Statistical Methods in Medical Research 2006; **15**: 611–625

Model-based estimation of the attributable risk in case–control and cohort studies

Christopher Cox Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, MD, USA

In a comprehensive review, Benichou recently discussed adjusted estimators of the attributable risk (AR). Among these are model-based estimates, where adjustment for confounding factors is based on a regression model. Different model-based approaches have been developed for case–control and cohort studies. The purpose of this article is to provide a detailed review and illustration of model-based methods for both types of sampling. For case–control studies, we show that two previously proposed approaches for the common case of a logistic regression model are in fact identical. This allows a unified approach to the estimation of the adjusted AR, which also accommodates stratified sampling. For cohort studies, a loglinear model is proposed for the case where cross-sectional sampling allows estimation of the prevalence of exposure; the approach can also be used for stratified sampling when the prevalence is known or can be estimated. For both designs, the standard error of the adjusted AR is estimated using the delta method. Estimation of the generalized AR is also discussed for both types of sampling. Examples show that for even fairly complex models, the computations are practical using standard statistical software. The bootstrap provides an easily implemented alternative to the delta method for the computation of standard errors.

1 Introduction

The attributable risk (AR) is widely used to provide a measure of the effect of exposure (E) to a risk factor on the occurrence of a disease (D). The AR is defined as the relative excess risk over the probability of disease in the unexposed ($\sim E$) population.

$$\mathrm{AR} = \frac{\Pr(D) - \Pr(D \mid \sim E)}{\Pr(D)} = 1 - \frac{\Pr(D \mid \sim E)}{\Pr(D)} \tag{1}$$

If $\mathrm{RR} = \Pr(D \mid E)/\Pr(D \mid \sim E)$ is the relative risk, then using Bayes' theorem it can be easily shown that

$$\mathrm{AR} = \frac{\Pr(E \mid D)(\mathrm{RR} - 1)}{\mathrm{RR}} = 1 - [\Pr(E \mid D)\mathrm{RR}^{-1} + \Pr(\sim E \mid D)] \tag{2}$$

This expression allows estimation of the AR in case–control studies of rare outcomes using the odds ratio, the standard estimate for this kind of sampling, as an approximation

Address for correspondence: Christopher Cox, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, E7642, MD 21205, USA. E-mail: ccox@jhsph.edu

 10.1177/0962280206071930

to the relative risk. A natural generalization of the basic definition is to allow several levels of exposure as well as adjustment for additional variables. The definition of the adjusted AR generalizes Equation (1) by conditioning on the covariates.[1–5] Consider a risk factor E with I levels (reference level $E = 1$), and a discrete set of distinct covariates $x_j (1 \leq j \leq J)$.[1,2] These may include several different variables, possibly with interaction terms. Interactions between covariates and exposure are also possible, but are not included in the covariate list. The definition of the adjusted AR is then as follows:

$$\begin{aligned} \mathrm{AR} &= 1 - \frac{\sum_j^J \Pr(x_j) \Pr(D \mid E = 1, x_j)}{\Pr(D)} \\ &= \frac{\sum_j \sum_i \Pr(E = i, x_j, D) - \sum_j \Pr(x_j) \Pr(D \mid E = 1, x_j)}{\Pr(D)} \\ &= \frac{\sum_j \sum_{i>1} \Pr(E = i, x_j) \left\{ \Pr(D \mid E = i, x_j) - \Pr(D \mid E = 1, x_j) \right\}}{\Pr(D)} \end{aligned} \tag{3}$$

The final expression can be used to define the risk attributable to the jth level of the covariates or, reversing the order of summation, the risk attributable to the ith level of exposure ($i > 1$).[2] Both sets of partial AR values sum to the adjusted AR. As discussed in the next section, if two sets of covariates (x_j, z_k) are available, then one can compute a partial AR_k by also summing over the levels of the x_j variables with z_k fixed. A number of different estimators of the adjusted AR have been proposed, including both weighted-sum and Mantel–Haenszel estimators based on stratification as well as estimators based on models. Benichou[1] provides a comprehensive review and comparison, with an extensive bibliography.

Model-based approaches to adjustment use a regression model to estimate the probabilities in Equation (3). As suggested by expressions (1) and (2), different approaches to estimation have been developed for cohort and case–control studies, although as one might expect logistic regression models can be used in both cases.[1,2,5] Both types of sampling are considered here, assuming a generalized linear model with the canonical link function. For case–control studies, two previously proposed approaches based on a logistic model are shown to be identical. For cohort studies, a loglinear model is proposed for the case of cross-sectional sampling; the approach can also be used for stratified cohort studies when additional data are available to estimate the prevalence of exposure. For both kinds of sampling, it is shown that the approach allows estimation of the adjusted AR from the model, and its variance using the delta method. This in turn allows estimation using standard statistical software, thus addressing the need for greater software availability.[1] The approach can also accommodate a number of generalizations of the basic definition in Equation (3), which have been proposed by several authors. These were not considered extensively by Benichou in his review,[1] and so are discussed in the following section. Examples illustrate the computations for both kinds of sampling. In recent years, the bootstrap has been proposed as an alternative to the delta method,[1] and results using the latter are compared with bootstrap standard errors.

