STATISTICAL METHODS FOR ESTIMATING ATTRIBUTABLE RISK
FROM RETROSPECTIVE DATA

ALICE S. WHITTEMORE

TECHNICAL REPORT NO. 54
JANUARY 1982

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
STATISTICAL METHODS FOR ESTIMATING ATTRIBUTABLE RISK
FROM RETROSPECTIVE DATA

by

Alice S. Whittemore
Stanford University

TECHNICAL REPORT NO. 54
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
Summary

Levin's measure of attributable risk\(^1\) is extended to adjust for confounding by etiologic factors other than the exposure of interest. This extended measure can be estimated from case-control data provided either (i) exposure prevalence within each stratum of the confounding factor can be estimated from the control data; or (ii) additional information is available concerning the confounder distribution and the stratum-specific disease rates. In both cases maximum likelihood estimates and their estimated asymptotic variances are given, and are shown to be independent of the sampling design (matched vs. random). Computer simulations are used to investigate the behavior of these estimates and of three types of confidence intervals when sample size is small relative to the number of confounder strata. The simulations indicate that attributable risk estimates tend to be too low. The bias is not serious except when exposure prevalence is high among controls. In this case the estimates and their standard error estimates are also highly unstable. In general, the asymptotic standard error estimates performed quite well, even in small samples, and even when the true asymptotic standard error was too small. By contrast, the bootstrap estimate\(^2\) tended to be too large. None of the three confidence intervals proved superior in accuracy to the other two. Thus there appears to be no advantage to using the log-based interval suggested by Walter\(^3,4\) which is always longer than the simpler symmetric interval.

Key words: Attributable risk; case-control studies; maximum likelihood estimates; small sample behavior.
1. Introduction

To evaluate disease prevention strategies, one must estimate the fraction of disease that might be avoided by reduction or elimination of population exposure to a given etiologic agent. Relative risk and odds-ratio measures are inadequate for this purpose. While the relative risk of pancreatic cancer due to familial relapsing pancreatitis is substantial, this hereditary condition is so rare that its elimination would prevent only a small fraction of pancreatic cancers. By contrast, the relative risk of pancreatic cancer associated with cigarette smoking is only about 2; nevertheless, elimination of this ubiquitous exposure would prevent a sizeable fraction of the disease.

Thus it is useful to have a measure of the proportion of disease attributable to a given exposure, one that accounts for correlations between the exposure and other etiologic factors. Levin's simple measure of attributable risk\(^1,3,4\), first proposed in 1953, does not account for such confounding. The purpose of this paper is (i) to extend Levin's definition to account for confounding by other etiologic factors; (ii) to present point estimates and confidence intervals for this extended measure based on matched or randomly sampled case-control data; and (iii) to investigate, using simulated data, the small sample properties of these estimates and confidence intervals. The extension of Levin's measure, defined below, has been described independently by Walter\(^5\).

2. Definitions and Theory

Consider a population that is classified into two groups according to the presence or absence of a characteristic \( E \) called exposure. Let
\( P(D|E), P(D|\bar{E}) \) and \( P(D) \) represent the incidence or prevalence rate of disease among the exposed, the unexposed, and the entire population, respectively. We assume that \( P(D|E) \geq P(D|\bar{E}) \). The ratio

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

represents the fraction of all disease that would persist if the entire population contracted disease at the rate of those unexposed to \( E \); this ratio shall be called the residual risk corresponding to \( E \). The remaining fraction,

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

(1)

represents the proportion of all disease that would be prevented if the rates for the unexposed prevailed in the population. Levin\(^1\) called this fraction the (population) risk attributable to \( E \).

If exposure is correlated with an etiologic factor \( C \), then elimination of \( E \) from the population need not produce the unexposed disease rate \( P(D|\bar{E}) \). Instead, the disease rate in the absence of \( E \) would be a weighted average of stratum-specific unexposed rates. In symbols, this rate is

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

(2)

where \( P(C_k) \) represents the fraction of the population in the \( k \)th stratum of \( C \), and \( P(D|\bar{E} \ C_k) \) is the stratum-specific disease rate among the unexposed. Substitution of (2) for \( P(D|\bar{E}) \) in (1) yields a measure of attributable risk which has been adjusted for \( C \):

\[
AR = 1 - \Sigma_k P(C_k) P(D|\bar{E} \ C_k)/P(D).
\]

(3)

Note that (3) reduces to (1) when \( C \) has only one level. By using in (1) the identity
\[ P(D|\bar{E}) = \sum_k P(D|\bar{E}, C_k) P(C_k|\bar{E}), \]

where the quantities \( P(C_k|\bar{E}) \) represent the confounder distribution among the unexposed, one can express the difference between the unadjusted measure (1) and the adjusted measure (3) as

\[ \{\sum_k [P(C_k) - P(C_k|\bar{E})] P(D|\bar{E}, C_k)\}/P(D). \]  

(4)

It is evident from (4) that either of the following is a sufficient condition for agreement between adjusted and unadjusted measures of attributable risk: (i) \( C \) is unrelated to disease among the unexposed, so that \( P(D|\bar{E}, C_k) = P(D|\bar{E}) \) for all strata \( C_k \); or (ii) \( E \) and \( C \) are uncorrelated, so that \( P(C_k|\bar{E}) = P(C_k) \) for all \( k \). These conditions, noted by Walter\(^5\), are also necessary when \( C \) has only two levels. They imply that confounding factors must be controlled in measuring risk attributable to \( E \).

