COMPUTERS AND BIOMEDICAL RESEARCH **29,** 63–75 (1996)
ARTICLE NO. 0006

Software for Attributable Risk and Confidence Interval Estimation in Case-Control Studies

MAURA MEZZETTI,* MONICA FERRARONI,* ADRIANO DECARLI,*,† CARLO LA VECCHIA,*,‡ AND JACQUES BENICHOU§

**Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Milano, Italy; †Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano, Italy; ‡Istituto di Ricerche Farmacologiche 'Mario Negri', Milano, Italy; and §Epidemiologic Methods Section, National Cancer Institute, Rockville, Maryland*

Received November 2, 1995

The increasing interest in obtaining model-based estimates of attributable risk (AR) and corresponding confidence intervals, in particular when more than one risk factor and/or several confounding factors are jointly considered, led us to develop a program based on the procedure described by Benichou and Gail for case-control data. This program is structured as an SAS-macro. It is suited to analysis of the relationship between risk factors and disease in case-control studies with simple random sampling of controls, in terms of relative risks and ARs, by means of unconditional logistic regression analysis. The variance of the AR is obtained by the delta method and is based on three components, namely, (i) the variance–covariance matrix of the vector of the estimated probabilities of belonging to joint levels of the exposure and confounding factors conditional on being a case, (ii) the variance–covariance matrix of the odds ratio parameter estimates from the logistic model, and (iii) the covariances between these probability and parameter estimates. Only a limited number of commands is requested from the user (i.e., the name of the work file and the names of the variables considered). The estimated relative risks for all the factors included in the model, the attributable risk for the exposure factor under consideration, and the corresponding 95% confidence intervals are given as outputs by the macro. Computational problems, if any, may arise for large numbers of covariates because of the resulting large size of vectors and matrices. The macro was tested for reliability and consistency on published data sets of case-control studies. © 1996 Academic Press, Inc.

Infectious diseases have long been reported to be caused by specific and ''necessary'' organisms or factors, and this was the basis of Koch's postulates for proof of causation. In contrast, for most major chronic diseases, a series of risk factors or a combination of factors are hypothesized to cause the disease (*1*). In this model, each risk factor can explain only a fraction of the disease. The attributable risk (AR), defined by Levin (*2*) as [(P(disease)-P(disease | no exposure))/P(disease)], gives a measure of the proportion of diseased people attributable to the exposure. AR estimates are mainly derived from observational

0010-4809/96 $18.00
Copyright © 1996 by Academic Press, Inc.
All rights of reproduction in any form reserved.

epidemiological studies. In this paper, we will focus on case-control studies. The attributable risk has major relevance on a public health scale, since it allows one to quantify the effect of various prevention strategies. This is because the AR is related not only to the strength of the association between risk factors and the disease, but also to the prevalence of exposure to the risk factors in the population. Exhaustive reviews of different aspects of interpretation, estimation, and application of AR are given, for example, by Benichou (*3*) and Coughlin *et al.* (*4*).

When estimating AR, it is useful to obtain not only a point estimate of AR, but also a variance estimate and hence the corresponding confidence interval (CI), in order to quantify the importance of various risk factors on a public health level. Over the last two decades, several authors have proposed new approaches to AR estimation from case-control studies and, correspondingly, have considered the issue of variance and interval estimation. Walter (*5*) proposed a variance estimate for an unadjusted AR estimate for a dichotomous exposure factor. Whittemore (*6*) extended Walter's work and derived point and variance estimates of AR for a dichotomous exposure, with possible adjustment for a (set of) counfounder(s) based on a weighted-sum approach. Denman and Schlesselman (*7*) provided point and variance estimates of AR for a risk factor with multiple levels but with no adjustment for confounding factors. Using Mantel–Haenszel estimates, both Greenland (*8*) and Kuritz and Landis (*9*) provided point and variance estimates of AR relative to a dichotomous exposure and with possible adjustment for confounding factors, both for matched and unmatched case-control studies.

Bruzzi *et al.* (*10*) derived model-based estimates that allow one to consider exposure factors with multiple levels and to control for confounders, which make this the most general approach. These model-based estimates can be obtained from the observed prevalence of various risk factors in cases and odds ratio estimates from logistic analysis provided that: (i) cases are representative of the diseased individuals in the population and (ii) odds ratio estimates are unbiased for relative risks (*10*). Benichou and Gail (*11*) presented variance estimates and confidence intervals for these model-based AR estimates.

There is however no available software to implement their work and obtain variance estimates and confidence intervals for model-based AR estimates. Given the increasing interest in estimating both the AR and the corresponding CIs (*12–16*), we developed an SAS-macro based on the procedure detailed by Benichou and Gail (*11*). The macro is general and requires only a limited number of commands to the user. The estimated relative risks (RR) for all the factors included in the model, the AR for the exposure factor under consideration, and the corresponding 95% CI are given as outputs by the macro, when the data are expressed as individual records.

The macro requires the Base SAS Module, the SAS/IML Software, and, obviously, SAS MACRO Language; it runs on version 6.04 of DOS and on version 6.10 of WINDOWS. The commented complete version of the macro is available on request from the authors by mail or e-mail (PGDUCA@IMICILEA.

