Interval estimation of the attributable risk in case-control studies with matched pairs

K-J Lui

Department of Mathematical and Computer Sciences, College of Sciences, San Diego State University, San Diego, CA 92182-7720, USA

Correspondence to:
Dr Lui (kjl@rohan.sdsu.edu)

Accepted for publication
25 May 2001

Abstract

***Objective*—The attributable risk (AR), which represents the proportion of cases who can be preventable when we completely eliminate a risk factor in a population, is the most commonly used epidemiological index to assess the impact of controlling a selected risk factor on community health. The goal of this paper is to develop and search for good interval estimators of the AR for case-control studies with matched pairs.**

***Methods*—This paper considers five asymptotic interval estimators of the AR, including the interval estimator using Wald's statistic suggested elsewhere, the two interval estimators using the logarithmic transformations: log(x) and log(1–x), the interval estimator using the logit transformation log(x/(1–x)), and the interval estimator derived from a simple quadratic equation developed in this paper. This paper compares the finite sample performance of these five interval estimators by calculation of their coverage probability and average length in a variety of situations.**

***Results*—This paper demonstrates that the interval estimator derived from the quadratic equation proposed here can not only consistently perform well with respect to the coverage probability, but also be more efficient than the interval estimator using Wald's statistic in almost all the situations considered here. This paper notes that although the interval estimator using the logarithmic transformation log(1–x) may also perform well with respect to the coverage probability, using this estimator is likely to be less efficient than the interval estimator using Wald's statistic. Finally, this paper notes that when both the underlying odds ratio (OR) and the prevalence of exposure (PE) in the case group are not large (OR ⩽2 and PE ⩽0.10), the application of the two interval estimators using the transformations log(x) and log(x/(1–x)) can be misleading. However, when both the underlying OR and PE in the case group are large (OR ⩾4 and PE ⩾0.50), the interval estimator using the logit transformation can actually outperform all the other estimators considered here in terms of efficiency.**

***Conclusions*—When there is no prior knowledge of the possible range for the underlying OR and PE, the interval estimator derived from the quadratic equation developed here for general use is recommended. When it is known that both the OR and PE in the case group are large (OR ⩾4 and PE ⩾0.50), it is recommended that the interval estimator using the logit transformation is used.**

(*J Epidemiol Community Health* 2001;**55**:885–890)

To assess the public health importance of controlling a selected risk factor, the attributable risk (AR), which represents the proportion of cases who could be preventable if we completely eliminated this risk factor in a population, is probably one of the most commonly used epidemiological indices.[1] When studying a rare disease in the presence of nuisance confounders, we may often use matched pair case-control study design to increase the efficiency. In fact, the estimation of the AR using the retrospective data has recently received intensive discussions.[2–19] There are, however, only a few papers that discuss estimation of the AR in matched case-control studies. Whittemore[18] included a brief discussion on estimation of the AR for frequency matching, but noted that her approach would not be appropriate for the matched pair study, in which each stratum consisted of only one case and one control. Using Wald's statistic, Kuritz and Landis[12] derived an asymptotic interval estimator of the AR. Kuritz and Landis[13] further extended their result to the case of more than one matched control per case, but found that the coverage probability of their interval estimator might be less than the desired confidence level by >2% even when the number of matched pairs was as large as 100.

The purpose of this paper is to search for other better alternative interval estimators of the AR to the one using Wald's statistic for the matched pair case-control study. This paper considers five interval estimators of the AR, including the estimator using Wald's statistic,[12] the two interval estimators using the logarithmic transformation[6 17]: log(x) and log(1–x), the interval estimator using the logit transformation[15]: log(x/(1–x)), and the interval estimator derived from a simple quadratic equation developed here. To compare the finite sample performance of these estimators, this paper calculates the exact coverage probability and the average length in a variety of situations. Finally, this paper includes an example taken from a study of oral conjugated oestrogens and endometrial cancer[20 21] to illustrate the use of these interval estimators.

Methods

Consider a case-control study, in which we take a random sample of n subjects from the case group and for each of these randomly selected cases, we match a control with respect to some nuisance confounders to form n matched pairs. We then classify each pair according to the status of exposure into one cell of the following fourfold table:

		Control		
		Exposed	Unexposed	
Case	Exposed	p_{11}	p_{12}	$p_{1\cdot}$
	Unexposed	p_{21}	p_{22}	$p_{2\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$	1

where $0 < p_{ij} < 1$ denotes the corresponding cell probability, $p_{i.} = p_{i1} + p_{i2}$, $p_{.j} = p_{1j} + p_{2j}$ for *i* and *j* = 1, 2. By definition, the AR is equal to[12 22]: P(E|D)(RR–1)/RR, where P(E|D) (=$p_{1.}$) denotes the prevalence of exposure (PE) in the case group, and the RR denotes the relative risk of possessing the underlying disease of interest between the exposed and the unexposed. When the underlying disease is rare, we can substitute the odds ratio (OR=p_{12}/p_{21}) for the RR and use $p_{1.}(p_{12}–p_{21})/p_{12}$ to approximate the AR. Thus, in the following discussion we assume that the underlying disease is so rare that the difference between the AR and $p_{1.}(p_{12}–p_{21})/p_{12}$ is indistinguishable.