3. Estimation

The attributable risk \( AR \) in (3) can be rewritten

\[ AR = \sum_k P(C_k|D)AR_k, \]  

(5)

where

\[ AR_k = 1 - [P(D|\bar{E}, C_k)/P(D|C_k)] = 1 - [P(\bar{E}|C_k, D)/P(\bar{E}|C_k)] \]

(6)

is the stratum-specific attributable risk. To estimate \( AR \) solely from case-control data one needs the "rare disease assumption" that within each stratum \( C_k \) the proportion \( P(\bar{E}|C_k) \) of unexposed can be approximated by the corresponding proportion \( P(\bar{E}|C_k, D) \) among the nondiseased. If so, then from (5) and (6) we have approximately
\[ \text{AR} = 1 - \sum_k P(C_k \mid D) \frac{P(\bar{E} \mid C_k D)}{P(\bar{E} \mid C_k \bar{D})}, \]  

(7)
a parameter which can be estimated from case-control data. Equation (7) holds only in the limit as the probability of disease tends to zero, a distinction that will be ignored. When \( C \) has only one level, (7) reduces to the unadjusted measure

\[ 1 - \left[ P(\bar{E} \mid D) / P(\bar{E} \mid \bar{D}) \right]. \]  

(8)

3.1. Random Sampling. The investigator using a random sampling design chooses random samples consisting of \( n \) diseased subjects (cases) and \( m \) disease-free subjects (controls). The case and control exposure and confounder data are then arranged into the \( 2 \times 2 \times K \) Table 1. The \( 2K \) dimensional vectors \( \mathbf{x} = (x_{11}, \ldots, x_{1k}, x_{21}, \ldots, x_{2k}) \) and \( \mathbf{y} = (y_{11}, \ldots, y_{1k}, y_{21}, \ldots, y_{2k}) \) of case and control frequencies each has a multinomial distribution, with respective parameters \( n, P(E \mid C_k D), P(\bar{E} \mid C_k D) \) and \( m, P(E \mid C_k \bar{D}), P(\bar{E} \mid C_k \bar{D}) \) \( k=1, \ldots, k \); moreover \( \mathbf{x} \) and \( \mathbf{y} \) are independent. The maximum likelihood estimate (MLE) of \( \text{AR} \) in (7) is

\[ \hat{\text{AR}} = 1 - \left( \sum_k y_{.k} x_{.2k} / n y_{.2k} \right), \]  

(9)

where the dot subscript is used to indicate summation. Standard large sample methods can be used to show that as \( m \) and \( n \) increase, \( \hat{\text{AR}} \) is normally distributed with asymptotic mean \( \text{AR} \) and with variance given by expression (A.1) in Appendix A. A consistent estimator for this variance is

\[ u^2 = \frac{1}{n} \left[ \sum_k (y_{.k} x_{.2k} / y_{.2k})^2 \left( \frac{1}{x_{.2k}} + \frac{y_{1k}}{y_{.k} y_{.2k}} \right) - n(1-\hat{\text{AR}})^2 \right]. \]  

(10)

Relatively accurate population estimates may exist both for the stratum-specific disease rates \( \phi_k = P(D \mid C_k) \) and for the confounder distribution.
\( P(C_k) \). If these probabilities can be assumed known, then \( AR \) can be estimated from case-control data without the rare disease assumption. To do so, rewrite the stratum-specific attributable risks \( AR_k \) of (6) as

\[
AR_k = 1 - \left[ \phi_k + (1-\phi_k) \frac{P(\bar{E}|C_k \bar{D})}{P(\bar{E}|C_k D)} \right]^{-1}.
\] (11)

The MLE of \( AR \) as given by (5) and (11) is

\[
\tilde{AR} = \sum \sigma_k \tilde{AR}_k.
\] (12)

In (12) \( \sigma_k \) is the known parameter \( P(C_k | D) = \phi_k P(C_k) / \sum_k \phi_k P(C_k) \), and \( \tilde{AR}_k \) is obtained by substituting in (11) sample moments \( x_{2k}/x_{*k} \) and \( y_{2k}/y_{*k} \) for the unknown probabilities \( P(\bar{E}|C_k D) \) and \( P(\bar{E}|C_k \bar{D}) \). The large sample variance of \( \tilde{AR} \) is given by appendix equations (A.7, A.8).

This variance is consistently estimated by

\[
u^2 = \sum_k [(1-\tilde{AR}_k)^2 \sigma_k (1-\phi_k) (y_{2k}/y_{*k} x_{2k}/x_{*k})] [\frac{x_{1k}}{x_{*k} x_{2k}} + \frac{y_{1k}}{y_{*k} y_{2k}}].
\] (13)

The estimates \( \hat{AR}, \tilde{AR}, \nu^2 \) and \( v^2 \) all reduce to the corresponding unadjusted expressions described by Walter\(^3\) when \( K=1 \).

3.2. Matched Sampling. The investigator using a matched design obtains information on the variables \( E \) and \( C \) for each of \( n \) randomly sampled cases. These data can again be arranged as in Table 1. The vector \( \tilde{x} \) of case frequencies has the multinomial distribution described for the random sampling design. Corresponding to the \( x_{*k} \) cases in stratum \( C_k \) the investigator randomly chooses \( R_k x_{*k} \) controls from the same stratum, where \( R_k \) need not be an integer. For this design the control frequency vector \( \gamma \) of Table 1 is not independent of \( \tilde{x} \).
Instead, given the $x_{1k}$, the $y_{1k}$ are independent binomial variates with parameters $R_k x_{1k}$ and $P(E|\bar{E} C_k)$, $k = 1, \ldots, K$.

