Epidemiologic Reviews
Copyright © 1994 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol. 16, No. 1
Printed in U.S.A.

Attributable Risk Estimation in Case-Control Studies

Steven S. Coughlin,[1] Jacques Benichou,[2] and Douglas L. Weed[3]

INTRODUCTION

The attributable risk, which has also been referred to as the etiologic fraction (1), attributable fraction (2), excess fraction (3), and population attributable risk percent (4), provides an estimate of the proportion of cases that are related to a given exposure. It is usually interpreted as the fraction of disease in a population that might be avoided by reducing or eliminating exposure to an etiologic agent, provided that it is causative (5). In contrast to the odds ratio and risk ratio, the attributable risk takes the number of exposed individuals in the population into account. The information gained from attributable risk estimates may contribute to the planning of public health programs that must choose between alternative disease prevention strategies (5–7). For example, a program that attempts to eliminate a rare exposure might prevent only a small fraction of cases in the target population, even when the odds of disease among exposed individuals are relatively great. On the other hand, elimination of a common exposure might prevent a much larger fraction of cases (assuming that the association is causal), even when the magnitude of the association is weak.

Statistical methods for estimating the attributable risk have undergone considerable development and refinement since Levin's measure of attributable risk was first proposed in 1953 (8). The pace of methodological advances has been particularly rapid over the past decade. Although adjustment procedures of potential interest to epidemiologists have been reviewed by Walter (9), Whittemore (10), and Benichou (11), existing reviews either predate recent developments in attributable risk estimation or were intended for those well-versed in statistical theory. This pragmatically oriented review is intended for a broad audience of health researchers. Methods of estimating the attributable risk from matched or unmatched case-control data are considered, including the weighted sum and Mantel-Haenszel approaches, along with more general regression methods for obtaining adjusted estimates of the attributable risk. The assumptions and limitations of alternative adjustment procedures will be discussed, along with special applications of the attributable risk in the design and interpretation of case-control studies.

UNADJUSTED ESTIMATION

The attributable risk (AR) is defined as

$$\mathrm{AR} = [P(D) - P(D \backslash \bar{E})]/P(D), \qquad (1)$$

where $P(D)$ is the probability of disease in the population and $P(D \backslash \bar{E})$ is the hypothetical probability of disease in that same population with the exposure eliminated (8, 11).

Received for publication June 11, 1993, and in final form February 1, 1994.

Abbreviations: AR, attributable risk; OR, odds ratio; PF, prevented fraction; RR, relative risk; SD, standard deviation.

[1] Department of Biostatistics and Epidemiology, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA.

[2] Biostatistics Branch, Division of Cancer Etiology, National Cancer Institute, Bethesda, MD.

[3] Preventive Oncology Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.

Reprint requests to Dr. Steven S. Coughlin, Department of Biostatistics and Epidemiology, Tulane University School of Public Health and Tropical Medicine, 1430 Tulane Avenue, New Orleans, LA 70112.

Estimating the attributable risk assumes that the exposure is causative (12). When no attempt is made to adjust for confounding variables, equation 1 may be rewritten as

$$\mathrm{AR} = \frac{P(E)(\mathrm{RR} - 1)}{1 + P(E)(\mathrm{RR} - 1)}, \qquad (2)$$

where $P(E)$ is the proportion of the population exposed to the factor and RR is the risk ratio. In estimating the attributable risk from case-control data, the RR can be replaced by the odds ratio and $P(E)$ can be replaced by $P(E\backslash\bar{D})$, the prevalence of exposure in the nondiseased population, as long as the rare disease assumption is valid (5, 11). If the controls represent a random sample of the population from which the cases have been obtained, $P(E)$ may be estimated using the exposure data from the control group. However, if the cases and controls have been matched with regard to one or more factors or if the controls have been selected by stratified random sampling, the estimate of the proportion of the population exposed to the factor may be biased. The alternative (and equivalent) formula

$$\mathrm{AR} = P(E\backslash D)\frac{\mathrm{RR} - 1}{\mathrm{RR}}, \qquad (3)$$

where $P(E\backslash D)$ is the proportion of cases exposed to the factor, must be used instead (13, 14).

Several alternative formulas for obtaining confidence intervals for the attributable risk have been proposed (15, 16), as shown in Appendix 1.

Example 1: Unmatched data and the impact on the attributable risk of changing the baseline level of exposure. To illustrate these estimation procedures for unmatched data, Benichou (11) analyzed data from a previously reported case-control study of esophageal cancer in which the controls had been randomly sampled from the general population (17). The crude attributable risk for alcohol consumption was estimated to be 0.395 (standard deviation (SD) = 0.042) when 0–79 g/day was chosen as the baseline level of alcohol consumption (11). When the baseline exposure level was defined to be 0–39 g/day, however, the attributable risk for alcohol consumption increased to 0.709 (SD = 0.051). The alternative definition of the baseline level results in a higher proportion of exposed individuals and a slightly larger odds ratio, and a corresponding increase in the attributable risk estimate (11). The interpretation of the attributable risk estimate is different, however. With the alternative definition of the baseline exposure level, an estimated 70.9 percent of the cases in the population could be prevented by eliminating alcohol consumption above 39 g/day (instead of 79 g/day). As this example illustrates, the point estimate of the attributable risk is greatly affected by the definition of the baseline level of exposure in estimating the risk attributable to exposures that have multiple levels or that are continuous (11, 18).

