BIOMETRICS 49, 865–872
September 1993

Maximum Likelihood Estimation of the Attributable Fraction from Logistic Models

Sander Greenland

Department of Epidemiology, UCLA School of Public Health,
Los Angeles, California 90024-1772, U.S.A.

and

Karsten Drescher

Institut für Statistik, FB Mathematik/Informatik, Universität Bremen,
D2800 Bremen 33, Germany

SUMMARY

Bruzzi et al. (1985, *American Journal of Epidemiology* **122,** 904–914) provided a general logistic-model-based estimator of the attributable fraction for case–control data, and Benichou and Gail (1990, *Biometrics* **46,** 991–1003) gave an implicit-delta-method variance formula for this estimator. The Bruzzi et al. estimator is not, however, the maximum likelihood estimator (MLE) based on the model, as it uses the model only to construct the relative risk estimates, and not the covariate-distribution estimate. We here provide maximum likelihood estimators for the attributable fraction in cohort and case–control studies, and their asymptotic variances. The case–control estimator generalizes the estimator of Drescher and Schill (1991, *Biometrics* **47,** 1247–1256). We also present a limited simulation study which confirms earlier work that better small-sample performance is obtained when the confidence interval is centered on the log-transformed point estimator rather than the original point estimator.

1. Introduction

An important measure of the public health impact of an exposure on disease burden is the population attributable fraction (Deubner et al., 1980; Kelsey, Thompson, and Evans, 1986; Last, 1983)

$$\frac{\Pr(\text{disease}) - \Pr(\text{disease} \mid \text{no exposure})}{\Pr(\text{disease})},$$

also known as the population etiologic fraction (Kleinbaum, Kupper, and Morgenstern, 1982; Schlesselman, 1982), or population attributable risk (Breslow and Day, 1980). (The term "population attributable risk" has been abandoned by many epidemiologists because the measure does not represent disease risk, and because the term "attributable risk" has traditionally been used to denote the risk difference

$$\Pr(\text{disease} \mid \text{exposure}) - \Pr(\text{disease} \mid \text{no exposure});$$

see, for example, MacMahon and Pugh (1970), Mausner and Bahn (1974), and Deubner et al. (1980).) The attributable fraction ranges from $-\infty$ to 1, although negative values are commonly transformed to the preventable fraction, in which "no exposure" is replaced by "exposure" in the first formula (Last, 1983).

Benichou (1991) reviews point and interval estimation methods for the population attributable fraction (hereafter denoted AF). Of interest here are those based on a logistic model for disease risk. Deubner et al. (1980) were apparently the first to apply such an estimate, but focused on cohort study data. Bruzzi et al. (1985) gave a general estimate for case–control data, and Kooperberg and Petitti (1991) used this to obtain a bootstrap interval for case–control data with known sampling fractions. Benichou and Gail (1990) gave variance estimators for the Bruzzi et al. point estimator. Drescher and

Key words: Attributable fraction; Epidemiologic methods; Logistic regression.

Schill (1991) provided another AF estimator and variance based on logistic modelling. In work as yet unpublished, Drescher and Osius have derived a maximum likelihood estimator (MLE) for case–control data and showed that the Drescher–Schill estimator is a special case of their MLE.

The present note provides both cohort and case–control maximum likelihood estimators for AF based on the logistic model, and corresponding variance estimators. A simulation study paralleling that in Benichou and Gail (1990) suggests that in large samples there is no practical difference between the MLE and the Bruzzi et al. estimator, or between confidence intervals based on the two estimators, but that for both estimators the log-transformed intervals are best in smaller samples.

2. Parameters

Consider first discrete covariates only. Let $\mathbf{x}_1, \ldots, \mathbf{x}_I$ represent I distinct values of a row K-vector of covariates (exposures, confounders, factors used for matching in the study, and products among these covariates); assume the list is exhaustive of all covariate patterns occurring in the sampled (source) population. Also, let $\mathbf{z}_1, \ldots, \mathbf{z}_I$ be a corresponding list of I not necessarily distinct values of $\mathbf{x}$, such that $\mathbf{z}_i$ is the covariate value a subject with actual covariate value $\mathbf{x}_i$ would have if not exposed. For example, if x_{i1} = smoking (in pack-years) is the exposure, and smoking occurs once more in $\mathbf{x}_i$ as the product $x_{i7} = x_{i1}x_{i5}$ with the confounder x_{i5} = beta-carotene, then each $\mathbf{z}_i$ is obtained from the corresponding pattern $\mathbf{x}_i$ by substituting 0 (no smoking) for the first and seventh components of $\mathbf{x}_i$. Finally, let

