STATISTICS IN MEDICINE, VOL. 10, 1753–1773 (1991)

METHODS OF ADJUSTMENT FOR ESTIMATING THE ATTRIBUTABLE RISK IN CASE-CONTROL STUDIES: A REVIEW

JACQUES BENICHOU*

Département de Biostatistique et Informatique Médicale, Hôpital Saint-Louis, 1 avenue Claude Vellefaux, 75475 Paris Cedex 10, France

SUMMARY

In the 1980's, progress was made in adjusting estimates of the attributable risk (AR) for confounding factors and in calculating associated confidence intervals. In this paper, methods of adjustment for estimation of the AR in case-control studies are reviewed. The limitations and problems associated with two methods based on stratification, the weighted-sum approach and the Mantel–Haenszel approach, are discussed. They include small-sample bias with the weighted-sum approach and the difficulty of taking interaction into account with the Mantel–Haenszel approach. A third method based on logistic regression is reviewed. It is argued that this latter method has the greatest generality and flexibility, and includes the two other approaches as special cases. Throughout the paper, an example of a case-control study of oesophageal cancer illustrates the use of the methods described.

1. INTRODUCTION

The attributable risk (AR) first proposed by Levin in 1953[1] is a widely used measure to assess the public health consequences of an association between an exposure factor and a disease. It is defined as:

$$\mathrm{AR} = \{P(D) - P(D|\bar{E})\}/P(D), \tag{1}$$

where $P(D)$ is the probability of disease in the population, which may have some exposed (E) and some unexposed ($\bar{E}$) individuals, and $P(D|\bar{E})$ is the hypothetical probability of disease in that same population but with all exposures eliminated. It is thought of as measuring the proportion of disease cases which can be related to the exposure[2] or the 'disease-producing role' of the exposure,[3] and thus is used to assess the potential impact of prevention programmes. The AR takes into account both the strength of the association between exposure and disease and the prevalence of exposure. Other terms have been used in the literature. These include population attributable risk,[4] population attributable risk per cent,[5] etiologic fraction or fraction of etiology[3] and attributable fraction.[6] It should not be mistaken for the less general attributable risk among the exposed[4,5] defined by replacing $P(D)$ by $P(D|E)$, the probability of disease among the exposed, in formula (1).

* Current address: National Cancer Institute, 6130 Executive Blvd., EPN/403, Rockville, Maryland 20892, U.S.A., where this work was carried out in part.

0277–6715/91/111753–21$10.50

Table I. Data from the case-control study of oesophageal cancer used for illustration (from Tuyns *et al.*[7])

Alcohol consumption (g/day)	Age (years)	Smoking (g/day)	Number of cases	Number of controls
0–39	25–44	0–9	0	100
		10–29	1	36
		30+	0	13
	45–64	0–9	1	45
		10–29	0	28
		30+	0	4
	65+	0–9	8	107
		10–29	14	47
		30+	5	6
40–79	25–44	0–9	0	62
		10–29	4	44
		30+	0	15
	45–64	0–9	6	32
		10–29	9	27
		30+	5	2
	65+	0–9	28	51
		10–29	19	44
		30+	4	3
80–119	25–44	0–9	0	13
		10–29	0	9
		30+	0	3
	45–64	0–9	3	13
		10–29	7	12
		30+	2	2
	65+	0–9	16	16
		10–29	18	19
		30+	5	0
120+	25–44	0–9	2	2
		10–29	3	6
		30+	0	2
	45–64	0–9	4	0
		10–29	5	2
		30+	4	0
	65+	0–9	10	6
		10–29	11	3
		30+	6	1

This paper reviews methods of adjustment for estimating the AR in case-control studies, namely methods of estimation that adjust for other factors than the exposure factor of interest. A lot of attention has been devoted to development of these methods in the recent statistical literature. Let us first consider, as a motivating example and for illustrative purposes, the case-control study on oesophageal cancer of Tuyns *et al.*[7] The data consist of 200 cases and 775 controls selected by simple random sampling. Like Breslow and Day[2] we restrict our attention to three factors: namely, alcohol consumption, smoking and age. The cross-classified data appear in Table I. Notice that while I retained the four levels of alcohol consumption (0–39, 40–79, 80–119

and 120+ g/day), I considered only three levels of smoking (0–9, 10–29 and 30+ g/day) and three 20-year age groups (25–44, 45–54 and 55+ years) for the sake of simplicity. The number of joint levels of the three factors is thus $4 \times 3 \times 3 = 36$.

