AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol 117, No. 1

Printed in U.S.A.

ESTIMATING ATTRIBUTABLE RISK FROM CASE-CONTROL STUDIES[1]

ALICE S. WHITTEMORE

Whittemore, A. S. (Stanford U. School of Medicine, Stanford, CA 94305). Estimating attributable risk from case-control studies. *Am J Epidemiol* 1983;117:76-85.

Levin's measure of attributable risk is extended to account for confounding. Maximum likelihood estimates and confidence intervals for this extended measure are presented. The estimates and confidence intervals apply both to matched and to stratified case-control studies. The statistical methods are illustrated by the use of data from a study of factors of womanhood as related to breast cancer, and data from a study of cigarette smoking as related to bladder cancer. The results of computer simulations are used to describe the behavior of the estimates and confidence intervals when sample size is small relative to the number of strata of the confounding factor.

epidemiologic methods; risk; statistics

Rational choice of disease control strategies requires some knowledge of the disease burden that could be prevented by modifying a given risk factor. Thus, it is useful to have a measure of the proportion of disease attributable to that factor. Moreover, the measure should be adjusted for correlations between the exposure of interest and other etiologic factors. Levin's measure of attributable risk (1–4), first proposed in 1953 and since used extensively by public health investigators, does not control for confounding by such factors. The purpose of this paper is to define an extension of Levin's measure that is adjusted for confounding; to present maximum likelihood estimates and confidence intervals for this extended measure based on data from case-control studies; and to investigate the behavior of the estimates and confidence intervals when the study size is small relative to the number of strata of the confounding factors.

Consider the case-control data of Paffenbarger et al. (5) for breast cancer among white parous women, shown in table 1. Controls (n = 2024) were matched to cases (n = 1051) on age at diagnosis. Suppose we wish to use these data to estimate the fraction of all breast cancers attributable to late age at first childbirth, while controlling for confounding by age at diagnosis and socioeconomic status. We are then confronted with several questions. 1) When and how must Levin's measure be adjusted for confounding? 2) Can adjusted attributable risk measures be estimated from case-control data? Are additional problems introduced when working with matched data? Can unbiased estimates be obtained from data that have been pooled over matching factors? 3) Must one use different estimates if adjusting for age at diagnosis (matching factor) than if adjusting for years of edu-

Received for publication May 3, 1982, and in final form August 30, 1982.

Abbreviations: AR, attributable risk; ML, maximum likelihood; SE, standard error.

[1] Department of Family, Community and Preventive Medicine, Stanford University School of Medicine, Stanford, CA 94305. (Reprint requests to Dr. Whittemore.)

This research was supported by NIH Grant Nos. ES00070 and CA23214 and by grants to SIMS from the Sloan Foundation, EPA, and NSF.

An earlier version of this paper was presented at the annual meeting of the Society for Epidemiologic Research, June 1981, Snowbird, UT.

TABLE 1

Case-control data and incidence rates for breast cancer among US white females

Age (years) at diagnosis	No. of white parous study subjects by age (years) at first childbirth*				US white female breast cancer incidence Rate per 10^5†	1970 US white female population in thousands‡	σ_k§
	>24 (E)		≤24 ($\bar{E}$)				
	Years of education		Years of education				
	>12	≤12	>12	≤12			
25–29							
Cases	1	0	0	1	8.7	5962.1	0.007
Controls	3	0	5	6			
30–34							
Cases	9	0	7	4	22.5	5042.4	0.015
Controls	8	0	13	13			
35–39							
Cases	6	7	16	16	52.5	4936.5	0.035
Controls	20	8	50	52			
40–44							
Cases	19	8	25	18	103.7	5412.3	0.075
Controls	31	16	65	74			
45–49							
Cases	42	19	46	36	159.2	5587.0	0.119
Controls	60	29	84	128			
50–54							
Cases	45	26	55	43	171.7	5169.3	0.119
Controls	68	44	116	111			
55–59							
Cases	72	34	36	38	191.8	4695.6	0.121
Controls	97	54	74	99			
60–64							
Cases	43	22	23	29	226.2	4157.5	0.126
Controls	70	40	27	84			
65–69							
Cases	38	18	19	36	234.2	3491.1	0 110
Controls	70	24	35	67			
70–74							
Cases	31	15	17	21	259.8	2874.5	0.100
Controls	28	24	24	39			
75+							
Cases	36	22	13	39	299.5	4319.1	0.173
Controls	39	41	14	70			
Total	836	451	764	1024	79.1	51647.4	1.00

* From Paffenbarger et al. (5). Cases were matched to controls on age at diagnosis.