$$p_i = \Pr(\mathbf{x}_i | d_i = 1) \quad \text{and} \quad s_i = \Pr(d_i = 1|\mathbf{z}_i)/\Pr(d_i = 1|\mathbf{x}_i),$$

where d is a disease indicator ($d = 1$ if disease occurs, 0 if not), p_i is the covariate distribution among cases, and s_i is the inverse of the risk ratio (relative risk) at covariate pattern i. The attributable fraction can then be expressed as

$$\text{AF} = 1 - \Pr(\text{disease}|\text{no exposure})/\Pr(\text{disease}) = 1 - \mathbf{p}'\mathbf{s},$$

where $\mathbf{p}$ and $\mathbf{s}$ are I-column vectors with elements p_i and s_i; this is equivalent to the formulation in Bruzzi et al. (1985). The generalized impact fraction of Morgenstern and Bursic (1982) is identical in form but, rather than substituting 0 for exposure to obtain $\mathbf{z}_i$ from $\mathbf{x}_i$, one may substitute another "reference" or "target" value of exposure, and this target value may vary with i.

For the general case, let $F(\mathbf{x})$ be the distribution of $\mathbf{x}$ among cases, let $\mathbf{z} = g(\mathbf{x})$ be a fixed function of $\mathbf{x}$ that maps each $\mathbf{x}$ into a reference value, and let $s(\mathbf{x}) = \Pr(d = 1|\mathbf{z})/\Pr(d = 1|\mathbf{x})$ be the inverse risk ratio. The attributable fraction under g is then $1 - \int s(\mathbf{x})\, dF$ (Bruzzi et al., 1985; Benichou and Gail, 1990).

3. Cohort Estimators

Let n_i be the total number of subjects observed at covariate level x_i, with m_x total cases observed, and suppose disease risk follows a logistic model

$$\Pr(d = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\theta}^*)}{1 + \exp(\mathbf{x}\boldsymbol{\theta}^*)} \equiv \text{expit}(\mathbf{x}\boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^* \equiv (\alpha^*, \beta)$ with α^* the intercept and $\boldsymbol{\beta}$ the exposure coefficients (so $x_1 \equiv 1$), and $\text{expit}(u) \equiv e^u/(1 + e^u)$ is the antilogit transform. Let $\mathbf{n}$, $\mathbf{r}_z$, and $\mathbf{r}_x$ be the column vectors with ith elements n_i, $\text{expit}(\mathbf{z}_i\boldsymbol{\theta}^*)$, and $\text{expit}(\mathbf{x}_i\boldsymbol{\theta}^*)$, respectively, and let $t_z = \mathbf{n}'\mathbf{r}_z$ and $t_x = \mathbf{n}'\mathbf{r}_x$ be the expected total cases under the reference and observed covariate levels. The attributable fraction for the cohort can then be written

$$\text{AF} = (t_x - t_z)/t_x = 1 - t_z/t_x.$$

Since the n_i, $\mathbf{x}_i$, and $\mathbf{z}_i$ are constants, under the ordinary logistic likelihood an MLE of AF can be found by substituting an MLE $\hat{\boldsymbol{\theta}}^*$ for $\boldsymbol{\theta}^*$ in $\mathbf{r}_z$ and $\mathbf{r}_x$ to obtain $\hat{\mathbf{r}}_z$, $\hat{\mathbf{r}}_x$, $\hat{t}_z$ and $\hat{t}_x$. Because $\hat{t}_x = m_x$ (Bishop, Fienberg, and Holland, 1975, Chap. 14), this MLE can be rewritten as

$$\widehat{\text{AF}} = 1 - m_z/m_x,$$

where $m_z \equiv \hat{t}_z = \mathbf{n}'\hat{\mathbf{r}}_z$. This estimator equals that proposed by Deubner et al. (1980).

Now let Z and X be the matrices with rows $\mathbf{z}_i$ and $\mathbf{x}_i$, let $\mathbf{w}_z$ and $\mathbf{w}_x$ be the vectors with elements $n_i r_{zi}(1 - r_{zi})$ and $n_i r_{xi}(1 - r_{xi})$, respectively, and let

$$C^* \equiv (X' \text{ diag}(\mathbf{w}_x)X)^{-1}$$

be the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}^*$. Noting that $\partial t_z/\partial\boldsymbol{\theta}^* = Z'\mathbf{w}_z$ and $\partial t_x/\partial\boldsymbol{\theta}^* = X'\mathbf{w}_x$, the

multivariate delta method (Bishop et al., 1975) yields

$$\text{var}^{\text{A}}(\widehat{\text{AF}}) = (t_z/t_x)^2[v_z/t_z^2 - 2c_{zx}/(t_z t_x) + v_x/t_x^2],$$