Throughout the paper I will focus on the estimation of the AR for alcohol consumption (Problem 1) and for alcohol consumption and smoking (Problem 2). The simplest approach would be to only consider the exposure of interest, for example, alcohol consumption in Problem 1, and to ignore the other factors, smoking and age. One would then obtain an unadjusted or crude estimate of the AR. Methods to obtain unadjusted estimates are briefly reviewed in Section 2 and applied to the example. The importance of the definition of the baseline level of exposure is underlined. However, it is in general necessary to take into account other factors than just the exposure of interest and to adjust for them, as it is when estimating the relative risk. In Section 3, the conditions under which adjustment is necessary to obtain unbiased estimates of the AR are recalled. In Section 4, the two main approaches based on stratification to obtain adjusted estimates of the AR are described, namely the Mantel–Haenszel approach and the weighted-sum approach. They are then applied to the example. In Section 5, a more flexible approach based on regression models is described and it is argued that it includes the two stratificaton approaches as special cases. This is also illustrated with the example. In order to be comprehensive, material on variance estimation, is given in Appendix I for each approach including methods of adjustment.

2. UNADJUSTED ESTIMATION

When a single binary exposure factor X is considered and no adjustment is attempted, formula (1) may be rewritten as:

$$\mathrm{AR} = \frac{P(E)(\mathrm{RR} - 1)}{1 + P(E)(\mathrm{RR} - 1)}, \tag{2}$$

where RR denotes the relative risk. Upon making the usual rare disease assumption, $P(E)$ can be replaced by $P(E|\bar{D})$, the prevalence of exposure in the non-diseased population and RR can be replaced by the odds ratio OR. All quantities are therefore estimable from case-control data and a point estimate of AR is given by:

$$\widehat{\mathrm{AR}} = \frac{n_1 m_0 - m_1 n_0}{m_0 n}, \tag{3}$$

where n_0 and n_1 are, respectively, the numbers of unexposed and exposed cases ($n_0 + n_1 = n$) and m_0 and m_1 the numbers of unexposed and exposed controls ($m_0 + m_1 = m$). Notice that the equivalent formula:

$$\mathrm{AR} = P(E|D)\frac{\mathrm{RR} - 1}{\mathrm{RR}} \tag{4}$$

yields the exact same point estimate. Finally, when X is polychotomous with $I + 1$ levels ($i = 0$ for the unexposed, $i = 1, \ldots, I$ for the I levels of exposure), the point estimate is obtained by collapsing the I exposed levels into a single exposed category and using (3) as noted by Walter.[8]

An estimate of the variance of $\widehat{\mathrm{AR}}$ can be obtained first by noting that the quantities n_1 and m_1 have independent binomial distributions when n and m are fixed, and then by applying the delta-method.[9] One obtains:

$$\widehat{\mathrm{var}}(\widehat{\mathrm{AR}}) = mn_0(n_1 m_0 m + nn_0 m_1)/n^3 m_0^3. \tag{5}$$