Let n_{ij} denote the observed frequency of pairs falling into the cell with the probability p_{ij}, where *i* and *j* = 1, 2. The random vector $\underline{n}'$ = (n_{11}, n_{12}, n_{21}, n_{22}) then follows the multinomial distribution with parameters n and $\underline{p}'$ = (p_{11}, p_{12}, p_{21}, p_{22}). Note that the sample proportion $\hat{p}_{ij} = n_{ij}/n$ is the maximum likelihood estimator (MLE) of p_{ij}, and so are $\hat{p}_{i.} = n_{i.}/n$ and $\hat{p}_{.j} = n_{.j}/n$, where $n_{i.} = n_{i1} + n_{i2}$, and $n_{.j} = n_{1j} + n_{2j}$, for $p_{i.}$ and $p_{.j}$, respectively. Therefore, the MLE of the AR is simply $\hat{AR} = \hat{p}_{1.}(\hat{p}_{12}–\hat{p}_{21})/\hat{p}_{12}$. Define the random vector $\underline{\hat{p}}' = (\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{21}, \hat{p}_{22})$. By the Central Limit Theorem, we know that the vector $\sqrt{n}(\underline{\hat{p}}–\underline{p})'$ asymptotically follows the normal distribution with mean vector $\underline{0}$ and the covariance matrix $\underline{D}(\underline{p})–\underline{p}\ \underline{p}'$, where $\underline{0}'$ = (0, 0, . . ., 0) and $\underline{D}(\underline{p})$ is a 4×4 diagonal matrix with diagonal elements equal to: p_{11}, p_{12}, p_{21}, and p_{22}. By use of the delta method, we obtain the asymptotic variance of $\hat{AR}$ to be $\mathrm{Var}(\hat{AR}) = \{(p_{12}–p_{21})^2p_{11} + (p_{12}^2 + p_{21}p_{11})^2/p_{12} + p_{1.}^2p_{21}–[p_{1.}(p_{12}–p_{21})]^2\}/(np_{12}^2)$, which we can estimate by simply substituting the MLE $\hat{p}_{ij}$ for the unknown parameter p_{ij}. We denote this estimated variance by $\hat{\mathrm{Var}}(\hat{AR})$. These lead us to obtain the asymptotic 100(1–α)% confidence interval proposed elsewhere[12] for the AR to be:

$$[\mathrm{AR}_l, \mathrm{AR}_u], \tag{1}$$

$$\text{where } \mathrm{AR}_l = \hat{\mathrm{AR}} - Z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{\mathrm{AR}})},\ \mathrm{AR}_u = \min\{\hat{\mathrm{AR}} + Z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{\mathrm{AR}})}, 1\},\ \text{and } Z_\alpha$$

is the upper 100(α)th percentile of the standard normal distribution

Attempting to improve the normal approximation to the statistic $\hat{AR}$, we follow Katz *et al*[23] and consider the logarithmic transformation. Using the delta method, we obtain the estimated asymptotic variance $\hat{\mathrm{Var}}(\log(\hat{AR}))$ = $(\hat{AR})^{-2}\ \hat{\mathrm{Var}}(\hat{AR})$. Hence, an asymptotic 100(1–α)% confidence interval for the AR is:

$$[\mathrm{AR}_l^*, \mathrm{AR}_u^*], \tag{2}$$

$$\text{where } \mathrm{AR}_l^* = \exp\{\log(\hat{\mathrm{AR}}) - Z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\log(\hat{\mathrm{AR}}))}\} \text{ and } \mathrm{AR}_u^* = \min\{\exp\{\log(\hat{\mathrm{AR}}) + Z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\log(\hat{\mathrm{AR}}))}\},\ 1\}.$$

Following Leung and Kupper,[15] we consider the logit transformation $\log(\hat{AR}/(1–\hat{AR}))$. By the delta method again, we can easily show that the estimated asymptotic variance $\hat{\mathrm{Var}}(\log(\hat{AR}/ (1–\hat{AR}))$ = $(\hat{AR}(1–\hat{AR}))^{-2}\ \hat{\mathrm{Var}}(\hat{AR})$. Hence, an asymptotic 100(1–α)% confidence interval for the AR using the logit transformation is:

$$[\mathrm{LT}_l/(\mathrm{LT}_l + 1),\ \mathrm{LT}_u/(\mathrm{LT}_u + 1)], \tag{3}$$

$$\text{where } \mathrm{LT}_l = \exp\{\log(\hat{\mathrm{AR}}/(1-\hat{\mathrm{AR}})) - Z_{a/2}\sqrt{\hat{\mathrm{Var}}[\log(\hat{\mathrm{AR}}/(1-\hat{\mathrm{AR}}))]}\},\ \text{and } \mathrm{LT}_u = \exp\{\log(\hat{\mathrm{AR}}/(1-\hat{\mathrm{AR}})) + Z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}[\log(\hat{\mathrm{AR}}/(1-\hat{\mathrm{AR}}))]}\}.$$