CILEA.IT). The kernel of the macro is given in the Appendix. In the next section, we describe methods for point and variance estimation that are implemented in the macro and refer to the corresponding lines of the macro as we proceed.

Model-Based Point and Variance Estimation of the Attributable Risk

Let X be the exposure factor with I levels ($i = 1, \ldots, I$) and C_l the lth confounding factor ($l = 1, \ldots, L$), characterized by J_l different modalities ($j_l = 1, \ldots, J_l$). The attributable risk is defined as (*10*)

$$AR = 1 - \sum_{i=1}^{I} \sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \rho_{ij_1 \ldots j_L} R^{-1}_{i|j_1 \ldots j_L}, \qquad [1]$$

where $\rho_{ij_1 \ldots j_L} = P(X = x_i, C_1 = c_{j_1}, \ldots, C_L = c_{j_L} | D = 1)$ represents the probability that a subject belonging to the ith category of the exposure factor and to the j_l category of the first counfounder factor $C_1, \ldots$, and to the $j_L{}^{\text{th}}$ category of the Lth counfounder factor C_L, conditional on being a case ($D = 1$); $R_{i|j_1 \ldots j_L}$ is the relative risk associated with level i of X, conditional on levels $j_1, \ldots, j_L{}^{\text{th}}$ of counfounders $C_l, \ldots, C_L$.

In case-control studies, $\rho_{ij_1 \ldots j_L}$ can be estimated by the corresponding observed proportion $\hat{\rho}_{ij_1 \ldots j_L}$. The relative risk $R_{i|j_1 \ldots j_L}$, is defined as

$$R_{i|j_1 \ldots j_L} = \frac{P(D = 1 | X = x_i, C_1 = c_{j_1}, \ldots, C_L = c_{j_L})}{P(D = 1 | X = x_1, C_1 = c_{j_1}, \ldots, C_L = c_{j_L})}$$

and can be estimated by the odds ratio, when $P(D = 1 | X = x_i, C_1 = c_{j_1}, \ldots, C_L = c_{j_L})$ is small. In the absence of interactions between the exposure factor X and the confounders $C_l, \ldots, C_L$, the relative risks $R_{i|j_1 \ldots j_L}$ do not depend on the confounder levels. Therefore, for all levels i of exposure ($i > 1$), the relative risk can be estimated by $e^{\hat{\beta}_i}$, where β_i is the parameter in the logistic model relative to the ith category of exposure (β_i is the log odds ratio for level i).

The variance of the AR can be obtained by the delta method (*11*). The delta method applied to formula [1], yields

$$\begin{aligned}
\operatorname{var}(1 - \widehat{AR}) &= \operatorname{var}\left(\sum_{i=1}^{I} \sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \hat{\rho}_{ij_1 \ldots j_L} \hat{R}^{-1}_{i|j_1 \ldots j_L} \right) \\
&= \sum_{i=1}^{I} \sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \sum_{i'=1}^{I} \sum_{j'_1=1}^{J_1} \cdots \sum_{j'_L=1}^{J_L} \qquad [2] \\
&\quad [R^{-1}_{i|j_1 \ldots j_L} R^{-1}_{i'|j'_1 \ldots j'_L} \operatorname{cov}(\hat{\rho}_{ij_1 \ldots j_L}, \hat{\rho}_{i'j'_1 \ldots j'_L}) \qquad \text{(P1)} \\
&\quad + \rho_{ij_1 \ldots j_L} \rho_{i'j'_1 \ldots j'_L} \operatorname{cov}(\hat{R}^{-1}_{i|j_1 \ldots j_L}, \hat{R}^{-1}_{i'|j'_1 \ldots j'_L}) \qquad \text{(P2)} \\
&\quad + 2 \times \rho_{ij_1 \ldots j_L} R^{-1}_{i'|j'_1 \ldots j'_L} \operatorname{cov}(\hat{R}^{-1}_{i|j_1 \ldots j_L}, \hat{\rho}_{i'j'_1 \ldots j'_L})], \qquad \text{(P3)}
\end{aligned}$$

where $\hat{R}^{-1}_{i|j_1 \ldots j_L}$ denotes the inverse of relative risk estimate, obtained as the odds

ratio estimate from logistic regression. If we use dummy variables to represent exposure to X, then $\hat{R}^{-1}_{i|j_1 \cdots j_L}$ is a function of a vector $(I-1)$ of parameter estimates.

The AR variance is hence given by summing three components. The first one (P1) is obtained from the variance–covariance matrix of the vector $\hat{\rho}$ of observed proportions, which has a multinomial structure; the second one (P2) involves the variance–covariance matrix of the odds ratio estimates and is obtained from the information matrix of the logistic model; the last component (P3) involves the covariance between observed proportions in each cell and odds ratio estimates and is derived through the implicit delta method (*17*).