2 Extensions of the adjusted AR

For case–control studies in which sampling of cases and controls was stratified by the levels of additional covariates z_k, Drescher and Schill[6] defined the AR specific to the kth covariate stratum. This definition can be stated in general terms as follows:

$$\mathrm{AR}s_k = 1 - \frac{\sum_j \Pr(x_j \mid z_k) \Pr(D \mid E = 1,\ x_j,\ z_k)}{\Pr(D \mid z_k)} \tag{4}$$

If the sampling was actually independent of the stratification variable, then the adjusted AR can be written as a weighted sum of the stratum specific attributable risks, with weights, $\Pr(z_k) \Pr(D \mid z_k) / \Pr(D) = \Pr(z_k \mid D)$.[6] Thus, the weights are determined by the distribution of the stratification variables among the diseased subjects, which is consistent with the sampling for a case–control study. An alternative definition can be obtained by conditioning only the numerator of Equation (3). The partially stratified AR is defined as follows:

$$\mathrm{AR}c_k = \frac{\sum_j \sum_i \Pr(D, E = i, x_j \mid z_k) - \sum_j \Pr(x_j \mid z_k) \Pr(D \mid E = 1, x_j, z_k)}{\Pr(D)} \tag{5}$$

The adjusted AR can again be written as a weighted sum; the weights in this case are the probabilities, $\Pr(z_k)$. This definition seems more useful for cohort studies. Note that if AR_k is the partial AR for stratum z_k, then $\mathrm{AR}c_k = \mathrm{AR}_k / \Pr(z_k)$.

The adjusted AR (3) is based on a comparison of disease risk among exposed individuals to that in the unexposed ($E = 1$) population. A generalization is to allow comparison with a population in which the exposure is not entirely absent (present only at the lowest level), but rather has a non-degenerate distribution, which is different from that of the original population. An example is when the exposure is reduced but not eliminated as a result of an intervention or education program. The definition of the generalized attributable risk (generalized impact fraction) given by Drescher and Becher[7] for such an alternative distribution, $\Pr^*(E = i, x_j)$, can be written as follows:

$$g\mathrm{AR} = \frac{\sum_j \sum_i \left\{ \Pr(E = i, x_j) \Pr(D \mid E = i, x_j) - \Pr^*(E = i, x_j) \Pr(D \mid E = i, x_j) \right\}}{\Pr(D)} \tag{6}$$

Although defined in more general terms, the alternative distribution would typically involve only the levels of the risk factor. For each value of the covariates, it is defined as a re-weighting of the original exposure probabilities by a specified probability density function $g(i \mid k) (1 \le i \le I)$, defined for each level of exposure, $k (1 \le k \le I)$.

$$\Pr^*(E = i, x_j) = \sum_k g(i \mid k) \Pr(E = k, x_j) \tag{7}$$

To illustrate this idea, an example considered by Drescher and Becher[7] is used. In this case $I = 4$, and we assume that a proportion q_1 of subjects change to the lowest risk

level, whereas an additional proportion q_2 change from the current level to the next lower level, with subjects already at the lowest level of risk remaining where they are. The family of density functions required in Equation (7) is given subsequently, and is used in Section 5 to illustrate the generalized AR.

i	k			
	1	2	3	4
1	1	$q_1 + q_2$	q_1	q_1
2	0	$1 - (q_1 + q_2)$	q_2	0
3	0	0	$1 - (q_1 + q_2)$	q_2
4	0	0	0	$1 - (q_1 + q_2)$

3 Case–control studies

As interactions between covariates and exposure are allowed, there will in general be different estimates of the relative risk ($\mathrm{RR}_{i|j} = \Pr(D \mid E = i, x_j)/\Pr(D \mid E = 1, xy)$) for different covariate categories. Extending formulation (2), it can be shown[2,3] that the adjusted AR (3) can be written as follows:

$$\mathrm{AR} = 1 - \sum_{i}^{I} \sum_{j}^{J} \Pr(E = i, x_j \mid D)\mathrm{RR}_{i|j}^{-1} = \sum_{i>1} \sum_{j} \Pr(E = i, x_j \mid D)(1 - \mathrm{RR}_{i|j}^{-1}) \quad (8)$$

The second expression, which does not involve the reference group, can be used to compute either set of partial AR values. For studies in which sampling of cases and controls was stratified by the levels of the covariate, z_k, it can similarly be shown that

$$\begin{aligned} \mathrm{AR}s_k &= \sum_{j,i} \Pr(E = i,\ x_j \mid D,\ z_k)(1 - \mathrm{RR}_{i|j,k}^{-1}) \\ &= \sum_{j,i} \frac{\Pr(E = i, x_j, z_k \mid D)}{\Pr(z_k \mid D)}(1 - \mathrm{RR}_{i|j,k}^{-1}) \end{aligned} \quad (9)$$

The partially stratified attributable risk, $\mathrm{AR}c_k$, is not useful as the probabilities $\Pr(z_k)$ cannot be estimated without additional data. Finally, Drescher and Becher,[7] show that the generalized attributable risk Equation (6) can be written as follows:

$$g\mathrm{AR} = 1 - \sum_{i} \sum_{j} \Pr(E = i, x_j \mid D)\mathrm{RR}_{i|j}^{-1} \sum_{k} g(k \mid i)\mathrm{RR}_{k|j} \quad (10)$$

To develop a model-based estimate of the AR and its various extensions for a case–control study, it is assumed that estimates of the relative risks (odds ratios) in expression