It is evident from (7) that under the rare disease assumption $AR$ is estimable from data obtained in the matched design; the MLE is again given by (9). It can also be seen that the pooled estimate $1 - (\sum_k R_k x_{1k} x_{2.})/(n y_{2.})$ for $AR$ is inconsistent. Indeed the pooled estimate converges in probability to the expression

$$1 - [P(\bar{E}|D) \sum_k R_k P(C_k|D) / \sum_k R_k P(C_k|D) P(\bar{E}|C_k \bar{D})],$$

(14)

which in general equals neither (7) nor (8). When the control:case ratio $R_k$ is the same for all strata, (14) reduces to

$$1 - [P(\bar{E}|D) / \sum_k P(\bar{E}|C_k \bar{D})].$$

(15)

Sufficient conditions for equality between (15) and (7) are either (i) the factor $C$ is independent of disease among both exposed and unexposed; or (ii) $C$ is independent of exposure among the controls. It is unlikely that either of these conditions will hold in practice, for if so, matching on $C$ would be unnecessary.

As $n$ becomes large the distribution for $\hat{AR}$ becomes normal with mean $AR$ and with variance given by appendix equation (A.5), an expression almost identical to the asymptotic variance (A.1) corresponding to the unmatched design. Use of sample moments for the unknown parameters in (A.5) again yields the consistent estimator $u^2$ of (10).

When the stratum-specific disease rates and the confounder distribution can be assumed known, the MLE for $AR$ obtained from the matched design is again $\tilde{AR}$ of (13), with asymptotic variance given by appendix equations (A.9, A.7), which is consistently estimated by $v^2$ of (13).
Thus both matched and random sampling designs produce the same point estimates \( \hat{AR} \) and \( \tilde{AR} \), and the same variance estimates \( u^2 \) and \( v^2 \).

3.3. Confidence Intervals. An approximate 100(1-\(\alpha\))% confidence interval for attributable risk based on the maximum likelihood methods of the preceding sections is

\[
C_{ML}: (\hat{AR} - w, \hat{AR} + w),
\]

where

\[
w = z_{1-\alpha/2} u.
\]

Various monotonic transformations \( g(AR) \) have been proposed to produce confidence intervals for attributable risk with smaller length or more accurate coverage probability. The log transformation \( g(AR) = \ln(1-AR) \) suggested by Walter\(^3,4\) gives rise to the 100(1-\(\alpha\))% confidence interval

\[
C_{LOG}: (1 - \hat{\beta}e^{w/\hat{\beta}}, 1 - \hat{\beta}e^{-w/\hat{\beta}}),
\]

where \( \hat{\beta} = 1-\hat{AR} \) is the residual risk estimate. The logit transformation \( g(AR) = \ln[AR/(1-AR)] \) (0 < AR < 1) proposed by Leung and Kupper\(^6\) yields

\[
C_{LT}: \left( [1 + (\hat{\beta}/\hat{AR})e^{w/\hat{\beta} \hat{AR}}]^{-1}, [1 + (\hat{\beta}/\hat{AR})e^{-w/\hat{\beta} \hat{AR}}]^{-1} \right).
\]

Confidence intervals for use when the investigator knows the stratum specific disease probabilities and confounder distribution are obtained by replacing \( \hat{AR} \) by \( \tilde{AR} \) and \( u \) by \( v \) in (16 - 19).

It is shown in Appendix B that the ML interval is always shorter than the log interval (except for the trivial case when both intervals have length zero). Leung and Kupper\(^6\) have also shown that the logit interval is shorter than the ML interval for attributable risk estimates between 21% and 79%, while the reverse tends to hold for estimates less than 21%.
4. Example

Table 2 gives data obtained from a case-control study of breast cancer conducted by Paffenbarger et al. The tabular data are restricted to white parous women. The 2024 controls were matched to the 1051 cases on age. The exposure of interest is age at first childbirth ($E \geq 24$ years). In estimating adjusted attributable risk, the zero cells in the table were assigned the value 0.5. Assuming that population exposure prevalence can be estimated from the controls, one finds from the study data alone that the age-adjusted attributable risk estimate is $\hat{AR} = 14.8\%$, with estimated standard error $u = 3.1\%$. Age-adjusted 95% confidence intervals for AR are $C_{LT} = (9.8, 22.0)$; $C_{ML} = (8.7, 20.9)$; and $C_{LOG} = (8.5, 20.7)$. The attributable risk estimate adjusted both for age (matching factor) and years of education (not a matching factor) is 10.5%, with $u = 3.8\%$. The unadjusted (inconsistent) attributable risk estimate is 17.1% with $u = 3.5\%$. The differences among these three estimates are not significant; nevertheless 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 by supplementing the case-control data with 1970 census data and with white female breast cancer incidence rates obtained from the Third National Cancer Survey, shown in Table 2. This yields the estimate $\tilde{AR} = 13.5\%$ with $v = 3.3\%$, which is consistent with the corresponding age-adjusted estimates obtained solely from the study data.

5. Simulation Results

Simulated data are used in this section to examine the small sample behavior of $\hat{AR}$ and $u$ and of confidence intervals based on log, ML
and logit procedures. Two simulation studies were undertaken. In the first study 1000 trials (i.e. "case-control studies") were conducted for each of 72 combinations of parameter values, numbers of strata and sample sizes, using both matched and unmatched designs, a total of $72 \times 2 \times 1000 = 144,000$ trials. For each trial conducted according to the unmatched design, two independent samples of $n$ and $m$ independent multinomial random variables were generated. For each trial conducted according to the matched design, $n$ independent multinomial random variables were generated. Corresponding to each such "case" variable, $R$ independent Bernoulli "control" variables were generated with exposure probability determined by the stratum of the case variable.