Note that the logarithmic function log(x) is defined only for x >0. When the resulting estimate $\hat{AR}$ <0, neither interval estimator (2) nor interval estimator (3) is applicable. Consider ϕ = 1–AR = $(p_{12}p_{2.} + p_{1.}p_{21})/p_{12}$, which is always >0. Thus, following Fleiss,[6] we consider the logarithmic transformation $\log(1–\hat{AR}) = \log(\hat{\phi})$ rather than $\log(\hat{AR})$ as used for deriving interval estimator (2). Note that $\hat{\mathrm{Var}}(1–\hat{AR}) = \hat{\mathrm{Var}}(\hat{AR})$. By use of the delta method, we obtain the estimated asymptotic variance $\hat{\mathrm{Var}}(\log(\hat{\phi}))$ to be $\hat{\mathrm{Var}}(\hat{AR})/\hat{\phi}^2$. Therefore, we obtain an asymptotic 100(1–α)% confidence interval of the AR to be:

$$[1-\phi_u,\ 1-\phi_l], \tag{4}$$

$$\text{where } \phi_l = \exp\{\log(\hat{\phi}) - Z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\log(\hat{\phi}))}\},\ \phi_u = \exp\{\log(\hat{\phi}) + Z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\log(\hat{\phi}))}\},\ \text{and } \hat{\phi} = 1 - \hat{\mathrm{AR}}$$

Recall that the asymptotic variance

$$\mathrm{Var}(\hat{\mathrm{AR}}) = \{(p_{12} - p_{21})^2 p_{11} + (p_{12}^2 + p_{21}p_{11})^2/ p_{12} + p_{1.}^2 p_{21}\}/(np_{12}^2) - \mathrm{AR}^2/_{\mathrm{n}}$$

As n is large, the probability

$$P((\hat{\mathrm{AR}} - \mathrm{AR})^2/\mathrm{Var}(\hat{\mathrm{AR}}) \le Z_{\alpha/2}^2) \doteq 1 - \alpha$$

These lead us to consider the following quadratic equation of

$$\text{AR: } \mathcal{A}\text{AR}^2 - 2\mathcal{B}\text{AR} + \mathcal{C} \leq 0,$$

$$\text{where } \mathcal{A} = 1 + Z_{\alpha/2}^2/\text{n},$$

$$\mathcal{B} = \hat{\text{AR}}, \text{ and } \mathcal{C} = \hat{\text{AR}}^2 - Z_{\alpha/2}^2\{(\hat{p}_{12} - \hat{p}_{21})^2\hat{p}_{11} + (\hat{p}_{12}^2 + \hat{p}_{21}\hat{p}_{11})^2/\hat{p}_{12} + \hat{p}_{1.}^2\hat{p}_{21}\}/(\text{n}\hat{p}_{12}^2).$$

An asymptotic 100(1–α;)% confidence interval of the AR is then

$$[\text{AR}_l^{**}, \text{AR}_u^{**}] \qquad (5)$$

$$\text{where AR}_l^{**} = (\mathcal{B} - \sqrt{\mathcal{B}^2 - \mathcal{AC}})/\mathcal{A}$$

$$\text{and AR}_u^{**} = \min\{(\mathcal{B} + \sqrt{\mathcal{B}^2 - \mathcal{AC}})/\mathcal{A}, 1\}.$$

Note that the coefficient A is >0 and hence the above quadratic equation is convex. Furthermore, when using the commonly used adjustment procedure for sparse data (which is described in the appendix and in the next section), we can show that the inequality that B^2–AC >0 holds for all samples (appendix) and thereby, the two distinct roots of confidence limits (5) always exist.

Evaluation of interval estimators

To compare the performance of interval estimators (1-5) of the AR, we calculate the exact coverage probability and the average length of the resulting confidence intervals on the basis of the multinomial probability mass function

$$\text{f}(\text{n}|\text{p}) = \frac{n!}{\text{n}_{11}!\text{n}_{12}!\text{n}_{21}!\text{n}_{22}!}\text{p}_{11}^{\text{n}_{11}}\text{p}_{12}^{\text{n}_{12}}\text{p}_{21}^{\text{n}_{21}}\text{p}_{22}^{\text{n}_{22}},$$

$$\text{where n}_{ij} \geq 0 \text{ and } \sum_i\sum_j \text{n}_{ij} = \text{n}.$$

By definition, we calculate the coverage probability of a given interval estimator [AR_l, AR_u] as

$$\text{as} \sum_{\underline{n}} 1(\text{AR} \in [\text{AR}_l, \text{AR}_u])\,\text{f}(\underline{\text{n}}|\underline{\text{p}}),$$

