On the estimation of relative risk from vital statistical data

VALERIE BERAL, CLAIR CHILVER, AND PATRICIA FRASER
From the Epidemiological Monitoring Unit, London School of Hygiene and Tropical Medicine

SUMMARY A method is described for the determination of a measure of relative risk from vital statistical data. If the frequency of disease in a population is linearly related to the level of exposure to a given factor, then a measure of the relative risk can be estimated from the slope and intercept of the regression line. For example, when the exposure is measured in terms of the proportion of the population exposed to the factor, then the relative risk is equal to \( \frac{\text{slope}}{\text{intercept}} + 1 \).

This offers an indirect but simple and inexpensive method for estimating relative risk. It should be used with caution, particularly where confounding factors may be responsible for the apparent association between disease and factor. Applications of the method to estimate the relative risk of (a) circulatory diseases in women using oral contraceptives and (b) ovarian cancer in women with different average family sizes, both yielded relative risk estimates comparable with those obtained from case-control and prospective studies.

Vital statistical data are widely used both to provide clues and to test existing hypotheses about disease causation. Many important relationships between disease frequency and level of exposure to a given factor in different populations have been described using such data. For example, Nielson and Clemmesen (1954) showed that lung cancer mortality rates in different countries were positively correlated with the average cigarette consumption per capita in that country; Dean (1946) showed that the amount of dental caries in different populations was negatively associated with the fluoride content of drinking water; and in addition Keys (1953) showed that the mortality from coronary heart disease in different countries was associated with the consumption of saturated fats. Correlation studies of this type are essentially descriptive: no indication of the magnitude of the risk associated with the factor under study is given. Several recent studies have used vital statistical data to estimate relative risk (Seigel and Markush, 1969; Beral, 1976; Beral et al., 1978). In each, the relative risk estimated was comparable with that found in case-control and prospective studies.

This paper derives simple formulae for the estimation of a measure of relative risk and its sampling error when the disease frequency in different populations is correlated with the average level of exposure or with the proportion of the population exposed to some factor. The method described was the basis of the estimations made by Beral (1976) and Beral et al. (1978).

DEFINITIONS Relative risk is the ratio of the risk of disease occurrence in a population exposed to a given factor to that in an unexposed population (Armitage, 1971). This definition requires that each individual in a population be classified as exposed or unexposed; but this distinction is often arbitrary. For example, where the factor is a continuous variable, such as blood pressure, individuals with a pressure greater than a specified value are classified as hypertensive or ‘exposed’, while those with a lower pressure are classified as normotensive or ‘unexposed’. This distinction is artificial because both the exposed and the unexposed populations are made up of individuals with varying levels of blood pressure. The size of the relative risk will depend on the difference in the average blood pressure between the two groups and on the distribution of the individual measurements. Even where exposure to a given factor in the ‘unexposed’ group is homogeneous, such as in non-smokers, or in those never exposed to a
certain drug or chemical, the 'exposed' population generally consists of individuals with varying degrees of exposure. In this paper 'exposed' and 'unexposed' populations are defined in terms of their average exposure to a given factor. The situation where the unexposed population has no exposure to the factor is regarded as a special case.

**Derivation of Relative Risk**

Let us assume that in n different populations the level of disease (y) is linearly related to the average level of exposure (x) to the factor being studied.

Consider two populations with different levels of exposure $x_0$ and $x_1$, and associated disease frequencies $y_0$ and $y_1$ respectively (Figure). The relative risk (R) in the two populations is defined as the frequency of disease in individuals with average exposure $x_i$, relative to that in individuals with average exposure $x_0$, or simply $y_i/y_0$. ... (1)

Suppose, as illustrated in the Figure, that the equation of the regression line fitted to the data from n populations is: $y = a + bx$ then $y_0 = a + bx_0$ and $y_1 = a + bx_1$ so that the relative risk estimate is

$$ R = \frac{a + bx_1}{a + bx_0} \ldots (2) $$

The sampling error of R is derived in the Appendix.

**Modifications and Extensions**

I. The level of disease (y) may be linearly related to the proportion of persons (x) exposed to the factor being studied. Suppose that in any given population a proportion x of that population is exposed to the factor and has a level of disease $y_1$, and that the proportion $(1-x)$ not exposed to the factor has a level of disease $y_0$.