Polychotomous risk factors

The concept of attributable risk has been extended to the situation in which the exposure is polychotomous (polytomous) and estimates of attributable risk at specified levels of exposure are of interest (9, 18, 19). Such estimates may have important policy implications for screening groups at highest risk of disease. When the case-control data are in the form of a 2 × $(I + 1)$ table, the numbers of cases and controls at exposure level i can be represented by n_i and m_i, respectively, where n_0 and m_0 are the respective numbers in the referent exposure level (19). For the ith level of exposure relative to the baseline level ($i = 0$), the odds ratio (OR) is estimated by $\mathrm{OR}_i = (n_i m_0 / n_0 m_i)$. Let n denote the total number of cases and m the total number of controls. If the proportion of controls exposed at the ith level ($p_i = m_i/m$) is used as an estimate of the proportion of the population exposed at that same level, then the risk attributable to the exposure at level i is

$$\mathrm{AR}_i = p_i(\mathrm{OR}_i - 1) \div [1 + \textstyle\sum p_i(\mathrm{OR}_i - 1)], \qquad (4)$$

which can be expressed as

$$\mathrm{AR}_i = (n_i m_0 - m_i n_0)/(n m_0). \qquad (5)$$

Confidence intervals for AR_i have been provided by Denman and Schlesselman (19).

In the esophageal cancer example (example 1), alcohol consumption was a polychotomous factor with four levels ($I + 1 = 4$), namely 0–39, 40–79, 80–119, and ≥120 g/day. We defined the baseline level as the lowest level (0–39 g/day) and estimated the attributable risk for each of the remaining three levels. We obtained the following estimates: 0.270 (SD = 0.044), 0.222 (SD = 0.033), and 0.217 (SD = 0.030) for alcohol consumptions of 40–79, 80–119, and ≥120 g/day, respectively. Note that these three attributable risk estimates sum to the overall attributable risk estimate for alcohol consumption of 0.709, as predicted by the theory for mutually exclusive levels of exposure (9, 11). The interpretation of these results is that by selectively targeting the heavy drinkers (consumption of ≥120 g/day) in the population and by reducing their consumption to less than 40 g/day, one could expect to prevent 22 percent of new cases. On the other hand, by selectively targeting the "moderate" drinkers (consumption of 40–79 g/day), one could expect to prevent 27 percent of new cases.

Effects of misclassification

The effects of nondifferential misclassification of exposure status on unadjusted estimates of the attributable risk were considered by Walter (20) and were more fully studied by Hsieh and Walter (21). Walter (20) noted that the attributable risk has a "canceling feature" in that misclassification sometimes results in compensatory effects. For example, reduced specificity biases the odds ratio toward the null, but exposure prevalence increases (20, 21). In general, the effect of nondifferential exposure misclassification is to bias the attributable risk estimate toward zero (20). It can also be shown that the sensitivity of the exposure classification has a greater influence on the magnitude of the bias than the specificity (21). For a perfectly sensitive exposure classification (sensitivity = 1), there is no bias due to exposure misclassification, even if the specificity is low (21). In addition, the bias caused by the misclassification of exposure status is larger when the prevalence of exposure is high (21). In view of this, it may be particularly important to minimize exposure misclassification when estimating the risk attributable to relatively common exposures and to ensure that exposed individuals not be classified as unexposed (21).

Prevented fraction

The attributable risk does not allow estimation of the fraction of disease in a population that is prevented by a protective exposure (18, 21). Miettinen (18) proposed an alternative index, the prevented fraction, for use when the disease rate is highest in the reference category (i.e., when the odds ratio is less than 1). The prevented fraction (PF) is defined as

$$\mathrm{PF} = [P(D\backslash\bar{E}) - P(D)]/P(D\backslash\bar{E}). \qquad (6)$$

It can be rewritten as

$$\mathrm{PF} = P(E)(1 - \mathrm{RR}), \qquad (7)$$

or equivalently as

$$\mathrm{PF} = \frac{P(E\backslash D)}{\mathrm{RR}[1 - P(E\backslash D)] + P(E\backslash D)} \times (1 - \mathrm{RR}). \qquad (8)$$

ADJUSTED METHODS OF ESTIMATION BASED ON STRATIFICATION

Estimates of attributable risk that are unadjusted for factors other than the exposure of interest may be biased, even if adjusted estimates of risk ratios are used in equation 2 (10). As a result, it is often desirable to adjust for the effects of such confounding factors. Walter (22) and Whittemore (5, 10) have identified sufficient conditions for a lack of bias due to an extraneous factor when estimating the attributable risk: Either 1) the factor is not correlated with disease status among unexposed individuals or 2) the factor is not correlated with exposure in the overall population. Since a true confounding factor meets neither of these two conditions, confounding factors should be adjusted for when estimating attributable risk (10). Walter (9) and Whittemore (5, 10) have extended Levin's measure of attributable risk using a weighted sum approach to account for such confounding. Alternative procedures have been proposed by Kuritz and Landis (13, 14) and Greenland (12) which make use of the Mantel-Haenszel estimator of the common odds ratio. These procedures allow one to estimate the public health impact of removing the exposure from the population while adjusting for the effects of other confounding factors.

The weighted sum approach

The adjusted measure of attributable risk proposed by Walter (9) and further developed by Whittemore (5, 10) consists of a weighted sum of the attributable risks (AR_j) formed over J strata defined by one or more polychotomous adjustment factors:

$$AR = \sum_{j=1}^{J} w_j AR_j, \qquad (9)$$

where w_j is the weight corresponding to level j. The weights (w_j) may be defined simply as the proportion of cases in level j. Alternatively, Ejigou (23) proposed to weight the individual attributable risks within each level of the adjustment factors by the inverse of their variance. However, this alternative choice of weighting can be shown to be inconsistent (11). In contrast to methods of adjustment based on Mantel-Haenszel procedures (to be discussed below), the weighted sum approach does not require the assumption of a common odds ratio, and it allows for the control of both confounding and effect modification (11). However, substantial bias can occur if the data are sparse and the probability of exposure is high (5). As a result, the weighted sum approach is inappropriate for case-control data that are individually matched, but it may be used when the controls have been frequency-matched (11). Confidence intervals for this adjusted measure of attributable risk (5) are shown in Appendix 2.