(8) are available from an appropriate generalized linear model with the canonical link function. The required exposure probabilities can be estimated empirically using the observed proportions among the cases.[1,2] Benichou and Gail[8,9] provided an extension of the delta method for implicit functions that can be used to compute a large sample variance for the adjusted attributable risk. Assuming a logistic model, Greenland and Drescher[10] described a maximum likelihood, that is model-based, approach generalizing work of Drescher and Schill.[6] For the case of categorical exposure variables it is shown that when the model is used with the canonical link function, the empirical approach can in general be viewed as model-based, and thus for the case of a logistic model is the same as the approach proposed by Greenland and Drescher.[10]

It is again assumed that the model has an ordinal exposure factor, E, with I levels, and J distinct covariates x_j. In the model matrix, a categorical variable with m categories is represented by $m - 1$ indicator variables, the omitted indicator corresponding to the reference level. With this formulation the model must include an intercept. Interaction terms are specified as collections of pairwise products of these indicators. Let w denote an indicator variable representing a particular level of exposure or exposure–covariate combination. It is well known from the form of the likelihood equations for a generalized linear model with the canonical link,[11,12] that if Y is the indicator variable for case status and $\hat{\mu}_l$ the predicted value for subject l from the model, then

$$\sum_l w_l y_l = \sum_l w_l \hat{\mu}_l$$

The left-hand side of this expression is the number of cases with the condition indicated by w. Thus, a model-based estimate of the probability is obtained.

$$\widehat{\Pr}(E = i, X = j \mid D) = \frac{\sum_l w_l y_l}{\sum_l y_l} = \frac{\sum_l w_l \hat{\mu}_l}{\sum_l \hat{\mu}_l} \tag{11}$$

The denominator in both right-hand side expressions (11) is the number of cases, as the model includes an intercept. Similar arguments apply to the reference category, whose indicator is a linear combination of columns of the model matrix. Now, model-based estimates of the exposure probabilities of Equation (11) are used in Equation (8) for the adjusted AR. The entire expression is a function of the parameters of the model, and the ordinary delta method can be applied. Both the stratified AR of Drescher and Schill[6] in Equation (9) and the generalized AR in Equation (10) can easily be accommodated as well.

To simplify the computations, only the distinct predicted values are summed up, that is over the unique patterns of exposure and covariates in the data. The number of subjects having each of these patterns can be used as weights, as these counts are marginal totals combining cases and controls and so are not random. If $n_l \geq 0$ denotes the number of subjects with predicted mean $\hat{\mu}_l$, then

$$\widehat{\Pr}(E = i,\ X = j \mid D) = \frac{\sum_{l|(w=1)} n_l \hat{\mu}_l}{\sum_l n_l \hat{\mu}_l} \tag{12}$$

The sum is now over the distinct predicted means. Comparison of expressions (12) and (8) with expressions (2) and (3) in Greenland and Dresher[10] shows that the two formulations are identical for the case of a logistic model.

4 Cohort studies

For the simple 2×2 case, Walter[13] discussed two different sampling designs for cohort studies. A stratified design is one in which sampling is stratified by exposure so that the marginal totals for the risk factor are fixed; for a cross-sectional design only the total number of subjects is fixed. The cross-sectional design allows estimation of the rate of exposure in the population, whereas the stratified design does not. In the latter case, an independent estimate of $\Pr(E)$ must be used to calculate the attributable risk. Walter[13] provided estimates and asymptotic standard errors using the delta method for both sampling designs. Greenland and Dresher[10] described an approach based on Equation (3), which uses a logistic regression model to provide adjustment for covariates. Basu and Landis[5] pointed out that this approach implicitly assumes that the estimates of the exposure probabilities from the data are fixed constants. They provided an approach for the case of cross-sectional sampling based on a logistic model and using the extended delta method of Benichou and Gail.[8]

Given a cross-sectional sample, consider the loglinear model corresponding to a logistic regression model having I levels of exposure, with the first category as the reference, a set of J distinct covariate values x_j for the observation specified by $(E = i, x_j)$, and possibly additional interactions between exposure and covariates.[12] Thus the model is saturated in the covariates and the exposure variable, and contains interaction terms with disease status corresponding to the terms in the logistic model. Note that for notational convenience both exposure and interactions with covariates are excluded from the covariate vector. Let $n_{ij} = n_{1ij} + n_{0ij}$ be the number of subjects (with n_{1ij} and without n_{0ij}, disease) having exposure level i and covariates x_j. Because the loglinear model corresponds to the logistic regression model, $n_{ij} = \hat{n}_{ij} = \hat{n}_{1ij} + \hat{n}_{0ij}$, where the right-hand side denotes the corresponding predicted values from the model. Denote the regression coefficients for the covariates (the interaction between covariates and disease status in the loglinear model) by γ. The adjusted AR in Equation (3) is written as follows, using the dot notation for summation over the levels of a subscript.

$$\begin{aligned} \mathrm{AR} &= 1 - \frac{\sum_j \sum_i \hat{n}_{ij}/(1 + \exp(-(\hat{\beta}_d + \hat{\gamma}' x_j)))}{\sum_j \sum_i \hat{n}_{1ij}} \\ &= 1 - \frac{\sum_j \hat{n}_{.j}/(1 + \exp(-(\hat{\beta}_d + \hat{\gamma}' x_j)))}{\hat{n}_{1..}} \end{aligned} \tag{13}$$