In the following sections the steps for the evaluation of the three components are described. From a computational point of view, the main problem in developing this software is related to the *a priori* unknown number of covariates and the increasing size of vectors and matrices as this number increases. It should be noted that the macro has been derived for a logistic model in which (i) the data are entered as individual records, (ii) the main effect of the I-level exposure factor of interest X is represented by $I-1$ dummy variables, (iii) the main effect of every confounder C_l is also represented by dummy variables, and (iv) there is no interaction term between exposure factor X and confounders $C_1, \ldots, C_L$ (assumption of no effect modification by the confounders). When one of these assumptions is used to compute the variance component, it is specified in the text below.

First Component (P1)

The kernel of component P1 is the variance matrix of the vector $\hat{\rho}$. It is obtained from the assumption of a multinomial structure. The computation of the $(I \times J_1 \times \cdots \times J_L) \times (I \times J_1 \times \cdots \times J_L)$ matrix Σ_ρ is based on the sum of two terms. One obtains

$$\Sigma_\rho = -\frac{\rho\rho^{\mathrm{T}}}{n} + \operatorname{diag}\left(\frac{\rho}{n}\right),$$

in which n represents the total number of cases. Since ρ is a $(I \times J_1 \times \cdots \times J_L) \times 1$ vector, the first term in the sum is a $(I \times J_1 \times \cdots \times J_L) \times (I \times J_1 \times \cdots \times J_L)$ matrix and the second term is a diagonal matrix of the same dimension.

This representation allows us to use simple vectors instead of matrices. We obtain P1 as

$$\sum_{i=1}^{I} \sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \sum_{i'=1}^{I} \sum_{j'_1=1}^{J_1} \cdots$$

$$\sum_{j'_L=1}^{J_L} R^{-1}_{i|j_1 \cdots j_L} R^{-1}_{i'|j'_1 \cdots j'_L} \operatorname{cov}(\hat{\rho}_{ij_1 \cdots j_L}, \hat{\rho}_{i'j'_1 \cdots j'_L}) = \mathbf{R}^{-1} \Sigma_{\hat{\rho}} (\mathbf{R}^{-1})^{\mathrm{T}},$$

where $\mathbf{R}^{-1}$ is the $1 \times (I \times J_1 \times \cdots \times J_L)$ vector with elements $R^{-1}_{i|j_1 \cdots j_L}$.

Lines 5 to 7 in the macro describe the SAS statements calculating component P1, with the notations *rr*1 and *rho* for $\mathbf{R}^{-1}$ and ρ, respectively.

Second Component(*P2*)

Observing that, in the absence of interaction between the exposure factor X and the counfounders $C_1, \ldots, C_L$, terms $R_{i|j_1 \ldots j_L}$ and $R_{i'|j'_1 \ldots j'_L}$ and their covariance only depend on modalities i and i' of the exposure factor, component P2 is rearranged as

$$\sum_{i=1}^{I} \sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \sum_{i'=1}^{I} \sum_{j'_1=1}^{J_1} \cdots \sum_{j'_L=1}^{J_L} \rho_{ij_1 \ldots j_L} \rho_{i'j'_1 \ldots j'_L} \operatorname{cov}(\hat{R}^{-1}_{i|j_1 \ldots j_L}, \hat{R}^{-1}_{i'|j'_1 \ldots j'_L})$$

$$= \sum_{i=1}^{I} \sum_{i'=1}^{I} \left(\sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \rho_{ij_1 \ldots j_L} \right) \left(\sum_{j'_1=1}^{J_1} \cdots \sum_{j'_L=1}^{J_L} \rho_{i'j'_1 \ldots j'_L} \right) \qquad [3]$$

$$\operatorname{cov}(\hat{R}^{-1}_{i|j_1 \ldots j_L}, \hat{R}^{-1}_{i'|j'_1 \ldots j'_L}).$$

The variance matrix of the vector $\mathbf{R}^{-1}$ can be obtained from the inverse of the information matrix $\mathbf{I}^{-1}$ from the logistic model, including as covariates the exposure factor and all the counfounding factors. By the delta method, the term for levels i and i' of exposure is given by

$$\operatorname{cov}(e^{-\hat{\beta}_i}, e^{-\hat{\beta}_{i'}}) = e^{-\beta_i} \times e^{-\beta_{i'}} \times \operatorname{cov}(\hat{\beta}_i, \hat{\beta}_{i'}),$$

where $\beta' = (\beta_2, \ldots, \beta_I)$ denotes the $(I-1) \times 1$ vector of parameters corresponding to dummy variables for levels 2, . . . , I of exposure to X, and $\beta_1 = 0$ by definition.

Only the $(I-1) \times (I-1)$ block of the information matrix concerning the exposure factor needs to be computed. In the macro, it is denoted by *cova1*. Line 11 describes the SAS statements used to compute P2. Terms *rhor1* and *I* represent $\rho \times \mathbf{R}^{-1}$ and $\mathbf{I}^{-1}$, respectively, and *de* represents the matrix of dummy variables describing levels of exposure to X.

