ESTIMATING ATTRIBUTABLE RISK FROM CASE-CONTROL STUDIES

by

Alice S. Whittemore
Stanford University

TECHNICAL REPORT NO. 53
JANUARY 1982

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM
ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ESTIMATING ATTRIBUTABLE RISK FROM CASE-CONTROL STUDIES

Alice S. Whittemore
Department of Family, Community and Preventive Medicine
Stanford University School of Medicine
Stanford, CA 94305

ABSTRACT

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 using 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. In addition, simulated data are used to study the behavior of the estimates and confidence intervals when sample size is small relative to the number of strata of the confounding factor.

Key words: Levin's attributable risk; case-control studies; confounding; maximum likelihood estimates; small sample behavior.

1This research was supported by NIH Grant Nos. ES00070 and CA23214. An earlier version of this paper was presented at the annual meeting of the Society for Epidemiologic Research, June 1981, Snowbird, Utah.
INTRODUCTION

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, one that has been adjusted for correlations between the factor 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. First, when and how must Levin's measure be adjusted for confounding? Second, 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? Third, must one use different estimates if adjusting for age at diagnosis (matching factor) than if adjusting
for years of education (not a matching factor)? Fourth, what is the sampling variability of adjusted attributable risk estimates, and what are appropriate confidence intervals? Finally, 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 using 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.

**POPULATION MEASURES OF ATTRIBUTABLE RISK**

Consider a population that is classified into those exposed (E) and unexposed (Ê) to an etiologic agent for disease. For the population of white parous women discussed above, "exposure" is delay of first full-term pregnancy until age 24 or later. 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

\[ \frac{P(D|\bar{E})}{P(D)} , \]

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

\[ 1 - \frac{P(D|\bar{E})}{P(D)} \] (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 \( R = \frac{P(D|E)}{P(D|\bar{E})} \) and on exposure prevalence \( P(E) \), as can be seen by rewriting expression (1) as

\[ \frac{P(E)(R-1)}{1 + P(E)(R-1)} . \] (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 U.S. 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 were 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:

*The methods described apply either to disease incidence rates or to disease prevalence rates.
\[ \Sigma_k P(C_k) P(D|EC_k). \] (3)

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

\[ AR = 1 - \Sigma_k P(C_k) P(D|EC_k)/P(D). \] (4)

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

Sufficient conditions for equality between the adjusted and unadjusted measures of attributable risk [1] and [4] are either: (i) the factor \( C \) is uncorrelated with disease status for unexposed individuals; or (ii) \( C \) 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 [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

\[ 1 - [(0.17 \times 0.17) + (0.83 \times 0.08)]/0.24 = 60\%. \] Thus adjusted and unadjusted measures differ by \( 60\% - 39\% = 21\%. \) This difference is due to variation in exposure prevalence among the strata. It shows that estimates of attributable risk obtained by using in [2] adjusted estimates of relative risk and unadjusted estimates of exposure prevalence can be seriously in error.
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, then 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 controlling it by stratifying in the analysis. This fortuitous simplification holds both for estimates obtained solely from case-control data, and for estimates obtained using 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 (7).

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

$$\hat{AR} = 1 - \left[ \sum_{k} y_{*k} x_{2k}/ny_{2k} \right].$$

(5)

*Actually, one estimates an approximation to attributable risk, one whose accuracy improves with the infrequency of disease in the population. This distinction will be ignored.
In [5] \( n \) is the total number of cases, \( y_{*k} \) is the number of controls in stratum \( k \), and \( x_{2k} \) and \( y_{2k} \) represent the number of unexposed cases and the number of unexposed controls in stratum \( k \). Standard large sample methods can be used to show that as the numbers of cases and controls in each stratum increase, \( \hat{\Delta}R \) becomes normally distributed, with asymptotic mean \( \Delta R \) of [4]. A consistent estimate of the asymptotic standard error of \( \Delta R \) is

\[
SE = \frac{1}{n} \left\{ \sum_k (y_{*k} x_{2k}/y_{2k})^2 \left( \frac{1}{y_{2k}} + \frac{y_{1k}}{y_{*k} y_{2k}} \right) - n(1-\hat{\Delta}R)^2 \right\}^{1/2}, \tag{6}
\]

where \( y_{1k} \) is the number of exposed controls in stratum \( k \). The unadjusted estimate obtained by pooling the data over a matching factor is inconsistent, except in the trivial case when the factor is not a confounder (7). Breslow (8) has noted that in practice the analogous odds-ratio estimation procedure frequently involves small bias, and the bias may also be small in this case.