For computational convenience, the second expression uses estimated frequencies rather than probabilities. Note that the individual terms $\hat{n}_{1ij}$ in the denominator will not equal the observed values unless the model is saturated; however, the sum will equal the total

number of diseased subjects, as the indicator variable for disease status is included in the loglinear model and again the canonical link function is used. The disease parameter β_{d} corresponds to the intercept of the logistic regression, and the parameters γ for the covariates are the same as for the logistic regression model. The number of probabilities $\Pr(D \mid E = 1, x_j)$ corresponds to the number of distinct covariate patterns (exclusive of interactions between exposure and covariates). Each probability is multiplied by a sum of $2I$ predicted values from the loglinear model. The estimate in Equation (13) is similar to that proposed by Greenland and Dresher,[10] which uses a logistic model with the observed totals n_{ij}. In the loglinear model these totals are random and not fixed. Using expression (13) with this model and applying the ordinary delta method gives the correct asymptotic variance for the adjusted AR.

The partial AR for the ith level ($i > 1$) of exposure can be written as follows:

$$\mathrm{AR}_i = \frac{\hat{n}_{1i\cdot} - \sum_j \hat{n}_{ij}/(1 + \exp(-(\hat{\beta}_d + \hat{\gamma}' x_j)))}{\hat{n}_{1\cdot\cdot}} \tag{14}$$

The risk attributable to the jth stratum of the covariates is similar. The partially stratified AR in Equation (5) also has a similar form.

$$\mathrm{AR}c_k = \frac{\hat{n}_{1\cdot\cdot k} - \sum_j \hat{n}_{\cdot jk}/(1 + \exp(-(\hat{\beta}_d + \hat{\gamma}'(x_j, z_k))))}{\hat{n}_{1\cdots}(\hat{n}_{\cdot\cdot k}/\hat{n}_{\cdots})} \tag{15}$$

Finally, the definition from Equations (6) and (7) can be applied to produce an expression for the generalized AR. Denote the regression coefficients for exposure in the logistic model by $\beta_i (1 \leq i \leq I)$, including any interactions between exposure and covariates as part of the exposure variable and noting that $\beta_1 = 0$. Then Equation (6) can be written as follows.

$$g\mathrm{AR} = 1 - \frac{\sum_j \sum_i \sum_k g(i \mid k)\hat{n}_{kj}/(1 + \exp(-(\hat{\beta}_d + \hat{\beta}_i + \hat{\gamma}' x_j)))}{\sum_j \sum_i \hat{n}_{1ij}} \tag{16}$$

For a cohort study with sampling stratified by the levels of exposure, additional information is required to estimate $\Pr(E = i)(1 \leq i \leq I)$. If this is available, then a modification of Equation (3) in the spirit of Equation (4) can be used to compute the adjusted AR.

$$\begin{aligned} \mathrm{AR} &= 1 - \frac{\sum_i \sum_j \Pr(E = i) \Pr(x_j \mid E = i) \Pr(D \mid E = 1, x_j)}{\sum_i \Pr(E = i) \Pr(D \mid E = i)} \\ &= 1 - \frac{\sum_i \sum_j \Pr(E = i)(\hat{n}_{ij}/\hat{n}_{i\cdot})/(1 + \exp(-(\hat{\beta}_d + \hat{\gamma}' x_j)))}{\sum_i \Pr(E = i)(\hat{n}_{1i\cdot}/\hat{n}_{i\cdot})} \end{aligned} \tag{17}$$

The conditional probabilities in Equation (17) do not involve the parameters that ensure the equality of the observed and expected exposure totals; in fact the same results would

be obtained if a multinomial regression model were used. If the exposure probabilities are known, then they can simply be substituted. If they must be estimated from data, then the independent multinomial likelihood can be included in the loglinear model, and the estimated probabilities used in Equation (17).

5 Illustrations of model-based estimates

The main advantage of the present formulation is that it can be implemented using standard statistical software. Particular packages are available which will estimate model parameters and their covariance matrix by maximum likelihood, and then compute non-linear functions of the parameter estimates and calculate standard errors using the delta method. One such package is SAS PROC NLMIXED (SAS Institute, Cary, NC, USA), which is the program used. The regression model can easily be specified using a flexible syntax. In addition, the program requires specification of the likelihood function; keywords are provided for standard distributions such as the binomial and Poisson. Similar programs are also available in other packages, including Stata (Stata Corp, College Station, TX, USA). For the bootstrap, sampling with replacement from the rows of the data matrix was used;[13] this is easily accomplished using the macro facility available in many statistical packages. For each example, a total of 200 bootstrap samples was generated, as recommended by Efron and Tibshirani.[14]

To illustrate estimation of both the AR and its variance for case–control studies, a data set considered by the number of authors is used. This is the Ille-et-Vilaine case–control study of esophageal cancer, which appears in Appendix 1 of the monograph by Breslow and Day.[15] A reduced data set was analysed extensively by Benichou,[2] and appears in his paper. There are 200 cases and 775 controls, chosen by random sampling. The ordinal exposure variable, alcohol consumption, has four levels, with the lowest level of consumption as the reference. Covariates are ordinal variables for age (six 10-year levels) and smoking (four levels), again with the lowest level as the reference category. This gives a total of 192 combinations and a count of the number of subjects having each combination, with 57 combinations having zero replicates. Benichou[2] collapsed both age and smoking to three levels (combining consecutive pairs of age and the middle two smoking categories) for a total of 72 observations, of which 13 have zero replicates.

