well known² and for which a yardstick exists in the form of several estimates of relative risk in both sexes. During this exercise we discovered a pitfall in the approach used by Beral *et al* and we propose an alternative method which avoids the problem.

Method

We abstracted age adjusted proportions of male and female smokers in the standard regions of England and Wales from the *General household survey 1972*.³ Average male and female SMRs from lung cancer (ICD8, 162) for 1969 to 1973 were taken from the Decennial Supplement on Area Mortality.⁴

A further set of male SMRs was obtained from the Decennial Supplement on Occupational Mortality for 25 occupational orders.⁵ This work also gives age adjusted smoking information but in the form of the proportional current smoking ratio. This is analogous to an SMR and is the number of smokers observed in a group divided by the number expected if the proportion of smokers in each age group was the same

as in a suitable standard population. In this example the standard was all men aged 15-64 years.

In Beral and Chilvers' original paper¹ the model adopted is a simple linear regression of a measure of disease frequency on a measure of exposure to a risk factor, the equation being:

$$y = a + bx$$

where y is a measure of disease frequency, x is a measure of exposure, and a and b are constants. From this they derive an estimate of relative risk R given by:

$$R = (a + bx)/(a + bx_0)$$

where x_0 is the baseline level of exposure. When x is a proportion and the relative risk of exposure v no exposure is to be found, $x_0 = 0$ and $x = 1$, so the equation simplifies to:

$$R = (b/a) + 1$$

A point that is omitted from their discussion is that the measure of disease frequency can equally well be an SMR instead of an indirectly standardised rate. This is because the indirectly standardised rate is the product of the SMR and the crude rate in the standard population. The relative risk is approximated by the ratio of the two rates (standardisation merely helps to avoid confounding) and the standard crude rates cancel, leaving a ratio of two SMRs. If the proportion exposed to the risk factor is expressed as a percentage, when all are exposed $x = 100$ and the relative risk is estimated by:

$$R = (100b/a) + 1$$

A disadvantage of this method is that the linear regression may estimate a negative value for a and

consequently for R, which is meaningless. This is especially likely when there are few data for exposures near zero, and a large dose response effect as measured by the slope, b.

The alternative method we propose requires the use of mortality rates as a measure of disease frequency. Regarding these as proportions with a range necessarily between 0 and 1 we avoid the problem of Beral's method by use of a logit transformation.⁶ We can then set up a logistic regression equation in which the logit of the mortality rate is the outcome variable:

$$\log_e(y/(1-y)) = c + dx$$

c is the value of the logit when x=0. Rearranging the equation we have:

$$\log_e(y/(1-y)) - c = dx \text{ [for c we can write}$$

$$\log(y_o/(1-y_o))]$$

$$\text{so } \log_e(y/(1-y)) - \log_e(y_o/(1-y_o)) = dx$$

$$\text{ie, } \log_e(y(1-y_o)/(y_o(1-y))) = dx$$

$$\text{ie, } \log_e(R) = dx$$

ie, $R = \exp(dx)$ where R cannot be negative. It is, in fact, the ODDS ratio which approximates to relative risk when incidence rates are small. As before, when x is a proportion and we wish to compare the effect of exposure with no exposure we may put x=1 and simplify, in this case to:

$$R = \exp(d)$$

An approximate estimate of the value d can be found by a weighted least squares regression. Suitable weights are proportional to the reciprocals of the variance of each logit,⁶ given by:

$$\text{ith weight} = w_i = (n_i)(y_i)(1-y_i)$$

n_i being in proportion to the size of the population for the ith pair of x and y values. The formula for d is then:

$$d = \frac{\sum w_i x_i L_i - ((\sum w_i x_i)(\sum w_i L_i)/\sum w_i)}{\sum w_i x_i^2 - ((\sum w_i x_i)^2/\sum w_i)}$$

where $L_i = \log_e(y_i/1-y_i)$

The calculation is rather tedious but can be done with patience or with the aid of a simple microcomputer program. Confidence limits on R are easily found by exponentiating the confidence limits on d, the standard error of which is:

$$SE(d) = \sqrt{\{ [Syy - ((Sxy)^2/Sxx)]/(n-2)(Sxx) \}}$$

where $Syy = \sum w_i L_i^2 - ((\sum w_i y_i)^2/\sum w_i)$

$$Sxx = \sum w_i x_i^2 - ((\sum w_i x_i)^2/\sum w_i)$$

$$Sxy = \sum w_i x_i L_i - ((\sum w_i x_i)(\sum w_i y_i)/\sum w_i)$$

Results

The data we abstracted are displayed in tables 1-3 together with estimated age standardised death rates