In the esophageal cancer example provided by Benichou (example 1), the weighted sum approach was used to obtain adjusted estimates of the risk attributable to alcohol consumption, taking into account potentially confounding variables (11). After adjustment was made for cigarette smoking and age using nine smoking × age strata, the risk of esophageal cancer attributable to alcohol consumption was estimated to be 0.380 (SD = 0.045) when 0–79 g/day was chosen as the baseline level of alcohol consumption. The adjusted attributable risk estimate obtained in this way is somewhat lower than the crude estimate (0.395, SD = 0.042), because both smoking and age influence risk of esophageal cancer. In order to obtain confidence intervals for the adjusted estimate of attributable risk (Appendix 2), Benichou (11) assigned values of 0.5 to all zero cells to avoid the need to divide by zero, as recommended by Whittemore (5).

The Mantel-Haenszel approach

Using equation 3, an adjusted estimate of the attributable risk may be obtained that utilizes the Mantel-Haenszel estimator of the common odds ratio (OR_{MH}):

$$AR = P(E \backslash D)(1 - 1/OR_{MH}), \qquad (10)$$

where $P(E\backslash D)$ is the proportion of cases that are exposed. Benichou (11) has shown that the Mantel-Haenszel approach can be regarded as a special case of the weighted sum approach, in the situation where one imposes the assumption of a common odds ratio (a restrictive assumption not generally required by the weighted sum approach). Confidence intervals have been provided by Kuritz and Landis (7, 13, 14) and Greenland (12), as shown in Appendix 3. The Greenland (12) method can also be used with the conditional maximum likelihood estimate of the odds ratio in place of the Mantel-Haenszel estimate.

For individual matching, the risk ratio may be estimated using a matched estimate of the odds ratio, which, if a single control has been matched to each case, is simply the ratio of the discordant cell frequencies from the cross-classification of the matched case-control pairs according to exposure status (13, 14). Since both the proportion of exposed cases in the population ($P(E\backslash D)$) and the odds ratio may be estimated from matched case-control data, the attributable risk is also estimable from matched data (13). An implicit assumption, however, is that the cases are representative of all cases in the target population with respect to the frequency of the exposure (13). Attributable risk estimates obtained from case-control studies that ascertain cases selectively—for example, from some but not all hospitals within a community—may be biased.

In contrast to the weighted sum approach, adjusted estimates of attributable risk obtained using the Mantel-Haenszel approach are suitable for sparse data and can be used for individually matched case-control data (11). Simulation results by Kuritz and Landis (7) and Greenland (12) show negligible bias when the attributable risk is estimated, and that the coverage of confidence intervals is close to the nominal (95 percent) rate. An important assumption in the Mantel-Haenszel approach is that there is a common odds ratio, i.e., a lack of interaction on a multiplicative scale between the adjustment factor and the exposure of interest. If more than mild heterogeneity is present, the adjusted estimates of attributable risk obtained using this approach may be biased (11, 24). Greenland (12) suggested using a hybrid odds ratio to overcome this problem, although the finite sample properties of this estimate have not been studied (11). Confidence intervals for adjusted estimates of attributable risk obtained using the Mantel-Haenszel approach are shown in Appendix 3.

Example 2: Matched estimation and the Mantel-Haenszel approach for obtaining adjusted estimates of the attributable risk. To illustrate these procedures for obtaining adjusted estimates of attributable risks from matched case-control studies, we analyzed data from a recent case-control study of breast cancer by Nasca et al. (25, 26). The methods of the study have been reported in detail elsewhere (25, 26). In brief, the cases consisted of 1,617 female residents of 18 contiguous counties of New York State who had been diagnosed with primary breast cancer between April 1, 1982, and March 31, 1984. One control was randomly selected for each case using records maintained by the New York State Department of Motor Vehicles. The subjects were matched with regard to sex, year of birth, and county of residence. For the purposes of this empirical example, we restrict our attention to the factors shown in table 1: age at first live birth, personal history of benign breast disease, history of breast cancer in one or more first-degree female relatives, and alcohol consumption.

Table 2 shows the matched pairs classified according to their risk factor status with respect to alcohol consumption. We defined the baseline level as a consumption of less than 1.5 g/day. We chose this much more restrictive definition than that used in example 1 because, in this population, the level of alcohol consumption is much lower than in the population considered in example 1. In this simplest case involving only a single control matched to each case, the maximum likelihood estimate of the odds ratio coincides with the Mantel-Haenszel odds ratio and is $371/316 = 1.17$ (SD = 0.086), a value significantly different from 1 at the 5 percent level. Assuming that the cases are representative of all cases in the target population, the proportion of cases exposed to the factor [$P(E\backslash D)$] is $952/1{,}490 = 0.639$. From equation 10, the risk attributable to having an alcohol consumption of at least 1.5 g/day is then estimated to be $0.639 \times [1 - (1/1.17)] = 0.095$ (SD = 0.041). Note that this estimate is adjusted for the matching factors, namely sex, year of birth, and county of residence. No further adjustment is possible with this approach unless one is willing to break the matching. From this analysis, an estimated 9.5 percent of breast cancer cases in the target population are attributable to an alcohol consumption of at least 1.5 g/day and could potentially be prevented provided that alcohol is causative.