As an example of a cross-sectional study, the data considered by Basu and Landis[5] is used. This is a data set from NHANESII (Second National Health and Nutrition Examination Survey) with 966 women, aged 18–24 years, classified by two racial categories and four ordered risk categories (0–3) based on body mass index (BMI). The outcome variable is high diastolic blood pressure, defined as at or above the 90th percentile. The data are given in Table 1 of their paper.

5.1 Case–control studies

5.1.1 Example 1

The first model has all four levels of exposure but no covariates. This model is number 11 in Table 4 of Benichou.[2] The model has a total of four parameters, so that it is necessary to compute three odds ratios and four exposure probabilities. For this example,

Table 1 Number of subjects in the Ille-et-Vilaine case–control study of the association between esophageal cancer and (any) alcohol exposure (*E*), controlling for three levels of age and smoking

	Case			Control			Total		
	Age 1	Age 2	Age 3	Age 1	Age 2	Age 3	Age 1	Age 2	Age 3
E Smoke 1	2	13	54	77	45	73	79	58	127
Smoke 2	7	21	48	59	41	66	66	62	114
Smoke 3	0	11	15	20	4	4	20	15	19
~*E* Smoke 1	0	1	8	100	45	107	100	46	115
Smoke 2	1	0	14	36	28	47	37	28	61
Smoke 3	0	0	5	13	4	6	13	4	11

one parameter was used for each exposure group (b_1–b_4), and the model had no intercept. The estimated relative risks are given by $RR_i = \exp(b_i - b_1)$. The total number of subjects in each of the four exposure groups (415, 355, 138, 67) is also needed. As there are no covariates, the numerator of expression (12) has only a single term in each case. The total number of cases is then

$$\frac{415}{1+\exp(-b_1)} + \frac{355}{1+\exp(-b_2)} + \frac{138}{1+\exp(-b_3)} + \frac{67}{1+\exp(-b_4)} = 200$$

The estimated AR (SE_{delta}, SE_{boot}) was 0.709 (0.0502, 0.0510) compared with 0.709 (0.0511) in Table 4 of Benichou.[2] The risks attributable to the three ordered categories of exposure were $AR_2 = 0.270$ (0.0423, 0.0446), $AR_3 = 0.222$ (0.0287, 0.0292) and $AR_4 = 0.217$ (0.0208, 0.0206). Bootstrap standard errors agree with those based on the delta method. Greenland and Drescher[9] demonstrated that approximate normality can be improved by estimating the transformed parameter, $\log(1 - AR)$. This transformation is easily programmed using the present approach. For this example, $\widehat{\log}(1 - AR) = -1.234(0.173, 0.177)$ is obtained. The next model in the list was also tried (model 12), having four levels of exposure as well as covariates smoking, age and their interaction, for a total of 12 parameters. In this case, the model did not fit well, and some of the interaction parameter estimates were very large, with extremely large standard errors. For this model, our results did not agree as well with those of Benichou, although there is a definite indication that the data are being over-fitted.

5.1.2 Example 2

Consider the reduced data set used by Benichou,[2] with a binary exposure in which the three highest levels of alcohol consumption are combined into a single exposure category; the data are summarized in Table 1. Following Benichou,[2] consider a model with three categories of age (parameters a_1, a_2) and smoking (s_1, s_2) as well as their interaction (as_{11}, as_{12}, as_{21}, as_{22}) as covariates. The model includes alcohol exposure and the interaction between alcohol exposure and age. For computational convenience, consider the exposure interaction as four categories of exposure [parameters aa_1 (the effect of exposure in the first age category), aa_2, aa_3]; the reference category is all unexposed subjects. Thus the model has a total of 12 parameters, including the intercept

(b_0). With this formulation, four different relative risks (odds ratios) are needed, one of which is one. For example for the second age category, $RR_2 = \exp(aa_2)$.

The second expression on the right-hand side of Equation (8) can be used to simplify the computations. Thus the AR is the sum of three terms, one for each exposure (age) category but the reference category, and the number of cases in these three groups (from Table 1: 9, 45, 117) is also needed. Each is obtained by summing three terms involving the number of exposed subjects in a given age category for each level of smoking. For example, the following is required for age 2 (Table 1).

$$\frac{58}{1+\exp(-(b_0+aa_2+a_1))}+\frac{62}{1+\exp(-(b_0+aa_2+a_1+s_1+as_{11}))}+\frac{15}{1+\exp(-(b_0+aa_2+a_1+s_2+as_{12}))}=45$$

Thus, the numerator of expression (12) has three different terms in each case. The expression for the total number of cases is even more complex, involving a number of terms equal to the number of cells defined by the model, which is $2 \cdot 3 \cdot 3 = 18$. Replacing these lengthy expressions by the actual numbers for simplicity, an expression is obtained that gives the correct estimate, but a standard error that is too small.

$$AR = \frac{9}{200}(1-\exp(-aa_1))+\frac{45}{200}(1-\exp(-aa_2))+\frac{117}{200}(1-\exp(-aa_3))$$