where

$$v_z \equiv \text{var}^{\text{A}}(m_z) = \mathbf{w}_z' ZC^* Z' \mathbf{w}_z,$$

$$c_{zx} \equiv \text{cov}^{\text{A}}(m_z, m_x) = \mathbf{w}_z' ZC^* X' \mathbf{w}_x,$$

$$v_x \equiv \text{var}^{\text{A}}(m_x) = \mathbf{w}_x' XC^* X' \mathbf{w}_x.$$

However, because $m_x - t_x$ is the efficient score component for α^* at $\boldsymbol{\theta}^*$, $\text{cov}^{\text{A}}(m_x, \hat{\alpha}^*) = 1$ and $\text{cov}^{\text{A}}(m_x, \hat{\boldsymbol{\beta}}) = \mathbf{0}$ (Cox and Hinkley, 1974, p. 256, eq. 16). Thus

$$c_{zx} \equiv \text{cov}^{\text{A}}(m_z, m_x) = \mathbf{w}_z' Z \, \text{cov}^{\text{A}}(\hat{\boldsymbol{\theta}}^*, m_x) = \mathbf{w}_z' Z(1, 0, \ldots, 0)' = \textstyle\sum_i w_{zi}$$

because the first column of Z comprises only ones. By a parallel argument, $v_x \equiv \text{var}^{\text{A}}(m_x) = \sum_i w_{xi}$. An estimator $\hat{v}_{\text{ML}}$ for $\text{var}^{\text{A}}(\widehat{\text{AF}})$ may then be obtained by substituting $\hat{\boldsymbol{\theta}}^*$ for $\boldsymbol{\theta}^*$ in the expressions.

For cohort studies with person-time denominators n_i, the logistic model is usually replaced by the Poisson rate model $\lambda(\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})$. In this case the above formulas change only in that r_{zi} and r_{xi} are replaced by $\lambda_{zi} = \exp(\mathbf{z}_i\boldsymbol{\beta})$ and $\lambda_{xi} = \exp(\mathbf{x}_i\boldsymbol{\beta})$, and w_{zi} and w_{xi} are replaced by $n_i\lambda_{zi}$ and $n_i\lambda_{xi}$. Unlike the binomial–logistic case, however, a rare-disease assumption must be invoked to avoid interpretational problems (Greenland and Robins, 1988).

4. Case–Control Estimators

As discussed by many previous authors (Anderson, 1972; Mantel, 1973; Farewell, 1979; Prentice and Pyke, 1979; Drescher and Schill, 1991), under simple random sampling of m_1 cases and m_0 controls, a logistic model for the sampled population implies that

$$\Pr(d = 1 | \mathbf{x}, \text{ selection into sample}) = \text{expit}(\mathbf{x}\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} \equiv (\alpha, \boldsymbol{\beta})$ and

$$\alpha = \alpha^* + \log(m_1/m_0) - \log[\Pr(d = 1)/\Pr(d = 0)].$$

Assuming no other constraints on the joint distribution of $(d, \mathbf{x})$ other than absolute continuity for each continuous component of $\mathbf{x}$, the case–control likelihood can then be written as

$$\prod_i \frac{\exp(\mathbf{x}_i\boldsymbol{\theta})^{a_i}}{[1 + \exp(\mathbf{x}_i\boldsymbol{\theta})]^{n_i}} \frac{\pi(\mathbf{x}_i)^{n_i}}{\mu^{m_1}(1 - \mu)^{m_0}}, \tag{1}$$

where a_i is the number of cases observed at covariate level $\mathbf{x}_i$, $\mu \equiv \Pr(d = 1 | \text{selection into sample}) = m_1/n_+$ with $n_+ = \sum_i n_i = m_1 + m_0$, and $\pi(\mathbf{x}) \equiv p(\mathbf{x} | \text{selection into sample})$ with p a discrete mass function or density according to whether $\mathbf{x}$ is discrete or not (Anderson, 1972).

Let $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\boldsymbol{\beta}})$ be the value of $\boldsymbol{\theta}$ that maximizes the logistic portion of the likelihood in expression (1) (the portion involving $\boldsymbol{\theta}$). Prentice and Pyke (1979) showed that $\hat{\boldsymbol{\theta}}$ is a $\sqrt{n_+}$-consistent asymptotically normal estimator of $\boldsymbol{\theta}$. See Drescher and Schill (1991) for a review of these results. Full maximization of expression (1) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\pi} \equiv (\pi(\mathbf{x}_1), \ldots, \pi(\mathbf{x}_I))$ yields the same $\hat{\boldsymbol{\theta}}$, with $\hat{\boldsymbol{\pi}} = (n_1, \ldots, n_I)/n_+$, μ being a known constant. Assuming disease rarity, one also obtains $s_i \approx \exp[(\mathbf{z}_i - \mathbf{x}_i)\boldsymbol{\theta}]$ and thus may employ $\hat{s}_i = \exp[(\mathbf{z}_i - \mathbf{x}_i)\hat{\boldsymbol{\theta}}]$.