$$\text{where } 1(\text{AR} \in [\text{AR}_l, \text{AR}_u])$$

where 1(ARε[AR_l, AR_u]]) is an indicator function and = 1 if the underlying AR falls into the interval [AR_l, AR_u], and = 0, otherwise, and where the summation is over all possible vectors $\underline{\text{n}}$ such that $\sum_i\sum_j \text{n}_{ij} = \text{n}.$ Similarly, we calculate the average length as $\sum_{\underline{\text{n}}}(\text{AR}_u - \text{AR}_l)\,\text{f}(\underline{\text{n}}|\underline{\text{p}}).$ Note that if n_{ij} were 0, the sample proportion $\hat{\text{p}}_{ij}$ would be on the boundary of 0. Thus, as noted in the appendix, whenever any n_{ij} is 0, we apply the commonly used adjustment procedure for sparse data by adding 0.50 to each cell and using (n_{ij} + 0.5)/(n + 2) to estimate p_{ij}. Recall that if the resulting estimate $\hat{\text{AR}}$ (or equivalently, the estimate $\hat{\text{OR}}$ ≤1) were ≤0, interval estimators (2 and 3) would be inapplicable. Thus, for interval estimators (2 and 3), we calculate the conditional coverage probability and average length of the resulting confidence intervals under the truncated multinomial distribution, excluding those random vectors $\underline{\text{n}}$ such that the corresponding interval estimate does not exist. For completeness, we also calculate the probability of failing to produce an interval estimate using (2 and 3).

KEY POINTS

- When the disease is rare, case-control studies with matched pairs is often used. However, the research on interval estimation of the AR under this design is limited.
- This paper considers and compares the performance of five asymptotic interval estimators of the AR, including the one proposed recently on the basis of Wald's statistic.
- This paper demonstrates that the interval estimator derived from a quadratic equation developed here is generally preferable to the one based on Wald's statistic.
- This paper provides a general guideline about the selection of better interval estimators with respect to the coverage probability and the average length of the confidence intervals.

Given the values of the underlying OR, $\text{p}_{1.}$, and p_{12}, we can uniquely determine all the other parameters through the following equations: p_{21} = p_{12}/OR; p_{11} = $\text{p}_{1.}$–p_{12}; p_{22} = 1–p_{11}–p_{12}–p_{21}; and AR = $\text{p}_{1.}$(p_{12}–p_{21})/p_{12}. We consider the situations, in which the OR = 1, 2, 4, 8, 32; the probabilities $\text{p}_{1.}$ and p_{12} equal: 0.01 and 0.005, 0.1 and 0.05, 0.5 and 0.25, 0.80 and 0.40, such that the combination of these parameters leads to a valid set of probability vector $\underline{\text{p}}$ for which p_{ij} >0 for all *i* and *j*; and n = 20, 50, 100, and 200. These cover the range of the AR from 0.0 to 0.775. We write programs in SAS[24] to enumerate the probability of the desired multinomial distribution in our calculation.

Results

Table 1 summarises the coverage probability and the average length of the 95% confidence interval in application of interval estimators (1-5). Firstly, note that when the underlying OR = 1 (that is, AR = 0), the coverage probability of the 95% confidence interval for both (2) and (3) is 0%. Note also that the coverage probability of the asymptotic 95% confidence interval using either (4) or (5) is almost always larger than or approximately equal to the desired confidence level in the situations considered in table 1, whereas the coverage probability of using (1) is occasionally less than this desired confidence level by >2% to 3% when n is not large (≤100). When comparing the average length of interval estimator (5) with that of (1), as shown in table 1, we find that the former is generally more efficient than the latter. When both the OR and the PE in the case group are moderate or large (OR ≥4, and $\text{p}_{1.}$ ≥0.50), we find that interval estimator (3) can even be slightly more efficient than (5), while maintaining the coverage probability ≥95%. Note that when the PE is small ($\text{p}_{1.}$ ≤0.10), the

Table 1 The coverage probability and the average length (in parentheses) of the 95% confidence interval in application of (1–5) for the situations, in which the OR=1, 2, 4, 8, 32; the probabilities p_1 and p_{12} equal: 0.01 and 0.005; 0.10 and 0.005; 0.50 and 0.25; 0.8 and 0.4; such that the combination of these parameters leads to a valid set of probabilities for which p_{ij} >0 for i and j=1, 2; and n=30, 50, 100, and 200