The computations for this model can be simplified by choosing exposure as the dependent variable in the logistic regression model, and replacing the indicator variable for exposure status by that for disease status. In this case, the total number of cases is no longer random, and it can then be used in the denominator of the exposure probabilities without being expressed as a function of the parameters of the model. The odds ratios and their covariance matrix will be the same, as can be seen by considering the corresponding Poisson regression model, having terms for both case status and exposure, as well as appropriate interactions with the covariates, and the effects and interaction for the covariates. Because the original model is saturated in the covariates, the regression coefficients for the odds ratios (which involve both exposure and case status) in the loglinear model are those for either logistic regression model.[11] And arguing as before, the exposure probabilities for this model are the same as the empirical exposure probabilities. This alternative model requires the number of cases (both exposed and unexposed), one of which is zero, for each of the nine combinations of the two covariates (Table 1). The second expression on the right-hand side of Equation (8) can again be used to further simplify the computations. The following is the complete expression for

the adjusted AR.

$$
\begin{aligned}
&(2/(1+\exp(-(b_0+aa_1)))\\
&\quad+8/(1+\exp(-(b_0+aa_1+s_1))))*(1-\exp(-aa_1))/200\\
&\quad+(14/(1+\exp(-(b_0+aa_2+a_1)))\\
&\quad+21/(1+\exp(-(b_0+aa_2+a_1+s_1+as_{11})))\\
&\quad+11/(1+\exp(-(b_0+aa_2+a_1+s_2+as_{12}))))*(1-\exp(-aa_2))/200\\
&\quad+(62/(1+\exp(-(b_0+aa_3+a_2)))\\
&\quad+62/(1+\exp(-(b_0+aa_3+a_2+s_1+as_{21})))\\
&\quad+20/(1+\exp(-(b_0+aa_3+a_2+s_2+as_{22}))))*(1-\exp(-aa_3))/200
\end{aligned}
$$

The original model with case status as the dependent variable is number eight in Table IV of Benichou.[2] The estimated AR (SE) using the alternative model was 0.723 (0.0493), which agrees well with the value of 0.723 (0.0502) reported by Benichou. The original model gave an estimate (SE_{delta}, SE_{boot}) of 0.723 (0.0491, 0.0481). The risks attributable to the three ordered categories of exposure defined by the interaction with age were $AR_1 = 0.0383$ (0.0157, 0.0156), $AR_2 = 0.219$ (0.0244, 0.0249) and $AR_3 = 0.466$ (0.0482, 0.0477). Thus when adjusted for age and smoking, using a model that includes an interaction between any exposure and age, the largest attributable risk occurs in the highest risk category, unlike the unadjusted values in the first example. A similar model (model nine) with an alternative exposure by smoking interaction gave AR = 0.703 (0.0545), compared with values of 0.703 (0.0544) reported by Benichou. A simpler model with no exposure by covariate interactions (model seven) had AR = 0.719 (0.0505) compared with 0.719 (0.0504). As both covariates are ordinal, scores can be assigned to the three categories and the two variables treated as continuous. The computations are similar to those for model seven. For this model, AR = 0.718 (0.0497).

5.1.3 Example 3

The Ille-et-Vilaine study was also considered by Drescher and Schill,[6] using the combination of alcohol and smoking with the original four categories, for a total of 16 exposure categories. The covariate was age with the original six categories, and the model did not include interactions. Drescher and Schill presented estimates and confidence intervals (CI) for each of the stratum-specific attributable risks and the adjusted AR, using both their method and the approach of Benichou and Gail.[8] Their method involves fitting the logistic regression model with separate intercepts for each age category and offset, $\log(n_{1j}/n_{0j})$, where n_{1j} and n_{0j} are the number of cases and controls in stratum j, assuming that these totals are fixed by the sampling design. The intercepts then estimate the stratum specific parameters $\log(1 - AR_j)$, and the standard errors can be obtained from a simple adjustment to the estimated covariance matrix. The overall AR is a weighted sum of the stratum-specific estimates, and the variance can be calculated using the delta method.

Table 2 shows a comparison of this approach with that based on Equation (9), similar to Table 3 of Drescher and Schill.[6] Results obtained using the approach of Drescher

Table 2 Transformed values $\log(1 - AR_i)$ and standard errors for age-specific and total attributable risks, together with the actual values and 95% CI

Age stratum	I				II				
	log	SE	AR_i	CI	log	SE_D	SE_B	AR_i	CI
25–34	−2.150	0.413	0.88	0.74–0.95	−2.151	0.334	0.350	0.88	0.78–0.94
35–44	−1.865	0.273	0.85	0.74–0.91	−1.865	0.253	0.257	0.85	0.75–0.91
45–54	−1.830	0.247	0.84	0.74–0.90	−1.830	0.234	0.227	0.84	0.75–0.90
55–64	−1.779	0.229	0.83	0.74–0.89	−1.779	0.218	0.218	0.83	0.74–0.89
65–74	−1.343	0.199	0.74	0.61–0.82	−1.343	0.185	0.186	0.74	0.62–0.82
75+	−1.120	0.242	0.70	0.52–0.81	−1.120	0.182	0.180	0.70	0.57–0.79
Total					−1.609	0.202	0.200	0.80	0.70–0.87

Note: Results in the left panel (I) are based on the approach of Drescher and Schill;[6] the right panel (II) is based on (9), with bootstrap standard errors (SE_B) for comparison.