The preceding results can be extended to stratified random sampling schemes with representative (equal probability of selection) sampling of cases, and to frequency-matched sampling schemes (representative sampling of cases, with stratified sampling of controls to match the case distribution). This involves replacing α^*, α, m_1, m_0, π, and μ by stratum-specific quantities, then taking the product of the stratum-specific likelihood contributions (Prentice and Pyke, 1979). In the remaining development, we will assume that the disease is rare and sampling is simple random, unless stated otherwise; for brevity, details for other schemes will be omitted. We also assume that m_1/m_0 remains constant as n_+ increases.

Bruzzi et al. (1985) set $\tilde{p}_i = a_i/\sum_i a_i$ and so obtained the model-based estimator

$$\widetilde{\text{AF}} = 1 - \tilde{\mathbf{p}}'\hat{\mathbf{s}}.$$

Since we assume a rare disease and simple random sampling of subjects, standard results (e.g., Ibragimov and Linnik, 1971) imply that the sum $\tilde{\mathbf{p}}'\mathbf{s}$ is a $\sqrt{n_+}$-consistent asymptotically normal estimator for the expected inverse risk ratio $\int s(\mathbf{x})\, dF$ (subject to mild regularity conditions on F).

Since $\hat{\boldsymbol{\theta}}$ is also $\sqrt{n_+}$-consistent and $\mathbf{s}$ is a smooth function of $\boldsymbol{\theta}$, $\tilde{\mathbf{p}}'\hat{\mathbf{s}}$ inherits the same convergence properties, and so $\widetilde{\mathrm{AF}}$ is $\sqrt{n_+}$-consistent asymptotically normal for AF. Using a delta method for implicit functions that they had previously developed, Benichou and Gail (1990) derived a variance estimator for $\widetilde{\mathrm{AF}}$, which we will denote $\tilde{v}_{\mathrm{BG}}$. Since $\tilde{p}_i$ does not incorporate the model constraints, it is not the MLE of p_i under the model; thus, $\widetilde{\mathrm{AF}}$ is not an MLE of AF.

To obtain an MLE of AF for discrete $\mathbf{x}$, note that random sampling and Bayes' theorem yield

$$p_i = \Pr(\mathbf{x}_i | d = 1) = \Pr(\mathbf{x}_i | d = 1, \text{ selection})$$

$$= \frac{\Pr(d = 1 | \mathbf{x}_i, \text{ selection}) \Pr(\mathbf{x}_i | \text{selection})}{\Pr(d = 1 | \text{selection})} = r_i \pi_i / \mu,$$

where $r_i \equiv \mathrm{expit}(\mathbf{x}_i\boldsymbol{\theta})$ and $\pi_i \equiv \pi(\mathbf{x}_i)$. It follows (Zehna, 1966) that the MLE of p_i based on the likelihood (1) is

$$\hat{p}_i = n_i \hat{r}_i / m_1, \tag{2}$$

where $\hat{r}_i \equiv \mathrm{expit}(\mathbf{x}_i\hat{\boldsymbol{\theta}})$. More generally, the MLE of $F(\mathbf{x})$ from expression (1) is the step function $\Sigma\{\hat{p}_i\colon \mathbf{x}_i \leqslant \mathbf{x}\}$, where $\hat{p}_i$ is as in equation (2), and the sum is over all i with $\mathbf{x}_i$ less than or equal to $\mathbf{x}$ in all components. Thus, an MLE of the attributable fraction is

$$\widehat{\mathrm{AF}} = 1 - \hat{\mathbf{p}}'\hat{\mathbf{s}}. \tag{3}$$

For discrete $\mathbf{x}$, $\sqrt{n_+}$-consistent asymptotic normality of $\widehat{\mathrm{AF}}$ follows immediately from standard theory (Bishop et al., 1975); for the general case, this follows from the fact that $\hat{\mathbf{p}}'\hat{\mathbf{s}}$ is a $\sqrt{n_+}$-consistent asymptotically normal estimator for the expectation $\int s(\mathbf{x})\, dF$ (again subject to regularity conditions on F).