Table 1 Regional SMRs, population, % smokers, and death rate (males)

Region	Age adjusted % smokers	SMR	Age adjusted death rate (annual proba- bility of death)	Population (thousands)
North	52.1	1.08	0.001154	1611.7
Yorks/H'side	53.3	1.01	0.001079	2343.1
North West	54.3	1.11	0.001186	3254.9
E Midlands	50.6	0.90	0.000962	1672.6
W Midlands	51.4	1.01	0.001079	2533.1
East Anglia	48.3	0.85	0.000908	828.5
GLC	52.2	1.18	0.001261	3561.7
Other SE	47.5	0.924	0.000987	4759.7
South west	47.6	0.82	0.000876	1829.8
Wales	56.1	0.84	0.000898	1324.5

Table 2 Regional SMRs, population, % smokers, and death rate (females)

Region	Age adjusted % smokers	SMR	Age adjusted death rate (annual proba- bility of death)	Population (thousands)
North	48	1.01	0.0002264	1689.0
Yorks/H'side	40.3	0.91	0.0002039	2465.5
North West	43.7	1.01	0.0002264	3483.6
E Midlands	36.5	0.84	0.0001883	1718.5
W Midlands	41.1	0.88	0.0001973	2585.6
East Anglia	37.5	0.87	0.0001950	852.4
GLC	41.2	1.33	0.0002982	3856.3
Other SE	36.4	1.034	0.0002309	5081.4
South west	36.9	0.85	0.0001906	1962.4
Wales	43.3	0.64	0.0001435	1400.7

and proportions of smokers. Table 4 presents relative risk estimates found by the two methods. The confidence intervals are wide because of the small number of data points available. In addition, the confidence limits estimated by our method are asymmetric because they are found by exponentiating the confidence limits on d and cannot therefore be negative.

For comparison, we weighted combined published values of the relative risk of lung cancer for different levels of cigarette consumption found in various prospective studies⁷ by the prevalence of current levels of smoking among men and women given in the *General household survey for 1972*.³ On this basis the overall relative risk of lung cancer in smokers was predicted to be 11.87 for men and 4.42 for women.

Discussion

Although Beral *et al* showed that with their method relative risks of some diseases for certain exposures were consistent with results from other studies, for one of the data sets we used this method gives a result for male lung cancer that is meaningless. The method we

Table 3 Death rate, No. in GHS sample, % smokers and SMR by occupational order (males)

Occupational Order	Age adjusted death rate (annual probability of death)	No. in GHS sample	Age adjusted % smokers	SMR
i	0.0005784	211	41.3	0.84
ii	0.0007988	153	74.1	1.16
iii	0.0008470	60	62.0	1.23
iv	0.0008814	31	50.6	1.28
v	0.0010674	103	62.6	1.55
vi	0.0006955	318	54.8	1.01
vii	0.0008126	1313	60.0	1.18
viii	0.0007782	186	50.0	1.13
ix	0.0007162	32	47.5	1.04
x	0.0006059	55	55.7	0.88
xi	0.0007162	35	48.1	1.04
xii	0.0008883	115	55.6	1.29
xiii	0.0005922	109	57.5	0.86
xiv	0.0006610	108	60.3	0.96
xv	0.0009916	311	60.9	1.44
xvi	0.0009572	153	59.0	1.39
xvii	0.0007782	180	67.0	1.13
xviii	0.0010054	383	71.6	1.46
xix	0.0008814	564	62.1	1.28
xx	0.0007919	302	56.1	1.15
xxi	0.0005440	533	46.8	0.79
xxii	0.0005853	621	49.1	0.85
xxiii	0.0008264	357	53.8	1.20
xxiv	0.0004132	371	41.0	0.60
xxv	0.0003512	894	35.7	0.51