† From the Third National Cancer Survey (9).

‡ From the National Center for Health Statistics (10).

§ σ_k is the product of column 5 and column 6, normalized so that $\sum_k \sigma_k = 1$.

cation (not a matching factor)? 4) What is the sampling variability of adjusted attributable risk estimates, and what are appropriate confidence intervals? 5) How well do these estimates and confidence intervals perform when sample size is small relative to number of confounder strata?

The first question has been addressed by Walter (6); his results are summarized in the next section. Questions two through four are answered in the following section, which presents adjusted maximum likelihood estimates and confidence intervals for attributable risk. These procedures are then applied to data

obtained from two case-control studies. This is followed by a discussion of the small-sample behavior of these new methods, which was investigated by the use of simulated data.

This paper does not discuss the impact on attributable risk estimates of selection, reporting, or recall bias. It does not include a lag time between exposure elimination and the resulting disease reduction. Instead, it assumes that cases and controls are randomly selected from the population of interest, and that exposure and confounder information are reported without bias. The hypothetical disease rates in the absence of exposure are evaluated after the latent period for disease has elapsed. The methods described apply either to disease incidence rates or disease prevalence rates. The paper also assumes that the confounding variables are categorical and that the exposure variable is dichotomous. The generalization to arbitrary confounding and exposure variables will be the subject of a separate investigation.

Population measures of attributable risk

Consider a population that is classified into those exposed (E) and unexposed ($\bar{E}$) to an etiologic agent for disease. For the population of white parous women discussed above, "exposure" is delay of first full-term pregnancy until after age 24 years. (The impact on attributable risk estimates of a different cutoff age for exposure has not been investigated.) We assume that the disease rate $P(D|E)$ among the exposed exceeds the corresponding rate $P(D|\bar{E})$ among the unexposed. The fraction of all disease that would persist if the entire population contracted disease at the unexposed rate is

$$P(D|\bar{E})/P(D),$$

where $P(D)$ is the disease rate in the entire population. This fraction shall be called the *residual risk* associated with E. The remaining fraction

$$1 - P(D|\bar{E})/P(D) \quad (1)$$

represents the proportion of disease that would be prevented if the unexposed rates applied to everyone. Levin called this factor the (population) attributable risk corresponding to E (1). Levin's measure depends only on the relative risk $R = P(D|E)/P(D|\bar{E})$ and on exposure prevalence $P(E)$, as can be seen by rewriting expression 1 as

$$\frac{P(E)(R-1)}{1+P(E)(R-1)}. \quad (2)$$

To see how this measure can be extended to control for confounding, let us return to the problem of determining breast cancer risk attributable to age at first childbirth. If one could study the entire US population of white parous women beyond age 24, one might find that those who became parous by age 24 differ in age from those who did not. If so, the breast cancer rate among the unexposed (those parous by age 24) would not reflect the prevailing rate if exposure had been eliminated by impregnating all women by their 24th year. Instead, the breast cancer rate after such intervention would be a weighted average of age-specific unexposed rates:

$$\Sigma_k P(C_k) P(D|\bar{E}C_k). \quad (3)$$

Here, $P(C_k)$ is the fraction of all women in the kth age group C_k, and $P(D|\bar{E}C_k)$ is the corresponding age-specific unexposed breast cancer rate. Substitution of expression 3 for the unadjusted rate $P(D|\bar{E})$ in formula 1 yields an adjusted measure of attributable risk (AR):

$$AR = 1 - \Sigma_k P(C_k) P(D|\bar{E}C_k)/P(D). \quad (4)$$

Definition 4 reduces to the unadjusted measure 1 if the confounder C (age in the above example) has only one level.

Sufficient conditions for equality between the adjusted and unadjusted measures 1 and 4 of attributable risk are:

1) The factor C is uncorrelated with disease status for unexposed individuals; or 2) C is uncorrelated with exposure for the whole population (6, 7). Since confounding factors are unlikely to satisfy either of these two conditions, they should be controlled when estimating attributable risk.