and Schill, in the four columns of the left-hand panel, agree closely with those reported by the authors. The stratified estimates from (9) in the right panel are very similar to those in the left, while the standard errors are slightly smaller. Bootstrap standard errors agree with those obtained from the delta method. These results are different from those reported by Drescher and Schill, which did not agree with their method, particularly in the youngest age category. Given our results, the two approaches are expected to show close agreement. The differences in the standard errors may be partly because of the use of the offset, which involves an additional approximation, in the method of Drescher and Schill. In this data set, the numbers of cases and controls in some of the age strata are fairly small; the largest numbers occur in the third, fourth and fifth age categories,

Table 3 Generalized attributable risks and standard errors for both a case–control (Ille-et-Vilaine) and a cohort (NHAINESII) study

q_1	q_2	Case–control study			Cohort study		
		*g*AR	SE	CI	*g*AR	SE	CI
0	0	0	–	–	0	–	–
0	0.2	0.115	0.008	0.10–0.13	0.052	0.011	0.03–0.07
0.2	0	0.145	0.010	0.13–0.16	0.058	0.014	0.03–0.09
0	0.4	0.229	0.016	0.20–0.26	0.105	0.021	0.06–0.15
0.2	0.2	0.260	0.017	0.23–0.29	0.111	0.023	0.07–0.15
0.4	0	0.290	0.019	0.25–0.33	0.117	0.028	0.06–0.17
0.2	0.4	0.374	0.025	0.32–0.42	0.163	0.033	0.10–0.22
0.4	0.2	0.404	0.027	0.35–0.45	0.169	0.036	0.10–0.24
0	0.8	0.459	0.033	0.39–0.52	0.210	0.042	0.12–0.29
0.4	0.4	0.519	0.034	0.45–0.58	0.222	0.045	0.13–0.31
0	1	0.574	0.041	0.49–0.65	0.262	0.053	0.15–0.36
0.8	0	0.579	0.039	0.50–0.65	0.234	0.055	0.12–0.33
0.2	0.8	0.604	0.041	0.51–0.68	0.268	0.053	0.16–0.37
0.8	0.2	0.694	0.046	0.59–0.77	0.286	0.063	0.15–0.40
1	0	0.724	0.048	0.61–0.80	0.292	0.069	0.14–0.42

Note: Following Drescher and Becher,[7] 95% CI are based on the transformation $\log(1 - g\text{AR})$. The distribution $\Pr^*$ is estimated using the data (7), and so is random.

where the agreement between the standard errors is a little better. The overall AR in the right panel is identical to the value reported by Drescher and Schill, as is the CI.

5.1.4 Example 4

Drescher and Becher[7] used the same model to illustrate the generalized AR (6) for alcohol consumption, with the alternative distribution, Pr* as given below (7). We used (10) to estimate the *g*AR, and its standard error using the delta method, for the same combinations of the probabilities q_1 and q_2 employed by Drescher and Becher, and followed their procedure for computing CI using the normal approximation to the transformed estimate $\log(1 - g\mathrm{AR})$. The results are shown in the three columns of the left hand panel of Table 3, and agree closely with results given in Table 1 of Drescher and Becher. The standard errors in Table 3 are very slightly (0.002 or less) smaller than those reported by Drescher and Becher, who used an alternative variance estimate to that of Benichou and Gail.[8] Note that the *g*AR is linear in each probability q_i when the other one is zero.

5.2 Cohort studies

5.2.1 Example 1

The same model as the first example in the previous section is considered, having four categories of exposure, the first of which is the reference category, and no covariates. The parameters for the exposure groups ($b_1 - b_4$) in the loglinear model are interactions between exposure group and case status, and the model has four additional nuisance parameters. Let 1_{e_i} be the indicator for the ith level of exposure and 1_d the indicator for disease status. With the convention $e_0 = 0$, the loglinear model is

$$\mu_{i1_d} = \exp(m_0 + e_i 1_{e_i} + b_i 1_{e_i} 1_d) \quad (i = 0, 3)$$

It follows that $\Pr(D \mid \sim E) = 1/(1 + \exp(-b_0))$. From this the following expression for the AR is obtained.

$$1 - \frac{\begin{array}{c}(\exp(b_1) + \exp(e_1 + b_1) + \exp(e_2 + b_2) + \exp(e_3 + b_3) + \\ 1 + \exp(e_1) + \exp(e_2) + \exp(e_3))/(1 + \exp(b_0))\end{array}}{1 + \exp(e_1 + b_1 - b_0) + \exp(e_2 + b_2 - b_0) + \exp(e_3 + b_3 - b_0)}$$

The result was AR $(\mathrm{SE}_{\mathrm{delta}}, \mathrm{SE}_{\mathrm{boot}}) = 0.301$ (0.0684, 0.0714). The estimate and standard error using the delta method are identical to the result in Table 3 of Basu and Landis.[5] The risks attributable to the three exposure categories (14) were, $\mathrm{AR}_2 = 0.00781$ (0.0320, 0.0345), $\mathrm{AR}_3 = 0.0184$ (0.0259, 0.0259) and $\mathrm{AR}_4 = 0.275$ (0.0493, 0.0462).