Attributable risk can be estimated 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 (A.1 - A.5) give the maximum likelihood estimate \( \tilde{\Delta}R \) and the estimated asymptotic standard error \( \tilde{SE} \) corresponding to this assumption.

**Confidence Intervals.** The upper and lower endpoints of an approximate 95% confidence interval for attributable risk, based on the maximum likelihood estimates [5, 6], are \( \Delta R \pm 1.96 \ SE \). This type of interval shall be called the maximum likelihood (ML) interval. The analogous confidence interval incorporating ancillary information is obtained by replacing \( \Delta R \) and \( SE \) by the corresponding estimates \( \tilde{\Delta}R \) 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 \( \log(1 - \text{AR}) \) suggested by Walter (2, 3); the other is the logit transformation \( \log(\text{AR}/(1 - \text{AR})) \) proposed by Leung and Kupper (4). The log and logit confidence intervals produced by these transformations are given in the Appendix.

The lengths of these three types of intervals are related in the following way: (i) the ML interval is always shorter than the log interval (except in the trivial case when both have length zero) (7); (ii) the logit interval is shorter than the ML interval whenever the attributable risk estimate is between 21% and 79% (5). 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. The zero cells in the table mandate some adjustment of formulae (5) and (6). Here and in the simulations such cells are assigned the value .5. 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 \( \hat{\text{AR}} = 14.8\% \), with estimated standard error \( \text{SE} = 3.1\% \). The age-adjusted 95% log, ML, and logit confidence intervals for \( \text{AR} \) are \( \text{LOG} = (8.5, 20.7); \text{ML} = (8.7, 20.9); \text{LOGIT} = (9.8, 22.0) \). The attributable risk estimate adjusted both for age and years of education is \( \hat{\text{AR}} = 10.5\% \) with \( \text{SE} = 3.8\% \). The unadjusted and inconsistent estimate obtained by pooling the data over age at diagnosis and years of education is \( 17.1\% \) with \( \text{SE} = 3.5\% \). Although the differences among the three estimates \( \hat{\text{AR}} \) are not statistically significant,
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, by using age-specific disease rates and population census data. Column 5 of Table 1 contains breast cancer incidence rates for white females obtained from the Third National Cancer Survey (9). Column 6 contains US 1970 census data (10). Ignoring statistical variability and nonparous women in these data, I have used them in Appendix formulae [A.1.-A.5] to obtain the age-adjusted attributable risk estimate $\tilde{AR} = 13.5\%$, with $\tilde{SE} = 3.3\%$. 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 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 for smoking the age-adjusted attributable risk estimate $\hat{AR} = 30.0\%$ with estimated standard error $SE = 19.7\%$. Thus the three 95% 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 $AR = 0\%$. The unadjusted estimate obtained by pooling the data over age is $\hat{AR} = 40.0\%$, with a smaller but still appreciable standard error of $SE = 10.5\%$.

The variability of the adjusted estimate is not reduced by incorporating into the estimate the incidence and population data shown in columns 5 and 6 of Table 3. The resulting estimates are $\tilde{AR} = 30.1\%$ with $\tilde{SE} = 19.2\%$, 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%, in contrast to a corresponding length of only 11.8% for the breast cancer data. The simulation studies described in the next section show that this difference is due less to differences in sample size than to the high smoking prevalence (75%) among the controls.

SIMULATION STUDIES

Methods. Two simulation projects were conducted. The first project was designed 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 do so, 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. For each combination, the following were calculated:

(i) 1000 $\hat{AR}$ values, one for each case-control study. Figure 1 shows the distribution of $\hat{AR}$ values for a typical combination; (ii) the mean and the standard deviation (SD) of these 1000 $\hat{AR}$ values. For each combination and for each of the three types of confidence interval* the following were also calculated: (i) 1000 confidence intervals, one for each study; (ii) the mean length of the 1000 intervals; (iii) the proportion of the 1000 intervals containing the true $AR$.