For simplicity, the stratum-specific exposure probabilities among cases and controls were taken to be:

$$
P(E|D C_k) = \frac{\pi}{K}, \quad P(\bar{E}|D C_k) = \frac{(1-\pi)}{K}$$

$$
P(E|\bar{D} C_k) = \frac{p}{K}, \quad P(\bar{E}|\bar{D} C_k) = \frac{(1-p)}{K}.$$  \hspace{1cm} (20)

This model, in which the true adjusted and unadjusted values of $AR$ are the same, was chosen to minimize the number of adjustable parameters. Of course, since $C$ is not a confounder, matching or stratification is unnecessary. A model in which $C$ is a confounding factor was also investigated. As the two sets of results were similar, discussion of the latter investigation is omitted. All possible combinations of the following sample sizes, numbers of strata and parameter values were considered: $n = 100, 500$; $K = 1, 5, 10$; control:case ratio $R(=m/n) = 1, 3$; $\pi = .2, .5, .8$; and $p = .2, .5, .8$ (subject to $p \leq \pi$). The cell frequencies were assigned the value $.5$ when they equalled zero.
Table 3 shows average values of attributable risk $\hat{AR}$ for the matched designs. Corresponding results (not shown) for the unmatched designs were similar. Table 3 indicates that attributable risk estimates are biased downward; this bias increases with increasing number of strata, with increasing values of $\pi$ and $p$, and with decreasing sample size. The bias is unusually severe for the case $\pi = p = 80\%$. This suggests that attributable risk estimates are unreliable when exposure prevalence is high among controls.

The bias is not surprising in view of the inconsistency of the maximum likelihood estimator for the common odds-ratio in several 2×2 tables, as the number of tables increases while the data within each table remain sparse. Indeed, it can be shown that in the extreme example of matched studies with one case and one control per stratum, the estimator $\hat{AR}$ for a true $AR$ of zero converges in probability to a negative number.

Observed confidence coefficients and average lengths for each of the three types of 95% confidence interval were also calculated. The logit interval was not calculated in trials for which the true $AR$ was zero, because the interval is undefined when $\hat{AR} < 0$, and this occurred in more than half of such trials. For trials involving 500 cases, all observed coefficients were between 94% and 96%, and none of the three methods proved uniformly more accurate than the others. The average lengths of the intervals differed from each other by no more than .3% and increased with the number of strata. The increase was slight except for the case $\pi = p = 80\%$, $R=1$, for which the average length increased from 50.4% (K=1) to 66.5% (K=10).
Observed coefficients and average lengths for trials with 100 cases are shown in Table 4 and Table 5. It is evident from the tables that, for all three types of intervals, average length and extent of disagreement between observed and nominal coefficients increase with both exposure prevalence and number of strata. In case $\pi = p = 80\%$ the ML and log intervals are conservatively long. According to Table 4, no one method is clearly superior in achieving the nominal coefficient. As expected, Table 5 shows a consistent gradient in length among the three types of interval, with the logit interval shortest and the log interval longest.

The second simulation study was designed to investigate the accuracy of $\hat{u}$ and Efron's bootstrap $^2$ as estimates of the standard error of $\hat{AR}$ when sample size is small relative to number of confounder strata. Ten trials or "case-control studies" were conducted for each combination of the parameter values, strata numbers, sample sizes and study designs described previously, a total of 1440 trials. Each trial yielded two estimates of standard error: (i) $\hat{u}$ of (10); (ii) a bootstrap estimate BSE. The computations used to obtain BSE are described in Appendix C.

Values of $\hat{u}$ and BSE, averaged over the ten trials, were compared with the standard deviation $SD$ of 1000 attributable risk estimates, defined by

$$SD = \left[ \frac{1}{999} \sum_{i=1}^{1000} (\hat{AR}_i - \overline{AR})^2 \right]^{1/2}, \quad (21)$$

where $\overline{AR}$ is the average of the 1000 estimates. SD was also compared with the true asymptotic standard error (A.2) or (A.7), which for the simple model (20) depends neither on $K$ nor on the study design. Except for the problem case $\pi = p = 80\%$, $\hat{u}$ agreed well with SD, even for trials involving only 100 cases and 100 controls. By contrast the bootstrap esti-
mate tended to be larger than SD, at times substantially so, while the
asymptotic standard error was often too small. These discrepancies were
quite pronounced in the troublesome case $\pi = p = 80\%$, as can be seen in
Table 6. The large and fluctuating standard errors shown in the table
provide further evidence of the instability of attributable risk estimates
when exposure is common among both cases and controls. The instability
is understood upon examination of $\hat{AR}$ in (9). When exposure is common
among controls, the denominators $y_{2k}$ have probability close to zero,
and in small samples these variables will often be zero (in practice,
.5), leading to very large values for the expression in brackets.

6. Conclusions

The results of Walter and of the preceding sections indicate that
adjustment for confounding is necessary when estimating attributable risk.
Adjusted estimates can be obtained from case-control studies provided that
stratum-specific population exposure prevalence can be estimated from con-
trol data. Estimation is still possible when this assumption fails, pro-
vided stratum-specific disease rates are available. In both cases, the
investigator obtains the same estimates and confidence intervals for
attributable risk when controlling confounding by matching as when con-
trolling it by stratifying during analysis.

Attributable risk estimates obtained using the $2 \times 2$ table of pooled
data from a matched study are inconsistent. Breslow has noted that in
practice the analogous odds-ratio estimation procedure involves small
bias, and the bias may also be small in this case.

The simulation studies described in Section 5 show that in small
samples the maximum likelihood estimate for attributable risk tends to
be too low. The bias is usually small, provided the number of strata is not large relative to sample size. An exception occurs when exposure is common among controls. In this case the estimates have severe downward bias and large standard errors, particularly when sample size is small relative to the number of confounder strata. Confidence intervals are conservatively long, and standard error estimates are themselves unstable.

In general the asymptotic standard error estimate showed good agreement with the standard deviation of $\hat{A}R$ from the simulation studies, even when sample sizes were small. The bootstrap estimate was frequently too large, and the true asymptotic standard error too small.

