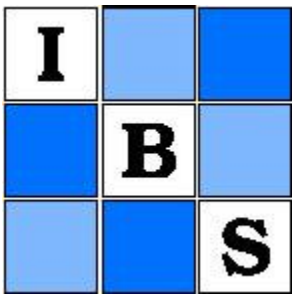

BIOMETRICS *32*, 829–849
DECEMBER 1976

The Estimation and Interpretation of Attributable Risk in Health Research

S. D. WALTER

Department of Epidemiology and Public Health, Yale University, 60 College Street, New Haven, Connecticut 06510, U.S.A.

Summary

Various measures of attributable risk are discussed together with a rationale for their use as an alternative to relative risk in health research. Methods of estimation are presented for use with three important kinds of epidemiological study design with one dichotomous risk factor for a dichotomous disease outcome; the study designs are then compared with respect to efficiency.

Procedures to analyse confounded, polytomous and interacting risk factors are proposed and it is shown that there is a simple relationship between two distinct estimators previously suggested for use with deleterious and beneficial (or preventive) factors. Finally the relevance of attributable risk to an assessment of the potential effects of risk factor modification is discussed in the preventive medicine framework.

1. Introduction

The odds ratio as an approximation to the relative risk of disease in a group of people exposed to a certain risk factor, compared to those not exposed, has been widely used since its introduction by Cornfield [1951]. It is quite simple to show that the odds ratio is invariant across three important types of epidemiological study, namely retrospective, cross-sectional and prospective (defined in Section 2), and there is now considerable literature concerning its sample properties in these situations.

Although the relative risk and odds ratio parameters have enjoyed widespread use as measures of association in etiologic research, neither takes into account the actual *number* of cases of disease which might be related to a given exposure to risk. For example, exposure of industrial workers to various chemicals often entails a high relative risk for carcinoma of the lung, with rates sometimes as much as 40 to 50 times the rate of similar workers not so exposed (e.g., Doll [1959]); smoking, with much lower relative risks (typically of the order of five to ten, depending on the definition of exposure) is however responsible for many more cases of the disease, simply because the fraction of the population exposed is much larger, perhaps 50 to 90 percent (depending on the definition of smoking, e.g., Levin [1953]), compared to a fraction of one percent exposed to a particular industrial chemical. For this kind of reason, when examining diseases with several risk factors varying both in their relative risks and prevalences, it seems inadequate to compare the epidemiological importance of these factors using relative risks alone.

Epidemiologists and public health practitioners, being concerned with issues of health policy and priorities in the population as a whole, have suggested some measures of attributable risk which attempt to circumvent these limitations of relative risk, and some are discussed in this paper. Lilienfeld [1973] draws attention to the general importance of

Key Words: Attributable risk; Health research; Epidemiologic study designs; Prevented fraction; Risk modification.

attributable risk in health research, and some examples of its use are given by Miettinen [1974], who also commented on the current paucity of results concerning the sampling behavior of attributable risk estimators, despite their potential utility. Approximate confidence intervals for one measure of attributable risk are derived by Walter [1975] for a retrospective study with one dichotomous risk factor. In the present paper, estimation procedures for other more complex situations are given. Maximum likelihood estimators and their asymptotic variances are obtained for three types of study design, which are then compared with respect to efficiency. Procedures to analyse confounded, polytomous and interacting risk factors are proposed, and a simple relationship between the attributable risk parameters for deleterious and beneficial factors is derived. Numerical examples of some of the calculations are given for three risk factors for ischaemic heart disease.

2. Definition of Study Designs and Risk Parameters

2.1 Study Designs.

Initially we consider the simplest problem, with exactly one dichotomous risk factor (present or absent), and an outcome measure which is also dichotomous (for example, the existence or otherwise of a particular disease). Suppose that a particular sample of N individuals produces a 2×2 contingency table with cell and marginal frequencies as shown in Figure 1. Three common methods of sampling used to generate this table will be discussed here; they might best be described using an example. Suppose the association between smoking and ischaemic heart disease is under study. In design I, a sample of n_1 known "cases" of the disease is compared with a sample of n_2 "controls" without the disease in question with respect to their previous exposure to the risk factor (smoking); sampling according to an effect (the disease) and investigating its causes leads to the term "retrospective" for this design. In design II, groups of m_1 smokers and m_2 non-smokers, both free of the disease, are followed up and compared with respect to their subsequent incidence of disease; this scheme is "prospective" in character, sampling according to cause and investigating effects. Finally in design III, a "cross-sectional" scheme, an unstratified sample of N individuals is taken from the entire population, each person being dichotomized on the presence or absence of both the disease and the risk factor.

There is occasional confusion in the literature over the use of the words "retrospective"

		Outcome (disease) Positive	Negative	Total
Risk Factor	Present	a	b	m_1 = a+b
	Absent	c	d	m_2 = c+d
	Total	n_1 = a+c	n_2 = b+d	N = a+b+c+d

Figure 1

GENERAL 2×2 SAMPLE OUTCOME CONTINGENCY TABLE

and "prospective." This usually occurs when the sampling design of a study may be described either with respect to the disease and risk factor or to the chronology of the data collection. One may envisage, for example, a study in which groups of persons exposed and not exposed to a certain risk at some time in the past are assembled through the use of records, and by follow-up the subsequent development of disease in the two groups is established; this study would be essentially prospective in nature, although the samples of study individuals would be drawn chronologically "in retrospect." It should be emphasised that the designs considered in this paper are defined according to their sampling schemes with respect to cause and effect, as described above. For further discussion on terminology in this area, see Feinstein [1973].

Throughout most of this paper, it is supposed that designs *I* and *II* involve stratified sampling procedures. This is almost always the case in retrospective studies, but there are some prospective studies where the initial sample is unstratified and individuals with similar exposure to a risk factor are grouped subsequent to the act of sampling. In the example of the previous paragraph, this would occur if an unstratified sample of individuals from the entire population was split into smokers and non-smokers, and here m_1 and m_2 should be regarded as random variables. It turns out that the parameters and their sample properties, which may be obtained from such unstratified prospective studies, are identical with those from cross-sectional studies; unless stated otherwise, however, it should be assumed that m_1 and m_2 in design *II* are fixed, i.e., a stratified sample is used.

All three designs have various advantages and disadvantages concerning their operational feasibility and the reliability of their results, including the following. Design *I* is perhaps the most commonly used in health research, possibly because of the relative ease of gathering disease cases in a short space of time. On the other hand, the retrospective design commonly seems to suffer from biases related to the recall of previous exposure to risk factors; patients suffering from a serious disease may be more motivated than healthy controls to provide such information, and unless the information is obtained carefully, patients may tend to misstate their exposure levels. The prospective design, while comparatively free of such problems, is often more costly to mount because of the time period required to allow the disease to become manifest; unless the follow-up period is substantial or the disease is quite common, the small yield of cases of disease (as compared to a retrospective study of similar size) implies higher standard errors in the resulting risk estimates. Problems such as loss to follow-up and censoring must also be dealt with. Design *III* also often suffers from a shortage of cases in samples of moderate size; other problems are the bias induced by the loss of cases by death shortly after the onset of disease, modification of risk exposure (e.g., smoking) as a result of the disease and the operational difficulties of obtaining a representative sample of the entire population including current cases of disease. The cross-sectional design is, however, unique in that it provides estimates of disease and risk factor prevalences simultaneously.

A common sequence in the study of a particular risk factor and disease is for the suggestion of an association to be made from subjective or non-systematic observations (or occasionally from a cross-sectional survey), followed by several retrospective studies before the more costly prospective approach is used.