TABLE 1. Distribution of breast cancer cases and their matched controls in all strata defined by the cross-classification of four risk factors*

Age (years) at first live birth	Family history of breast cancer	Alcohol consumption (g/day)	History of benign breast disease	No. of cases	No. of controls
<20	Yes	<1.5	Yes	0	2
			No	2	4
		≥1.5	Yes	2	0
			No	4	6
	No	<1.5	Yes	8	6
			No	34	41
		≥1.5	Yes	14	14
			No	28	44
20–24	Yes	<1.5	Yes	9	8
			No	20	28
		≥1.5	Yes	24	10
			No	36	21
	No	<1.5	Yes	48	42
			No	121	157
		≥1.5	Yes	92	71
			No	185	232
25–29	Yes	<1.5	Yes	5	4
			No	23	10
		≥1.5	Yes	26	10
			No	23	29
	No	<1.5	Yes	34	28
			No	79	110
		≥1.5	Yes	87	49
			No	128	169
≥30	Yes	<1.5	Yes	5	0
			No	10	5
		≥1.5	Yes	18	2
			No	9	10
	No	<1.5	Yes	15	14
			No	53	62
		≥1.5	Yes	43	21
			No	84	87
Nulliparous	Yes	<1.5	Yes	7	2
			No	17	5
		≥1.5	Yes	9	2
			No	19	11
	No	<1.5	Yes	22	15
			No	45	74
		≥1.5	Yes	49	33
			No	108	122

* Source of data: Nasca et al. (25, 26).

ADJUSTED METHODS OF ESTIMATION BASED ON REGRESSION MODELS

Multivariable adjustment procedures that utilize the logistic regression model have also been proposed for estimation of attributable risks from case-control data (11, 27, 28). These regression methods may be used to estimate the risk attributable to an indi-

TABLE 2. Frequency of increased alcohol consumption (≥1.5 g/day) among breast cancer cases and their matched controls*

Risk factor status of case	Risk factor status of control		
	Alcohol consumption of ≥1.5 g/day	Alcohol consumption of <1.5 g/day	Total
Alcohol consumption of ≥1.5 g/day	581	371	952
Alcohol consumption of <1.5 g/day	316	222	538
Total	897	593	1,490

* Source of data: Nasca et al. (25, 26).

vidual factor while simultaneously adjusting for the other factors included in the model.

Walter (9) and Sturmans et al. (29) first suggested the use of the logistic model in the estimation of attributable risk. Deubner et al. (27) subsequently used the logistic model to obtain adjusted estimates of the risk attributable to cardiovascular disease risk factors in a prospective mortality study. However, Bruzzi et al. (28) provided the first detailed description of how the logistic regression model can be used to obtain adjusted estimates of attributable risk from case-control data. Assuming that the disease is rare in the population, stratum-specific relative risks are estimated using adjusted odds ratios derived from the regression coefficients (11, 28). Adjusted attributable risk estimates can then be obtained using the distribution of the exposures of interest among the cases alone, provided that the cases are representative of all cases in the population (28). When the exposures of interest are categorical, a baseline stratum j^* can be defined for each exposure stratum j which has the same levels of the confounding factor as does stratum j but baseline levels of the factor of interest (28, 30). Adjusted estimates of attributable risk can then be obtained using the following formula provided by Bruzzi et al.:

$$\mathrm{AR} = 1 - \frac{\sum(n_j/\mathrm{RR}_j)}{n}, \qquad (11)$$

where n is the total number of cases, n_j is the number of cases within stratum j, and RR_j is the relative risk for stratum j compared with the baseline stratum (28). This is equivalent to equations 2 and 3 in the simplest case in which only a single covariate is included in the model and the variable is binary (11). The summary attributable risk for all of the factors considered jointly may be similarly derived (28). Under the logistic model, the overall attributable risk for all of the exposures combined is generally neither the sum nor the product of the adjusted attributable risks for the individual factors, and the sum of the adjusted attributable risks for each factor may exceed 100 percent (28, 30).

Confidence intervals for attributable risk estimates obtained using the logistic model have been derived by Benichou and Gail (31) from Taylor series expansions applied to implicitly related random variables. Using this method, confidence intervals can be obtained for attributable risks that are adjusted for one or more discrete variables and for simple random sampling, stratified random sampling, frequency matching, and 1:1 individual matching. Simulations showed that these confidence intervals have coverages very close to the nominal (95 percent) level for these four kinds of sampling (31). Kooperberg and Petitti (32) used a bootstrap method to obtain confidence intervals for adjusted attributable risk estimates derived from unmatched case-control data using the logistic model. Confidence intervals for adjusted attributable risks obtained using this computer-intensive method have not been compared with those obtained using formulas provided by Benichou and Gail (33). Drescher and Schill (34) proposed a further approach for obtaining confidence intervals for stratum-specific or summary attributable risks obtained under an unconditional logistic model. The method is applicable only to unmatched or moderately stratified case-control data, and confounding can only be controlled for by stratification using this approach (34). However, the method proposed by Drescher and Schill may be computationally easier than alternative approaches (34).

The method of Drescher and Schill (34) has recently been extended to allow for the control of multiple and continuous confounders (35). A simulation study by

Greenland and Drescher (35) suggested that in large samples the estimators proposed by Drescher and Schill (34) and Bruzzi et al. (28) perform equally well, and that there is no practical difference between confidence intervals based on the two methods.

The main appeal of the regression approach is its flexibility and generality. It allows one to control for confounding and effect modification by introducing the relevant terms in the logistic model. It can be used regardless of whether or not the data have been matched. It includes two stratification approaches, namely the weighted sum approach and the Mantel-Haenszel approach, as special cases and provides a unified framework for estimation and hypothesis testing (11). Illustrations based on example 1 are provided by Benichou (11). We provide further illustration below based on example 2.

Despite their appeal, regression methods for the estimation of attributable risk are susceptible to general problems associated with regression methods, such as overmodeling the data or fitting models to data that are too sparse (30). In addition, if all relevant factors are not included in the model or the model does not have the correct parametric form, the adjusted estimates of attributable risk may be biased (30). The failure to account for significant interaction may also introduce bias (11, 30). Strategies for selecting a final model must balance parsimony against biases that result from model misspecification (11).