For large samples there was little difference among the three types of confidence intervals in coverage probability or in average length. For small samples there were often sizeable differences in length, with the log interval length exceeding that of the ML interval, and, when $0.21 < \hat{A}R < 0.79$, with the ML interval length exceeding that of the logit interval. The distributions of $\hat{A}R$ obtained from 1000 trials involving 100 cases and 100 controls were plotted for selected parameter combinations. These distributions were slightly skewed to the left, while the distributions of $\ln(1-\hat{A}R)$ were more symmetric. Although this suggests improved accuracy for the log interval, such improved accuracy was not supported by the simulation studies. Thus there appears to be no advantage to using the longer log interval. Instead one should use the logit interval for attributable risk estimates in the range $0.21$ to $0.79$, and the ML interval for attributable risk estimates outside this range.
REFERENCES


2. Efron, B. "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods", Biometrika (in press).


10. Breslow, N. "Odds ratio estimators when the data are sparse", Biometrika 68. 73-84 (1981).

### TABLE 1

Numbers of Cases and Controls by Exposure Status and Confounder Stratum

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Exposure Status</th>
<th>Confounder Stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>Exposed</td>
<td>$x_{11}$</td>
</tr>
<tr>
<td></td>
<td>Unexposed</td>
<td>$x_{21}$</td>
</tr>
<tr>
<td>controls</td>
<td>Exposed</td>
<td>$y_{11}$</td>
</tr>
<tr>
<td></td>
<td>Unexposed</td>
<td>$y_{21}$</td>
</tr>
</tbody>
</table>
TABLE 2
Case-Control Data and Incidence Rates for Breast Cancer Among US White Females

<table>
<thead>
<tr>
<th>AGE AT DIAGNOSIS</th>
<th>AGE AT FIRST CHILDBIRTH</th>
<th>YEARS OF EDUCATION</th>
<th>US WHITE FEMALE BREAST CANCER INCIDENCE RATE PER $10^5$ (9)</th>
<th>1970 US WHITE FEMALE POPULATION IN THOUSANDS (8)</th>
<th>$\sigma_k^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 24</td>
<td>&gt; 12</td>
<td>&lt; 24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-29</td>
<td>Cases</td>
<td>1</td>
<td>0</td>
<td>8.7</td>
<td>5962.1</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>30-34</td>
<td>Cases</td>
<td>9</td>
<td>0</td>
<td>22.5</td>
<td>5042.4</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>35-39</td>
<td>Cases</td>
<td>6</td>
<td>7</td>
<td>52.5</td>
<td>4936.5</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>20</td>
<td>8</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>40-44</td>
<td>Cases</td>
<td>19</td>
<td>8</td>
<td>103.7</td>
<td>5412.3</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>31</td>
<td>16</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>45-49</td>
<td>Cases</td>
<td>42</td>
<td>19</td>
<td>159.2</td>
<td>5587.0</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>60</td>
<td>29</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>50-54</td>
<td>Cases</td>
<td>45</td>
<td>26</td>
<td>171.7</td>
<td>5169.3</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>68</td>
<td>44</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>55-59</td>
<td>Cases</td>
<td>72</td>
<td>34</td>
<td>191.8</td>
<td>4695.6</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>97</td>
<td>54</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>60-64</td>
<td>Cases</td>
<td>43</td>
<td>22</td>
<td>226.2</td>
<td>4157.5</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>70</td>
<td>40</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>65-69</td>
<td>Cases</td>
<td>38</td>
<td>18</td>
<td>234.2</td>
<td>3491.1</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>70</td>
<td>24</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td>Cases</td>
<td>31</td>
<td>15</td>
<td>259.8</td>
<td>2874.5</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>28</td>
<td>24</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>75+</td>
<td>Cases</td>
<td>36</td>
<td>22</td>
<td>299.5</td>
<td>4319.1</td>
</tr>
<tr>
<td></td>
<td>Controls</td>
<td>39</td>
<td>41</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>836</td>
<td>451</td>
<td>764</td>
<td>1024</td>
<td>51647.4</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 $\Sigma_k \sigma_k = 1$. 
| P(E|D) (%) | P(E|D) (%) | AR (%) | Number of Strata | 100 Case:Control Ratio | 500 Case:Control Ratio |
|-----------|-----------|--------|------------------|------------------------|------------------------|
| 20        | 20        | 0      | 1                | -0.5 0.0               | -0.1 -0.1              |
|           |           |        | 5                | -1.6 -0.4              | -0.3 -0.1              |
|           |           |        | 10               | -3.0 -0.9              | -0.6 -0.2              |
| 50        | 20        | 37.5   | 1                | 37.9 37.0              | 37.5 37.5              |
|           |           |        | 5                | 36.7 37.0              | 37.3 37.3              |
|           |           |        | 10               | 35.6 36.9              | 37.3 37.6              |
| 50        | 0         | 0      | 1                | -1.0 -0.1              | -0.5 -0.3              |
|           |           |        | 5                | -5.4 -2.6              | -1.1 -0.2              |
|           |           |        | 10               | -14.8 -4.1             | -2.1 -0.4              |
| 80        | 20        | 75     | 1                | 74.8 74.8              | 75.0 75.0              |
|           |           |        | 5                | 74.4 74.9              | 74.9 75.0              |
|           |           |        | 10               | 73.3 73.9              | 74.9 75.0              |
| 50        | 60        | 1      | 1                | 59.2 59.5              | 59.6 59.8              |
|           |           |        | 5                | 58.2 59.1              | 59.6 59.8              |
|           |           |        | 10               | 52.3 57.0              | 59.0 59.8              |
| 80        | 0         | 0      | 1                | -4.7 -1.7              | -1.1 -0.2              |
|           |           |        | 5                | -35.2 -9.8             | -4.8 -1.1              |
|           |           |        | 10               | -48.6 -23.7            | -10.0 -2.2             |
TABLE 4
Observed Coefficients (%) of 95% Confidence Intervals for AR
Using 1000 Case-Control Simulations; Matched Design; n = 100 Cases