2.2 Risk Parameters.

Suppose the joint distribution of the disease and risk factor in the population is determined by the probabilities $\{\pi_{ij}; i, j = 1, 2\}$, where $i = 1$ indicates the presence of the risk factor and $i = 2$ the absence, and $j = 1$ and 2 indicate the presence or absence of the disease,

respectively, the π_{ij}'s thus corresponding to the cells of the contingency table in Figure 1. Only design *III* provides estimates of the π_{ij}'s directly. For design *I*, let the conditional probabilities of having been exposed to the risk factor be θ_1 and θ_2 for cases and controls, respectively, i.e. $\theta_1 = \pi_{11}/(\pi_{11} + \pi_{21})$ and $\theta_2 = \pi_{12}/(\pi_{12} + \pi_{22})$; these parameters are also known as the factor prevalence. For design *II*, let the conditional probabilities of developing the disease be ϕ_1 and ϕ_2 among the exposed and unexposed, i.e. $\phi_1 = \pi_{11}/(\pi_{11} + \pi_{12})$ and $\phi_2 = \pi_{21}/(\pi_{21} + \pi_{22})$; these parameters are thus the disease rates in the two groups. Finally we will require the overall rate of disease in the population (denoted by ϕ) and the overall prevalence of the risk factor (denoted by θ). Thus

$$\phi = \pi_{11} + \pi_{21} = \theta\phi_1 + (1 - \theta)\phi_2$$

and

$$\theta = \pi_{11} + \pi_{12} = \phi\theta_1 + (1 - \phi)\theta_2 .$$

The relative risk of exposed compared to non-exposed individuals is $\psi = \phi_1/\phi_2$, and we will assume initially that $\psi \geq 1$, i.e., that the factor is neutral or deleterious with respect to the disease. (Beneficial factors are considered in Section 6.) If the proportion developing the disease in the exposed and non-exposed groups are small (as is often the case in practice), then ψ may be approximated by the odds ratio $\psi' = \pi_{11}\pi_{22}/(\pi_{12}\pi_{21})$; the parameter ψ' may be estimated by $ad/(bc)$ for all three sampling schemes, while ψ is estimable directly from designs *II* and *III* only.

We now turn to the discussion of parameters concerned with the absolute number, rather than a relative number, of cases related to a particular risk factor. Berkson [1958] noted, for example, that the difference in incidence rates of coronary thrombosis for males between heavy smokers (with an age adjusted rate of 599 per 100,000 per year) and non-smokers (422 per 100,000 per year) was a little higher than the corresponding difference for lung cancer (166 versus 7 per 100,000 per year), whereas the relative risks were 1.42 and 23.7, respectively. Arguing that the higher relative risk for lung cancer was in part a consequence of its lower overall death rate (about one sixth that of coronary thrombosis), making a given number of deaths from lung cancer among smokers apparently six times more important than the same number of deaths from coronary thrombosis, he therefore proposed using the difference in rates

$$\delta = \phi_1 - \phi_2 \tag{1}$$

as a measure of the effect of smoking; δ is now known as Berkson's simple difference. Sheps [1959] proposed that

$$\phi_1 = \phi_2 + \delta'(1 - \phi_2) \tag{2}$$

would be a reasonable model for the comparison of the two groups, with δ' (equal to some constant, assumed positive) representing an additional factor applied to those individuals in the exposed group not otherwise destined to contract the disease; the term $\delta'(1 - \phi_2)$ is thus the component of the rate of disease among the exposed ascribed to the exposure *per se*. Equation (2) leads to

$$\delta' = (\phi_1 - \phi_2)/(1 - \phi_2) \tag{3}$$

with δ' known as Sheps' relative difference. If the common assumption that the disease rates are low is invoked, then $\delta \simeq \delta'$.

Two further parameters to be discussed concern the proportion, percentage, or number

of cases which may be attributed to the risk factor, termed attributable risk. MacMahon and Pugh [1970] consider the proportion of cases *among the exposed* attributed to the factor as follows: the excess risk among the exposed is $\phi_1 - \phi_2$, yielding an expected $n(\phi_1 - \phi_2)$ excess cases from a sample of n exposed persons. The total expected cases among the exposed is $n\phi_1$ and thus the proportion (γ) explained by the exposure is

$$\gamma = n(\phi_1 - \phi_2)/(n\phi_1) = (\psi - 1)/\psi. \tag{4}$$

Of greater interest is the parameter developed by Levin [1953] as the proportion of *all* cases attributable to the factor; this will be denoted by λ, and from (4) it may be seen that

$$\lambda = \frac{\theta\phi_1(\psi - 1)/\psi}{\theta\phi_1 + (1 - \theta)\phi_2} = \frac{\theta(\psi - 1)}{\theta(\psi - 1) + 1} \tag{5}$$

or equivalently that

$$\xi = 1 - \lambda = 1/[\theta(\psi - 1) + 1]. \tag{6}$$

The parameter λ has also been termed the etiologic fraction by Miettinen [1974]. There exist two indices used in other disciplines which have algebraic similarities to λ; these are Abbott's formula [1925] to adjust for the natural response rate (or placebo effect) in comparative bioassay experiments, and Cohen's index [1960] (or kappa statistic) used to measure agreement between nominal scales in psychology and education. Attributable risk does, however, present a number of special problems in its estimation not generally encountered in these other areas.

In the following sections, some uses and sample properties of the parameters δ, δ', γ and λ are discussed in the contexts of the three sample designs described earlier. The emphasis will be on the parameter λ which appears to be the most useful measure in general, being the only one of these quantities which reflects changes in *both* the magnitude of the risk factor effect (through ψ) and the risk factor prevalence (through θ).

3. Estimation of δ, δ' and γ

The parameters δ and δ' are functions of ϕ_1 and ϕ_2, the disease rates in the exposed and non-exposed groups, and thus are only estimable directly from designs *II* and *III*. In design *II*, δ may be estimated by

$$\hat{\delta} = a/m_1 - c/m_2$$

Regarding the two strata as independent gives

$$E(\hat{\delta}) = \delta; V(\hat{\delta}) = \phi_1(1 - \phi_1)/m_1 + \phi_2(1 - \phi_2)/m_2.$$

Kitagawa [1955] discusses the decomposition of δ-type parameters into components related to the main effects of risk factors and to confounding; the methods given are deterministic. Edwards [1957] also discusses δ in the context of fourfold tables with no fixed margins (corresponding to cross-sectional designs). The sample properties of $\hat{\delta}'$ may be investigated by noting that

$$\ln(1 - \delta') = \ln(1 - \phi_1) - \ln(1 - \phi_2).$$

There is a direct algebraic correspondence between δ' in prospective studies and λ in retrospective studies, and the results for the latter (discussed in Section 4) may be applied to δ' by a simple interchange of symbols.

The attributable risk among the exposed, γ, is a function of the relative risk ψ which may be estimated in designs *II* and *III*. If the incidence of disease is low, then ψ may be replaced by the odds ratio ψ', which is estimable from all three designs. Writing $1 - \gamma \simeq \psi'^{-1}$, the usual theory for the sample moments of $\hat{\psi}'$ may be used to derive the approximate moments of $\hat{\gamma}$. (For example, see Gart [1962], Gart and Thomas [1972]).

Occasionally a retrospective study will use an entire series of cases from a completely enumerated population from which the controls also originate. In this fortunate (but rather rare) situation, estimates of ϕ_1 and ϕ_2 may be obtained, giving in turn estimates of δ and δ'. An example of this kind of study is given by Cole and MacMahon [1971]; the estimation of relative risk in such "hybrid" study designs has been discussed recently by Kupper, McMichael and Spirtas [1975].

4. Estimation of λ *in* 2×2 *Tables*

In order to estimate λ, different approaches must be adopted for the three study designs so that the particular type of information that each provides may be utilised. Estimation methods will first be presented for the 2×2 tables generated in each design, which are then compared with respect to efficiency.