Table 4 Relative risks by different methods (95% confidence limits)*

Method	Lung cancer relative risk: smokers v non smokers		
	Area based (M)	Area based (F)	Occupation based (M)
Beral <i>et al</i>	20.71 (-28.57 to 69.99)	1.78 (-4.99 to 8.56)	-116.65 (-3566.8 to 3333.5)
Logistic regression	10.68 (0.50 to 227.24)	1.83 (0.03 to 123.31)	13.72 (7.06 to 26.69)

*Calculated on the basis of *t* for (n-2) degrees of freedom.

propose (weighted logistic regression) is slightly more complex, requires the outcome variable in the form of a proportion and some knowledge of population size, but avoids the problem of negative relative risk values. In addition, confidence limits are easily found. Another advantage of weighted regression is that less emphasis is given to exposure or mortality data based on few individuals, for example, the proportion of smokers in occupational units iv and ix (table 3). Both methods may be biased if, for instance, a man's occupation order is coded differently when exposure is measured compared with when death occurs—a well recognised problem with unlinked data.

How well do the relative risk estimates found with these methods compare with the results of specifically

designed studies of the effect of smoking? For men, the method we propose gave point estimates close to those expected on the basis of established relative risks and cigarette consumption in 1972. With occupational data, Beral's method gave a meaningless value and a much higher than predicted estimate with area data. For women, both methods gave similar results which were rather lower than that predicted. To some extent this may be expected because the relative risks estimated here were for current smokers compared with current non-smokers. Because the latter group includes ex-smokers, who are at increased risk of lung cancer, estimates of relative risk based on present smoking status will be rather lower than estimates based on a comparison of smokers with those who never smoked. This bias is insignificant, however, compared with the width of the confidence limits. For area based mortality these included relative risk estimates of less than unity. On the other hand, occupational data—at the time readily available only for men—gave a highly significant estimate of relative risk using the weighted logistic regression approach. The fact that information may be limited or unavailable is an obstacle common to all methods using secondary data sources.

A further point deserves mention and that is the use of age standardised data to remove an important source of confounding though adjustment for other variables such as social class would also be possible if suitable data were available. In a study of persons with lung cancer, smoking status is ascertained in the same individual as disease status and so the age of the subject is automatically taken into account. For groups of individuals disease and smoking status are expressed as rates. Since we are testing a hypothesis by comparing different groups, allowance may have to be made for differing age structures in the populations. This can be done by age standardisation which should be applied to both the measure of disease and to the measure of exposure if, as in this case, both are related to age. In practice, of course, it may be impossible to fulfil this requirement using readily available data. This, together with the fact that grouped data may be subject to the ecological fallacy,⁸ may severely limit the ability of these methods to give definitive answers. Nevertheless some information is better than none at all, and when carrying out preliminary work at the start of an investigation such techniques provide a "cheap and cheerful" alternative to the "quick and dirty" school of epidemiology.

Address for correspondence: Department of Clinical Epidemiology & Social Medicine, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE.

References

- ¹ Beral V, Chilvers C, Fraser P. On the estimation of relative risk from vital statistical data, *J Epidemiol Comm Health* 1979; **33**: 159–62.
- ² Doll R, Peto R, *Causes of cancer*. Illinois: Oxford University Press; 1982.
- ³ *General household survey 1972*. OPCS Social Survey Division. London: HMSO; 1975.
- ⁴ *Area mortality: the Registrar General's decennial supplement for England and Wales 1969–73*. OPCS series DS no. 4. London: HMSO; 1981.
- ⁵ *Occupational mortality: the Registrar General's decennial supplement for England and Wales 1970–1972*. OPCS series DS no. 1 London: HMSO; 1978; 148–159
- ⁶ Armitage P. *Statistical methods in medical research* Oxford and Edinburgh; Blackwell Scientific Publications; 1971.
- ⁷ Various studies cited in: *The health consequences of smoking: cancer: a report of the Surgeon General* US Department of Health & Human Services. Rockville; Maryland: 1982.
- ⁸ Robinson WS. Ecological correlations and the behaviour of individuals. *Am Sociol Rev* 1950 **15**: 351–7.