| P(E|D) (%) | P(\overline{E}|\overline{D}) (%) | AR (%) | Case: Control Ratio | Number of Strata |
|----------|------------------|--------|---------------------|-----------------|
|          |                  |        | LOGIT   | ML | LOG | LOGIT | ML | LOG | LOGIT | ML | LOG |
| 20       | 20               | 0      | 1:1      | a  | 94.4 | 94.2 | -   | 95.8 | 95.6 | -   | 96.9 | 96.5 |
|          |                  |        | 1:3      | -  | 94.9 | 94.8 | -   | 94.7 | 94.8 | -   | 94.5 | 94.4 |
| 50       | 20               | 37.5   | 1:1      | 96.5 | 95.2 | 95.8 | 97.0 | 94.6 | 94.4 | 97.8 | 95.8 | 94.6 |
|          |                  |        | 1:3      | 96.1 | 94.7 | 93.5 | 95.6 | 94.8 | 96.0 | 96.9 | 95.1 | 94.4 |
| 50       | 0                |        | 1:1      | -   | 95.6 | 95.9 | -   | 96.7 | 96.5 | -   | 98.4 | 96.3 |
|          |                  |        | 1:3      | -   | 94.7 | 94.7 | -   | 96.6 | 95.7 | -   | 97.1 | 95.2 |
| 80       | 20               | 75     | 1:1      | 95.8 | 94.4 | 95.0 | 96.2 | 94.6 | 95.1 | 96.0 | 96.6 | 94.7 |
|          |                  |        | 1:3      | 95.1 | 94.7 | 94.6 | 95.3 | 93.8 | 94.8 | 97.4 | 97.4 | 96.4 |
| 50       | 60               |        | 1:1      | 96.4 | 93.8 | 94.3 | 97.7 | 96.2 | 96.4 | 99.3 | 98.6 | 96.0 |
|          |                  |        | 1:3      | 94.7 | 93.1 | 96.1 | 96.8 | 95.5 | 95.0 | 98.3 | 96.6 | 95.3 |
| 80       | 0                |        | 1:1      | -    | 94.1 | 95.6 | -   | 98.7 | 98.7 | -   | 99.7 | 99.1 |
|          |                  |        | 1:3      | -    | 95.3 | 94.3 | -   | 95.7 | 94.7 | -   | 99.3 | 96.1 |

a) Not calculated (see text)
### TABLE 5
Average Length (%) of 95% Confidence Intervals for AR Using 1000 Case-Control Simulations
Matched Design; n = 100 Cases

| P(E|D) (%) | P(E|δ) (%) | AR (%) | Case: Control Ratio | LOGIT | Number of Strata |
|----------|-----------|--------|---------------------|-------|-----------------|
|          |           |        |                     | ML    | 1               | 5            | 10            |
|          |           |        |                     | LOG   | 5               | 10           |
|          |           |        |                     | LOGIT | 5               | 10           |
| 20       | 20        | 0      | 1:1                 | -     | 28.0            | 30.0         | 33.6          | 33.8          |
|          |           |        | 1:3                 | -     | 22.6            | 23.0         | 23.6          | 23.7          |
| 50       | 20        | 37.5   | 1:1                 | 25.0  | 25.4            | 28.0         | 28.9          | 30.4          | 31.2          | 31.5          |
|          |           |        | 1:3                 | 12.2  | 12.2            | 25.2         | 25.7          | 25.9          | 25.6          | 26.1          | 26.3          |
| 50       | 0         |        | 1:1                 | -     | 56.8            | 66.8         | 100.2         | 107.1         |
|          |           |        | 1:3                 | -     | 45.3            | 47.6         | 50.8          | 51.3          |
| 80       | 20        | 75     | 1:1                 | 20.0  | 20.1            | 20.9         | 21.5          | 22.0          | 22.2          | 22.9          |
|          |           |        | 1:3                 | 19.6  | 19.7            | 19.9         | 20.3          | 20.2          | 20.3          | 20.9          |
| 50       | 60        |        | 1:1                 | 34.2  | 35.6            | 39.6         | 41.1          | 47.6          | 55.9          | 61.5          |
|          |           |        | 1:3                 | 31.6  | 32.7            | 33.7         | 34.7          | 34.4          | 35.9          | 37.1          |
| 80       | 0         |        | 1:1                 | -     | 118.4           | 253.6        | 304.7         | 375.5         |
|          |           |        | 1:3                 | -     | 92.1            | 110.2        | 156.3         | 173.2         |

a) Not calculated (see text)
<table>
<thead>
<tr>
<th>Number of Cases</th>
<th>Case: Control Ratio</th>
<th>Asymptotic Standard Error</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD</td>
<td>u</td>
<td>BSE</td>
</tr>
<tr>
<td>100</td>
<td>1:1</td>
<td>28</td>
<td>32</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>1:3</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>500</td>
<td>1:1</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>1:3</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

a) From (20) and either (A.1) or (A.5)  
b) Standard deviation of 1000 attributable risk estimates; see formula (21)  
c) Formula (10), averaged over 10 trials  
d) Bootstrap estimate of standard error using 200 replications, averaged over 10 trials
Appendix A. Asymptotic Variance of Attributable Risk Estimates

Let the probability \( P(E C_k | D) \) that a case falls in cell \( E C_k \) be denoted by \( \pi_{1k} \), and set \( \pi_{2k} = P(\bar{E} C_k | D) \), \( k = 1, \ldots, K \). Let \( \pi_1 \) and \( \pi_2 \) represent the column vectors whose \( k \)th components are \( \pi_{1k} \) and \( \pi_{2k} \) respectively, with \( \pi' = (\pi_1, \pi_2) \). Let \( p_1, p_2 \) and \( p \) represent the corresponding probability vectors for the controls. Similarly, let \( x_1 = (x_{11}, \ldots, x_{1k}) \) and \( x_2 = (x_{21}, \ldots, x_{2k}) \) be the sample vectors of cell frequencies for the cases, with \( x' = (x_1', x_2') \). The control frequency vectors \( x_1, x_2 \) and \( y \) are defined analogously. It is convenient to obtain the asymptotic variance of the residual risk estimate \( \hat{\beta} = 1 - \hat{AR} \), using the delta method.