No formula analogous to formula 2 is available to express the adjusted measure of attributable risk in terms of relative risk and exposure prevalence, even when the relative risk associated with E is invariant across confounder strata. For the hypothetical example shown in table 2, the relative risk is 3 at each level of C, yet the stratum-specific attributable risks vary, and the adjusted and unadjusted attributable risk measures differ substantially. The adjusted measure of attributable risk is $1 - \{[(0.17 \times 0.17) + (0.83 \times 0.08)]/0.24\} = 60$ per cent. Thus, adjusted and unadjusted measures differ by $60 - 39 = 21$ per cent. This difference is due to variation in exposure prevalence among the strata. It shows that estimates of attributable risk obtained from formula 2 can be seriously in error, even if one uses adjusted estimates of relative risk.

Estimating attributable risk

Maximum likelihood estimates and estimated standard error. If the disease is so rare that exposure prevalence among all individuals in stratum C_k can be approximated by the corresponding prevalence among disease-free individuals, case-control data alone (either matched or stratified) can be used to estimate attributable risk. Otherwise, additional information is needed concerning the stratum-specific disease rates and confounder distribution in the population. Accordingly, two sets of maximum likelihood procedures will be described, depending on whether or not ancillary information is available.

It is shown elsewhere (7) that one obtains the same maximum likelihood estimates and the same confidence intervals for attributable risk when controlling confounding by matching as when con-

Table 2

Hypothetical disease rates by exposure and confounder (e.g., smoking)

	Smoker		Nonsmoker		Total		Total
	Unexposed	Exposed	Unexposed	Exposed	Unexposed	Exposed	
No. of diseased individuals	6	6	1	57	7	63	70
Size of source population in 10^3	36	12	12	228	48	240	288
Disease rate $\cdot 10^3$	0.17	0.50	0.08	0.25	0.15	0.26	0.24
Relative risk for exposure	1.0	3.0	1.0	3.0	1.0	1.78	
Exposure prevalence (%)	25		95		83		
Attributable risk (%) for exposure	33		66		39		
% of population in smoking stratum	17		83		100		

trolling it by stratifying in the analysis. This fortuitous simplification holds both for estimates obtained solely from case-control data and for estimates obtained from case-control data augmented with population-based disease rates. It means that the same statistical procedures can be used when adjusting for confounding, regardless of the adjustment method. Therefore, no design distinctions are made in the following discussion, which summarizes the results derived in reference 7.

The maximum likelihood estimate of attributable risk based solely on case-control data is

$$\hat{AR} = 1 - [\Sigma_k\, y_{\cdot k}\, x_{2k}/n y_{2k}]. \qquad (5)$$

Here, n is the total number of cases, $y_{\cdot k}$ is the number of controls in stratum k, and x_{2k} and y_{2k} represent the numbers of unexposed cases and controls in stratum k. (Actually, formula 5 estimates an approximation to attributable risk, one whose accuracy improves with the infrequency of disease in the population. This distinction will be ignored.)

Standard large-sample methods show that as the numbers of cases and controls in each stratum increase, $\hat{AR}$ becomes normally distributed, with mean AR of formula 4. An asymptotically unbiased estimate of the large-sample standard error (SE) of $\hat{AR}$ is

$$SE = \frac{1}{n}\left\{ \sum_k (y_{\cdot k}\, x_{2k}/y_{2k})^2 \left[\frac{1}{y_{2k}} + \frac{y_{1k}}{y_{\cdot k} y_{2k}} \right] - n(1 - \hat{AR})^2 \right\}^{1/2}, \qquad (6)$$

where y_{1k} is the number of exposed controls in stratum k. Formulae 5 and 6 require adjustment when the control frequencies y_{2k} are zero. In this paper, such frequencies are assigned the value 0.5. The simulations suggest that other small values can be used with essentially identical results.

The unadjusted attributable risk estimate obtained from a matched study by pooling the data over the matching factor is asymptotically biased, except in the trivial case when the factor is not a confounder (7). In this sense, attributable risk estimates are similar to odds ratio estimates obtained by pooling matched data. Breslow (8) has noted that the bias in the pooled odds ratio estimate is often small in practice, and this may also be true for the pooled attributable risk estimate.

One can estimate attributable risk from case-control data without the rare disease assumption, provided estimates exist both for the stratum-specific disease rates $P(D|C_k)$ and for the confounder distribution $P(C_k)$. For simplicity, I assume that these quantities are known with negligible standard error. Appendix formulae A1–A4 give the maximum likelihood estimate $\hat{AR}$ and the estimated asymptotic standard error $\tilde{SE}$ corresponding to this assumption.