The corresponding $100(1-\alpha)$ per cent confidence interval (CI) for AR is given by:

$$\widehat{AR} - z_{1-\alpha/2}\{\widehat{\text{var}}(\widehat{AR})\}^{1/2}, \quad \widehat{AR} + z_{1-\alpha/2}\{\widehat{\text{var}}(\widehat{AR})\}^{1/2}, \tag{6}$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution. Alternatively, given the asymptotic normality of $\log(1-\widehat{AR})$, Walter[10] suggested using the log-transform and obtained:

$$\widehat{\text{var}}\{\log(1-\widehat{AR})\} = n_1/nn_0 + m_1/mm_0. \tag{7}$$

The corresponding $100(1-\alpha)$ per cent CI for AR is then:

$$1-(1-\widehat{AR})\exp \text{EF}, \quad 1-(1-\widehat{AR})/\exp \text{EF}, \tag{8}$$

where the error factor EF is equal to $z_{1-\alpha/2}[\widehat{\text{var}}\{\log(1-\widehat{AR})\}]^{1/2}$. Whittemore[11] noted that this interval is wider than the interval in (6) for $\widehat{AR} > 0$. Finally, Leung and Kupper[12] suggested using the logit transform to obtain:

$$\widehat{\text{var}}\{\text{logit}(\widehat{AR})\} = (n_1/nn_0 + m_1/mm_0)\{m_0 m/(n_1 m - m_1 n)\}^2. \tag{9}$$

The corresponding $100(1-\alpha)$ per cent CI for AR is:

$$1+\{(1-\widehat{AR})/\widehat{AR}\}/\exp \text{EF}, \quad 1-\{(1-\widehat{AR})/\widehat{AR}\}/\exp \text{EF}, \tag{10}$$

where EF is now equal to $z_{1-\alpha/2}[\widehat{\text{var}}\{\text{logit}(\widehat{AR})\}]^{1/2}$. They showed that the logit CI is narrower than that given in (6) for $0{\cdot}21 < \widehat{AR} < 0{\cdot}79$ whereas the reverse holds outside this range. They provided simulation results showing that the coverage probability is close to nominal even for samples as small as 50 cases and 50 controls. However, no comparison of the three CIs (6), (8) and (10) has been carried out.

These results are now applied to the case-control data on oesophageal cancer. Our purpose is first to estimate the AR for alcohol consumption (Problem 1), a factor known to strongly influence the risk of oesophageal cancer. Since there are four levels of alcohol consumption, there are several ways of defining the baseline level. We first define it by an alcohol consumption of 0–79 g/day and this will be termed Problem 1*a* throughout the rest of the paper. The odds ratio (OR) for alcohol consumption is estimated at 5·64. The AR is estimated at 0·395 with a standard deviaton (SD) estimated at 0·0420 (Table II) and a corresponding 95 per cent CI based on (6) of [0·313, 0·477]. The interpretation is that 39·5 per cent of all oesophageal cancers can be attributed to alcohol consumption and could potentially be prevented by reducing alcohol consumption to less than 80 g/day.

An alternative way is to use a more restrictive and perhaps more sensible definition of the baseline level, namely 0–39 g/day. This will be termed Problem 1*b* throughout the rest of the paper. With this definition, we obtain a slightly higher value of $\widehat{OR}$, namely 5·85 and a much higher, almost doubled, value of $\widehat{AR}$, namely 0·709, ($\widehat{SD} = 0{\cdot}0511$, Table II). The corresponding 95 per cent CI based on (6) is [0·609, 0·809]. The interpretation is now that 70·9 per cent of all oesophageal cancers can be attributed to alcohol consumption and could potentially be prevented by reducing alcohol consumption to less than 40 g/day. The steep increase in the value of $\widehat{AR}$ is due to the more restrictive definition of the baseline level, resulting in an increase in the proportion of exposed. This shows the crucial importance of the definition of the baseline when estimating the AR. Notice, however, that the prevention interpretation of the two figures is different. In this example, the influence of the baseline definition on the AR estimate is more important than on the OR estimate. As a practical consequence, $\widehat{AR}$ values can only be appropriately interpreted relative to the definition of the baseline. Of course, this problem is not specific to the unadjusted approach and is encountered with all methods of estimation as will be seen further in this paper.