The retrospective design yields estimates of θ_1 and θ_2 ; in order to estimate λ in this case, additional information on the value of ϕ is required. Frequently the value of ϕ may not be known explicitly, but may reasonably be assumed to be small; if this is done, then the control group may be regarded as approximately representative of the total population with respect to the prevalence of the risk factor. It may be shown that

$$\xi = 1 - \lambda = (1 - \theta_1)/(1 - \theta)$$

and so replacing θ by θ_2 , we have in this case that

$$\xi = 1 - \lambda \simeq (1 - \theta_1)/(1 - \theta_2). \tag{7}$$

Walter [1975] has shown that the estimator $\ln \hat{\xi}'$, where

$$\hat{\xi}' = (c + 1/2)(n_2 + 1/2)/[(d + 1/2)(n_1 + 1/2)]$$

is unbiased for $\ln \xi$ except for terms $0(n_1^{-2})$ and $0(n_2^{-2})$; because zero values of c and d occur with non-zero probability in a finite sample, the addition of the 1/2's also avoids the possibility of using the logarithm of zero. The asymptotic variance of $\ln \hat{\xi}'$ is

$$V(\ln \hat{\xi}') = \frac{\theta_1}{(1 - \theta_1)n_1} + \frac{\theta_2}{(1 - \theta_2)n_2}. \tag{8}$$

To estimate the variance, a device of Goodman [1964] may be employed to again avoid the problems arising from zero values of c and d[1]. Taking only the first term in (8), one may consider $a/(c + 1)$ as an estimator of $\theta_1/(1 - \theta_1)$. Then

$$\begin{aligned} E[a/(c + 1)] &= \sum_{x=0}^{n_1} \frac{x}{n_1 - x + 1} \binom{n_1}{x} \theta_1^x (1 - \theta_1)^{n_1 - x} \\ &= \theta_1(1 - \theta_1^{n_1})/(1 - \theta_1). \end{aligned}$$

Thus $a/(c + 1)$ is a negatively biased estimator of $\theta_1/(1 - \theta_1)$, in which the bias is small for large n_1 and/or small θ_1 . The corresponding estimator of $V(\ln \hat{\xi}')$ is

[1] I am grateful to a referee for pointing this out.

$$\hat{V}(\ln \hat{\xi}') = \frac{a}{(c+1)n_1} + \frac{b}{(d+1)n_2}$$

which takes finite values even when c or d is zero.

Even if ϕ is not assumed to be small, often vital statistics will provide an estimate with high precision. In particular, if the value of ϕ is regarded as a known constant, then a maximum likelihood estimator of λ may be obtained with its asymptotic variance as derived in the appendix. The estimator is given by

$$(1 - \hat{\lambda}) = n_1 c / \{\phi(n_1 c - n_2 a)\}$$

with variance given by (A2). Note that if $\phi \ll 1$, and using equation (7), then (A2) reduces to an equivalent form of (8).

At this stage the similarity between (7) and (3) should be noticed. It follows that an estimator of Sheps' relative difference δ' may be obtained in a prospective study from

$$1 - \hat{\delta}' = (b + 1/2)(m_2 + 1/2)/[(d + 1/2)(m_1 + 1/2)].$$

Then $\ln \hat{\delta}'$, as an estimator of $\ln \delta'$, is unbiased except for terms $0({m_1}^{-2})$, $0({m_2}^{-2})$, and

$$V(\hat{\delta}') = \frac{\phi_1}{(1 - \phi_1)m_1} + \frac{\phi_2}{(1 - \phi_2)m_2}$$

$$\simeq \frac{\phi_1}{m_1} + \frac{\phi_2}{m_2}$$

for diseases with low incidence rates. Sheps [1959] gives a maximum likelihood estimation approach for δ'.

Turning now to the prospective design, we have estimates of ϕ_1 and ϕ_2 from the study. The maximum likelihood approach requires additional knowledge of θ. As discussed earlier, it will often be the case that many retrospective studies are carried out before the completion of a more expensive and difficult prospective investigation, and thus the previous retrospective work may provide a good estimate of θ for use in the prospective analysis. For simplicity, in the appendix, θ is taken as a known constant, and the maximum likelihood estimator derived; any reasonable prior distribution for θ might be adopted, however. From (A4), and substituting $\hat{\phi}_1 = a/m_1$, $\hat{\phi}_2 = c/m_2$, we have

$$1 - \hat{\lambda} = m_1 c / \{\theta(m_2 a - m_1 c) + m_1 c\}$$

with variance given by (A5). Recall that in the special case of an unstratified prospective design, an estimate $\hat{\theta}$ is available from the study itself; the estimation of λ here proceeds as in a cross-sectional study.

Finally, in the cross-sectional study, estimates of all the relevant parameters are available. The maximum likelihood estimator of λ is derived by substituting $\hat{\theta} = (a + b)/N$ and $\hat{\psi} = am_2/(cm_1)$ into (5); the asymptotic variance, derived in the appendix, is given by (A7).

A summary of the sampling methods used and the parameters estimable from each of the designs discussed is given in Table 1.

4.1 Comparison of Study Efficiencies.

It is of interest to investigate which of the three study designs yields the most efficient estimator of λ. Consider the null variances of the three estimators, given in (A3), (A6) and (A9); comparing first the cross-sectional with the retrospective design, we have that

Table 1

SAMPLING SCHEMES AND PARAMETERS ESTIMATED FOR THREE HEALTH RESEARCH DESIGNS

	I Retrospective	II Prospective	III Cross-sectional
Sampling Scheme	Stratified sample of n_1 cases and n_2 controls[1]	Stratified sample of m_1 exposed and m_2 non-exposed persons	Unstratified sample of N persons from entire population
Parameters estimated	(i) Factor prevalences among cases and controls, θ_1 and θ_2	(i) Disease rates among exposed and non-exposed ϕ_1 and ϕ_2	(i) Probabilities in joint distribution of disease and risk factor π_{11}, π_{12}, π_{21} and π_{22}
		(ii) Relative risk $\psi = \phi_1/\phi_2$	(ii) Relative risk $[\psi = \pi_{11}(\pi_{21}+\pi_{22})/[(\pi_{11}+\pi_{12})\pi_{21}]$
	(ii) Odds ratio $\psi' = \theta_1(1-\theta_2)/[\theta_2(1-\theta_1)]$	(iii) Odds ratio $\psi' = \phi_1(1-\phi_2)/[\phi_2(1-\phi_1)]$	(iii) Odds ratio $\psi' = \pi_{11}\pi_{22}/(\pi_{12}\pi_{21})$
		(iv) Berkson's simple difference $\delta = \phi_1 - \phi_2$	(iv) Berkson's simple difference $\delta = \frac{\pi_{11}}{(\pi_{11}+\pi_{12})} - \frac{\pi_{21}}{(\pi_{21}+\pi_{22})}$
		(v) Sheps' relative difference $\delta' = (\phi_1 - \phi_2)/[1-\phi_2]$	(v) Sheps' relative difference $\delta' = \delta(\pi_{21}+\pi_{22})/\pi_{22}$
	(iii) Attributable risk among exposed $\gamma = (\psi' - 1)/\psi'$	(vi) Attributable risk among exposed $\gamma = (\psi-1)/\psi$	(vi) Attributable risk among exposed $\gamma = (\psi-1)/\psi$
	(iv) Attributable risk: λ as in (7) if ϕ small, or as in (A1) if ϕ known	(vii) Attributable risk (if θ known) $\lambda = \theta(\psi-1)/[\theta(\psi-1)+1]$	(vii) Attributable risk $\lambda = 1-[(\pi_{11}+\pi_{12})(\psi-1)+1]^{-1}$

$$\sigma_{III}^2/\sigma_I^2 = \{4\phi(1-\phi)\}^{-1} \geq 1. \quad (9)$$

Considering only non-trivial cases where $\theta \neq 0, 1$ and $\phi \neq 0, 1$, equality occurs in (9) if and only if $\phi = 1/2$. Similarly it may be shown that $\sigma_{III}^2 = \sigma_{II}^2$ if and only if $\theta = 1/2$. Thus designs *I* and *II* are always at least as efficient as design *III*. To compare designs *I* and *II*, note that

$$\sigma_{II}^2/\sigma_I^2 = \theta(1-\theta)/\{\phi(1-\phi)\},$$

which quantity is > 1 if $|\phi - 1/2| \geq |\theta - 1/2|$, with equality if $\phi(1-\phi) = \theta(1-\theta)$. Similarly $\sigma_{II}^2 \leq \sigma_I^2$ if $|\theta - 1/2| \geq |\phi - 1/2|$.