Confidence intervals. The upper and lower endpoints of an approximate 95 per cent confidence interval for attributable risk, based on the maximum likelihood estimates of formulae 5 and 6, are $\hat{AR} \pm 1.96$ SE. This type of interval shall be called the maximum likelihood (ML) interval. The analogous confidence interval incorporating ancillary information is obtained by replacing $\hat{AR}$ and SE by the corresponding estimates $\tilde{AR}$ and $\tilde{SE}$ provided in the Appendix.

Two types of transformations have been proposed to produce confidence intervals of greater accuracy or shorter length. One is the log transformation log(1 − AR) suggested by Walter (2, 3); the other is the logit transformation log(AR/(1 − AR)) proposed by Leung and Kupper (4). The log and logit confidence intervals pro-

duced by these transformations are given in the Appendix.

The lengths of these three types of intervals are related in the following way: 1) The ML interval is always shorter than the log interval (except in the trivial case when both have length zero) (7); 2) The logit interval is shorter than the ML interval whenever the attributable risk estimate is between 21 per cent and 79 per cent (4). The accuracy of the three intervals in simulation studies is described below.

Examples

We now apply the methods of the preceding sections to the breast cancer data in table 1. By invoking the rare disease assumption, one finds from the case-control data alone that the estimate of risk attributable to late age at first childbirth, adjusted for age at diagnosis, is $\hat{AR}$ = 14.8 per cent, with estimated SE = 3.1 per cent. The age-adjusted 95 per cent log, ML, and logit confidence intervals for AR are LOG = (8.5, 20.7); ML = (8.7, 20.9); and LOGIT = (9.8, 22.0). The attributable risk estimate adjusted both for age and years of education is $\hat{AR}$ = 10.5 per cent with SE = 3.8 per cent. The unadjusted and asymptotically biased estimate obtained by pooling the data over age at diagnosis and years of education is 17.1 per cent with SE = 3.5 per cent. Although the differences among the three estimates $\hat{AR}$ are small with respect to their associated standard errors, they suggest that both age at diagnosis and years of education should be controlled when estimating risk attributable to age at first childbirth.

Age-adjusted attributable risk can also be estimated without the rare disease assumption, when age-specific breast cancer rates and population census data are used. Table 1 shows breast cancer incidence rates for white females obtained from the Third National Cancer Survey (9) and data from the 1970 US census (10). Ignoring statistical variability and nonparous women in these data, I have used them in Appendix formulae A1–A4 to obtain the age-adjusted attributable risk estimate $\hat{AR}$ = 13.5 per cent, with $\hat{SE}$ = 3.3 per cent. These estimates are consistent with the corresponding results obtained solely from the study data.

The methods of the preceding sections are also applied to data concerning bladder cancer and smoking among males in eastern Massachusetts. These data, shown in table 3, were obtained from a

Table 3

*Case-control data and incidence rates for bladder cancer among males in eastern Massachusetts**

Age (years) at diagnosis	No. of study subjects				Incidence rates ·10⁵	Size of source population in 10³	σ_k‡
	Smoker		Nonsmoker				
	Case	Control†	Case	Control			
50–54	24	22	1	4	30	77.4	0.086
55–59	35	35	2	4	51	68.4	0.130
60–64	31	38	5	3	56	61.5	0.128
65–69	46	42	7	15	120	47.4	0.212
70–74	60	51	13	28	180	38.0	0.254
75–79	39	32	14	20	220	23.2	0.190
Total	235	220	42	74	85	315.9	1.000

* As published previously (12, table 1). All data are from studies by Cole et al. (11).

† Controls were matched to cases by year of birth.

‡ σ_k is the product of column 5 and column 6, normalized so that $\sum_k \sigma_k = 1$.

case-control study and an incidence study by Cole et al. (11) as reported by Miettinen (12). Controls ($n = 294$) were matched to cases ($n = 277$) on year of birth. Use of the case-control data alone yields the age-adjusted attributable risk estimate $\hat{AR} = 30.0$ per cent for smoking with estimated SE = 19.7 per cent. Thus, the three 95 per cent confidence intervals are quite wide: LOG = (−21.5, 59.6); ML = (−8.6, 68.5); and LOGIT = (6.4, 72.9). Only the logit confidence interval excludes the null value AR = 0 per cent. The unadjusted estimate obtained by pooling the data over age is $\hat{AR} = 40.0$ per cent, with a smaller but still appreciable SE = 10.5 per cent.