				n=																			
				20					*50*					*100*					*200*				
OR	$p_{1.}$	p_{12}	*AR*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*
1	0.01	0.005	0.00	1.00 (0.443)	0.000* (0.821)	0.000* (0.508)	1.00 (0.449)	1.00 (0.406)	1.00 (0.225)	0.000* (0.344)	0.000* (0.272)	1.00 (0.226)	1.00 (0.217)	0.999 (0.147)	0.000* (0.192)	0.000* (0.164)	1.00 (0.148)	0.999 (0.145)	0.993 (0.102)	0.000* (0.124)	0.000* (0.108)	0.993 (0.103)	0.993 (0.101)
	0.10	0.05	0.00	0.994 (0.958)	0.000* (0.817)	0.000* (0.570)	0.998 (1.12)	0.997 (0.881)	0.956 (0.580)	0.000* (0.593)	0.000* (0.448)	0.963 (0.646)	0.963 (0.559)	0.947 (0.332)	0.000* (0.549)	0.000* (0.440)	0.955 (0.347)	0.951 (0.326)	0.955 (0.194)	0.000* (0.520)	0.000* (0.423)	0.960 (0.195)	0.956 (0.192)
	0.50	0.25	0.00	0.938* (1.58)	0.000* (0.932)	0.000* (0.817)	0.976 (1.98)	0.950 (1.45)	0.952 (0.839)	0.000* (0.916)	0.000* (0.752)	0.961 (0.865)	0.958 (0.809)	0.953 (0.571)	0.000* (0.855)	0.000* (0.693)	0.956 (0.579)	0.957 (0.560)	0.952 (0.397)	0.000* (0.779)	0.000* (0.633)	0.953 (0.400)	0.954 (0.394)
2	0.01	0.005	0.005	1.00 (0.423)	0.866* (0.821)	0.866* (0.508)	1.00 (0.428)	1.00 (0.388)	1.00 (0.203)	0.860* (0.336)	0.861* (0.269)	1.00 (0.204)	1.00 (0.196)	1.00 (0.125)	0.892* (0.178)	0.892* (0.155)	1.00 (0.125)	1.00 (0.122)	0.998 (0.079)	0.858* (0.102)	0.858* (0.092)	0.999 (0.080)	0.999 (0.079)
	0.10	0.05	0.05	0.998 (0.752)	0.906* (0.790)	0.910* (0.540)	1.00 (0.828)	1.00 (0.691)	0.985 (0.407)	0.871* (0.471)	0.875* (0.359)	0.993 (0.430)	0.992 (0.393)	0.967 (0.230)	0.870* (0.366)	0.881* (0.295)	0.974 (0.235)	0.975 (0.226)	0.962 (0.140)	0.903* (0.269)	0.907* (0.229)	0.964 (0.140)	0.966 (0.139)
	0.50	0.25	0.25	0.943 (1.08)	0.840* (0.848)	0.908* (0.707)	0.982 (1.28)	0.983 (0.994)	0.945 (0.596)	0.893* (0.707)	0.924* (0.566)	0.953 (0.612)	0.961 (0.576)	0.947 (0.411)	0.921* (0.511)	0.940 (0.426)	0.951 (0.416)	0.957 (0.404)	0.949 (0.288)	0.936* (0.324)	0.949 (0.294)	0.951 (0.290)	0.954 (0.285)
4	0.01	0.005	0.008	1.00 (0.413)	0.866* (0.821)	0.866* (0.508)	1.00 (0.417)	1.00 (0.379)	1.00 (0.192)	0.881* (0.332)	0.881* (0.267)	1.00 (0.193)	1.00 (0.185)	1.00 (0.113)	0.892* (0.171)	0.948 (0.151)	1.00 (0.113)	1.00 (0.111)	0.999 (0.068)	0.944 (0.091)	0.944 (0.084)	0.999 (0.068)	0.999 (0.068)
	0.10	0.05	0.075	0.999 (0.648)	0.961 (0.776)	0.964 (0.524)	1.00 (0.691)	1.00 (0.595)	0.994 (0.324)	0.945 (0.394)	0.955 (0.311)	0.995 (0.334)	0.995 (0.313)	0.979 (0.181)	0.944 (0.254)	0.949 (0.214)	0.980 (0.183)	0.981 (0.178)	0.965 (0.113)	0.944 (0.152)	0.948 (0.139)	0.966 (0.113)	0.969 (0.112)
	0.50	0.25	0.375	0.979 (0.813)	0.946 (0.755)	0.986 (0.621)	0.986 (0.927)	0.992 (0.759)	0.947 (0.454)	0.923* (0.505)	0.951 (0.432)	0.955 (0.465)	0.971 (0.442)	0.945 (0.316)	0.936* (0.333)	0.953 (0.309)	0.949 (0.320)	0.956 (0.312)	0.947 (0.223)	0.947 (0.227)	0.956 (0.219)	0.949 (0.224)	0.953 (0.221)
	0.80	0.40	0.600	0.955 (0.832)	0.931* (0.678)	1.00 (0.675)	0.967 (1.06)	0.996 (0.825)	0.922* (0.524)	0.910* (0.549)	0.964 (0.481)	0.947 (0.565)	0.957 (0.514)	0.936* (0.370)	0.930* (0.378)	0.963 (0.354)	0.950 (0.384)	0.953 (0.366)	0.943 (0.262)	0.943 (0.264)	0.956 (0.256)	0.950 (0.266)	0.952 (0.260)
8	0.01	0.005	0.009	1.00 (0.408)	0.866* (0.821)	0.949 (0.507)	1.00 (0.412)	1.00 (0.374)	1.00 (0.187)	0.882* (0.330)	0.882* (0.266)	1.00 (0.188)	1.00 (0.180)	1.00 (0.107)	0.948 (0.167)	0.948 (0.149)	1.00 (0.108)	1.00 (0.105)	1.00 (0.063)	0.955 (0.086)	0.955 (0.081)	1.00 (0.063)	1.00 (0.062)
	0.10	0.05	0.088	0.999 (0.596)	0.979 (0.768)	0.987 (0.516)	1.00 (0.624)	1.00 (0.547)	0.993 (0.285)	0.972 (0.351)	0.980 (0.286)	0.991 (0.290)	0.981 (0.275)	0.976 (0.159)	0.965 (0.199)	0.969 (0.177)	0.974 (0.160)	0.973 (0.156)	0.963 (0.100)	0.960 (0.113)	0.965 (0.108)	0.964 (0.100)	0.965 (0.099)
	0.50	0.25	0.438	0.984 (0.686)	0.979 (0.688)	0.998 (0.568)	0.976 (0.758)	0.977 (0.646)	0.963 (0.377)	0.957 (0.398)	0.977 (0.361)	0.967 (0.385)	0.970 (0.369)	0.949 (0.261)	0.948 (0.266)	0.960 (0.256)	0.953 (0.264)	0.959 (0.259)	0.947 (0.184)	0.948 (0.186)	0.953 (0.182)	0.949 (0.185)	0.952 (0.183)
	0.80	0.40	0.700	0.991 (0.654)	0.988 (0.581)	1.00 (0.591)	0.967 (0.841)	0.988 (0.671)	0.937* (0.399)	0.926* (0.407)	0.983 (0.381)	0.966 (0.431)	0.979 (0.399)	0.930* (0.281)	0.927* (0.284)	0.952 (0.275)	0.946 (0.292)	0.956 (0.281)	0.940 (0.200)	0.939* (0.200)	0.950 (0.198)	0.948 (0.203)	0.952 (0.200)
32	0.01	0.005	0.010	1.00 (0.405)	0.949 (0.821)	0.949 (0.507)	1.00 (0.408)	1.00 (0.371)	1.00 (0.183)	0.967 (0.329)	0.967 (0.266)	1.00 (0.183)	1.00 (0.176)	1.00 (0.103)	0.948 (0.165)	0.967 (0.147)	1.00 (0.103)	1.00 (0.101)	1.00 (0.059)	0.971 (0.082)	0.971 (0.078)	1.00 (0.059)	1.00 (0.058)
	0.10	0.05	0.097	1.00 (0.557)	0.979 (0.762)	0.992 (0.510)	1.00 (0.575)	1.00 (0.511)	0.979 (0.257)	0.981 (0.316)	0.985 (0.268)	0.980 (0.259)	0.980 (0.248)	0.973 (0.145)	0.979 (0.165)	0.983 (0.155)	0.974 (0.145)	0.972 (0.142)	0.954 (0.092)	0.972 (0.096)	0.974 (0.094)	0.954 (0.092)	0.954 (0.091)
	0.50	0.25	0.484	0.983 (0.598)	0.993 (0.627)	0.999 (0.526)	0.973 (0.642)	0.957 (0.568)	0.968 (0.323)	0.979 (0.331)	0.983 (0.312)	0.962 (0.328)	0.954 (0.319)	0.960 (0.218)	0.965 (0.220)	0.967 (0.215)	0.961 (0.220)	0.949 (0.217)	0.955 (0.152)	0.955 (0.152)	0.957 (0.150)	0.954 (0.152)	0.951 (0.151)
	0.80	0.40	0.775	0.994 (0.534)	1.00 (0.505)	0.992 (0.519)	0.963 (0.676)	0.928* (0.564)	0.982 (0.297)	0.985 (0.299)	0.979 (0.294)	0.959 (0.318)	0.951 (0.307)	0.965 (0.199)	0.965 (0.199)	0.970 (0.198)	0.967 (0.205)	0.962 (0.203)	0.949 (0.138)	0.950 (0.138)	0.956 (0.138)	0.955 (0.140)	0.956 (0.140)