To illustrate the logistic modeling approach to attributable risk estimation, we fitted a conditional logistic model to the breast cancer data (example 2). This analysis was based on the same 1,490 pairs (for which no information was missing) as for the Mantel-Haenszel analysis presented above. Seven terms were included in the conditional logistic model to fit the main effects of age at first live birth (in five categories: <20, 20–24, 25–29, and ≥30 years, and nulliparous), family history of breast cancer (in two categories: yes or no), personal history of benign breast disease (in two categories: yes or no), and alcohol consumption (in two categories: <1.5 g/day and ≥1.5 g/day). It allowed us to adjust for the confounding effects of the matching factors, namely sex, year of birth, and county of residence, as well as the confounding effects of age at first live birth, family history of breast cancer, and personal history of benign breast disease. The odds ratio for alcohol consumption was estimated to be 1.15 (SD = 0.079), slightly smaller than the odds ratio obtained previously. This value was not quite significant at the 5 percent level ($p = 0.079$). The corresponding adjusted attributable risk estimate was 0.083 (SD = 0.045), again smaller than the estimate from the Mantel-Haenszel approach. Therefore, from this analysis, 8.3 percent rather than 9.5 percent of breast cancer cases are attributable to alcohol consumption in excess of 1.5 g/day and could potentially be prevented, provided that alcohol is causative. Furthermore, it is possible to introduce interaction terms into the model. When using (four) additional parameters to fit the interaction of age at first live birth and family history of breast cancer, similarly to what was done by Gail et al. (36), we obtained a somewhat higher odds ratio for alcohol consumption of 1.16 (SD = 0.073) and an attributable risk estimate of 0.088 (SD = 0.045). However, the increase in the log-likelihood was not significant at the 5 percent level ($p = 0.086$), which suggested that the simpler model might be preferable.

Additive relative risk models

Additive relative risk models have also been applied to the analysis of case-control data (30, 37, 38). They assume an additive effect of the covariates on the relative risk. Once the regression coefficients have been obtained under the additive model, adjusted attributable risks may be estimated using equation 11. The adjusted estimates of the risk attributable to each factor included in the additive model sum to the overall estimate for all of the factors considered jointly, when no interaction terms are included in the model (30). Furthermore, the sum of the adjusted attributable risks for each factor may not exceed 1 (30). Thus, the additive model may provide better estimates of the risk attributable to multiple exposures, which are more easily interpretable, when there is an absence of interaction on an additive scale between the exposures of interest. The conditional likelihood function is influenced by the baseline that is chosen for the covariates, even when they are dichotomous, and a change in the way the independent variables are coded may result in a change in the beta coefficients and the relative and attributable risk estimates (30, 37). As a consequence, results obtained with the additive model may be difficult to interpret. To overcome this difficulty, Breslow and Storer (38) have recommended that the level of each factor with the lowest risk be used as the baseline level, although this ad hoc rule has been criticized (37). Nonetheless, both additive and multiplicative regression methods may be used to obtain adjusted estimates of attributable risk from case-control data, with choices between alternative models based principally upon goodness of fit (30).

OTHER APPLICATIONS OF ATTRIBUTABLE RISK ESTIMATES

Attributable risk estimates obtained from case-control studies have been applied to a variety of other uses, including sample size and power calculations, estimation of the proportion of cases that would be prevented *if* a disease-risk factor association is later judged to be causal, and the awarding of compensatory damages by courts of law for injuries resulting from environmental exposures (39–42). These novel applications have challenged traditional conceptual frameworks for attributable risk estimation and highlight the need for further methodological and theoretical work in this area.

Use of the attributable risk to determine an appropriate sample size

Sample size calculations for case-control studies have traditionally been based on the odds ratio as a measure of the strength of an association. Browner and Newman (39) derived formulas for sample size estimation that are based instead upon the attributable risk. This approach may be more appropriate when the public health importance of an association is of interest, such as when a case-control study is undertaken to estimate the impact of reducing exposure to an etiologic agent. For the situation where the exposure is dichotomous, Browner and Newman (39) compared sample size and power estimates based on the detection of a given attributable risk with conventional estimates based on the detection of a given risk ratio. Their findings suggest that, when the exposure is rare, case-control studies having little power to detect a small risk ratio may still have adequate power to detect a small attributable risk (39). On the other hand, even relatively large case-control studies may have inadequate power to detect a small attributable risk when the exposure is common (39).

Adams et al. (40) further considered the relation between the attributable risk, the risk ratio, and exposure frequency as an aid to the design and interpretation of epidemiologic studies. They found that arbitrary risk ratios specified during the design of epidemiologic studies (e.g., an odds ratio of 2) may not provide adequate statistical power to detect the associations of interest, since low risk ratios may exist in the presence of etiologic heterogeneity and interactions between risk factors (40). Thus, when carrying out sample size calculations based on the detection of a given risk ratio, it may be important to consider how large the risk attributable to the exposure is likely to be in specifying the size of the odds ratio that one wishes to detect (40).

Use of the attributable risk to estimate the proportion of potentially preventable cases when the causality of the association is uncertain

When attributable risk estimates are used in the planning of public health programs that must choose between alternative disease prevention strategies, the disease-risk factor association is generally assumed to be causal (12). However, attributable risk estimates have sometimes been calculated prior to a determination of causality and have been used to argue for the potential public health significance of an exposure and for continued research on that exposure. For example, it was argued that despite evidence of weak associations, further studies of alcohol and breast cancer in women were needed, because the risk attributable to this common exposure would be appreciable *if* the association was later judged to be causal (43, 44).