The variability of the adjusted estimate is not reduced by incorporating into the estimate the incidence and population data shown in table 3. The resulting estimates are $\hat{AR} = 30.1$ per cent with $\hat{SE} = 19.2$ per cent, in close agreement with the results based solely on the case-control data.

The wide confidence intervals for adjusted attributable risk in this example are disturbing. The length of the logit confidence interval reported above is 66.5 per cent, in contrast to a corresponding length of only 11.8 per cent for the breast cancer data. The simulation results described in the next section show that this difference is due less to differences in sample size than to the high smoking prevalence (75 per cent) among the controls.

Simulation studies

Computer simulations were performed to examine the mean value of $\hat{AR}$ and the accuracy and length of the three types of confidence intervals when sample size is small relative to number of strata. To perform these simulations, 1000 "case-control studies" were generated for each combination of true parameter values, sample sizes, numbers of strata, and study designs shown in table 4, a total of $6 \times 4 \times 3 \times 2 = 144$ combinations. In all, 144,000 case-control studies were generated. Each study provided a value for $\hat{AR}$, for SE, and for the three types of confidence intervals. (The logit interval was not calculated from studies for which the true AR was zero.) Figure 1 shows the distribution of the 1000 $\hat{AR}$ values for a typical combination. The mean of the 1000 $\hat{AR}$ values corresponding to each combination was calculated. Also calculated for each combination and each type of confidence interval were the mean length of the 1000 intervals and the "coverage probability," i.e., the proportion of the

Table 4

True parameter values, sample sizes, numbers of confounder strata, and types of study design used in simulations

Exposure prevalence (%) among cases (controls)*	Total no. of case (controls)*	No. of confounder strata	Study design
20 (20)	100 (100)	1	Matched
50 (20)	100 (300)	5	Stratified
50 (50)	500 (500)	10	
80 (20)	500 (1500)		
80 (50)			
80 (80)			

* For the unmatched designs, cases and controls were randomly assigned to exposure-confounder cells EC_k and $\bar{E}C_k$ according to the probabilities $P(EC_k) = p/K$, $P(\bar{E}C_k) = (1 - p)/K$, where p is the appropriate exposure prevalence and K is the number of confounder strata. For the matched designs, cases were randomly assigned according to this scheme. Controls were matched to cases in the ratios 1:1 or 3:1, and were randomly "exposed" with probability p.