*Denotes the coverage probability is less than the desired 95% confidence level by more than 1.0%.

Table 2 The probability of failing to produce an interval estimate in application of interval estimators (2) and (3) for the situations, in which the RR=1, 2, 4, 8, 32; the probabilities $p_{1.}$ and p_{11} equal: 0.01 and 0.005; 0.10 and 0.05; 0.50 and 0.25; 0.80 and 0.40; such that the combination of these parameters leads to a valid set of probabilities for which p_{ij} >0 for i and j=1, 2; and n=20, 50, 100, and 200

				n=			
OR	$p_{1.}$	p_{12}	*AR*	*20*	*50*	*100*	*200*
1	0.01	0.005	0.00	0.91	0.82	0.73	0.65
	0.10	0.05	0.00	0.65	0.59	0.56	0.55
	0.50	0.25	0.00	0.56	0.54	0.53	0.52
2	0.01	0.005	0.005	0.91	0.80	0.68	0.53
	0.10	0.05	0.05	0.52	0.35	0.23	0.12
	0.50	0.25	0.25	0.23	0.09	0.02	0.00
4	0.01	0.005	0.008	0.91	0.79	0.64	0.45
	0.10	0.05	0.075	0.45	0.21	0.09	0.02
	0.50	0.25	0.375	0.08	0.01	0.00	0.00
	0.80	0.40	0.600	0.03	0.00	0.00	0.00
8	0.01	0.005	0.009	0.91	0.78	0.62	0.41
	0.10	0.05	0.088	0.40	0.14	0.04	0.00
	0.50	0.25	0.438	0.03	0.00	0.00	0.00
	0.80	0.40	0.700	0.01	0.00	0.00	0.00
32	0.01	0.010	0.010	0.91	0.78	0.61	0.38
	0.10	0.05	0.097	0.37	0.09	0.01	0.00
	0.50	0.25	0.484	0.01	0.00	0.00	0.00
	0.80	0.40	0.775	0.00	0.00	0.00	0.00

probability of failing to produce an interval estimate using (2) and (3) can be substantial (table 2). When the OR is large (⩾4), this probability diminishes, however, to approximately 0 as $p_{1.}$ increases to 0.80.