The second project was designed to examine the accuracy of $SE$ and of Efron's bootstrap (13) as estimates of the standard deviation of $\hat{AR}$.

*The logit interval was not calculated for combinations in which the true AR was zero.
Accordingly, 10 studies were generated for each of the 144 combinations shown in Table 4, a total of 1440 studies. For each combination the following were calculated: (i) 10 values of SE, one for each study; (ii) the mean of the 10 SE values; (iii) 10 bootstrap estimates; (iv) the mean of the 10 bootstrap estimates. The two means and the true asymptotic standard error were compared with the value SD obtained from the first project. A detailed description of these simulations and of their results can be found in (7).

**Results.** All estimates and confidence limits performed well in large samples (500 cases, 500 or more controls). The values \( \hat{AR} \) were close to the true AR. The three types of confidence interval 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% 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 percent confidence intervals were very long and overly conservative, with actual coverage probabilities in excess of 99%. 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 ML interval, and for \( .21 < \hat{AR} < .79 \), the ML interval length exceeding that of the logit interval. The standard error estimate \( SE \) showed good agreement with \( SD \). By contrast, the true asymptotic standard error was often too small, and the bootstrap often too large.

**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 (14). Unfortunately, no simple, robust, Mantel-Haenzel 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, then even in moderate samples the unexposed control frequencies \( y_{2k} \) will often be zero (in practice, .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% suggested by McMahon et al. for coffee consumption as related to pancreatic cancer (15) 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 .21 to .79, and the simple ML interval for estimates outside this range.
REFERENCES


<table>
<thead>
<tr>
<th>AGE AT DIAGNOSIS</th>
<th>CASES</th>
<th>CONTROLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>25-29</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>30-34</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>35-39</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>40-44</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>45-49</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>50-59</td>
<td>43</td>
<td>44</td>
</tr>
<tr>
<td>55-59</td>
<td>72</td>
<td>74</td>
</tr>
<tr>
<td>60-64</td>
<td>35</td>
<td>38</td>
</tr>
<tr>
<td>65-69</td>
<td>38</td>
<td>35</td>
</tr>
<tr>
<td>70-74</td>
<td>31</td>
<td>38</td>
</tr>
<tr>
<td>75+</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>836</td>
<td>764</td>
</tr>
</tbody>
</table>

**TABLE 1**

Number of White Parous Study Subjects (5) and US White Breast Cancer Incidence Rates, Years of Education; Female Population in thousands (10).

<table>
<thead>
<tr>
<th>YRS OF EDUCATION</th>
<th>CASES</th>
<th>CONTROLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;= 12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&gt; 12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1970 US WHITE FEMALE BREAST CANCER INCIDENCE RATE PER 105 (9) IN THOUSANDS (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.7 5962.1 0.007 1 22.5 5042.4 0.015 13 52.5 4936.5 0.035 16 103.7 5412.3 0.075 18 159.2 5587.0 0.119</td>
</tr>
</tbody>
</table>

a) Controls were matched to cases on age at diagnosis.

b) $\sigma_k$ is the product of column 5 and column 6, normalized so that $\sum \sigma_k = 1$. 

16
<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Nonsmoker</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unexposed</td>
<td>Exposed</td>
<td>Unexposed</td>
</tr>
<tr>
<td>No. of diseased individuals</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Size of source population in $10^3$</td>
<td>36</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Disease rate $\cdot 10^3$</td>
<td>0.17</td>
<td>0.50</td>
<td>0.08</td>
</tr>
<tr>
<td>Relative risk for exposure</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Exposure prevalence (%)</td>
<td>30</td>
<td>95</td>
<td>83</td>
</tr>
<tr>
<td>Attributable risk (%) for exposure</td>
<td>38</td>
<td>66</td>
<td>39</td>
</tr>
<tr>
<td>% of population in smoking stratum</td>
<td>17</td>
<td>83</td>
<td>100</td>
</tr>
</tbody>
</table>
TABLE 3
Case-Control Data and Incidence Rates for Bladder Cancer Among Males in Eastern Massachusetts a)