5.2.2 Example 2

The binary covariate for race is added to the loglinear model, which now has 13 parameters, including eight nuisance parameters. There are two exposure probabilities, one for each racial category. In the numerator of expression (13) each probability is multiplied by a sum of eight terms, and the denominator, the (estimated) number of cases, requires eight terms. The result was again identical to that of Basu and Landis,[5] AR = 0.292 (0.0692, 0.0673); the category-specific risks were similar to those in the first example. The risk (SE) attributable (14) to blacks, $\mathrm{AR}_B = 0.078$ (0.022) and whites

$AR_W = 0.214$ (0.054), and the partially stratified risks (15), $ARc_B = 0.568$ (0.155) and $ARc_W = 0.248$ (0.063) were also calculated. The former reflect the racial distribution of the sample, whereas the latter are adjusted for the relatively small proportion of blacks (13.8%). For this reason, as well as the fact that the model is not saturated, and the prevalence of disease is greater in blacks than whites, both overall (18.0 versus 10.1%) and in the reference group (13.7 versus 7.1%), the partially stratified attributable risks will be different from those for the individual races. For comparison these were, for blacks, AR = 0.241 (0.167) and for whites, AR = 0.300 (0.076). The stratified attributable risk for blacks, $ARs_B = 0.352$ (0.082) was also greater than for whites $ARs_W = 0.275$ (0.069).

The generalized attributable risk is also illustrated using this example, with the same distribution Pr* as for Section 5.1.4, and CI based on $\log(1 - g\text{AR})$. The three columns in the right hand panel of Table 3 give the gAR and its standard error based on the delta method, as well as large sample 95% CI. In contrast to the case–control example, when the distribution of risk is shifted one category ($q_2 = 1$), the AR is nearly as great as when the risk factor is eliminated ($q_1 = 1$). Again the estimate is linear in each of the two probabilities when the other one is zero.

Finally, the case of stratified sampling is illustrated using this same example. Basu and Landis[5] considered stratification on both the exposure and covariate categories. They observed that in this case the standard error of the AR can be computed using only the covariance matrix of the parameter estimates from the logistic regression. Treating the eight exposure × race totals as constants and modifying (17) to include the probabilities $\Pr(E_i, x_j)$, the loglinear model gave AR = 0.292 (0.0673), the same as in Table 3 of Basu and Landis. We can also consider the less restrictive case where sampling is stratified by exposure alone, so that only the four exposure totals are fixed. The standard error from (17) was nearly identical, although this will not always be the case. Of course, any set of known exposure probabilities can be used, or they can be estimated from additional data.

6 Discussion

Model-based estimates of the adjusted attributable risk and its standard error for both case–control and cohort studies have been reviewed. For the former, it was shown that for a generalized linear model with the canonical link, the empirical approach to the computation of the exposure probabilities used in the calculation is in fact model-based, and for logistic models, the same as the model-based approach proposed by Greenland and Drescher.[9] This allows a consistent approach to estimation of the adjusted AR from a logistic model. For cohort studies a loglinear approach was proposed, which can be used for both cross-sectional and stratified sampling. For both types of studies the delta method was applied to estimate the standard error. The bootstrap provides an alternative approach that requires additional computational effort. For additional discussion and references see Benichou.[1] Stratified and partially stratified attributable risks were also considered, which may be used either to accommodate the sampling design of the study, or to compute an attributable risk for a given exposure or covariate

category. The generalized AR allows the reference population to have a non-degenerate exposure distribution, which may be more realistic in many instances.

The examples show that the calculations can be easily performed using standard statistical software, so that the approach is practical. For both types of studies the agreement with previously reported results is quite good. Results using the bootstrap agree with those provided by the delta method, as might be expected with fairly large samples. For smaller sample sizes the bootstrap would be preferred. The computations do require some programming, and one must be careful to count the number of replicates correctly. This should be done using summary tables prepared from the data. It is helpful in practice to program separate estimates of intermediate quantities such as the total number of cases if this must be estimated from the model. As these values are positive integers it is very easy to check that the programming has been done correctly.

References

1 Benichou J. A review of adjusted estimators of attributable risk. *Statistical Methods in Medical Research* 2001; **10**: 195–216.

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

3 Benichou J. Reply to comment on: methods of adjustment for estimating the attributable risk in case–control studies: a review (1991; **10**: 1753–73). *Statistics in Medicine* 2001; **20**: 981–82.

4 Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine* 1982; **1**: 229–43.

5 Basu S, Landis JR. Model-based estimation of population attributable risk under cross-sectional sampling. *American Journal of Epidemiology* 1995; **142**: 1338–43.

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

7 Drescher K, Becher H. Estimating the generalized impact fraction from case–control data. *Biometrics* 1997; **53**: 1170–76.

8 Benichou J, Gail MH. A delta-method for implicitly defined random variables. *The American Statistician* 1989; **43**: 41–44.

9 Benichou J, Gail MH. Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. *Biometrics* 1990; **46**: 991–1003.

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

11 Agresti A. *Categorical data analysis*. Wiley, 1990: 451–52.

12 McCullagh P, Nelder JA. *Generalized linear models*, second edition. Chapman & Hall, 1989: 115–17 and 211–13.

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

14 Efron B, Tibshirani RJ. *An introduction to the bootstrap*. Chapman & Hall, 1993: 65, 113–14 and 50–53.

15 Breslow N, Day NE. *Statistical methods in cancer research Volume 1: the analysis of case–control studies*. International Agency for Research on Cancer, Scientific Publications No. 32; 1980.