Third Component (*P3*)

From the delta method for implicit functions, it is possible to derive the covariance matrix between the estimates of the $(p \times 1)$ vector θ of parameters from the logistic model of which β is a subvector and the estimate of vector $(\pi : \rho)$, where $\pi_{ij_1 \ldots j_L} = P(X = x_i, C_1 = c_{j_1}, \ldots, C_L = c_{j_L} | D = 0)$ represents the probability of belonging to the cell $X = x_i$, $C_1 = c_{j_1}$, . . . , $C_L = c_{j_L}$ when $D = 0$ (nondiseased subject). This covariance matrix is (*11*)

$$\operatorname{cov}\begin{pmatrix} \hat{\theta}, \begin{matrix} \hat{\pi} \\ \hat{\rho} \end{matrix} \end{pmatrix} = \mathbf{I}^{-1}\mathbf{H}\Sigma, \qquad [4]$$

where $\mathbf{I}^{-1}$ is the inverse of the information matrix, $\mathbf{H}$ is a $(p) \times (2 \times I \times J_1 \times \cdots \times J_L)$ matrix, and Σ is the estimated covariance matrix of vector $(\hat{\pi} : \hat{\rho})$. If only main effects of X and confounders are considered and if dummy variables are used, p is equal to $I + J_1 + \cdots + J_L - L$. The matrix Σ is structured as

$$\Sigma = \begin{bmatrix} \Sigma_{\hat{\pi}} & 0 \\ 0 & \Sigma_{\hat{\rho}} \end{bmatrix},$$

with $\Sigma_{\hat{\pi}}$ and $\Sigma_{\hat{\rho}}$ the variance matrices for vectors $\hat{\rho}$ and $\hat{\pi}$, respectively (both with multinomial structure). This form is suitable for controls selected by means of simple random sampling as well as stratified random sampling. In the macro, we considered the case of simple random sampling of the controls. Finally, the elements h_{sm} of $\mathbf{H}$ are given by

$$\frac{\partial^2 l}{\partial \theta_s \partial \pi_{ij_1 \cdots j_L}} \quad \text{for } 1 \le m = (i-1) \times J_1 \times \cdots \times J_L + (j_1 - 1) \times J_2 \times \cdots \times J_L + \cdots + (j_{L-1} - 1) \times J_L + j_L \le I \times J_1 \times \cdots \times J_L$$

$$\frac{\partial^2 l}{\partial \theta_s \partial \rho_{ij_1 \cdots j_L}} \quad \text{for } I \times J_1 \times \cdots \times J_L + 1 \le m = (i-1) \times J_1 \times \cdots \times J_L + (j_1 - 1) \times J_2 \times \cdots \times J_L + \cdots + (j_{L-1} - 1) \times J_L + j_L + I \times J_1 \times \cdots \times J_L \le 2 \times I \times J_1 \times \cdots \times J_L, \qquad [5]$$

where l represents the logistic log-likelihood of the observed data.

An estimate is obtained by replacing all matrices by their estimated matrices. The element h_{sm} of matrix $\mathbf{H}$ is estimated as

$$\begin{cases} -v p_{ij_1 \cdots j_L} & \text{if } s = i - 1 \text{ or } s = I + j_1 - 1 \text{ or } \ldots \text{ or } s = I + J_1 + \cdots + J_{L-1} + j_L - L \\ 0 & \text{otherwise} \end{cases}$$

$$\text{for } 1 \le m = (i-1) \times J_1 \times \cdots \times J_L + (j_1 - 1) \times J_2 \times \cdots \times J_L + \cdots + (j_{L-1} - 1) \times J_L + j_L \le I \times J_1 \times \cdots \times J_L$$

$$\begin{cases} n q_{ij_1 \cdots j_L} & \text{if } s = i - 1 \text{ or } s = I + j_1 - 1 \text{ or } \ldots \text{ or } s = I + J_1 + \cdots + J_{L-1} + j_L - L \\ 0 & \text{otherwise} \end{cases}$$

$$\text{for } I \times J_1 \times \cdots \times J_L + 1 \le m = (i-1) \times J_1 \times \cdots \times J_L + (j_1 - 1) \times J_2 \times \cdots \times J_L + \cdots + (j_{L-1} - 1) \times J_L + j_L + I \times J_1 \times \cdots \times J_L \le 2 \times I \times J_1 \times \cdots \times J_L,$$

where v is the total number of controls, n is the total number of cases, $p_{ij_1 \cdots j_L}$ is the predicted probability from the logistic model that a subject belonging to the $(ij_1 \cdots j_L)$ cell be a case, and $q_{ij_1 \cdots j_L}$ is the predicted probability that a subject belonging to the $(ij_1 \cdots j_L)$ cell be a control. The element h_{sm} is equal to zero

if the parameter corresponding to the sth position in θ (that is, the parameter with respect to which we are differentiating) is not relative to a joint category of the exposure factor X and the confounders. In the macro, lines 12–23 deal with the computation of matrix $\mathbf{H}$.