Expression (3) is equivalent to the MLE derived by Drescher and Osius (unpublished manuscript). They also derive an asymptotic variance formula similar to that of Benichou and Gail. A convenient computing formula for the asymptotic variance of $\widehat{\mathrm{AF}}$ can also be obtained by the ordinary multivariate delta method. Let $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{r}$ be the column vectors of a_i, $b_i = n_i - a_i$, and r_i; let $\mathbf{w}$ and $\mathbf{q}$ be the column vectors with elements $n_+\pi_i r_i(1 - r_i)$ and $n_+\pi_i(1 - r_i)/m_0$, respectively; and define

$$D_\theta \equiv \partial(\mathbf{p}'\mathbf{s})/\partial\boldsymbol{\theta} = (Z - X)'\mathrm{diag}(\mathbf{s})\mathbf{p} + X'\mathrm{diag}(\mathbf{w})\mathbf{s}/m_1,$$

$$D_\pi \equiv n_+\partial(\mathbf{p}'\mathbf{s})/\partial\boldsymbol{\pi} = \mathrm{diag}(\mathbf{s})\mathbf{r}/m_1,$$

$$V_a \equiv \mathrm{cov}(\mathbf{a}) = m_1(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}'),$$

$$V_b \equiv \mathrm{cov}(\mathbf{b}) = m_0(\mathrm{diag}(\mathbf{q}) - \mathbf{q}\mathbf{q}'),$$

$$V_n \equiv \mathrm{cov}(\mathbf{n}) = V_a + V_b,$$

$$C^* \equiv (X'\mathrm{diag}(\mathbf{w})X)^{-1} = [c^*_{jk}],$$

$$C \equiv \mathrm{cov}^{\mathrm{A}}(\hat{\boldsymbol{\theta}}) = [c_{jk}],$$

where $c_{11} = c^*_{11} - 1/m_1 - 1/m_0$, $c_{jk} = c^*_{jk}$ otherwise. Then, using $\hat{\boldsymbol{\theta}} \approx C^*X'(\mathbf{a} - \mathrm{diag}(\mathbf{r})\mathbf{n})$, we obtain

$$U \equiv \mathrm{cov}^{\mathrm{A}}(\hat{\boldsymbol{\theta}}, n) \approx C^*X(V_a - \mathrm{diag}(\mathbf{r})V_n) \tag{4}$$

and

$$\mathrm{var}^{\mathrm{A}}(\widehat{\mathrm{AF}}) = \mathrm{var}^{\mathrm{A}}(\mathbf{p}'\mathbf{s}) \approx D'_\theta C D_\theta + 2D'_\theta U D_n + D'_n V_n D_n. \tag{5}$$

An estimator $\hat{v}_{\mathrm{ML}}$ for this expression may be obtained by substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\pi}} = \mathbf{n}/n_+$ for $\boldsymbol{\pi}$.

For frequency-matched sampling schemes with m_{1k} cases and m_{0k} controls in stratum k, a separate intercept term α_k must be entered for each matching stratum (Prentice and Pyke, 1979). If $\boldsymbol{\alpha}$ is the vector of the α_k, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, and q_k is the vector with elements $n_+\pi_i(1 - r_i)/m_{0k}$ when i is in stratum k, V_a becomes block diagonal with blocks $m_{1k}(\mathrm{diag}(\mathbf{p}_k) - \mathbf{p}_k\mathbf{p}'_k)$, V_b becomes block diagonal with blocks $m_{0k}(\mathrm{diag}(\mathbf{q}_k) - \mathbf{q}_k\mathbf{q}'_k)$, and the variance c_{kk} of $\hat{\alpha}_k$ in C becomes $c^*_{kk} - 1/m_{1k} - 1/m_{0k}$ (Prentice and Pyke, 1979; see also Drescher and Schill, 1991).

5. Simulations

The first set of simulations presented here parallels that given by Benichou and Gail (1990). The sampled population follows a distribution based on a study of breast cancer (Brinton, Hoover, and

Fraumeni, 1982). Disease risk is given by

$$\Pr(d = 1 | x_1 = k, s_2 = h) = [1 + \exp(-\alpha_h - k\beta)]^{-1},$$

where $\alpha_1 = -6.24$, $\alpha_2 = -6.10$, $\alpha_3 = -6.01$ are effects for age ranges 50–54, 55–59, 60–64; $\beta = .155$ is a trend effect for category 0, 1, 2, 3 of age at first live birth ($\leq$19, 20–24, 25–29, $\geq$30 yr); for k = 0, 1, 2, 3, $\Pr(x_1 = k) = .117, .370, .418, .095$; for $h = 1, 2, 3$, $\Pr(x_2 = h) = .440, .335, .225$; and x_1 and x_2 are assumed to be independent. The true AF under $\mathbf{z} = (0, x_2)$ is then .2127. Each column of Tables 1 and 2 summarizes 7,600 simulation trials; thus, there is a 95% probability that a simulated coverage rate falls between .945 and .955 if the true coverage rate is .95. No convergence failures or infinite MLEs occurred in any trial.