For the unmatched design, the asymptotic variance of \( \hat{\beta} = \frac{1}{n} \sum y_k x_{2k} / y_{2k} \) is given by the variance of its first order Taylor series expansion

\[
\hat{\beta} = \rho + \frac{1}{n} x_2 - \frac{\pi_2}{\pi_1} \rho_{x_2} + \left( \frac{1}{m} y - \frac{p}{p} \right) \rho_y
\]

Here \( \rho = \sum \frac{P_k \pi_{2k}}{p_{2k}} \), and the vectors \( \rho_{x_2} \) and \( \rho_y \) of first partials are \( \rho_{x_2} = \left< \frac{P_k}{p_{2k}} \right> \) and \( \rho_y' = \left( \frac{\pi_{21}}{p_{21}}, \ldots, \frac{\pi_{2k}}{p_{2k}}, -\frac{\pi_{21} P_{11}}{p_{21}}, \ldots, -\frac{\pi_{2k} P_{1k}}{p_{2k}} \right) \). The resulting asymptotic variance is

\[
\text{A var } \hat{\beta} = \frac{\rho_y'}{\rho_{x_2}} \left[ D(\pi_2') - \frac{\rho_y'}{\rho_{x_2}} D(p) \right] \frac{\rho_y'}{\rho_{x_2}}
\]

where \( D(z) \) is the diagonal matrix whose entries are the components of the vector \( z \). This gives

\[
\text{A var } \hat{\beta} = \frac{1}{n} \left\{ \sum_k \left( \frac{\pi_{2k} P_k}{p_{2k}} \right)^2 \left( \frac{1}{\pi_{2k}} + \frac{np_{1k}}{mp_k p_{2k}} \right) - \rho^2 \right\}
\]

\[
= \frac{1}{n} \left\{ \sum_k \left( \frac{P(\bar{E} C_k | D)}{P(\bar{E} C_k | D)} \right)^2 \left[ \frac{1}{P(\bar{E} C_k | D)} + \frac{n p(\bar{E} C_k | \bar{D})}{m p(\bar{E} C_k | \bar{D})} \right] - (1 - \hat{AR})^2 \right\} \quad \text{.(A.1)}
\]

The expression \( u^2 \) of (10) is obtained by substituting sample moments for the unknown probabilities in (A.1).
For the matched design the conditional distribution of \( y_{2k} \) given \( R_k x'k \) is binomial with parameters \( R_k x'k \) and \( p_{2k}/p_k \). The variance of \( \beta \) is approximated by the variance of the first order expansion

\[
\beta = \rho + \left( \frac{1}{n} x' - \eta \right)' \rho_x + \left( \left< y_{2k}/n - R_k \pi_{k*} p_{2k}/p_k \right> \right)' \rho_y_2 , \tag{A.2}
\]

where the vectors of first partials are \( \rho' = (\pi_{21} p_{1}/p_{21} \pi_{1} , \ldots , \pi_{2k} p_{k}/p_{2k} \pi_{k} , (1 + \pi_{21} / \pi_{1}) p_{1}/p_{21} , \ldots , (1 + \pi_{2k} / \pi_{k}) p_{k}/p_{2k}) \), and

\[
\rho_y_2 = \left< \left( \frac{\pi_{2k}}{\pi_{k}} \right) (p_{k}/p_{2k})^2 / R_k \right> . \tag{A.3}
\]

where \( S \) is the \( 3K \times 3K \) covariance matrix of \( (x' , y_{2}') \). The entries of \( S \) can be calculated from the joint moment generating function for \( (x' , y_{2}') \), given by

\[
E[ \exp(\Sigma_j \Sigma_j^t x'k + \Sigma_k s_{2k} y_{2k})] = \left[ \Sigma_k (\pi_{1k} e^1 + \pi_{2k} e^2) (p_{1k}/p_k + p_{2k}/p_k) s_{2k}/R_k \right]^n ,
\]

where \( t_{2k} = 0 \). This gives

\[
n^{-1}S = \begin{pmatrix}
D(\pi) & D(\gamma_1) \\
D(\gamma_1') & D(\gamma_2) \\
D(\gamma_2') & D(\gamma + \eta)
\end{pmatrix} - \begin{pmatrix}
\pi \\
\gamma
\end{pmatrix} \begin{pmatrix}
\pi' \\
\gamma'
\end{pmatrix} , \tag{A.4}
\]

where \( \gamma_j = \left< \pi_{jk} R_k p_{2k}/p_k \right> , j = 1, 2, \gamma = \gamma_1 + \gamma_2 \), and \( \eta = \left< R_k (R_k - 1) \cdot p_{2k}/p_k \right> \). Straightforward (and tedious) computation from (A.3) and (A.4) yields
A \text{ var } \hat{\beta} = \frac{1}{n} \left[ \sum_k \left( \frac{\pi_{2k} p_{*k}}{p_{2k}} \right)^2 \left( \frac{1}{\pi_{2k}} + \frac{p_{1k}}{R_k \pi_{*k} p_{2k}} \right) - \rho^2 \right] \nonumber \\
= \frac{1}{n} \left\{ \sum_k \left[ \frac{P(E|C_k D)}{P(E|C_k \bar{D})} \right]^2 \left[ \frac{1}{P(E|C_k D)} + \frac{P(E|C_k \bar{D})}{R_k P(C_k|D) P(E|C_k \bar{D})} \right] - (1-AR)^2 \right\}. \quad (A.5) 

Substitution of sample moments for the unknown parameters in (A.5) yields $u^2$ of (10).