The matrix in [4] contains $(p) \times (2 \times I \times J_l \times \cdots \times J_L)$ elements, so it can have a huge number of columns, but by partitioning the matrices we can simplify the computation and optimize the memory size requested. The variance–covariance matrix between $\hat{\rho}$ and the vector of estimated exposure parameters, $\hat{\beta}$, is the principal concern. Therefore only a block of the matrix in [4] is of interest. If the matrices $\mathbf{I}^{-1}$, $\mathbf{H}$, and Σ are considered in blocks, one obtains

$$\mathbf{I}^{-1}\mathbf{H}\Sigma = \begin{bmatrix} \mathbf{I}_1^{-1} & \mathbf{I}_2^{-1} \\ \mathbf{I}_3^{-1} & \mathbf{I}_4^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_3 & \mathbf{H}_4 \end{bmatrix} \begin{bmatrix} \Sigma_\pi & 0 \\ 0 & \Sigma_\rho \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{I}_1^{-1}\mathbf{H}_1\Sigma_\pi + \mathbf{I}_2^{-1}\mathbf{H}_3\Sigma_\pi & \mathbf{I}_1^{-1}\mathbf{H}_2\Sigma_\rho + \mathbf{I}_2^{-1}\mathbf{H}_4\Sigma_\rho \\ \mathbf{I}_3^{-1}\mathbf{H}_1\Sigma_\pi + \mathbf{I}_4^{-1}\mathbf{H}_3\Sigma_\pi & \mathbf{I}_3^{-1}\mathbf{H}_2\Sigma_\rho + \mathbf{I}_4^{-1}\mathbf{H}_4\Sigma_\rho \end{bmatrix}.$$

The submatrix of interest is the upper right block, $\mathbf{I}_1^{-1}\mathbf{H}_2\Sigma_\pi + \mathbf{I}_2^{-1}\mathbf{H}_4\Sigma_\rho$. The required components are: $\mathbf{I}_1^{-1}$, the $(I-1) \times (I-1)$ block in the inverse of information matrix containing the variances and covariances of the estimated elements of the vector β representing the effect of X; $\mathbf{I}_2^{-1}$, the second $(I-1) \times (J_1 + \cdots + J_L - L - 1)$ block in the inverse of the information matrix containing the covariances between the elements of vector $\hat{\beta}$, and the intercept estimate $\hat{\alpha}$ and the elements of the vector of confounding parameters estimates $\hat{\gamma}$; $\mathbf{H}_2$, the second $(I-1) \times (I \times J_1 \times \cdots \times J_L)$ block of the matrix $\mathbf{H}$, obtained by differentiating l with respect to β then ρ, $\mathbf{H}_4$, the fourth $(J_1 + \cdots + J_L L + 1) \times (I \times J_1 \times \cdots \times J_L)$ block of the matrix $\mathbf{H}$, obtained by differentiating l with respect to α and γ then ρ. Thus, the third component, P3, can be written as

$$2 \times \sum_{i=1}^{I} \sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \sum_{i'=1}^{I} \sum_{j'_1=1}^{J_1} \cdots \sum_{j'_L=1}^{J_L} \rho_{ij_1 \ldots j_L} R^{-1}_{i'|j'_1 \ldots j'_L} \operatorname{cov}(\hat{R}^{-1}_{i|j_1 \ldots j_L}, \hat{\rho}_{i'j'_1 \ldots j'_L})$$

$$= 2 \times \sum_{i=1}^{I} \left(\sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \rho_{ij_1 \ldots j_L} \right) \sum_{i'=1}^{I} \sum_{j'_1=1}^{J_1} \cdots \sum_{j'_L=1}^{J_L} R^{-1}_{i'|j'_l \ldots j'_L} \operatorname{cov}(\hat{R}^{-1}_{i|j_l \ldots j_L}, \hat{\rho}_{i'j'_l \ldots j'_L})$$

$$= 2 \times (e^{-\beta_2}\rho_{2\cdot}, \ldots, e^{-\beta_I}\rho_{I\cdot})(\mathbf{I}_1^{-1}\mathbf{H}_2 + \mathbf{I}_2^{-1}\mathbf{H}_4)\Sigma_\rho(\mathbf{R}^{-1})^{\mathrm{T}},$$

where

$$\rho_{i\cdot} = \sum_{j_1=1}^{J_1} \cdots \sum_{j_L=1}^{J_L} \rho_{ij_1 \ldots j_L}, \qquad \text{for } i = 1, \ldots, I.$$

In the macro, the evaluation of component P3 is given by lines 24–27. From line 28 to line 33 we compute three sets of confidence intervals (*6*) based on

TABLE 1

MODEL DEFINITION, CORRESPONDING LIKELIHOOD, $\hat{\beta}$, AND $\widehat{AR}$ (WITH $\widehat{SD}$ VALUES IN PARENTHESES) FOR ALCOHOL CONSUMPTION IN THE ILLE-ET-VILAINE STUDY

Model	log (odds ratio)	−2 log-likelihood	$\hat{\beta}(\widehat{SD})$	$\widehat{AR}(\widehat{SD})$
I[a,d]	β_i	893.506	1.7299(.1752)	0.39489(.04203)
II[a,c,d]	$\alpha_j + \gamma_k + \beta_i$	764.744	1.5849(.1920)	0.38161(.04393)
III[b,d]	β'_i	842.991	1.2712(.2323) 2.0545(.2611) 3.3042(.3237)	0.70887(.05108)
IV[b,c,d]	$\alpha_j + \gamma_k + \beta'_i$	709.417	1.3939(.2464) 1.9264(.2791) 3.5135(.3815)	0.71811(.05016)

[a] The exposure factor is espressed in two levels, (i = 1, 2), 1 = 0–79 g/day and 2 = 80 + g/day.