Table 1 presents a series of cohort study simulations from this population. Cases were generated from fixed denominators using binomial variates generated from the uniform random number generator in GAUSS (Aptech, 1992). Two ML intervals were examined: the untransformed interval $\widehat{\text{AF}} \pm 1.96\hat{v}_{\text{ML}}^{1/2}$ and the log(1 − $\widehat{\text{AF}}$)-based ("log AF") interval

$$1 - (1 - \widehat{\text{AF}})\exp[\pm 1.96\hat{v}_{\text{ML}}^{1/2}/(1 - \widehat{\text{AF}})],$$

which is suggested by the observation that AF is bounded above by 1 but has no lower bound (Walter, 1975). The point estimator tends to be downward biased with fewer expected cases, leading to undercoverage of the untransformed confidence intervals in smaller samples, but the standard error estimator and log-transformed interval appear reasonably accurate.

Table 2 presents a series of case–control simulations from the same population. Both cases and controls were simple random samples from the $d = 1$ and $d = 0$ subpopulations, using multinomial variates generated from the uniform random number generator in GAUSS (Benichou and Gail

Table 1

Simulation results for MLE ($\widehat{\text{AF}}$) *of the attributable fraction: Cohort study; exposure with four categories, modelled by single ordinal variable*

Expected no. cases	50	200	300	250	1,000	4,000
Cohort size	1,000	4,000	16,000	1,000	4,000	16,000
True AF	.204	.204	.204	.166	.166	.166
Mean point est.	.185	.196	.204	.164	.164	.166
Sample std. dev.	.219	.108	.0550	.0913	.0459	.0228
Mean SE est.[a]	.218	.108	.0540	.0914	.0456	.0228
Coverage[b]						
Untransformed	.937	.949	.943	.948	.947	.948
Log transform	.947	.949	.946	.952	.946	.949
Mean length						
Untransformed	.856	.424	.212	.358	.179	.089
Log transform	.897	.429	.212	.361	.179	.089

[a] Simulation means of $\hat{v}_{\text{ML}}^{1/2}$.

[b] 7,600 trials each; simulation standard errors of coverages range from .002 to .003.

Table 2

Simulation results for MLE ($\widehat{\text{AF}}$) *of the attributable fraction: Case–control study; exposure with four categories, modelled by single ordinal variable; true* AF = .213

No. cases	100	100	300	300	600	3,000
No. controls	100	500	300	1,200	600	3,000
Mean point est.	.183	.202	.206	.205	.208	.212
Sample SD	.227	.170	.125	.100	.0880	.0386
Mean SE est.[a]	.223	.169	.124	.099	.0870	.0386
Coverage[b]						
Untransformed	.943	.940	.944	.945	.945	.950
Log transform	.952	.952	.950	.950	.949	.951
Mean length						
Untransformed	.876	.663	.485	.389	.341	.151
Log transform	.918	.683	.492	.393	.343	.152

[a] Simulation means of $\hat{v}_{\text{ML}}^{1/2}$.

[b] 7,600 trials each; simulation standard errors of coverages range from .002 to .003.

simulated frequency-matched control samples). The two "ML" intervals were examined, as well as the two "BG" intervals studied by Benichou and Gail: the untransformed interval $\widetilde{AF} \pm 1.96\tilde{v}_{BG}^{1/2}$ and the log(1 − AF)-based interval

$$1 - (1 - \widetilde{AF})\exp[\pm 1.96\tilde{v}_{BG}^{1/2}/(1 - \widetilde{AF})].$$

The ML and BG point and variance estimators differed only trivially from trial to trial; their mean differences were less than .001 and their correlations exceeded .999 in all but two of the present simulations. This resulted in nearly identical coverage behavior. Consequently, only the ML results are presented here. Also included in the simulation were confidence intervals based on using only the leading term $D_\theta' C D_\theta$ in formula (5) to estimate the variance of $\widehat{AF}$. This simplification also yielded results nearly identical to those based on the entire formula.