AN EXAMPLE

To illustrate use of interval estimators (1-5), we consider the data that are consisted of 183 pairs taken from a case-control study of the use of oral conjugated oestrogens and the endometrial cancer.[12 20 21] We match each case with a control on race, age (within five years), date of admission (within 6 months), and hospital of admission. We then classify these 183 matched pairs according to their exposure status (ever versus never) with regard to use of the estrogens. We obtain $n_{11} = 12$, $n_{12} = 43$, $n_{21} = 7$, and $n_{22} = 121$. Suppose that we are interested in estimation of the AR of endometrial cancer attributable to the use of the oestrogens. As given elsewhere,[12] we obtain the estimate $\hat{AR}$ to be 0.252. Furthermore, when using interval estimators (1-5), we obtain the asymptotic 95% confidence intervals to be: [0.172, 0.331], [0.183, 0.345], [0.181, 0.339], [0.168, 0.327], and [0.167, 0.325], respectively. We see that the resulting 95% confidence intervals using (2) and (3) tends to slightly shift to the right as compared with the other three resulting interval estimates (1), (4), and (5), which are all similar to one another.

Discussion

To evaluate whether it is appropriate to apply interval estimators (1-5) in the particular configuration given by the example, we consider the situations in which the parameters are determined by the empirical estimates from the data: $\hat{AR} = 6.14$, $\hat{p}_{1.} = 0.30$, $\hat{p}_{12} = 0.24$, and n = 183. In application of interval estimators (1-5), we obtain the coverage probability and the average length (in parentheses) of the corresponding asymptotic 95% confidence intervals to be: 0.948 (0.158), 0.953 (0.161), 0.956 (0.158), 0.949 (0.159), and 0.950 (0.157). We can see that all interval estimators (1-5) perform reasonably well with respect to the coverage probability and interval (5) seems to be slightly more efficient than the others in terms of the average length. This is certainly consistent with the finding that interval estimator (5) is generally more efficient than the others unless both the RR and the PE are moderate or large (RR ⩾4, $p_{1.}$ ⩾0.50) as presented in table 1.

Note that the functions exp(x) and exp(x)/[1 + exp(x)] are always positive and so are both the lower limits of interval estimators (2) and (3). Thus, if the underlying AR were 0, the coverage probability of these interval estimator would obviously be 0. This explains the reason why the coverage probability of (2) and (3) is 0 when the underlying OR = 1 regardless of the sample size n (table 1). Furthermore, if the PE were small ($p_{1.}$ ⩽0.10), then both the probabilities p_{12} and p_{21} (=p_{12}/OR, where OR ⩾1) would even be close to 0. Thus, the probability that the difference between the estimates $\hat{p}_{12}$ and $\hat{p}_{21}$ is ⩽0 (and hence the resulting estimate $\hat{AR}$ ⩽1) can be substantial. This accounts for the finding that the probability of failing to produce an interval estimate using (2) and (3) can be quite large in this case (table 2).

We find that except for a few cases where the PE in the case group is large ($p_{1.}$ = 0.80), interval estimator (1) using Wald's statistic does perform reasonably well. While applying interval estimator (4) using the transformation log(1–x) can improve the coverage probability of applying (1), using the former is likely to lose efficiency as compared with the latter. By contrast, applying interval estimator (5) can generally not only improve the coverage probability of (1) but also increase the efficiency. Thus, we recommend interval estimator (5) for general use. When we know that both the underlying RR and $p_{1.}$ are not small (RR ⩾4 and $p_{1.}$ ⩾0.50) from our prior studies, however, we may wish to use interval estimator (3) as well, especially when n is not large.

Finally, note that although interval estimators considered here are derived on the basis of large sample theory, we note that interval estimators (1, 4, and 5) can generally, as shown here, perform well with respect to the coverage probability even when the number of matched pairs n is as small as 20 (table 1). Furthermore, interval estimators (3 and 4) could also perform well for n = 20 if the underlying OR and PE were not small (OR ⩾4 and $p_{1.}$ ⩾0.10). Because it would be quite rare for public health administrators to estimate the AR based on data with less than 20 cases, the situations considered here should cover most cases encountered in practice.

In summary, this paper considers five interval estimators of the AR for the matched pair case-control studies. This paper includes a discussion that provides an insight into the characteristics of the performance of these five interval estimators. This paper shows that the interval estimator derived from a quadratic equation suggested here can outperform the interval estimator using Wald's statistic proposed elsewhere. This paper further notes that interval estimators using the two logarithmic transformations log(x) and log(1–x) generally causes the loss of efficiency. Finally, this paper notes that the interval estimator using the logit transformation can be useful when both the underlying RR and the PE in the case group are large. The discussion and the findings presented here should have use for biostatisticians and epidemiologists when they want to estimate the AR using a matched pair case-control study.