This situation serves to illustrate the need for further research and refinement of the methods currently used to determine causality from associations found in epidemiologic research. Often there is a lack of certainty about causality (45, 46), and therefore it is unclear what evidence would be sufficient to satisfy the assumption of causality required by attributable risk estimation. It is beyond the scope of this paper to examine this issue further, but it is unlikely to be resolved by a call for more meta-analyses or randomized trials.

Estimability of the attributable risk in science and the law

The attributable risk (attributable fraction) has traditionally referred to the fraction of all cases in the population that are *excess* cases, i.e., cases that would not have occurred if the exposure had not occurred (3). From a biologic or legal perspective, however, the attributable risk may refer instead to cases for which exposure played an *etiologic* role, i.e., cases for which the exposure was a contributory cause of the outcome (3). Thus, more than one concept has been identified as the attributable risk (3, 41). In addressing this problem, Greenland and Robins (3) argued for the adoption of more precise definitions and the use of the terms "excess fraction" and "etiologic fraction" to refer to these attributable risk concepts. The estimability of these two measures is quite different. Whereas the excess fraction can be estimated under the same conditions that are often cited for general validity of an epidemiologic study (e.g., lack of systematic error or biases), estimation of the etiologic fraction requires nonidentifiable biologic assumptions about exposure action and interactions (41).

A related problem arises in courts of law when there is a need to assign shares of responsibility to two or more exposures. Cox (47, 48) proposed a method of partitioning the increase in disease risk that can be attributed to each risk factor which has potential applications in tort law liability cases. In contrast to the attributable risk, the measure proposed by Cox does not have a clear population interpretation (49). However, it does offer the attractive property of additivity; i.e., the sum of the separate increases in disease risk assigned to each of two factors is equal to the joint risk assigned to both factors (49, 50).

CONCLUSIONS AND FUTURE DIRECTIONS

The modeling approaches to attributable risk estimation discussed in this review overcome important limitations of adjustment procedures based on stratification. They also offer several advantages (11, 28, 30). For example, regression methods may be used to obtain adjusted estimates of attributable risk even when individual matching has been carried out and unbiased estimates cannot be obtained using the weighted sum approach (11). Furthermore, modeling allows for both confounding factors and effect modifiers to be taken into account (11, 28, 30). A common odds ratio, as required by the Mantel-Haenszel approach, need not be assumed (11). The regression approach includes both the Mantel-Haenszel and weighted sum approaches as special cases (11). By including one or more interaction terms in the model, adjusted estimates of attributable risk may be obtained in the presence of significant interaction between the adjustment factor and the exposure of interest.

Further methodological advances in the estimation of attributable risks from case-control data are likely to depend in part upon refinements in the way in which biologic relations are modeled using epidemiologic data. In view of the strong assumption of causality generally required by attributable risk estimation (12) and the conceptual problems identified by Greenland and Robins (3), theoretical and pragmatic advances in causal inference methodology will probably improve our ability to interpret and use attributable risk estimates.

ACKNOWLEDGMENTS

This work was supported in part by grants 1P30-CA51008 and 1P50-CA58185 from the National Cancer Institute. Dr. Coughlin is the current recipient of a First Independent Research Support and Transition Award (R29 HL44904) from the National Heart, Lung, and Blood Institute.

The authors are grateful to Dr. Philip Nasca for providing the breast cancer data used in the empirical example.

REFERENCES

1. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.
2. Ouellet BL, Romeder JM, Lance JM. Premature mortality attributable to smoking and hazardous drinking in Canada. Am J Epidemiol 1979;109: 451–63.
3. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. Am J Epidemiol 1988;128:1185–97.
4. Cole P, MacMahon B. Attributable risk percent in case-control studies. Br J Prev Soc Med 1971; 25:242–4.
5. Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. Stat Med 1982;1:229–43.
6. Walter SD. Calculation of attributable risks from epidemiological data. Int J Epidemiol 1978;7: 175–82.
7. Kuritz SJ, Landis JR. Summary attributable risk estimation from unmatched case-control data. Stat Med 1988;7:507–17.
8. Levin ML. The occurrence of lung cancer in man. Acta Unio Int Contra Cancrum 1953;9: 531–41.
9. Walter SD. The estimation and interpretation of attributable risk in health research. Biometrics 1976;32:829–49.
10. Whittemore AS. Estimating attributable risk from case-control studies. Am J Epidemiol 1983; 117:76–85.
11. Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: a review. Stat Med 1991;10:1753–73.
12. Greenland S. Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data. Stat Med 1987;6:701–8.
13. Kuritz SJ, Landis JR. Attributable risk ratio estimation from matched-pairs case-control data. Am J Epidemiol 1987;125:324–8.
14. Kuritz SJ, Landis JR. Summary attributable risk estimation from matched case-control data. Biometrics 1988;44:355–67.
15. Walter SD. The distribution of Levin's measure of attributable risk. Biometrika 1975;62:371–4.
16. Leung HM, Kupper LL. Comparisons of confidence intervals for attributable risk. Biometrics 1981;37:293–302.
17. Tuyns AJ, Pequignot G, Jensen OM. Le cancer de l'oesophage en Ille-et Vilaine en fonction des niveaux de consommation d'alcool et de tabac. (In French). Bull Cancer (Paris) 1977;64:45–60.
18. Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. Am J Epidemiol 1974;99:325–32.
19. Denman DW, Schlesselman JJ. Interval estimation of the attributable risk for multiple exposure levels in case-control studies. Biometrics 1983; 39:185–92.
20. Walter SD. Effects of interaction, confounding and observational error on attributable risk estimation. Am J Epidemiol 1983;117:598–604.
21. Hsieh CC, Walter SD. The effect of nondifferential exposure misclassification on estimates of the attributable and prevented fraction. Stat Med 1988;7:1073–85.
22. Walter SD. Prevention for multifactorial diseases. Am J Epidemiol 1980;112:409–16.
23. Ejigou A. Estimation of attributable risk in the presence of confounding. Biometric J 1979;21: 155–65.
24. Greenland S. Bias in methods for deriving standardized morbidity ratio and attributable fraction estimates. Stat Med 1984;3:133–58.
25. Nasca PC, Baptiste MS, Field NA, et al. An epidemiological case-control study of breast cancer and alcohol consumption. Int J Epidemiol 1990;19:532–8.
26. Nasca PC, Baptiste MS, Field NA, et al. An epidemiologic case-control study of breast cancer and exposure to hair dyes. Ann Epidemiol 1992;2:577–86.
27. Deubner DC, Wilkinson WE, Helms MJ, et al. Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. Am J Epidemiol 1980;112: 135–43.
28. Bruzzi P, Green SB, Byar DP, et al. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 1985;122:904–14.
29. Sturmans F, Mulder PGH, Valkenburg HA. Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage. Am J Epidemiol 1977;105:281-9.
30. Coughlin SS, Nass CC, Pickle LW, et al. Regression methods for estimating attributable risk in