It is apparent both in the tables and in earlier simulations (Benichou and Gail, 1990; Whittemore, 1982) that, even for fairly large samples, the case–control AF estimators are downward biased (Whittemore's estimator is equivalent to that based on a saturated logistic model). This results in statistically (but not substantively) significant undercoverage of the untransformed intervals at even fairly substantial sample sizes. This problem is, however, rectified by the log transformation, although this transform invariably lengthens the interval (Whittemore, 1982).

To further investigate the performance of the point and interval estimators, the sequence of simulations was repeated using the same population and sampling structure, but with an analysis model that treated age at first birth as a categorical factor, so that the exposure was represented by three parameters instead of one. With this richer exposure parameterization one should expect slower convergence to limiting behavior, as well as larger variance of the AF estimators. This was observed, but otherwise results were the same. In particular, the log-transformed intervals again had better small-sample coverage than their untransformed counterparts.

Finally, case–control simulations were conducted with x_1 and x_2 bivariate Gaussian among noncases, with means μ and 0, unit variances, and correlation ρ among noncases. Under a logistic model with $\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2)$, this implies that the case distribution $F(x_1, x_2)$ will be bivariate Gaussian with the same covariance matrix, but with means shifted to $\mu + \beta_1 + \rho\beta_2$ and $\rho\beta_1 + \beta_2$; the AF parameter for $\mathbf{z} = (0, \mathbf{x}_2)$ then has closed form $1 - \exp[-\beta_1(\mu + \beta_1/2 + \rho\beta_2)]$. Because of the much longer run time of the continuous simulations (due to the need to fit ungrouped data at each trial), only 1,900 trials were used for each combination. This yields a 95% probability that a simulated coverage rate falls between .94 and .96 if the true coverage rate is .95. Table 3 shows the results of these simulations for $\rho = 0, .5$, $\mu = 2$, $\beta_1 = .28$, $\beta = .82$; these imply relative risks of 3 and 25 when comparing the 97.5 and 2.5 percentiles of the exposure and covariate distributions, and AF of .4507 and .5103 for $\rho = 0, .5$. The results for the Benichou–Gail approach are again omitted because in nearly all cases the mean difference of $\widetilde{AF}$ and $\widehat{AF}$ was below .001 and the correlation of $\widetilde{AF}$ and $\widehat{AF}$ exceeded .999; also, confidence intervals based on $\tilde{v}_{BG}$ or $D_\theta' C D_\theta$ performed in a manner nearly identical to the intervals based on $\widetilde{AF}\hat{v}_{ML}$. As can be seen from Table 3, correlation of x_1 and x_2

Table 3

Simulation results for MLE ($\widehat{AF}$) *in case–control study, bivariate Gaussian exposure and covariate with unit variances, exposure mean* = 2, *covariate mean* = 0 *among noncases,* $\beta_1 = .28$, $\beta_2 = .82$. *For reference exposure level of* 0, *true* AF = .451 *when exposure–covariate correlation is* 0, .510 *when correlation is* .5.

No. cases, controls	100, 100		250, 250		500, 500	
Correlation	0	.5	0	.5	0	.5
Mean point est.	.436	.461	.438	.494	.446	.502
Sample SD	.192	.267	.122	.149	.080	.104
Mean SE est.[a]	.189	.256	.118	.148	.082	.103
Coverage[b]						
Untransformed	.928	.926	.938	.937	.949	.951
Log-transform	.957	.957	.943	.948	.961	.951
Mean length						
Untransformed	.739	1.001	.463	.582	.322	.404
Log-transform	.794	1.155	.476	.615	.323	.415

[a] Simulation means of $\hat{v}_{ML}^{1/2}$.

[b] 1,900 trials per experiment; simulation standard errors of coverages range from .004 to .006.

worsened the small-sample behavior of the confidence intervals, although the log-transformed interval exhibited satisfactory coverage in all situations examined.

Other simulations were conducted, varying the distribution of the covariates and the magnitude of their effects, and examining continuous covariates for cohort studies. These simulations revealed patterns no different from those in Tables 1–3, and so are omitted here. In sum, considering the systematic biases present in typical epidemiologic studies, there appears to be little practical difference among the estimators, although the log-transformed intervals seem preferable for smaller samples.

6. Discussion

Both Benichou and Gail (1990) and Drescher and Schill (1991) provide example results of applying their methods to a well-known study of oesophageal cancer in Brittany. This example is not repeated here, because the ML method yields results numerically identical to the Drescher and Schill method for such data (this fact was verified numerically in a suite of checks on the simulation program). The ML method may in fact be viewed as a natural generalization of the Drescher and Schill method to allow control of multiple and continuous confounders.