[b] The exposure factor is expressed in four levels, (i = 1, 2, 3, 4), 1 = 0–39 g/day; 2 = 40–79 g/day, 3 = 80–119 g/day, and 4 = 120 + g/day.

[c] α_j (j = 1, 2, 3, 4) are the parameters for age, 1 = 25–34 years, 2 = 35–44 years, 3 = 45–54 years, and 4 = 55 + years. γ_k (k = 1, 2, 3) are the parameters for tobacco consumption, 1 = 0–9 g/day, 2 = 10–29 g/day, and 3 = 30 + g/day.

[d] By convention, $\beta_{i=1} = 0$ in models I and II, $\beta'_{i=1} = 0$ in models III and IV, and $\alpha_{j=1} = 0$ and $\gamma_{k=1} = 0$ in models II and IV.

$\widehat{AR}$, log $(1 - \widehat{AR})$, and $\log\{\widehat{AR}/(1 - \widehat{AR})\}$ and labeled inf and sup 1, logl, and logitl, respectively.

EXAMPLE

To illustrate the use of the Macro, we applied it to the data reported by Breslow and Day (*18*, Appendix 1) relative to a case-control study on esophageal cancer, conducted in the French department of Ille-et-Vilaine (*19*). We considered alcohol consumption as the exposure of interest. The odds ratios for alcohol consumption and their standard deviation, estimated using unconditional logistic regression, the model-based AR estimate, and its standard deviation estimate are shown in Table 1 for four models. Considering the strong influence that age and tobacco consumption have on the risk of esophageal cancer, these variables were included as confounding factors (without terms for interaction) in models II and IV, while they were ignored in models I and III for means of comparison. In models I and II the exposure factor has two levels, while in models III and IV it is split into four levels.

The attributable risk estimates and the corresponding SD estimates from the first two models (I and II) are lower than the ones in the other two models (III and IV). This is due to the more restrictive definition of the baseline level (*11*) in models III and IV (≤39 g/day) than in models I and II (≤79 g/day). Moreover, models I and III only include alcohol while models II and IV also include age and tobacco consumption, therefore yielding adjusted estimates of the effect of alcohol consumption. For the first two models, the SD for the adjusted AR estimate for alcohol consumption (model II) is greater than the SD for the crude

AR estimate (model I), while the inverse is true with models III and IV. However the ratio $\widehat{SD}(1 - \widehat{AR})/(1 - \widehat{AR})$ is bigger for adjusted than for crude estimates in both cases, as noted by Benichou (*3*) for other models applied to the same data.

Results in Table 1 are identical to those in Benichou and Gail (*11*) for models I and III, except for a typo on the standard deviation of $\widehat{AR}$ in that paper for model III. When the interaction between age and tobacco consumption, as well as main effects, was considered in models II and IV, our macro also yielded results identical to those reported by Benichou and Gail (*11*) and Benichou (*3*).

In order to illustrate the implementation of the macro, we now describe, as an example, how to run model II. The data utilized are published in Breslow and Day (*18*, Appendix 1) as grouped data. After reassembling the data following the categories indicated in Table 1, we proceeded to create the data set as individual records.

The macro only accepts one exposure factor but accepts several counfounding factors. Moreover, it is assumed that:

(1) The dependent variable (disease status) is coded 1 for cases and 2 for controls.
(2) The independent variables must have the same prefix (for example C, MAU, or MICKEY), followed by consecutive numbers (i.e., for model II, we have three independent variables, alcohol, age, and tobacco consumption, hence C1, C2, and C3, respectively). The first one must be the exposure factor (i.e., C1 = alcohol). All the variables can assume only consecutive integer values, without 0 (i.e., C2 = age has four categories that assume values from 1 to 4).
(3) The exposure factor must be coded in such a way that the risk increases with the categories.

The macro is called by means of the instruction %attrib(file,v,name,l), in which the macro parameters are defined as follows:

file = name of the data set containing the response variable, exposure variable, and all the confounding variables, each categorized following the above indications;
v = name of dependent variable;
name = prefix name for the independent variables (see point 2 above);
l = total number of independent variables (exposure + confounding factors).

In the example, the name of the data set containing individual data was ''esopha,'' the dependent variable was ''cascon,'' and the $l = 3$ independent variables (one exposure variable, alcohol, and two confounders, age and tobacco) were prefixed by the letter C. Therefore, the complete macro was called by %attrib(esopha,cascon,c,3).

In the Appendix we report only the kernel of the macro, i.e., the evaluation of the variance of the attributable risk and three different types of confidence intervals. To run this kernel the macro first generates the following components from the data:

—A row vector with dimension given by the number of cells (*&t2*) determined by the modalities of all the variables, including the dependent one (i.e., for model II *&t2* = 24). This vector containing the values $\exp(-\hat{\beta}_i)$, $i = 1, \ldots, I$(i.e. for model II the first 12 components of *rr*1 are equal to 1 corresponding to the reference category, while the last 12 components are equal to $\exp(-1.5849)$) and is denoted by *rr*1 in the macro.