Next we calculate the asymptotic variance of $\tilde{\sigma} = 1 - \tilde{AR} = \Sigma \sigma_k [\phi_k + (1-\phi_k) y_{2k} x_{*k} / x_{2k} y_{*k}]^{-1}$, the residual risk estimate obtained using known stratum-specific disease probabilities $\phi_k = P(D|C_k)$ and confounder distribution $\sigma_k = P(C_k|D)$. For the unmatched design we use the first order expansion

$$\tilde{\sigma}^2 \approx \sigma_k^2 (1 - \phi_k) P(E|C_k \bar{D}) / P(E|C_k D) \quad (A.6)$$

where $\rho' = (\rho'_x, \rho'_y)$ and $\sigma' = (\sigma'_1, \sigma'_2)$, and where the kth components, respectively, of $\rho'_x, \rho'_y, \sigma'_1, \sigma'_2$ are $-\tau_k / \pi_{*k}, \tau_k \pi_{1k} / (\pi_{2k} \pi_{*k})$, $\tau_k / p_{*k}$, and $-\tau_k p_{1k} / (p_{2k} p_{*k})$. Here

$$\tau_k = (1-AR)^2 \sigma_k (1 - \phi_k) P(E|C_k \bar{D}) / P(E|C_k D) \quad (A.7)$$

with $AR_k$ given by (11). The approximation (A.6) yields

$$\text{ A \text{ var } } \tilde{\beta} = \frac{1}{n} \left( D(p) - \frac{\pi p}{m^2} \right) \rho_x / n + \frac{\rho_y}{m} \left( D(p') - \frac{p p'}{m} \right) \rho_y / m \nonumber \\
= \sum_k \tau_k^2 \left( \frac{\pi_{1k}}{m \pi_{*k} p_{2k}} + \frac{p_{1k}}{m p_{*k} p_{2k}} \right) \nonumber \\
= \sum_k \tau_k^2 \left( \frac{P(E|D C_k)}{nP(E|C_k D)} + \frac{P(E|\bar{D} C_k)}{mP(E|C_k \bar{D})} \right). \quad (A.8)$$

24
Substitution of sample moments in (A.8) gives the consistent estimate \( \nu^2 \) of (13).

The estimate \( \tilde{\rho} = 1 - \tilde{\mathcal{A}} \tilde{\mathcal{R}} \) obtained from the matched design depends on the data only through the vectors \( x_2, y_2 \) and the known control:case ratios \( R_k, k=1, \ldots, K \). Thus the linear expansion of \( \tilde{\rho} \) appropriate for this design is given by (A.2) with \( z \) and \( \pi \) replaced by \( x_2 \) and \( \pi_2 \), and with \( \rho_{x_2} = \langle \tau_k / \pi_{2k} \rangle \) and \( \rho_{y_2} = \langle -\tau_k \rho_k / (R_k \pi_{2k} \pi_{2k}) \rangle \). This gives

\[
A \text{ var } \tilde{\rho} = n^{-2} \left( \rho_{x_2}^{i'}, \rho_{y_2}^{i'} \right) S_* \left( \rho_{x_2}^{i'}, \rho_{y_2}^{i'} \right) ',
\]

where \( S_* \) is the \( 2K \times 2K \) submatrix obtained by deleting the first \( K \) rows and columns of \( S \) in (A.4). Substituting sample moments in the resulting expression

\[
A \text{ var } \tilde{\rho} = n^{-1} \sum_k \tau_k^2 \left( \frac{\pi_{1k}}{\pi_{2k}} + \frac{P_{1k}}{R_k P_{2k}} \right) = n^{-1} \sum_k \tau_k^2 \left( \frac{P(E\mid D \ C_k)}{P(E\mid \bar{D} \ C_k)} + \frac{P(E\mid \bar{D} \ C_k)}{R_k P(E\mid \bar{D} \ C_k)} \right), \tag{A.9}
\]

again gives the consistent estimate \( \nu^2 \).

Appendix B.

We prove that length of the ML interval is less than the length of the log interval whenever these intervals have nonzero length. It is evident from the interval definitions (16) and (18) that the desired result is equivalent to the inequality

25
\[
\beta [e^{w/\beta} - e^{-w/\beta}] - 2w > 0 . \tag{A.10}
\]

When \( \beta = 0 \) expression (A.10) is infinite for all finite \( w \). When \( \beta > 0 \) the left-hand side of (A.10) can be rewritten

\[
f(x) = e^x - e^{-x} - 2x ,
\]

where \( x = w/\beta \geq 0 \). It remains to show that \( f(x) \geq 0 \), with equality holding only in the trivial case \( x = w = 0 \). But this follows immediately from the fact that \( f(0) = 0 \) and that

\[
f'(x) = e^x + e^{-x} - 2 = 2[\cosh x - 1] ,
\]

which is positive for all \( x > 0 \).

Appendix C.

For the random sampling design the bootstrap estimate BSE was obtained as follows:

1. \( n \) of the case observations were sampled with replacement. Thus one "case" may have been chosen 3 times, while another was not chosen at all.

2. \( m \) of the control observations were sampled with replacement.

3. The bootstrap sample of cases and controls obtained in steps 1 and 2 was used in (9) to calculate attributable risk \( AR^* \).

4. Steps 1-3 were independently repeated 200 times.

5. \( BSE = \left[ \frac{1}{199} \sum_{i=1}^{200} (AR^*_i - AR^*_*)^2 \right]^{1/2} \) was calculated, where \( AR^*_* \) is the average of the 200 attributable risk values \( AR^*_i \). For the matched design, step 2 was eliminated and step 1 was modified so that each time a case observation was sampled, the \( R \) control observations matched to the case were selected as well.