Appendix

Firstly, note that if we obtained the estimate $\hat{p}_{ij}$ to be 0 for some cell (i, j) from a given sample, then $\hat{p}_{ij}$ would be on the boundary. To avoid this concern, if $\hat{p}_{ij}$ should be 0 for some cell (i, j), we would recommend using the commonly used adjustment procedure for sparse data by adding 0.50 to each cell and using $(n_{ij} + 0.50)/(n + 2)$ to estimate p_{ij}. Thus, we may assume that the resulting estimate $\hat{p}_{ij}$ always falls in $0 < \hat{p}_{ij} < 1$.

Note that $B^2–AC = Z^2_{\alpha/2}(G^\star– \hat{AR}^2)/n + Z^4_{\alpha/2} G^\star/n^2$, where $G^\star = \{(\hat{p}_{12}–\hat{p}_{21})^2\hat{p}_{11} + (\hat{p}^2_{12} + \hat{p}_{21}\hat{p}_{11})^2/\hat{p}_{12} + \hat{p}^2_{1}\hat{p}_{21}\}/p^2_{12}$. Note that the asymptotic variance $Var(\hat{AR})$, that equals $\{[(p_{12}–p_{21})^2p_{11} + (p^2_{12}+ p_{21}p_{11})^2/p_{12} + p^2{}_{1.}p_{21}]/p^2_{21}–AR^2\}/n$, is always $\geqslant 0$, for any vector $\underline{p}' = (p_{11}, p_{12}, p_{21}, p_{22})$, that satisfies $0 < p_{ij} < 1$ and $\sum_i \sum_j p_{ij} = 1$. Because we can easily show that we can obtain $G^\star– \hat{AR}^2$ by simply substituting the particular estimate $\hat{p}_{ij}$ (which obviously satisfies $0 < \hat{p}_{ij} < 1$ and $\sum_i \sum_j \hat{p}_{ij} = 1$) for p_{ij} in $nVar(\hat{AR})$, the inequality: $G^\star– \hat{AR}^2 \geqslant 0$ is always true. Furthermore, when $0 < \hat{p}_{ij} < 1$, we can easily see that $G^\star > 0$. These results suggest that the condition $B^2–AC = Z^2_{\alpha/2} (G^\star– \hat{AR}^2)/n + Z^4_{\alpha/2} G^\star/n^2$ should be >0 for all samples.

1 Levin ML. The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum* 1953;**9**:531–1.

2 Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Stat Med* 1991;**10**:1753–73.

3 Bruzzi P, Green SB, Byar DP, *et al*. Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol* 1985;**122**:904–14.

4 Denman DW III, Schlesselmann JJ. Interval estimation of the attributable risk for multiple exposure levels in case-control studies. *Biometrics* 1983;**39**:185–92.

5 Drescher K, Schill W. Attributable risk estimation from case-control data via logistic regression. *Biometrics* 1991; **47**:1247–56.

6 Fleiss JL. Inference about population attributable risk from cross-sectional studies. *Am J Epidemiol* 1979;**110**:103–4.

7 Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: Wiley, 1981.

8 Gefeller O. An annotated bibliography on the attributable risk. *Biometrical Journal* 1992;**34**:1007–12.

9 Gefeller O, Windeler J. Risk factors for cervical cancer: comments on attributable risk calculations and the evaluation of screening in case-control studies. *Int J Epidemiol* 1991;**20**:1140–1.

10 Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics* 1993;**49**:865–72.

11 Kooperberg C, Petitti DB. Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study. *Epidemiology* 1991;**2**:363–6.

12 Kuritz SJ, Landis JR. Attributable risk ratio estimation from matched-pairs case-control data. *Am J Epidemiol* 1987;**125**: 324–8.

13 Kuritz SJ, Landis JR. Attributable risk estimation from matched case-control data. *Biometrics* 1988a;**44**:355–67.

14 Kuritz SJ, Landis JR. Summary attributable risk estimation from unmatched case-control data. *Stat Med* 1988b;**7**:507–17.

15 Leung HM, Kupper LL. Comparisons of confidence intervals for attributable risk. *Biometrics* 1981;**37**:293–302.

16 Taylor JW. Simple estimation of population attributable risk from case-control studies. *Am J Epidemiol* 1977;**106**:260.

17 Walter SD. The estimation and interpretation of attributable risk in health research. *Biometrics* 1976;**32**:829–49.

18 Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. *Stat Med* 1982;**1**:229–43.

19 Whittemore AS. Estimating attributable risk from case-control studies. *Am J Epidemiol* 1983;**117**:76–85.

20 Antunes CMF, Stolley PD, Rosenshein NB, *et al*. Endometrial cancer and estrogen use: report of a large case-control study. *N Engl J Med* 1979;**300**:9–13.

21 Schlesselman JJ. *Case-control studies*. New York: Oxford University Press, 1982.

22 Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol* 1974;**99**:325–32.

23 Katz D, Baptista J, Azen SP, *et al*. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 1978;**34**:469–74.

24 SAS Institute, Inc. *SAS language, reference version 6*. 1st ed. Cary, NC: SAS Institute, Inc, 1990.