**Valerie Beral, Clair Chilvers and Patricia Fraser**

The overall level of disease in the population is given by:

$$ y = xy_1 + (1-x)y_0 $$

$$ = y_0 \left( \frac{y_1}{y_0} - l \right) x + l \ldots (3) $$

From equations (3) and (1) $y = y_0 + (y_0(R - l)) x$ which is of the form of the regression line $y = a + bx$ where the intercept $a = y_0$ ... (4) and the slope $b = y_0(R - l) \ldots (5)$

From equations (4) and (5) $R = \frac{b}{a} + 1$

that is, relative risk $= \frac{\text{slope}}{\text{intercept}} + 1$

This formula for relative risk is a special case of equation (2) obtained by setting $x_0 = O, x_1 = I$.

II. The disease frequency (y) may be a modification of a simple mortality rate. For example, Beral (1976) used the percentage change in mortality over a 10-year period as a measure of disease frequency (y) to estimate the relative risk of death from cardiovascular disease associated with oral contraceptive use.

III. In a given population, the overall disease frequency and the proportion of the population exposed to the factor are known. Thus the absolute risk of disease, that is, mortality, incidence, or prevalence rate, in exposed and unexposed individuals can be calculated if the relative risk is known.

If the overall mortality, incidence, or prevalence rate of the disease in a given population is m and the rate in the unexposed is $m_o$, then the rate in the exposed is $Rm_o$, and

$$ m = xRm_o + (1-x)m_o $$

Therefore $m_o = \frac{m}{x(R - l) + 1}$

and $Rm_o = \frac{Rm}{x(R - l) + 1}$

The attributable risk, as described by Levin (1953), is given by $m_o(R - l)$.

**Applications of the Method**

The technique described above has already been applied in two studies. The first used mortality data from 21 countries to estimate the relative risk of certain circulatory diseases in women taking oral contraceptives (Beral, 1976). Table 1 compares the risk estimates derived in that study with those subsequently obtained in a large prospective study (Royal College of General Practitioners, 1977). The second related the international pattern of ovarian cancer mortality and the mortality trends in successive generations of women to the completed
On the estimation of relative risk from vital statistical data

Table 1 Relative risk of mortality from (a) non-rheumatic heart disease and hypertension and (b) cerebrovascular disease in oral contraceptive users. Comparison of risk estimates derived from vital statistical data with those from prospective data

<table>
<thead>
<tr>
<th>ICD 8th revision†</th>
<th>Cause of death</th>
<th>Relative risk (oral contraceptive users: non-users)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400-429</td>
<td>Non-rheumatic heart disease and hypertension</td>
<td>4·0</td>
</tr>
<tr>
<td>430-438</td>
<td>Cerebrovascular disease</td>
<td>4·7</td>
</tr>
</tbody>
</table>

† World Health Organisation (1967).
* Beral (1976).
* Royal College of General Practitioners (1977).

Table 2 Relative risk of ovarian cancer by family size. Comparison of risk estimates derived from vital statistical data with those from case-control studies

<table>
<thead>
<tr>
<th>Comparison groups</th>
<th>Relative risk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>From vital statistical data*</td>
</tr>
<tr>
<td></td>
<td>Cohort trends</td>
</tr>
<tr>
<td>Women with 2 children:</td>
<td>1·3</td>
</tr>
<tr>
<td>women with 3 children</td>
<td></td>
</tr>
<tr>
<td>Women with 1 or 2 pregnancies:</td>
<td>1·8†</td>
</tr>
<tr>
<td>women with 3 or 4 pregnancies</td>
<td></td>
</tr>
</tbody>
</table>

† Assuming that women with 1 or 2 pregnancies had, on average, 1·5 children; and that women with 3 or 4 pregnancies had, on average, 3·5 children.
† Newhouse et al. (1977).
* Beral et al. (1978).
** Joly et al. (1974).

family size of women in each population (Beral et al., 1978). These data were used to estimate relative risk of ovarian cancer according to number of live-born children. Table 2 compares the risk estimates derived from that study with those obtained in earlier retrospective studies by Joly et al. (1974) and Newhouse et al. (1977). The comparison with the findings of Joly et al. (1974) is approximate, because their data were tabulated by number of pregnancies, rather than number of live-born children, and were presented in a grouped form. Furthermore, obtaining comparable risk estimates from our vital statistical data involved a small extrapolation of the regression line to an average family size of 1·5, although the lowest recorded average family size was 1·8 (Beral et al., 1978).