Thus the general conclusion in large samples is that one should choose between sampling schemes *I* and *II* such that the stratification is made on the attribute which is less well dichotomised. If either θ or ϕ is not equal to 1/2, then scheme *III* will be inferior to either *I* or *II* or both. Only when $\theta = \phi = 1/2$ does scheme *III* achieve the same Type *I* error rates as schemes *I* and *II*, and in this case all three are equally efficient. These recommendations, together with that of having equal strata when schemes *I* and *II* are used, correspond exactly to those given by Lehmann [1959] when testing for independence in a 2×2 table; Lehmann points out that the results were originally conjectured by Berkson and subsequently proved by Neyman. The choice between *I* and *II* made on the basis of the degree of dichotomisation is also valid in small samples (see Lehmann [1959], p. 76). In epidemiological research, it is usually true that the disease is the less well dichotomised attribute, which implies that retrospective studies would be preferred to prospective studies from considerations of efficiency.

The result that design *III* is almost always less efficient than *I* or *II* might be expected intuitively since both ϕ and θ are estimated in *III*, whereas in *I* and *II*, the estimation of λ is made conditionally on a fixed value of either ϕ or θ, thus removing one source of variation.

5. Standardisation, the Estimation of λ for Confounded and Interacting Risk Factors, and Some Numerical Examples

In this section some practical examples are given of some of the above calculations on λ in fourfold tables, together with some extensions to more complex problems. The data to be used come from Dick and Stone [1973], who present the results of a retrospective study of ischaemic heart disease; 146 men aged 30–69 years who had ischaemic heart disease (referred to as "cases") are compared to 283 male controls taken from the general population in the same age range with respect to three important risk factors. No matching of cases and controls was used. The risk factors, all dichotomous, were as follows: *A*—hyperlipoproteinaemia; *B*—smoking at least 15 cigarettes per day; *C*—diastolic blood pressure of 95 mm. Hg or more. Further operational details are given in the original paper.

Table 2 gives the distribution of cases and controls over all eight combinations of the three risk factors. The marginal 2×2 tables for each risk factor are shown in Table 3. It may be noted that *B* (smoking) has the highest attributable risk (34 percent) even though it has an odds ratio somewhat smaller than that of *A* (hyperlipoproteinaemia); this is because *B* is more than twice as prevalent as *A* (affecting 48 percent vs. 19 percent of the controls). In the absence of additional information on ϕ here, the estimates $\hat{\theta}$ come from the control groups (effectively assuming ϕ to be small). Confidence limits may be put on λ as described by Walter [1975]. The final row of Table 3 gives risk estimates when all three factors are considered together; the estimates formed in this way are typically rather imprecise because

Table 2

DISTRIBUTION OF CASES AND CONTROLS WITH RESPECT TO THREE RISK FACTORS FOR ISCHAEMIC HEART DISEASE

Factor A	Factor B	Factor C	Number of Cases	Number of Controls	Total
-	-	-	15	82	97
-	-	+	10	37	47
-	+	-	39	81	120
-	+	+	23	28	51
+	-	-	18	16	34
+	-	+	7	12	19
+	+	-	19	19	38
+	+	+	15	8	23
		Total	146	283	429

A: Hyperlipoproteinaemia; B: Smoking; C: High diastolic blood pressure.
+ denotes presence of factor, - denotes absence.

of the difficulty of obtaining a reasonable number of cases and controls exposed to none of the factors. (This of course becomes even more evident when more factors are considered.)

5.1 Standardisation.

It may sometimes be required to adjust for a nuisance risk factor which is confounded with a risk factor of interest. In the example discussed here, the age distributions of the cases and controls are quite different and this should be allowed for. The entry into the analysis of such a nuisance risk factor may sometimes be avoided by individual matching of cases and controls. The net benefit of matching in terms of the efficiency and validity of the study depends on the degree of confounding, and on the relationship between the nuisance factor and the disease outcome; for example, efficiency may suffer by matching if the nuisance factor is unrelated to the disease, and so-called overmatching occurs. In other situations

Table 3

MARGINAL FOURFOLD TABLES FOR THREE RISK FACTORS

Factor		Cases	Controls	$\hat{\psi}'$	$\hat{\theta}$(%)	$\hat{\lambda}$(%)
A	+	59	55	2.81	19	26
	-	87	228			
B	+	96	136	2.08	48	34
	-	50	147			
C	+	55	85	1.41	30	11
	-	91	198			
One or more factors	+	131	201	3.56	71	65
	-	15	82			

the practical difficulties of obtaining matched controls may outweigh the expected gain in efficiency (Billewicz [1965], McKinlay [1974]). For a discussion of the relative merits of individual matching see Worcester [1964], Cochran [1965], Miettinen [1968], Hardy and White [1971] and Rubin [1973].

Table 4 gives the distribution of cases and controls by age group and smoking status; also shown are the estimates of ψ', γ, θ and λ for each age group. It may be noted that the case group is somewhat older than the control group and this fact distorts the effects of risk factors (e.g., smoking) which are not independent of age. Denoting by λ_i the proportion of cases attributable to B in age group i, a standardized estimate of the proportion of all cases explained by B is defined as

$$\hat{\lambda}_s = \sum_i w_i \lambda_i / \sum_i w_i$$

where the weights w_i are appropriately chosen; in this example with no stratification on age in the selection of cases, it is appropriate to take w_i as the number of cases found in age group i. Then we may calculate $\hat{\lambda}_s = 0.3231$ a value slightly lower than the 34 percent obtained in Table 2 without regard to age.

5.2 Confounded Risk Factors.

It is often required to differentiate the effects of two or more risk factors which are correlated in their occurrences in the population under study. By extending a procedure used with relative risk measures (Miettinen [1972]), it is possible to decompose the attributable risk λ into components representing the effect of one risk factor of interest and the risk due to confounding with a second, confounded factor. Let sample frequencies in the fourfold table corresponding to level i of the confounded factor be denoted by suffix i, and the frequencies in the aggregate table (formed over all levels of the confounder) be denoted in the normal way without suffix. Then the crude estimate of λ, formed without regard to the confounding, is obtained from the aggregate table as

$$\hat{\lambda}_{\text{crude}} = 1 - cn_2/(dn_1).$$

Table 4

DISTRIBUTION OF CASES AND CONTROLS BY AGE AND SMOKING STATUS

	AGE GROUPS									
	30-39		40-49		50-59		60-69		Total	
Factor	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
B +	8	27	36	43	36	45	16	21	96	136
B -	2	34	15	41	19	37	14	35	50	147
Total	10	61	51	84	55	82	30	56	146	283
$\hat{\psi}'$	5.04		2.23		1.56		1.90		2.07	
$\hat{\gamma}$	0.80		0.56		0.36		0.48		0.52	
$\hat{\theta}$	0.44		0.51		0.55		0.38		0.48	
$\hat{\lambda}$	0.64		0.40		0.23		0.25		0.34	

Within the table for level i of the confounder, we may calculate, for example, a value $b_i^* = a_i d_i / c_i$, which substituted in the table in place of b_i (but with the other cell entries unchanged) would correspond to an odds ratio of one for the factor of interest. Then $b^* = \sum b_i^*$ is the value which would appear in the aggregate table on the null assumption that exposure to the factor of interest has no effect (in terms of the odds ratio) at any level of the confounder. Hence an estimate of the component of attributable risk due to confounding is given by

$$\hat{\lambda}_{\text{conf}} = 1 - cn_2^*/(dn_1)$$

where $n_2^* = b^* + d$; also an attributable risk estimate, adjusted for confounding, is given by the difference

$$\hat{\lambda}_{\text{adj}} = \hat{\lambda}_{\text{crude}} - \hat{\lambda}_{\text{conf}} = c(b^* - b)/(dn_1).$$

This equation is analogous to the multiplicative equivalent for relative risks (equation (4) of Miettinen [1972]). As an example, consider the confounding of factors B and C in Table 1. It might be required to estimate the proportion of cases attributable to high blood pressure (factor C) after adjusting for the effect of smoking (factor B) which is also known to affect blood pressure. Forming the two fourfold tables for each level of smoking, we have the data arranged as in Table 5.