—A vector (*rho*) which contains the probabilities that a case belongs to each cell (i.e., for model II the fifth component of the vector *rho* is $\hat{\rho}_{122} = 5/200 = 0.025$, because 5 cases belong to the first category of exposure and the second category of each confounder and 200 is the total number of cases).

—A vector containing the elements of submatrices $\mathbf{H}_2$ and $\mathbf{H}_4$ of $\mathbf{H}$, which can all be expressed as products of the total number of cases by the predicted probability from the unconditional logistic model to be a control conditional on the fact that a subject belongs to a particular cell, called *h* (i.e., the fifth component of the vector *h* in the model II is 200×0.96724).

—A data set containing the matrix $\mathbf{I}_1^{-1}$ previously defined, called *cova1* (variance and covariance matrix of the exposure parameter estimates; in model II *cova1* is a 1×1 matrix).

—A data set containing the matrix $\mathbf{I}_2^{-1}$ previously defined, called *cova2* (covariance matrix between exposure parameter estimates, and intercept and the confounder parameter estimates; in model II *cova2* is a 1×6 matrix).

As already pointed out, the previous components are automatically generated by the macro.

Discussion

We developed a computer program to obtain model-based estimates of the AR and corresponding CIs. It relies on methods presented by Bruzzi *et al.* (*10*) and Benichou and Gail (*11*), in which odds ratios are obtained through unconditional logistic regression and prevalence of exposure and confounding factors in cases is directly obtained from observed counts. Greeland and Drescher (*21*) proposed a modification of this approach in which prevalence of exposure and confounding factors in cases is estimated from the model. We did not implement their approach in this version of the program because simulations showed that the two approaches yield nearly identical results (*21*), and Greenland and Drescher's approach cannot be extended to conditional logistic regression, which we plan to consider in a future version of the program.

The program allows one to take into account exposure factors with multiple levels and to adjust for confounders. Although it does not set any limit to the number of confounders that can be included, in practice that number depends on RAM capacity and the number of free bytes on the hard disk. Moreover, execution time substantially increased with the number of confounding factors. The running time of the complete macro for the four models in Table 1 were checked on two different personal computers and one mainframe. The results are reported in Table 2.

TABLE 2

TOTAL MACRO RUNNING TIME (IN SECONDS) USING TWO DIFFERENT PCS AND ONE MAINFRAME

Model	First PC[a]	Second PC[b]	Mainframe[c]
I	161	32	26
II	229	39	30
II	170	33	26
IV	253	40	30

[a] 386dx, 20 MHz, 4 Mb of RAM, SAS for DOS version 6.04.
[b] 486dx, 66 MHz, 16 Mb of RAM, SAS for WINDOWS version 6.08.
[c] IBM 3090/200s, virtual machine 4 Mb of RAM, SAS for VM/CMS version 6.08.

In addition to the data on esophageal cancer, the program has been tested on stomach (*13*) and bladder cancer (*20*) data against a less general program written in Fortran. Identical results were found.

This current version only handles simple random sampling of the controls when data are given as individual records and does not allow for interaction between exposure and confounding factors. Interactions between confounders can, however, be introduced by creating new variables by cross-classifying confounding factors. In future versions of the program, we plan to (i) allow for stratified random sampling, frequency-matching, and individual matching of the controls, the latter option requiring use of conditional rather than unconditional logistic regression, (ii) allow for interactions between exposure and confounding factors, and (iii) allow for grouped data as input.

APPENDIX

```
%ATTRIB
 1  proc iml;
 2  use dummy var {rr1 rho h};
 3  read all into m1;
 4  rr1=m1[,1]; rho=m1[,2]; h=m1[,3]; rhor1=rr1#rho;
 5  prot=(rho)`/&n;          *where &n is the number of cases;
 6  pro=(rho)/&n;
 7  p1=-(rr1)`*(rho)*(prot)*(rr1)+(rr1)`*(pro#(rr1));
 8  use coval; read all into i1; use cova2; read all into i2; use cova1; read all into I;
 9  use dummy var {list of the names of the dummy variables relative to the exposure factor}
10  read all into de;
11  p2=(rhor1)`*de*I*((rhor1)`*de)`;
12  *building of matrix H (gg and gg1 to become matrix H2, ggg and ggg1 to become matrix H4);
13  ql=h#pro;   g=h`; gl=ql`;
14  gg=j(&ii1,&t2,0);       *where &ii1=number of modalities of exposure factor -1;
15  gg1=gg;
16  ggg=j(&t3,&t2,0);       *where &t3=number of parameters of confounders +1;
17  ggg1=ggg;
18  %do k= 1 %to &ii1; gg1[&k,]=g1;   gg[&k,]=g; %end;
19  %do k=1 %to &t3; ggg1[&k,]=gl; ggg[&k,]=g; %end;
```

```
20  use dummy var {list of the dummy variables relative to the all confounding factors};
21  read all into dc;
22  det=de`;  u=j(1,&t2,1);  dct=dc`;  u1=u//dct;
23  gg=gg#det;  gg1=gg1#det;  ggg=ggg#dct;  ggg1=ggg1#dct;
24  cc1=-i1*gg*(rho)*(prot);  cc2=i1*gg1; ccc1=-i2*ggg*(rho)*(prot); ccc2=i2*ggg1;
25  cc=cc1+cc2+ccc1+ccc2;      *matrix  I1^-1*H2*Σρ + I2^-1*H4*Σρ;
26  v1=((rhor1)`*de)`;
27  p3=2*v1`*cc*(rr1);
28  var=pl+p2-p3;   *attributable risk variance;
29  sd=sqrt(var);        *attributable risk standard deviation;   print sd;
30  ar=1-(rho)`*(rr1); infl=ar-1.96*sd; supl=ar+1.96*sd;
31  inflogl=1-(1=ar)*exp(1.96*sd/(1-ar)); suplogl=1-(1-ar)*exp(-1.96*sd/(1-ar));
32  inflogil=1/(1+((1-ar)/ar)*exp(+1.96*sd/(ar*(1-ar))));
    suplogil=1/(1+((1-ar)/ar)*exp(-1.96*sd/(ar*(1-ar))));
33  print ar infl supl inflogl suplogl inflogitl suplogitl;
34  quit;
%MEND;
```

