Statistical Methods in Medical Research

P. Armitage
MA, PhD
Emeritus Professor of Applied Statistics
University of Oxford

G. Berry
MA, PhD
Professor in Epidemiology and Biostatistics
University of Sydney

J.N.S. Matthews
MA, PhD
Professor of Medical Statistics
University of Newcastle upon Tyne

FOURTH EDITION

b
Blackwell
Science

19.6 Attributable risk

The full implications of an excess risk depend not only on the size of the relative risk but also on the proportion of the population positive for the aetiological factor. A moderate relative risk applicable to a high proportion of the population would produce more cases of disease than a high relative risk applicable to just a small proportion of the population. A measure of association, due to Levin (1953), that takes account of the proportion of the population at risk is the *attributable risk*. Terminology is not completely standard and other names for this measure are *aetiological fraction* and *attributable fraction*. Also these measures may be used to refer either to the whole population or just to the exposed subgroup and, to avoid ambiguity, it is preferable to be specific and to use the terms *population attributable risk* and *attributable risk among the exposed*.

The population attributable risk is defined as the proportion of cases in the total population that are attributable to the risk factor. It is usually calculated in circumstances in which it is considered justifiable to infer causation from an observed association. Then it may be interpreted as the proportion of cases in the population that are due to the factor, and hence as a measure of the importance of eliminating the factor as part of a disease-prevention strategy. The circumstances in which this interpretation is justified are probably few and there are particular difficulties when more than one cause is operating.

Suppose I_P is the incidence of the disease in the population and, as in §19.5, I_E and I_{NE} are the incidences in the exposed and non-exposed, respectively. Then the excess incidence attributable to the factor is $I_P - I_{NE}$ and dividing by the population incidence gives the population attributable risk,

$$\lambda_P = \frac{I_P - I_{NE}}{I_P}. \tag{19.30}$$

Now suppose a proportion θ_E of the population are exposed to the factor, then

$$\begin{aligned} I_P &= \theta_E I_E + (1 - \theta_E) I_{NE} \\ &= \theta_E \phi I_{NE} + (1 - \theta_E) I_{NE} \\ &= I_{NE}[1 + \theta_E(\phi - 1)], \end{aligned} \tag{19.31}$$

where ϕ is the relative risk and the second line is obtained using (19.13). Substituting in (19.30) gives

$$\lambda_P = \frac{\theta_E(\phi - 1)}{1 + \theta_E(\phi - 1)}, \tag{19.32}$$

and the attributable risk can be estimated from the relative risk and the proportion of the population exposed. Often the attributable proportion is multiplied by 100 to give a result in percentage terms and this may be referred to as the *population attributable risk per cent*.

An alternative expression may be derived using the notation of (19.14). If the probability of disease in those positive for the factor were the same as for those negative, then the proportion of the population positive for both factor and disease would be

$$(P_1 + P_3) \times \frac{P_2}{P_2 + P_4}.$$

Subtracting this from the actual proportion, P_1, gives the excess proportion related to the factor, and, dividing by the proportion with disease, the population attributable risk is given by

$$\begin{aligned}
\lambda_P &= \left[P_1 - \frac{P_2(P_1 + P_3)}{P_2 + P_4}\right] \div (P_1 + P_2) \\
&= \frac{P_1}{P_1 + P_2}\left[1 - \frac{P_2(P_1 + P_3)}{P_1(P_2 + P_4)}\right] \\
&= \frac{P_1}{P_1 + P_2}\left(1 - \frac{I_{NE}}{I_E}\right) \\
&= \theta_1 \frac{\phi - 1}{\phi},
\end{aligned} \tag{19.33}$$

where θ_1 is the proportion of cases exposed to the factor.

It follows from (19.32) that the population attributable risk can be estimated from any study that provides estimates both of relative risk and the proportion of the population exposed to the factor. Thus, it may be estimated from a case–control study, since the relative risk can be estimated approximately and the proportion of the population exposed can be estimated from the controls (again assuming that the disease is rare). Alternatively, (19.33) could be used, since θ_1 can be estimated in a case–control study. Clearly, population attributable risk may be estimated from a cohort study of a random sample of the total population, but only from a sample stratified by the two levels of exposure if the proportion of the population exposed is known.

The attributable risk among the exposed may be defined as the proportion of exposed cases attributable to the factor. This measure is

$$\lambda_E = \frac{I_E - I_{NE}}{I_E} = \frac{\phi - 1}{\phi}. \tag{19.34}$$

The above formulation has been in terms of a factor which increases risk, i.e. relative risk greater than unity. For a factor which is protective, i.e. relative risk less than unity, a slightly different formulation is necessary; this is discussed by Kleinbaum *et al.* (1982, Chapter 9), although often the simplest approach is to reverse the direction of the relative risk, e.g. 'exercise is protective against heart disease' is equivalent to 'lack of exercise is a risk factor'.