In the B-(non-smokers) table, $b_1^* = 17 \times 98/33 = 50.5$; similarly $b_2^* = 65.5$ giving $b^* = 116.0$, $n_2^* = 314.0$, $\hat{\lambda}_{\text{conf}} = 0.0116$ and $\hat{\lambda}_{\text{adj}} = 0.1092 - 0.0116 = 0.0977$. The effect here of confounding is thus rather weak, with factor C still accounting for a small proportion of cases after adjustment for B.

It is possible to regard age as a confounded factor in these data because of the lack of individual case-control matching and the resulting difference in age distributions. If a value b_i^* is calculated for the tables corresponding to each age group in Table 4, it may be calculated that the component of the attributable risk due to smoking which is confounded with age is $\hat{\lambda}_{\text{conf}} = -0.1451$; hence $\hat{\lambda}_{\text{adj}} = 0.3407 + 0.1451 = 0.4858$, somewhat higher than the crude attributable risk of 34 percent. The negative sign in $\hat{\lambda}_{\text{conf}}$ arises from the underrepresentation of older aged controls (compared to cases); at the older ages the odds ratio (and hence also the attributable risk) is in general somewhat lower than at younger ages.

5.3 Interaction.

Attributable risk measures may be used to assess the interaction between two or more risk factors. First suppose, for simplicity, that there are two dichotomous risk factors;

Table 5

DISTRIBUTION OF CASES AND CONTROLS WITH RESPECT TO SMOKING (B) AND HIGH DIASTOLIC BLOOD PRESSURE (C)

		B −			B +	
		Cases	Controls		Cases	Controls
C	+	17	49	+	38	36
	−	33	98	−	58	100

let (i, j) indicate the stratum with factor 1 at level i and factor 2 at level j; further denote the presence of a factor by 1 and absence by 0. Denote by ψ_{ij} the relative risk for individuals in stratum (i, j), compared to the referent stratum (0, 0); thus $\psi_{00} = 1$. The proportion of cases in (i, j) explained by the factors present in the stratum is $(\psi_{ij} - 1)/\psi_{ij} = \gamma_{ij}$. If the two factors have independent biological actions, then we will have

$$(1 - \gamma_{11}) = (1 - \gamma_{10})(1 - \gamma_{01}) \tag{10}$$

or equivalently that $\psi_{11} = \psi_{10}\psi_{01}$. Thus an approximate test of statement (10) is given by using $\tau = \ln \hat{\psi}_{11}' - \ln \hat{\psi}_{10}' - \ln \hat{\psi}_{01}'$; on the null hypothesis of independent biological actions, $E(\tau) = 0$ and

$$V(\tau) = \mathrm{V}(\ln \hat{\psi}_{11}') + V(\ln \hat{\psi}_{10}') + V(\ln \hat{\psi}_{01}') - 2\,\mathrm{Cov}\,(\ln \hat{\psi}_{11}', \ln \hat{\psi}_{10}') - 2\,\mathrm{Cov}\,(\ln \hat{\psi}_{11}', \ln \hat{\psi}_{01}') + 2\,\mathrm{Cov}\,(\ln \hat{\psi}_{10}', \ln \hat{\psi}_{01}').$$

Let a_{ij}, b_{ij} be the number of cases and controls in stratum (i, j); all odds ratio estimates are calculated using the frequencies c, d in the referent stratum (0, 0), e.g., $\hat{\psi}_{10}' = a_{10}d/(b_{10}c)$. For a large sample

$$V(\ln \hat{\psi}_{ij}') = 1/a_{ij} + 1/b_{ij} + 1/c + 1/d$$

$$\mathrm{Cov}\,(\ln \hat{\psi}_{11}', \ln \hat{\psi}_{10}') = \mathrm{Cov}\,(\ln \hat{\psi}_{11}', \ln \hat{\psi}_{01}') = \mathrm{Cov}\,(\ln \hat{\psi}_{10}', \ln \hat{\psi}_{01}') = 1/c + 1/d.$$

Hence $V(\tau) = \sum_{i,j} (1/a_{ij} + 1/b_{ij}) + 1/c + 1/d$. Thus $\{\tau/(V(\tau))^{1/2}\}$ may be used as a test statistic with approximately a standard normal distribution.

As an example consider again the data of Table 5. Letting the first suffix correspond to factor B (smoking), and the second to factor C (high diastolic blood pressure), we have, for example $\hat{\psi}_{01}' = (17 \times 98)/(33 \times 49) = 1.03$; similarly $\hat{\psi}_{10}' = 1.72$ and $\hat{\psi}_{11}' = 3.13$, leading to $\tau = 0.57$, $V(\tau) = 0.20$, $z = 1.27$, indicating a non-significant synergy between these two factors. (It should be remembered that this calculation has ignored the differential effects of age in the two groups.) A similar test for departures from the additive model $\psi_{11} = \psi_{10} + \psi_{01} - 1$ is given by Rothman [1974] and a practical example using this model may be found in Rothman and Keller [1972].

A different approach to the testing of interactions has been presented by Miettinen [1974] who suggests that the statement

$$(1 - \lambda) = (1 - \lambda_1)(1 - \lambda_2) \tag{11}$$

is true if two factors, with attributable risks λ_1 and λ_2, and a combined attributable risk λ, are independent both statistically and biologically. Synergy, or positive interaction is indicated if

$$1 - \lambda < (1 - \lambda_1)(1 - \lambda_2)$$

and antagonism or negative interaction if

$$1 - \lambda > (1 - \lambda_1)(1 - \lambda_2).$$

It is not clear from Miettinen's paper, however, if λ is to be calculated from the group of persons exposed to both or either factor; similarly, there is an ambiguity as to whether λ_1, for example, refers to the groups exposed to factor 1 only, or to all individuals exposed to factor 1 (including those exposed to factor 2 also). It does appear that departures from

equality in (11) may be due either to dependent physical factor actions (better tested by (10) or the additive model equivalent) or equally well to a correlation in the factor distributions. For example, two completely confounded risk factors (i.e., factors always found together and never apart) would have equal marginal attributable risks (λ_1 , say), and a combined attributable risk also equal to λ_1 . Thus $(1 - \lambda_1)$ on the left-hand side of (11) would be compared to $(1 - \lambda_1)^2$ on the right-hand side; the departure from equality in (11) would in this case be entirely due to the correlation in the risk factor distributions. More usually, inequality would arise from a combination of statistical and biological dependencies, making significance tests formulated in this way rather difficult to interpret. (This is true however the parameters of (11) are defined.) The joint distribution of risk factors is, nevertheless, important when considering the potential impact of changes in exposure levels; a highly significant (statistically) biological synergy of risk factors would be of little practical import if the risk factors were rarely found in the same individuals.

There are obvious generalisations of the above test procedures to the case of several factors. As the number of factors increases, however, the chance that an individual will have some missing data also increases. The methods used here would require the deletion from the analysis of all the data from individuals with any missing information; the results may be invalid if the pattern of missing data is not a random one, and potential relationships of the occurrence of missing data to the outcome measure (disease status) should be investigated in the usual ways.

The examination of factors and their interactions individually (or at most in small sets) is a common strategy in health research, but in large data sets with many factors, it may sometimes be preferable to adopt a model such as the loglinear (discussed by Cox [1970], Ku and Kullback [1974], Bishop, Fienberg and Holland [1974]) on the multidimensional contingency table so that as many main effects and interactions as are required may be estimated simultaneously.