The presentation here has considered only "large-stratum" estimation, in which the number of matching strata (and hence the dimension of θ^*) is fixed. For a single binary risk factor, Greenland (1987) provides attributable-fraction estimators for "sparse data," in which the number of strata increases with sample size, as occurs when one matches on neighborhood, sibship, etc. Benichou and Gail provide an extension of their method to sparse data by using conditional logistic regression to eliminate all but a finite number of parameters from the estimation problem. In contrast, the unconditional methods discussed here and in Drescher and Schill (1991) and Drescher and Osius (unpublished manuscript) inherently depend on estimated intercepts, and so do not readily generalize to sparse data without further modelling of the intercepts to keep the number of parameters fixed.

Acknowledgment

This work was supported by National Institutes of Health Contract N01-AI-72631.

Résumé

Bruzzi et al. (1985, *American Journal of Epidemiology* **122,** 904–914) ont présenté, à partir d'un modèle logistique général, un estimateur de la fraction étiologique pour des études cas–témoins, et Benichou et Gail (1990, *Biometrics* **46,** 991–1003) ont donné une formule de la variance de cet estimateur implicitement déduite de la delta méthode. Cependant, l'estimateur de Bruzzi et al. n'est pas l'estimateur du maximum de vraisemblance fondé sur le modèle, puisque le modèle est utilisé uniquement pour construire les estimateurs du risque relatif, et non l'estimation de la distribution de la covariable. Nous fournissons les estimateurs du maximum de vraisemblance de la fraction attribuable pour des études cohortes et cas–témoins ainsi que leurs variances asymptotiques. L'estimateur cas–témoin généralise l'estimateur de Drescher et Schill (1991, *Biometrics* **47,** 1247–1256). Nous présentons aussi une étude de simulation limitée qui confirme un travail récent indiquant que le meilleur résultat pour de petits échantillons est obtenu quand l'intervalle de confiance est centré sur l'estimateur en échelle log plutôt que sur l'estimateur sur données brutes.

References

Anderson, J. A. (1972). Separate-sample logistic discrimination. *Biometrika* **59,** 19–35.

Aptech Systems (1992). *The GAUSS Programming Language, Version 3.0.* Kent, Washington: Aptech Systems.

Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case–control studies: A review. *Statistics in Medicine* **10,** 1753–1773.

Benichou, J. and Gail, M. H. (1990). Variance calculations and confidence intervals for estimates of attributable risk based on logistic models. *Biometrics* **46,** 991–1003.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: MIT Press.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Vol. I: The Analysis of Case–Control Data.* Lyon: International Agency for Research on Cancer.

Brinton, L. A., Hoover, R., and Fraumeni, J. F., Jr. (1982). Interaction of familial and hormonal risk factors for breast cancer. *Journal of the National Cancer Institute* **69,** 817–822.

Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A., et al. (1985). Estimating the population attributable risk for multiple risk factors using case–control data. *American Journal of Epidemiology* **122,** 904–914.

Cox, D. R. and Hinkley, D. R. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Deubner, D. C., Wilkinson, W. E., Helms, M. J., Tyroler, H. A., and Hames, C. G. (1980). Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. *American Journal of Epidemiology* **112,** 135–143.

Drescher, K. and Schill, W. (1991). Attributable risk estimation from case–control data via logistic regression. *Biometrics* **47,** 1247–1256.

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66,** 27–32.

Greenland, S. (1987). Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. *Statistics in Medicine* **6,** 701–708.

Greenland, S. and Robins, J. M. (1988). Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology* **128,** 1185–1197.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables.* Groningen: Wolters-Noordhoff.

Kelsey, J. L., Thompson, W. D., and Evans, A. S. (1986). *Methods in Observational Epidemiology.* New York: Oxford University Press.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, California: Lifetime Learning.

Kooperberg, C. and Petitti, D. B. (1991). Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case–control study. *Epidemiology* **2,** 363–366.

Last, J. M. (1983). *A Dictionary of Epidemiology.* New York: Oxford University Press.

MacMahon, B. and Pugh, T. F. (1970). *Epidemiology: Principles and Methods.* Boston: Little, Brown.

Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics* **29,** 479–486.

Mausner, J. S. and Bahn, A. K. (1974). *Epidemiology: An Introductory Text.* Philadelphia: W. B. Saunders.

Morgenstern, H. and Bursic, E. S. (1982). A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *Journal of Community Health* **7,** 292–309.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66,** 403–411.

Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis.* New York: Oxford University Press.

Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika* **62,** 371–375.

Whittemore, A. S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine* **1,** 229–243.

Zehna, P. W. (1966). Invariance of maximum likelihood. *Annals of Mathematical Statistics* **37,** 744.

Received March 1992; *revised September* 1992; *accepted January* 1993.