We now consider the sampling variation of an estimate of population attributable risk from a case–control study, and give an example.

Consider data of the form (19.17) from a case–control study. Since $a/(a+b)$ estimates $P_1/(P_1+P_2)$ or θ_1, and ad/bc estimates ψ, which is approximately ϕ, substitution in (19.33) gives as an estimate of λ_P

$$\begin{aligned}\hat{\lambda}_P &= \frac{a}{a+b}\left(1-\frac{bc}{ad}\right)\\ &= \frac{ad-bc}{d(a+b)}.\end{aligned} \tag{19.35}$$

Exactly the same expression may be obtained substituting $c/(c+d)$ for θ_E in (19.32). It is convenient to work with $\ln(1-\hat{\lambda}_P)$ since, as shown by Walter (1975), this variable is approximately normally distributed.

$$\begin{aligned}1-\hat{\lambda}_P &= \frac{bd+bc}{d(a+b)}\\ &= \frac{b}{a+b} \div \frac{d}{c+d}\\ &= \frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\end{aligned}$$

and

$$\ln(1-\hat{\lambda}_P) = \ln(1-\hat{\theta}_1) - \ln(1-\hat{\theta}_2),$$

where $\hat{\theta}_1$ is an estimator of $P_1/(P_1+P_2)$ and $\hat{\theta}_2$ of $P_3/(P_3+P_4)$. $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent estimators of proportions, so, using (4.13) and (5.19), we obtain

$$\text{var}[(1-\hat{\lambda}_P)] = \frac{a}{b(a+b)} + \frac{c}{d(c+d)}. \tag{19.36}$$

Example 19.7

Consider the data from study 1 of Table 19.5, in which 83 out of 86 lung cancer patients were smokers compared with 72 out of 86 controls. Then, using (19.35) and (19.36),

$$\begin{aligned}\hat{\lambda}_P &= \frac{83\times 14 - 3\times 72}{14\times 86}\\ &= 0{\cdot}7857,\\ \ln(1-\hat{\lambda}_P) &= -1{\cdot}540,\\ \text{var}[\ln(1-\hat{\lambda}_P)] &= \frac{83}{3\times 86} + \frac{72}{14\times 86}\\ &= 0{\cdot}3815.\end{aligned}$$

Therefore approximate 95% confidence limits for $\ln(1 - \hat{\lambda}_P)$ are

$$-1{\cdot}540 \pm (1{\cdot}96)\sqrt{0{\cdot}3815} = -2{\cdot}751 \text{ and } -0{\cdot}329,$$

and the corresponding limits for $\hat{\lambda}_P$ are 0·281 and 0·936.

Whittemore (1983) extended Levin's measure to account for a qualitative confounding variable by forming a weighted average of the attributable risks in each stratum of the confounder with weights equal to the proportion of cases estimated to be within each stratum in the whole population. This does not require any assumption of a uniform relative risk across strata, and so also adjusts for interaction effects. Where the confounders are accounted for by a stratified analysis, then a pooled relative risk is often estimated using the Mantel–Haenszel approach, and this estimate can be used to estimate the attributable risk, using (19.33). Kuritz and Landis (1988) used this approach for a matched case–control study and proposed a method of calculating the confidence interval. Greenland (1987) gave variance estimators for the attributable risk based on the sampling variability of R_{MH} (19.24); these estimators are satisfactory for both large strata and sparse data—that is, data divided into a large number of strata with small numbers within each stratum. The interpretation of these adjusted measures of population attributable risk as the proportion of cases that could be eliminated by removing exposure to the factor depends on all other factors remaining unchanged. In most situations it would be impossible to modify one factor without influencing other factors and so this interpretation would be invalid.

Further details of the estimation of attributable risk and the problems of interpretation are given in a review article by Benichou (1998).

19.7 **Subject-years method**

As noted in §19.4, a commonly used research method is the *cohort study*, in which a group is classified by exposure to some substance, followed over time and the vital status of each member determined up to the time at which the analysis is being conducted. A review of methods of cohort study design and application was given by Liddell (1988). It may be possible to use existing records to determine exposure in the past, and this gives the *historical prospective cohort* study, used particularly in occupational health research. Such studies often cover periods of over 20 years. The aim is to compare the mortality experience of subgroups, such as high exposure with low exposure, in order to establish whether exposure to the agent might be contributing to mortality. As such studies cover a long period of time, individuals will be ageing and their mortality risk will be changing. In addition, there may be period effects on mortality rate. Both the age and period effects will need to be taken