6. *Estimation of* λ *for Protective and Polytomous Factors*

6.1 Protective Factors.

In the previous sections, the risk factors under consideration were assumed to be deleterious. When studying factors associated with decreased risk, it is of course quite possible to adopt the definition of attributable risk given by (5), using a value of ψ less than unity. Recently Miettinen [1974] has suggested an alternative definition of attributable risk or "prevented fraction" for beneficial factors, being the proportion of the *potential* disease experience prevented by the factor and/or other associated factors. Thus the risk "denominator," being the existing disease load for deleterious factors, is replaced by a hypothetical disease load which would exist in the absence of a protective factor. The derivation of a suitable parameter in this case is as follows.

The disease rates ϕ_1 and ϕ_2 here correspond to groups of individuals respectively protected and not protected by the factor; thus $\phi_1 \leq \phi_2$. If θ is the factor prevalence, then the disease rate for the whole population is $\theta\phi_1 + (1 - \theta)\phi_2$. The risk if the factor were not present would be ϕ_2 , and thus the prevented fraction of the risk is

$$\lambda_p = \{\phi_2 - [\theta\phi_1 + (1 - \theta)\phi_2]\}/\phi_2 = \theta(1 - \psi).$$

This gives $1 - \lambda_p = (1 - \lambda)^{-1}$, where λ is defined as in (5); if $\phi \ll 1$, then in a retrospective study we may take as usual $\hat{\theta} = b/n_2$, $\hat{\psi}' = ad/bc$, and hence

$$\hat{\xi}_p = 1 - \hat{\lambda}_p = dn_1/cn_2 = \hat{\xi}^{-1}.$$

This simple relationship may easily be exploited to develop the sample properties of $\hat{\lambda}_p$ from the corresponding theory of Section 4 for deleterious factors.

6.2 Polytomous Risk Factors.

Suppose now that a single risk factor is found at $(k + 1)$ distinct exposure levels; we let level 0 be the referent and presume for the moment that all other levels are associated with increased risk. Let the number of cases and controls at level i be a_i and b_i, respectively, and the number in the referent exposure level be c and d. Also let θ_i be the proportion of the population exposed at level i, and κ_i the proportion of cases exposed at level i. Then the proportion of all cases which are found at level i and are attributable to the risk factor is $\lambda_i = \kappa_i(\psi_i - 1)/\psi_i$, where ψ_i is the relative risk for exposure at level i. Using the obvious estimators

$$\hat{\kappa}_i = a_i/n_1 ; \qquad \hat{\psi}_i' = a_i d/b_i c$$

we may obtain

$$\hat{\lambda}_i = (a_i d - b_i c)/(n_1 d).$$

It may be noted that

$$1 - \sum_{i=1}^{k} \hat{\lambda}_i = 1 - cn_2/(dn_1) = 1 - \hat{\lambda}$$

where $\hat{\lambda}$ is the estimate of attributable risk which would be derived from the marginal 2×2 table for the risk factor, not distinguishing between different exposure levels. An alternative, but equivalent definition of λ_i is given by

$$\lambda_i = \theta_i(\psi_i - 1) \Big/ \left\{1 + \sum_{i=1}^{k} \theta_i(\psi_i - 1)\right\}.$$

Subject to the factor-disease relationship being causal rather than merely a statistical association, it is of interest to examine the effect of changes in the distribution of exposure in the population. For example, we might require the estimated impact of a change in the maximum permissible exposure to some occupational hazard (e.g., X-radiation for radiographers or atmospheric dust levels in factories), an alteration of concentrations of potentially harmful food additives or the adoption of a new "safe" level of exposure (i.e., redefining the referent). Suppose that the exposure distribution following the changes is given by a set $\{\theta_i'; i = 0, 1 \cdots k\}$ such that $\sum \theta_i' = 1$. Then the proportional change in the disease load would be

$$\sum_{i=0}^{k} (\theta_i - \theta_i')\phi_i \Big/ \left(\sum_{i=0}^{k} \theta_i\phi_i\right) = \sum_{i=0}^{k} (\theta_i - \theta_i')\psi_i \Big/ \left(\sum_{i=0}^{k} \theta_i\psi_i\right). \tag{12}$$

Similarly the new attributable risk for the altered factor would be

$$\lambda' = \sum_{i=0}^{k} \theta_i'(\phi_i - \phi_0) \Big/ \left(\sum_{i=0}^{k} \theta_i'\phi_i\right) = \left\{\left(\sum_{i=0}^{k} \theta_i'\psi_i\right) - 1\right\} \Big/ \left(\sum_{i=0}^{k} \theta_i'\psi_i\right). \tag{13}$$

Note that, as in the analysis of confounded and interacting risk factors the referent level sample is important because it contributes to the estimates of attributable and relative risk at all exposure levels; stratification of the sample may be desirable in order to ensure referent groups of adequate size.

A special case in the use of formulae (12) and (13) arises if $\theta_i' = 0$, $i = 1, 2, \cdots k$, $\theta_0' = 1$, corresponding to a complete removal of the risk factor; if $\phi = \sum \theta_i \phi_i$ is the overall disease rate, then from (12) the proportional change in the disease load is $1 - \phi_0/\phi$, and from (13) the new attributable risk is zero.

Occasionally one might have a polytomous risk factor whose exposure levels include associated risks both higher and lower than that for the referent level. An example is blood pressure as a risk factor for heart disease; the referent might here be taken as a range of "normal" (or common) values, with values above and below this range representing higher and lower risks, respectively. Miettinen [1974] advocates adopting at all levels the same definition of attributable risk (i.e., that for a deleterious or protective factor) depending on whether the factor is overall deleterious or protective.

An example of the calculation of attributable risks for polytomous factors is given in Rothman and Keller [1972], who examine the effects of alcohol and tobacco, each at several exposure levels, on the risk of oral cancer.

7. *Discussion*

This paper proposes some procedures for the estimation of attributable risk in study situations commonly used in health research, and many of the methods have analogies in the theory of relative risk. As remarked previously, however, the interpretations of the two risk measures are quite different, and attributable risk should in no way be regarded as a substitute for relative risk, but rather as an alternative or additional dimension of health hazard appraisal. The identification of a particular risk factor with a high relative risk may yield important clues to the disease mechanism. On the other hand, if the same factor were found only rather rarely in the population, and hence had low attributable risk, it would be of little immediate interest to health administrators more concerned with preventive strategy than definitive etiology.

In this context of identifying risk factors important to the population as a whole (rather than to subpopulations exposed to fairly rare factors), attributable risk appears useful when attempting to estimate the effect of modifications in risk factor exposure. Equating attributable risk with the expected change in disease load following the reduction or complete removal of a risk factor is only valid, however, if the factor is indeed a causal agent, rather than an agent merely associated statistically with the disease. The establishment of causality in health problems is a large subject not to be discussed here, but the reader may usefully consult Susser [1973] for an overview of the topic.

Even if a definite causal link has been established between factor and disease, the estimate of attributable risk still may not represent the actual change in disease incidence which would occur as a result of risk modifications for a variety of reasons. First, it may be difficult to alter the level of exposure of one factor independently of other risk factors. For example, suppose inactivity were a causal risk factor in the development of heart disease; the introduction of a daily exercise program as a preventive measure might so alter lifestyle in other ways as to significantly change exposure to other factors, such as hypertension. Second, the disease entity itself may change as a result of induced changes in the environment, such as occurs when a new strain of virus develops following mass immunisation. Finally, we must also consider changes to be expected in other diseases by iatrogenesis; the administration of a protective or therapeutic drug for one disease may induce adverse side effects or other diseases more serious than the original problem.

Despite these limitations, it should be remembered that even if the causal mechanisms

are not clearly understood, modifications of environmental risks may still lead to reductions in disease rates. This happened, for example, with the discovery that the consumption of fresh fruit prevented scurvy, long before the present knowledge of vitamins. Some contemporary medical problems may also require this kind of solution, based on community observation rather than laboratory experimentation. Several current medical problem areas, notably heart disease and cancer, present a plethora of risk factors and associated options for preventive strategy. A rational choice between alternative programs of, for example, lifestyle intervention and environmental modification studies, must be based at least in part on an assessment of their *potential* benefits; it appears that attributable risk is a suitable measure in this context.