population-based case-control studies: a comparison of additive and multiplicative models. Am J Epidemiol 1991;133:305–13.

31. Benichou J, Gail MH. Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. Biometrics 1990;46:991–1003.
32. Kooperberg C, Petitti DB. Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study. Epidemiology 1991;2:363–6.
33. Greenland S. The bootstrap method for standard errors and confidence intervals of the adjusted attributable risk. (Letter). Epidemiology 1992;3: 271.
34. Drescher K, Schill W. Attributable risk estimation from case-control data via logistic regression. Biometrics 1991;47:1247–56.
35. Greenland S, Drescher K. Maximum-likelihood estimation of the attributable fraction from logistic models. Biometrics 1993;49:865–72.
36. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81: 1879–86.
37. Moolgavkar SH, Venzon DJ. General relative risk regression models for epidemiologic studies. Am J Epidemiol 1987;126:949–61.
38. Breslow NE, Storer BE. General relative risk functions for case-control studies. Am J Epidemiol 1985;122:149–62.
39. Browner WS, Newman TB. Sample size and power based on the population attributable fraction. Am J Public Health 1989;79:1289–94.
40. Adams MJ, Khoury MJ, James LM. The use of attributable fractions in the design and interpretation of epidemiologic studies. J Clin Epidemiol 1989;42:659–62.
41. Robins JM, Greenland S. Estimability and estimation of excess and etiologic fractions. Stat Med 1989;8:845–59.
42. Robins JM, Greenland S. The probability of causation under a stochastic model for individual risk. Biometrics 1989;45:1125–38.
43. Hiatt RA, Bawol RD. Alcoholic beverage consumption and breast cancer incidence. Am J Epidemiol 1984;120:676–83.
44. Hiatt RA. Alcohol consumption and breast cancer. Med Oncol Tumor Pharmacother 1990;7: 145–51.
45. Susser M. What is a cause and how do we know one? A grammar for pragmatic epidemiology. Am J Epidemiol 1991;133:635–48.
46. Weed DL. On the logic of causal inference. Am J Epidemiol 1986;123:965–79.
47. Cox LA Jr. Probability of causation and the attributable proportion of risk. Risk Anal 1984;4: 221–30.
48. Cox LA Jr. A new measure of attributable risk for public health applications. Management Sci 1985;7:800–13.
49. Benichou J. Re: "Methods of adjustment for estimating the attributable risk in case-control studies: a review." (Letter). Stat Med 1993;12: 94–6.
50. Gefeller O, Eide GE. Re: "Methods of adjustment for estimating the attributable risk in case-control studies: a review." (Letter). Stat Med 1993;12:91–4.
51. Robins JM, Breslow NE, Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. Biometrics 1986;42:311–23.
52. Cox DR, Hinkley DV. Theoretical statistics. London, England: Chapman and Hall Ltd, 1974: 254.

APPENDIX 1

Confidence Intervals for Unadjusted Attributable Risk (AR) Estimates

Let n_0 and n_1 equal the numbers of unexposed and exposed cases, respectively, and m_0 and m_1 the numbers of unexposed and exposed controls, with $n_0 + n_1 = n$ and $m_0 + m_1 = m$. Text equations 2 and 3 may then be expressed as

$$\mathrm{AR} = \frac{n_1 m_0 - m_1 n_0}{m_0 n}.$$

A maximum likelihood estimate of the variance of the unadjusted attributable risk may be obtained by applying the delta method (11):

$$\mathrm{var(AR)} = m n_0 (n_1 m_0 m + n n_0 m_1)/n^3 m_0^3.$$

The corresponding symmetric confidence interval for AR is then

$$\mathrm{AR} - z_{1-\alpha/2}[\mathrm{var(AR)}]^{1/2},\ \mathrm{AR} + z_{1-\alpha/2}[\mathrm{var(AR)}]^{1/2},$$

where $z_{1-\alpha/2}$ is the 100(1 − α/2)th percentile of the standard normal distribution. An alternative variance estimate was obtained by Walter (15) using the log-transform:

$$\mathrm{var}[\log(1 - \mathrm{AR})] = n_1/n n_0 + m_1/m m_0.$$

The corresponding $100(1 - \alpha)$ percent confidence interval for AR is then

$$1 - (1 - \mathrm{AR}) \exp \mathrm{EF},\ 1 - (1 - \mathrm{AR})/\exp \mathrm{EF},$$

where the error factor EF is equal to $z_{1-\alpha/2}\{\mathrm{var}[\log(1 - \mathrm{AR})]\}^{1/2}$. Whittemore (5) noted that this asymmetric interval is wider than the interval based on the maximum likelihood. Finally, Leung and Kupper (16) suggested using the logit transform and obtained

$$\mathrm{var}[\mathrm{logit}(\mathrm{AR})] = (n_1/nn_0 + m_1/mm_0)[m_0 m/(n_1 m - m_1 n)]^2.$$