It can be seen from Tables 1 and 2 that the risk estimates obtained using vital statistical data are similar to those obtained in more direct studies.

Discussion

This paper describes a simple method for estimating a measure of relative risk when the frequency of disease in different populations is correlated with the average exposure of a population or the proportion of the population exposed to a given factor. The approach differs from conventional methods for measuring relative risk in that the exposed and unexposed individuals in the populations are not actually identified. Instead, it is assumed that each population consists of a mixture of individuals with varying degrees of exposure and with different risks of disease.

Certain statistical and epidemiological assumptions underlie this approach. In the least squares regression analysis used here, it is assumed that the y variable (disease frequency) is linearly related to the x variable (exposure), which is measured without error. If there were some error in the measurement of exposure, this would lead to underestimation of the slope of the regression line, and hence of relative risk. Unless there were major errors in the measurement of exposure this underestimate would be negligible.

A relationship between disease frequency and exposure, although real, may be due to confounding factors, diagnostic differences, or other sources of bias. In practice, it may be impossible to exclude all alternative explanations for an observed association. For this reason, relative risk estimates obtained from vital statistical data should not be considered in isolation but examined for consistency with risk estimates obtained from more conventional studies. If this approach is being used to test an existing hypothesis, and if the relative risk has already been estimated, the consistency of the various risk estimates can be compared. If, however, an
hypothesis is being suggested solely on the basis of vital statistical data, then estimates of relative and absolute risk based on this method may be biased and must be regarded with extreme caution. In this situation, the risk estimates may most usefully serve as the basis for calculations of sample size for formal studies designed to test the hypothesis.

Past applications of this technique have demonstrated that it is useful, and does yield reasonable estimates of risk. The estimates of relative and absolute risks of death from non-rheumatic cardiovascular disease among women using oral contraceptives were similar to those subsequently obtained in a prospective study. Similarly, the estimated relative risk of ovarian cancer in women with different completed family sizes is comparable with those found in case-control studies. In both instances strong correlations were noted between the factor and the level of disease. It is possible that this method would not yield such comparable risk estimates if a weak correlation existed between factor and disease. Future applications of the method to different sets of data should be valuable for further evaluation of the approach.

The method described here is of practical value, provided that it is used with caution. It offers a quick, simple, and inexpensive method for estimating relative risk which may be particularly useful in the absence of other relevant data. It extends the use of descriptive correlation studies, enabling the potential magnitude of the effect of environmental factors on disease frequency to be predicted.

Reprints from Valerie Beral, Epidemiological Monitoring Unit, Department of Medical Statistics and Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT.

Appendix

Sampling error of $R$

The usual linear regression model is:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where the $\epsilon_i$ are independent.

Normally distributed errors with zero mean and variance $\sigma^2$. $\alpha$ and $\beta$ are estimated by $a$ and $b$ in equation (2), which may be re-written:

$$y = \hat{y} + b(x - \bar{x})$$

From equations (1) and (6) $\hat{R} = \frac{\hat{\beta} + b(x_0 - \bar{x})}{\hat{\beta} + b(\bar{x} - \bar{x})}$

The relative risk is thus a function of $\hat{\beta}$ and $\hat{b}$ and the variance of the function is given approximately by:

$$\text{Var}(\hat{R}) = \left( \frac{\delta \hat{R}}{\delta \hat{\beta}} \right)^2 \text{Var}(\hat{\beta}) + \left( \frac{\delta \hat{R}}{\delta \hat{b}} \right)^2 \text{Var}(\hat{b}) + 2 \left( \frac{\delta \hat{R}}{\delta \hat{\beta}} \right) \left( \frac{\delta \hat{R}}{\delta \hat{b}} \right) \text{Cov}(\hat{\beta}, \hat{b})$$

So $\text{Var}(\hat{R}) = \frac{\sigma^2}{y_0^2} \frac{\hat{y}_0^2}{n} + \frac{\hat{y}_0^2}{n} \frac{\Sigma(x_i - \bar{x})^2}{i=1}$

References