Finally, it may be noted that attributable risk type measures may be utilised in disciplines other than health. For example, suppose that university education is thought to be a precedent factor influencing subsequent career "success" (defined in some formal manner); it may be of interest to examine what proportion of successful individuals achieved their success as a result of university education. Groups of individuals with and without university education would correspond to persons exposed and not exposed to a risk factor in the health context, and the outcome measure (previously health status) is career success or lack of success. In situations such as this, where the concept of "risk" is not appropriate, the terms "attributable fraction" or "attributability" are suggested as preferable alternatives to attributable risk. Thus, in the example cited, one might speak of the attributability of success to university education, or equivalently the fraction of success attributable to university education. The former seems more concise, but on the other hand the term "prevented fraction" has already gained some acceptance in the health field for factors negatively associated with the outcome, and attributable fraction for positively associated factors is consistent with this terminology.

Acknowledgments

I am grateful to several members of Long Range Health Planning, Health and Welfare, Canada for many useful discussions on the practical implementation of the ideas presented here. My thanks are also due to the editor, an associate editor and the referees for helpful suggestions on an earlier draft of the work.

L'estimation et l'Interprétation du Risque Imputable en Recherche Médicale

Résumé

Diverses mesures du risque imputable sont discutées en même temps qu'un raisonnement justifiant son utilisation comme alternative à celle du risque relatif en recherche médicale. On présente des méthodes d'estimation sur trois sortes importantes de plan d'étude épidémiologique à facteur de risque dichotomique pour une manifestation dichotomique de la maladie; les plans d'études sont alors comparés en fonction de leur efficacité.

On propose des procédures pour analyses des facteurs de risque confondus, polytomiques et en interaction, et l'on montre qu'il existe une relation simple entre deux estimateurs suggérés précédemment pour les facteurs nuisibles et bénéfiques (ou préventifs).

Finalement, on discute l'intérêt du risque imputable pour l'évaluation des effets potentials de la modification du facteur de risque dans le cadre de la médecine préventive.

References

Abbott, W. S. [1925]. A method of computing the effectiveness of an insecticide. *Journal of Economic Entomology 18*, 265–7.

Billewicz, W. Z. [1965]. The efficiency of matched samples: an empirical investigation. *Biometrics 21*, 623–44.

Berkson, J. [1958]. Smoking and lung cancer. Some observations on two recent reports. *Journal of the American Statistical Association 53*, 28–38.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. [1975]. *Discrete Multivariate Analysis MIT.* Press, Boston.

Cochran, W. G. [1965]. The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A, 128*, 234–65.

Cohen, J. [1960]. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, 37–46.

Cole, P. and MacMahon, B. [1971]. Attributable risk percent in case-control studies. *British Journal of Preventive and Social Medicine 25*, 242–4.

Cornfield, J. [1951]. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute 11*, 1269–75.

Cox, D. R. [1970]. *The Analysis of Binary Data.* Methuen, London.

Dick, T. D. S. and Stone, M. C. [1973]. Prevalence of three cardinal risk factors in a random sample of men and in patients with ischaemic heart disease. *British Heart Journal 35*, 381–5.

Doll, R. [1959]. Occupational lung cancer: a review. *British Journal of Industrial Medicine 16*, 181–90.

Edwards, J. H. [1957]. A note on the practical interpretation of 2×2 tables. *British Journal of Preventive and Social Medicine 11*, 73–8.

Feinstein, A. R. [1973]. Clinical biostatistics, XX the epidemiologic trohoc, the ablative risk ratio, and "retrospective" research *Clinical Pharmacology and Therapeutics 14*, 291–307.

Gart, J. J. [1962]. Approximate confidence limits for the relative risk. *Journal of the Royal Statistical Society. Series B, 24*, 454–63.

Gart J. J. and Thomas, D. G. [1972]. Numerical results on approximate confidence limits for the odds ratio. *Journal of the Royal Statistical Society, Series B. 34*, 441–7.

Goodman, L. A. [1964]. Interactions in multidimensional contingency tables. *Annals of Mathematical Statistics 35*, 632–46.

Hardy, R. H. and White, C. [1971]. Matching in retrospective studies. *American Journal of Epidemiology 93*, 75–6.

Kitagawa, E. S. [1955]. Components of a difference between two rates. *Journal of the American Statistical Association 50*, 1168–94.

Ku, H. H. and Kullback, S. [1974]. Loglinear models in contingency table analysis. *The American Statistician 28*, 115–22.

Kupper, L. L , McMichael, A. J. and Spirtas, R. [1975]. A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association 70*, 524–8.

Lehmann, E. L. [1959]. *Testing Statistical Hypotheses.* Wiley, Inc., New York.

Levin M. L. [1953]. The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum 19*, 531–41.

Lilienfeld, A. M. [1973]. Epidemiology of infectious and non-infectious disease: some comparisons. *American Journal of Epidemiology 97*, 135–47.

MacMahon, B. and Pugh, T. F. [1970]. *Epidemiology: Principles and Methods.* Little, Brown and Co., Boston.

McKinlay, S. M. [1974]. The expected number of pairs and its variance for matched-pair designs. *Applied Statistics 23*, 372–83.

Miettinen, O. S. [1968]. The matched-pairs design in the case of all-or-none responses. *Biometrics 24*, 339–52.

Miettinen, O. S. [1972]. Components of the crude risk ratio. *American Journal of Epidemiology 96*, 168–72.

Miettinen, O. S. [1974]. Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology 99*, 325–32.

Rothman, K. J. [1974]. Synergy and antagonism in cause-effect relationships. *American Journal of Epidemiology 99*, 385–8.

Rothman, K. J. and Keller, A. [1972]. The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx. *Journal of Chronic Diseases 25*, 711–16.
Rubin, D. B. [1973]. Matching to remove bias in observational studies. *Biometrics 29*, 159–83.
Sheps, M. C. [1959]. An examination of some methods of comparing several rates or proportions. *Biometrics 15*, 87–97.
Susser, M. [1973]. *Causal Thinking in Health Sciences; Concepts and Strategies in Epidemiology*. Oxford University Press.
Walter, S. D. [1975]. The distribution of Levin's measure of attributable risk. *Biometrika 62*, 371–5.
Worcester, J. [1964]. Matched samples in epidemiologic studies. *Biometrics 20*, 840–8.

Received March 1975, *Revised February* 1976

Appendix

Maximum Likelihood Estimation of λ

(*a*) *Design I: Retrospective.*

The parameterisation required here is in terms of θ_1, θ_2 and ϕ, the last being taken as a known constant. Rearrangement of (5) gives

$$\lambda = \frac{(1 - \phi)(\theta_1 - \theta_2)}{1 - \theta_2 - \phi(\theta_1 - \theta_2)}. \tag{A1}$$

The log likelihood L is given by

$$L = \ln K + a \ln \theta_1 + c \ln (1 - \theta_1) + b \ln \theta_2 + d \ln (1 - \theta_2)$$

where K is some constant. From (A1), we have that

$$\theta_1 = \frac{\theta_2(1 - \lambda)(1 - \phi) + \lambda}{1 - \phi(1 - \lambda)}.$$