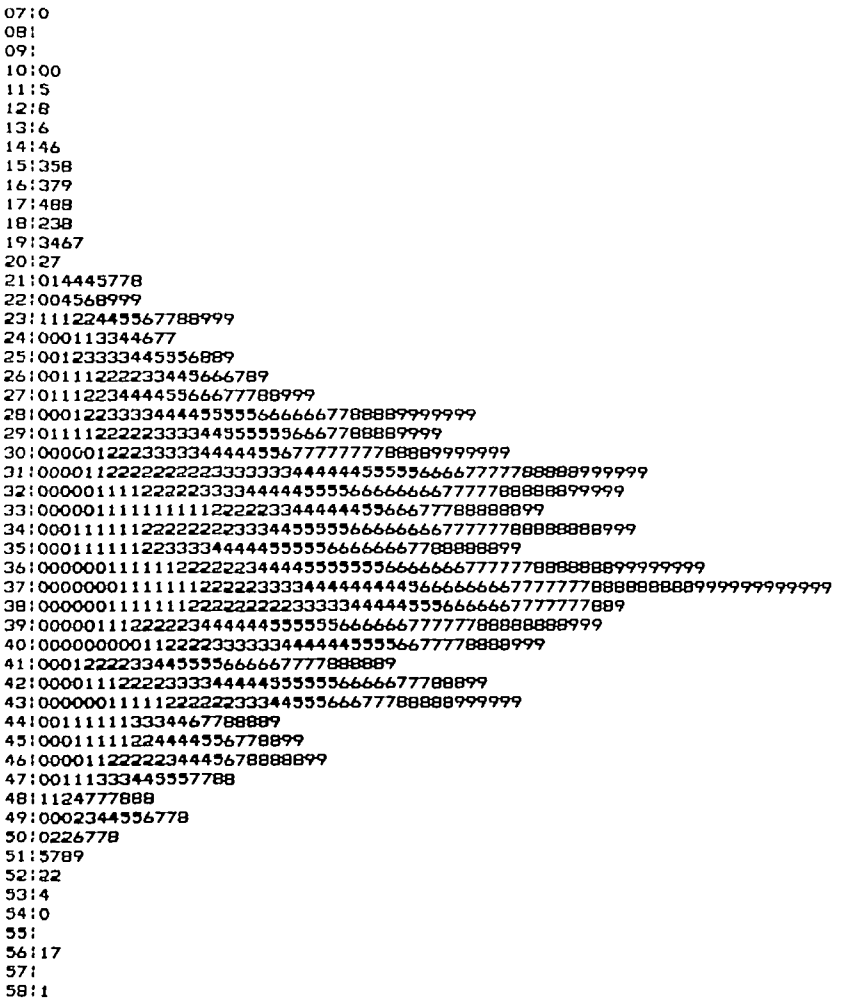

FIGURE 1. Distribution of $\hat{AR}$ values for 1000 matched case-control studies, each with 100 cases, 100 controls, and 10 strata of the matched factor. Stratum-specific exposure prevalence was 50 per cent for cases and 20 per cent for controls. The true attributable risk was 37.5 per cent. The graph indicates that one $\hat{AR}$ value was 7.0 per cent, two values were 10.0 per cent, one value was 11.5 per cent, etc. When held horizontally, the graph shows some left-skewness in the distribution.

1000 intervals containing the true AR. A detailed description of these simulations can be found in reference 7.

All estimates and confidence limits performed well in large samples (500 cases, 500 or more controls). The values for $\hat{AR}$ were close to the true AR. The three types of confidence intervals differed little in length, and all had close to the correct coverage probability.

In small samples (100 cases, 100 controls), the estimator $\hat{AR}$ tended to be too small. The bias increased with increasing number of strata and with increasing exposure prevalence among cases and controls. The bias was severe when 80 per cent of cases and controls were exposed to the risk factor of interest. In this situation (high exposure prevalence in the entire population), all estimates performed badly. Standard error estimates were large and were themselves unstable. Ninety-five per cent confidence intervals were very long and overly conservative,

with actual coverage probabilities in excess of 99 per cent. Apart from this exceptional situation, the bias in $\hat{AR}$ was not serious. For all three types of intervals, average length and extent of disagreement between actual and nominal coverage probabilities increased both with exposure prevalence and number of strata. No one type of confidence interval proved superior in achieving the nominal coverage probability. On the other hand, there were often appreciable differences in length, with (as expected) the log interval length exceeding that of the maximum likelihood interval, and for $0.21 < \hat{AR} < 0.79$, the maximum likelihood interval length exceeding that of the logit interval.

Discussion

The downward bias of attributable risk estimates obtained from small studies with large numbers of strata is not surprising, in view of similar results for the maximum likelihood estimate of a common odds ratio (13). Unfortunately, no simple, robust, Mantel-Haenszel type estimator for attributable risk seems to exist.

The large variability of $\hat{AR}$ when exposure prevalence is high in the control population can be understood by examining the attributable risk estimate of formula 5. If exposure is common among controls, even in moderate samples the unexposed control frequencies y_{2k} will often be zero (in practice, 0.5), leading to very large values for the expression in brackets. This fact explains the large standard errors and long confidence intervals for bladder cancer risk attributable to smoking obtained from the data in table 3. It indicates that attributable risk estimates for common characteristics such as coffee consumption are unreliable. Thus, the attributable risk estimate of 50 per cent suggested by MacMahon et al. (14) for coffee consumption as related to pancreatic cancer must have wide confidence intervals.

The simulation results suggest that confidence intervals for attributable risk are more reliable than are point estimates. In situations when the estimates have severe downward bias, the corresponding intervals, although too long, have higher than the nominal coverage probability. The simulations suggest no advantage to using the longer log interval. Instead, one should use the logit interval for attributable risk estimates in the range $0.21-0.79$, and the simple maximum likelihood interval for estimates outside this range.