<table>
<thead>
<tr>
<th>Age at Diagnosis</th>
<th>Number of Study Subjects</th>
<th>Incidence Rates $\cdot 10^5$</th>
<th>Size of Source Population in $10^5$</th>
<th>$\sigma_k$ c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Smoker</td>
<td>Control b</td>
<td>Nonsmoker</td>
<td>Case</td>
</tr>
<tr>
<td>50-54</td>
<td>24</td>
<td>22</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>55-59</td>
<td>35</td>
<td>35</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>60-64</td>
<td>31</td>
<td>38</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>65-69</td>
<td>46</td>
<td>42</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>70-74</td>
<td>60</td>
<td>51</td>
<td>13</td>
<td>28</td>
</tr>
<tr>
<td>74-79</td>
<td>39</td>
<td>32</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>220</td>
<td>42</td>
<td>74</td>
</tr>
</tbody>
</table>

a) As published in (12), Table 1. All data are from studies by Cole et al. (11).
b) Controls were matched to cases by year of birth.
c) $\sigma_k$ is the product of column 5 and column 6, normalized so that $\Sigma_k \sigma_k = 1$. 
### TABLE 4

True Parameter Values, Sample Sizes, Numbers of Confounder Strata, and Types of Study Design Used in Simulations

<table>
<thead>
<tr>
<th>Exposure Prevalence (%) Among Cases (Controls) *</th>
<th>Number of Cases (Controls)</th>
<th>Number of Confounder Strata</th>
<th>Study Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 (20)</td>
<td>100 (100)</td>
<td>1</td>
<td>Matched</td>
</tr>
<tr>
<td>50 (20)</td>
<td>100 (300)</td>
<td>5</td>
<td>Stratified</td>
</tr>
<tr>
<td>50 (50)</td>
<td>500 (500)</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>80 (20)</td>
<td>500 (1500)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80 (50)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80 (80)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Stratum specific exposure prevalence was assumed constant across strata.*
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% for cases and 20% for controls. The true attributable risk was 37.5%. The graph indicates that one $\hat{AR}$ value was 7.0%, two values were 10.0%, one value was 11.5%, etc.
APPENDIX

Attributable Risk Estimation with Ancillary Information. When the confounder distribution \( P(C_k) \) and the stratum-specific disease rates \( P(D|C_k) = \phi_k \) are known with negligible standard error, the maximum likelihood estimate of residual risk is

\[
\tilde{\mathcal{R}} = \sum_k \sigma_k \tilde{\mathcal{R}}_k.
\]

(A.1)

In (A.1) the quantities \( \tilde{\mathcal{R}}_k \) are estimates of stratum-specific attributable risk:

\[
\tilde{\mathcal{R}}_k = 1 - \left\{ \frac{1}{\phi_k + (1-\phi_k)(x_{k}\gamma_{2k}/y_{k}\gamma_{2k})} \right\},
\]

(A.2)

and the weights \( \sigma_k \) are products of the known disease rates \( \phi_k \) and confounder distribution \( P(C_k) \) normalized to sum to 1:

\[
\sigma_k = \phi_k P(C_k)/\sum_k \phi_k P(C_k).
\]

(A.3)

As the numbers of cases and controls increase \( \tilde{\mathcal{R}} \) becomes normally distributed about \( \mathcal{R} \). The asymptotic standard error of \( \tilde{\mathcal{R}} \) is consistently estimated by

\[
\tilde{SE} = \left\{ \sum_k [(1-\tilde{\mathcal{R}}_k)^2 \sigma_k (1-\phi_k) y_{2k} x_{k}/x_{2k} y_{k}]^2 \left[ \frac{x_{1k}}{x_{k} x_{2k}} + \frac{y_{1k}}{y_{k} y_{2k}} \right] \right\}^{1/2}.
\]

(A.4)

When \( C \) has only one level the formulae for \( \tilde{\mathcal{R}} \) 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 $100(1-\alpha)\%$ confidence interval for attributable risk based on the estimates [7 - 9] is

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

where

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

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

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