Acknowledgments

This work was conducted within the framework of CNR (Italian National Research Council) Applied Project ''Clinical Application of Oncological Research'' (Contracts 94.01119.PF39 and 94.01321.PF39) and with the contribution of the Italian Association for Cancer Research. The authors thank Mrs. Angela R. Simm for editorial assistance. Dr M. Mezzetti is supported by the Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Rome.

References

1. Wynder, A. L. Principles of disease prevention from discover to application. *Soz Präventivmed* **39,** 267 (1994).
2. Levin, M. L. The occurrence of lung cancer in man. *Acta Un. Intern. Cancer* **9,** 531 (1953).
3. Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: A review. *Statist. Med.* **10,** 1753 (1991).
4. Coughlin, S., Benichou, J. and Weed, D. Attributable Risk Estimation in Case-Control Studies. *Epidem. Rev.* **16,** 51 (1994).
5. Walter, S. D. The distribution of Levin's measure of attributable risk. *Biometrika* **62,** 371 (1975).
6. Whittemore, A. S. Statistical methods for estimating attributable risk from retrospective data. *Statist. Med.* **1,** 229 (1982).
7. Denman, D. W., Schlesselman, J. J. Interval estimation of the attributable risk for multiple exposure levels in case-constrol studies. *Biometrics* **39,** 185 (1983).
8. Greenland, S. Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. *Statist. Med.* **6,** 701 (1987).
9. Kurtiz, S. J., and Landis, J. R. Attributable risk ratio estimation from matched-pairs case-control data. *Am. J. Epidemiol.* **175,** 324 (1987).
10. Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A., *et al.* Estimating the population attributable risk for multiple risk factors using case-control data. *Am. J. Epidemiol.* **122,** 904 (1985).
11. Benichou, J., and Gail, M. H. Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. *Biometrics* **46,** 991 (1990).
12. Gao, Y. T., McLaughlin, J. K., Blot, W. J., Ji, B. T., Benichou, J., Dai, Q., and Fraumeni, J. F. Risk factors for esophageal cancer in Shangai, China. I. Role of cigarette smoking and alcohol drinking. *Int. J. Cancer* **58,** 192 (1994).
13. La Vecchia, C., D'Avanzo, B., Negri, E., Decarli, A., Benichou, J. Attributable risks for stomach cancer in northern Italy. *Int. J. Cancer* **60,** 1748 (1995).
14. Caiaffa, W. T., Chiari, C. A., Figueiredo, A. R., Orefice, F., and Antunes, C. M. Toxoplas-

mosis and mental retardation-report of a case control study. *Mem. Int. Oswaldo Cruz* **88,** 253 (1993).

15. MacDonald, T. M., Beardon, P. H., McGillchrist, M. M., Duncan, I. D., McKendrick, A. D., and McDevitt, D. G. The risks of symptomatic vaginal candidiasis after oral antibiotic therapy. *Q. J. Med.* **86,** 419 (1993).
16. DeVos Irvine, H., Lamont, D. W., Hole, D. J., and Gillis, C. R. Asbestos and lung cancer in Glasgow and the west of Scotland. *Br. Med. J.* **306,** 1503 (1993).
17. Benichou, J., Gail, M. H. A delta-method for implicitly defined random variables. *Am. Statistician* **42,** 41 (1989).
18. Breslow, N. E., and Day, N. E. "Statistical Methods in Cancer Research. Volume I: The Analysis of Case-Control Studies" International Agency for Research on Cancer, Lyon, 1980.
19. Tuyns, A. J., Péquignot, G., and Jensen, O. M. Le cancer de l'oesophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac. *Bull. Cancer* **64,** 45 (1977).
20. D'Avanzo, B., La Vecchia, C., Negri, E., Decarli, A., and Benichou, J. Attributable risks for bladder cancer in northern Italy. *Ann. Epidemiol.* **5,** 427 (1995).
21. Greenland, S., and Drescher, K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics* **49,** 865 (1993).