Differentiating L with respect to λ and θ_2, and using

$$\partial\theta_1/\partial\lambda = (1 - \phi)(1 - \theta_2)/\{1 - \phi(1 - \lambda)\}^2$$

$$\partial\theta_1/\partial\theta_2 = (1 - \phi)(1 - \lambda)/\{1 - \phi(1 - \lambda)\}$$

yields maximum likelihood estimates $\hat{\theta}_1 = a/n_1$, $\hat{\theta}_2 = b/n_2$, and in turn the maximum likelihood estimate $\hat{\lambda}$ after appropriate substitution of $\hat{\theta}_1$ and $\hat{\theta}_2$. Further differentiation, and noting that $E(a) = n_1\theta_1$, $E(b) = n_2\theta_2$, gives

$$E\left(-\frac{\partial^2 L}{\partial\lambda^2}\right) = \frac{n_1(1 - \phi)^2(1 - \theta_2)^2}{\{1 - \phi(1 - \lambda)\}^4\theta_1(1 - \theta_1)}$$

$$E\left(-\frac{\partial^2 L}{\partial\theta_2^2}\right) = \frac{n_1(1 - \phi)^2(1 - \lambda)^2}{\{1 - \phi(1 - \lambda)\}^2\theta_1(1 - \theta_1)} + \frac{n_2}{\theta_2(1 - \theta_2)}$$

$$E\left(-\frac{\partial^2 L}{\partial\lambda\,\partial\theta_2}\right) = \frac{n_1(1 - \phi)^2(1 - \lambda)(1 - \theta_2)}{\{1 - \phi(1 - \lambda)\}^3\theta_1(1 - \theta_1)}.$$

The determinant of the information matrix is

$$\frac{n_1n_2(1 - \phi)^2(1 - \theta_2)}{\{1 - \phi(1 - \lambda)\}^4\theta_1\theta_2(1 - \theta_1)}$$

and hence the asymptotic variance of $\hat{\lambda}$ is

$$V(\hat{\lambda}) = \sigma_I^2 = \frac{(1 - \phi)^2(1 - \lambda)^4(1 - \theta_2)^2}{(1 - \theta_1)^2}\left[\frac{\theta_1}{n_1(1 - \theta_1)} + \frac{\theta_2}{n_2(1 - \theta_2)}\right]. \tag{A2}$$

In the null case, where $\lambda = 0$ and $\theta_1 = \theta_2 = \theta$, the variance is minimised (for fixed N) when $n_1 = n_2 = N/2$; in this situation the null variance is

$$\sigma_I^2 = 4\theta(1 - \phi)^2/N(1 - \theta). \tag{A3}$$

(*b*) *Design II: Prospective.*

It is convenient here to recast λ as

$$\lambda = \frac{\theta(\phi_1 - \phi_2)}{\theta(\phi_1 - \phi_2) + \phi_2}. \tag{A4}$$

The log likelihood is now

$$L = \ln K + a \ln \phi_1 + b \ln (1 - \phi_1) + c \ln \phi_2 + d \ln (1 - \phi_2).$$

Similarly to the method for retrospective designs, ϕ_1 may be expressed in terms of λ and ϕ_2 ; regarding θ as a known constant, the maximum likelihood estimates in this case are $\hat{\phi}_1 = a/m_1$, $\hat{\phi}_2 = c/m_2$. Also

$$E\left(-\frac{\partial^2 L}{\partial \lambda^2}\right) = \frac{m_1 \phi_2^2}{\theta^2(1 - \lambda)^4 \phi_1 (1 - \theta_1)}$$

$$E\left(-\frac{\partial^2 L}{\partial \phi_2^2}\right) = \left[\frac{m_1 \phi_1}{(1 - \phi_1)} + \frac{m_2 \phi_2}{(1 - \phi_2)}\right]/\phi_2^2$$

$$E\left(-\frac{\partial^2 L}{\partial \lambda \, \partial \theta_2}\right) = \frac{m_1}{\theta(1 - \lambda)^2(1 - \phi_1)}$$

giving

$$V(\hat{\lambda}) = \sigma_{II}^2 = \frac{\theta^2(1 - \lambda)^4 \phi_1^2}{\phi_2^2}\left[\frac{(1 - \phi_1)}{\phi_1 m_1} + \frac{(1 - \phi_2)}{\phi_2 m_2}\right] \tag{A5}$$

with a corresponding null variance for $m_1 = m_2 = N/2$

$$\sigma_{II}^2 = 4\theta^2(1 - \phi)/N\phi. \tag{A6}$$

(*c*) *Design III: Cross-sectional.*

The log likelihood is that of the appropriate multinomial distribution, i.e.,

$$L = \ln K + a \ln \pi_{11} + b \ln \pi_{12} + c \ln \pi_{21} + d \ln \pi_{22} .$$

One may write $\pi_{12} = 1 - \pi_{11} - \pi_{21}/f$ and $\pi_{22} = -\pi_{12} + \pi_{21}/f$, where $f = (\pi_{11} + \pi_{21})(1 - \lambda)$, in order to express the likelihood in terms of π_{11}, π_{21} and λ. The maximum likelihood estimates of the π_{ij}'s are the sample equivalents (e.g. $\hat{\pi}_{11} = a/N$, etc), and the algebraically distinct elements of the information matrix are given by

$$N^{-1}E\left(-\frac{\partial^2 L}{\partial \lambda^2}\right) = \frac{\pi_{21}^2(\pi_{12} + \pi_{22})}{f^2(1 - \lambda)^2 \pi_{12}\pi_{22}}$$

$$N^{-1}E\left(-\frac{\partial^2 L}{\partial \pi_{11}^2}\right) = \frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} - \frac{2\pi_{21}}{f\lambda\pi_{12}} + \frac{\pi_{21}^2(\pi_{12} + \pi_{22})}{f^2\lambda^2\pi_{12}\pi_{22}}$$

$$N^{-1}E\left(-\frac{\partial^2 L}{\partial \pi_{21}^2}\right) = \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}} - \frac{2\pi_{11}}{f\lambda\pi_{22}} + \frac{\pi_{11}^2(\pi_{12} + \pi_{22})}{f^2\lambda^2\pi_{12}\pi_{22}}$$

$$N^{-1}E\left(-\frac{\partial^2 L}{\partial \lambda \, \partial \pi_{11}}\right) = \frac{\pi_{21}}{f(1 - \lambda)\pi_{12}} - \frac{\pi_{21}^2(\pi_{12} + \pi_{22})}{f^2\lambda(1 - \lambda)\pi_{12}\pi_{22}}$$

$$N^{-1}E\left(-\frac{\partial^2 L}{\partial \lambda \, \partial \pi_{21}}\right) = \frac{-\pi_{21}}{f(1 - \lambda)\pi_{22}} + \frac{\pi_{11}\pi_{22}(\pi_{12} + \pi_{22})}{f^2\lambda(1 - \lambda)\pi_{12}\pi_{22}}$$

$$N^{-1}E\left(-\frac{\partial^2 L}{\partial \pi_{11} \, \partial \pi_{21}}\right) = \frac{(\pi_{11}\pi_{22} + \pi_{12}\pi_{21})}{f\lambda\pi_{12}\pi_{22}} - \frac{\pi_{11}\pi_{12}(\pi_{12} + \pi_{22})}{f^2\lambda^2\pi_{12}\pi_{22}}.$$

The determinant of the information matrix, after simplification, is

$$\frac{(\pi_{21} + \pi_{22})^4 N^3}{(\pi_{11} + \pi_{21})^2 \pi_{11}\pi_{12}\pi_{21}{}^3\pi_{22}}$$

and

$$V(\hat{\lambda}) = \sigma_{III}{}^2 = \frac{(1 - \lambda)^4(\pi_{11} + \pi_{21})(\pi_{21} + \pi_{22})\{\pi_{21}(\pi_{12}\pi_{21} - \pi_{11}\pi_{22}) + \pi_{11}\pi_{22}\}}{N\pi_{21}{}^3}. \tag{A7}$$

In the null case $\lambda = 0$, and $\pi_{12}\pi_{21} = \pi_{11}\pi_{22}$, giving a null variance

$$\sigma_{III}{}^2 = \frac{(\pi_{11} + \pi_{21})(\pi_{21} + \pi_{22})\pi_{11}\pi_{22}}{N\pi_{21}{}^3}. \tag{A8}$$

Noting that in the null situation $\pi_{11} = \theta\phi$, $\pi_{12} = \theta(1 - \phi)$, and so on, (A8) reduces to

$$\sigma_{III}{}^2 = \theta(1 - \phi)/\{N\phi(1 - \theta)\}. \tag{A9}$$