The corresponding $100(1 - \alpha)$ percent confidence interval for AR is

$$1/\{1 + [(1 - \mathrm{AR})/\mathrm{AR}] \exp \mathrm{EF}\},\ 1/\{1 + [(1 - \mathrm{AR})/\mathrm{AR}]/\exp \mathrm{EF}\},$$

where EF is equal to $z_{1-\alpha/2}\{\mathrm{var}[\mathrm{logit}(\mathrm{AR})]\}^{1/2}$. This approach results in an asymmetric confidence interval with better coverage properties than those of a maximum likelihood-based interval (16).

APPENDIX 2

Confidence Intervals for the Weighted Sum Approach

Text equation 9 showing the attributable risk (AR) written as a weighted sum of the attributable risks across J strata, with weights equal to the proportion of cases for each stratum, may be rewritten as

$$\mathrm{AR} = 1 - \sum_{j=1}^{J} m_j n_{0j}/nm_{0j},$$

where m_j is the number of controls in adjustment level (stratum) j and m_{0j} and n_{0j} are the numbers of unexposed controls and cases in stratum j, respectively. Whittemore (5) obtained the following variance estimate using the delta method:

$$\mathrm{var}(\mathrm{AR}) = \left[\sum_{j=1}^{J} (m_j n_{0j}/m_{0j})^2 (1/n_{0j} + m_{1j}/m_j m_{0j}) - n(1 - \mathrm{AR})^2\right] / n^2,$$

where m_{1j} is the number of exposed controls in stratum j. This variance estimate applies to simple random sampling and frequency matching of the controls. It can then be used to obtain confidence intervals as outlined in Appendix 1.

(Appendix 3 follows)

APPENDIX 3

Confidence Intervals for the Mantel-Haenszel Approach

Let us denote by n_{ij} and m_{ij} the respective numbers of cases and controls with level i of exposure ($i = 0{,}1$) and j of adjustment ($j = 1, \ldots, J$). Moreover, we denote by $n_{.j}$ and $m_{.j}$ the respective numbers of cases and controls with level j of adjustment. Finally, let us use the notation "^" to denote estimators.

Variance estimators have been developed by Kuritz and Landis (7, 14) and by Greenland (12). Here we briefly outline Greenland's approach. More details can be found in the paper by Benichou (11). Greenland (12) used the variance estimator of the Mantel-Haenszel odds ratio estimator, $\widehat{\mathrm{OR}}_{\mathrm{MH}}$, which is known to be valid even for sparse data—for example, even in the case of individual matching (when j indexes the matched sets)—and is thus termed a dually consistent estimator (51). He wrote:

$$\mathrm{var}[\log(\widehat{\mathrm{AR}})] = \mathrm{var}[\log \hat{P}(E \mid D)] + \mathrm{var}\{\log[(\widehat{\mathrm{OR}}_{\mathrm{MH}} - 1)/\widehat{\mathrm{OR}}_{\mathrm{MH}}]\}$$
$$+\ 2\,\mathrm{cov}\{\log \hat{P}(E \mid D), \log[(\widehat{\mathrm{OR}}_{\mathrm{MH}} - 1)/\widehat{\mathrm{OR}}_{\mathrm{MH}}]\}.$$

The first term can be estimated from the binomial distribution of $n_{1.}$, the total number of exposed cases. The second term is a function of $\mathrm{var}(\widehat{\mathrm{OR}}_{\mathrm{MH}})$. By use of the delta method, it can be seen that the covariance term is asymptotically equivalent to

$$\frac{1 - P(E \mid D)}{\mathrm{OR}_{\mathrm{MH}} - 1}[\mathrm{cov}(\log \widehat{\mathrm{OR}}_{\mathrm{MH}}, \log n_{1.}) - \mathrm{cov}(\log \widehat{\mathrm{OR}}_{\mathrm{MH}}, \log n_{0.})].$$

Since $\widehat{\mathrm{OR}}_{\mathrm{MH}}$ is an asymptotically unbiased estimator of OR, the true odds ratio, then (from the book by Cox and Hinkley (52)), asymptotically,

$$\mathrm{cov}[\log(\widehat{\mathrm{OR}}_{\mathrm{MH}}), U(\mathrm{OR})] = 1,$$

in which $U(\mathrm{OR}) = n_{1.} - E(n_{1.} \mid \mathrm{OR})$, is the score calculated in OR. Thus, from the delta method,

$$\mathrm{cov}(\log \widehat{\mathrm{OR}}_{\mathrm{MH}}, \log n_{1}) = 1/E(n_{1})$$

and

$$\mathrm{cov}(\log \widehat{\mathrm{OR}}_{\mathrm{MH}}, \log n_{0.}) = 1/E(n_{0.}).$$

Hence, Greenland (12) obtained the following variance estimate:

$$\widehat{\mathrm{var}}[\log(\widehat{\mathrm{AR}})] = \widehat{\mathrm{var}}(\log \widehat{\mathrm{OR}}_{\mathrm{MH}})/(\widehat{\mathrm{OR}}_{\mathrm{MH}} - 1)^2 + n_{0.}/nn_{1.} + 2/n_{1.}(\widehat{\mathrm{OR}}_{\mathrm{MH}} - 1).$$

Note that this estimator is valid even for sparse data and can therefore be used for simple random sampling, frequency matching, stratified random sampling, and individual matching of the controls.