References

1. Levin ML. The occurrence of lung cancer in man. Acta Un Intern Cancer 1953;19:531-41.
2. Walter SD. The estimation and interpretation of attributable risk in health research. Biometrics 1976;32:829-49.
3. Walter SD. The distribution of Levin's measure of attributable risk. Biometrika 1975;62:371-4.
4. Leung HM, Kupper LL. Comparisons of confidence intervals for attributable risk. Biometrics 1981;37:293-302.
5. Paffenbarger RS Jr, Kampert JB, Chang H-G. Characteristics that predict risk of breast cancer before and after the menopause. Am J Epidemiol 1980;112:258-68.
6. Walter SD. Prevention for multifactorial diseases. Am J Epidemiol 1980;112:409-16.
7. Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. Stat Med 1982:1(3).
8. Breslow N. Design and analysis of case-control studies. Annu Rev Public Health 1982;3:29-54.
9. Third National Cancer Survey. Incidence data. Washington, DC: National Cancer Institute Monograph 41 (DHEW publication no. (NIH) 75-87), 1975.
10. Vital Statistics of the United States, 1970. Volume II. Mortality, Part A. Rockville, MD: National Center for Health Statistics (DHEW publication no. (HRA) 75-1101), 1974:6-17.
11. Cole P, Monson RR, Haning H, et al. Smoking and cancer of the lower urinary tract. N Engl J Med 1971;284:129-34.
12. Miettinen O. Estimability and estimation in case-referent studies. Am J Epidemiol 1976; 103:226-35.
13. Breslow N. Odds ratio estimators when the data are sparse. Biometrika 1981;68:73-84.
14. MacMahon B, Yen S, Trichopoulos D, et al. Coffee and cancer of the pancreas. N Engl J Med 1981;304:630-3.

Appendix

Attributable risk estimation with ancillary information. When the confounder distribution $P(C_k)$ and the stratum-spe-

cific disease rates $P(D|C_k) = \phi_k$ are known with negligible standard error, the maximum likelihood estimate of residual risk is

$$\tilde{AR} = \Sigma_k \sigma_k \tilde{AR}_k. \quad \text{(A1)}$$

In formula A1, the quantities $\tilde{AR}_k$ are estimates of stratum-specific attributable risk:

$$\tilde{AR}_k = 1 - \{1/[\phi_k + (1 - \phi_k)(x_{\cdot k} y_{2k}/y_{\cdot k} x_{2k})]\}, \quad \text{(A2)}$$

and the weights σ_k are products of the known disease rates ϕ_k and confounder distribution $P(C_k)$ normalized to sum to 1:

$$\sigma_k = \phi_k P(C_k)/\Sigma_i \phi_i P(C_i). \quad \text{(A3)}$$

As the numbers of cases and controls increase, $\tilde{AR}$ becomes normally distributed about AR. The asymptotic standard error of $\tilde{AR}$ is consistently estimated by

$$\tilde{SE} = \left\{ \sum_k [(1 - \tilde{AR}_k)^2 \sigma_k (1 - \phi_k) y_{2k} x_{\cdot k} / x_{2k} y_{\cdot k}]^2 \left[\frac{x_{1k}}{x_{\cdot k} x_{2k}} + \frac{y_{1k}}{y_{\cdot k} y_{2k}} \right] \right\}^{1/2}. \quad \text{(A4)}$$

When C has only one level, the formulae for $\tilde{AR}$ and $\tilde{SE}$ agree with Walter's results (2).

Confidence intervals. The confidence intervals for attributable risk described below are based entirely on case-control data. Analogous confidence intervals incorporating ancillary information can be obtained by replacing $\hat{AR}$ by $\tilde{AR}$ and SE by $\tilde{SE}$.

An approximate $100(1 - \alpha)$ per cent confidence interval for attributable risk based on the estimates of formulae 7–9 is

$$(\hat{AR} - w, \hat{AR} + w),$$

where

$$w = z_{1-\alpha/2} SE.$$

The confidence interval produced by the log transformation $\log(1 - AR)$ is

$$[1 - \hat{RS} \exp(w/\hat{RS}), 1 - \hat{RS} \exp(-w/\hat{RS})],$$

where $\hat{RS}$ is the residual risk estimate $1 - \hat{AR}$. The logit transformation $\log[AR/(1 - AR)]$ yields

$$(\{1 + \hat{RS} \exp[w/(\hat{RS} \cdot \hat{AR})]/\hat{AR}\}^{-1}, \{1 + \hat{RS} \exp[-w/(\hat{RS} \cdot \hat{AR})]/\hat{AR}\}^{-1}).$$