Table II. Attributable risk of alcohol consumption for two choices of the baseline (Problems 1*a* and 1*b*), and of alcohol consumption and smoking (Problem 2) in oesophageal cancer. Comparison of three methods of estimation

	Problem 1*a* $\widehat{AR}(\widehat{SD})$	Problem 1*b* $\widehat{AR}(\widehat{SD})$	Problem 2 $\widehat{AR}(\widehat{SD})$
Unadjusted approach	0·395 (0·0420)	0·709 (0·0511)	0·862 (0·0456)
Mantel–Haenszel approach	0·380 (0·0441)	0·716 (0·0508)	0·869 (0·0436)
Weighted-sum approach	0·380	0·700	0·868
	0·379 (0·0445)	0·682 (0·0579)	0·861 (0·0452)

In Problem 1*a*, the baseline level of alcohol consumption is 0–79 g/day, while it is 0–39 g/day in Problem 1*b*. In Problem 1a and 1*b*, the Mantel–Haenszel and weighted-sum estimates are adjusted for smoking and age (both three levels). In Problem 2, the baseline level of alcohol consumption × smoking is 0–39 g/day × 0–9 g/day. The Mantel–Haenszel and weighted-sum estimates are adjusted for age (three levels). For the weighted-sum approach, the first value of $\widehat{AR}$ is obtained from the original data while the second value is obtained after assigning the value 0·5 to all zero cells

Finally, the AR is estimated for alcohol consumption and smoking (Problem 2). We retain the same definition of exposure for alcohol consumption as in Problem 1*b* and, for smoking, the baseline is defined by level 1 (0–9 g/day). Therefore, I define the joint baseline level for alcohol consumption and smoking by joint levels of 0–39 g/day and 0–9 g/day, respectively. As could be anticipated from the more restrictive definition of the baseline compared to Problem 1, a higher value of $\widehat{OR}$ is obtained, namely 10·23 and a higher value of $\widehat{AR}$, namely 0·862 ($\widehat{SD}$ = 0·0456, Table II). The corresponding 95 per cent CI is [0·773, 0·951]. The interpretation is that while 70·9 per cent of all oesophageal cancers can be attributed to alcohol consumption alone, 86·2 per cent can be attributed to alcohol consumption and smoking. Computationally, it is useful to notice that by restricting the definition of the baseline, by going from Problem 1*a* to Problem 2, both $\widehat{OR}$ and the proportion of exposed increase. As a result, so does $\widehat{AR}$ since from (2) $\widehat{AR}$ is an increasing function of both $\widehat{RR}$ (when $\widehat{RR} > 1$) and $\hat{P}(E)$.

3. THE IMPORTANCE OF ADJUSTMENT

When estimating the RR for a given exposure factor, it is common practice to adjust for stratification or matching factors that have been used in the design of the study and for other confounders or secondary exposure factors. Otherwise, one obtains crude estimates of the RR that are likely to be biased. The need for adjustment also arises when estimating the AR. Crude estimates are in general biased. This problem has been studied byMiettinen,[3] Walter[8,13,14] and Whittemore.[11,15] If we assume X_1 and X_2 are two binary factors and one is interested in estimating the AR for X_1, then Walter[13] showed that the crude AR estimate that is ignoring X_2, is unbiased if and only if at least one of the two conditions (*a*) and (*b*) is true:

(*a*) X_1 and X_2 are independently distributed in the population, that is:

$$P(X_1 = 0, X_2 = 0)\,P(X_1 = 1, X_2 = 1) = P(X_1 = 0, X_2 = 1)\,P(X_1 = 1, X_2 = 0),$$

where level 0 denotes the absence of exposure.

(*b*) Exposure to X_2 alone does not increase disease risk, that is:

$$P(D|X_1 = 0, X_2 = 1) = P(D|X_1 = 0, X_2 = 0).$$

It appears therefore that, if X_2 truly confounds the associaton between X_1 and the disease, then the crude estimate of the AR is biased, as is the unadjusted OR. If neither (*a*) nor (*b*) is true, the direction of the bias can be determined. If X_2 alone increases risk, then the bias is positive leading to an overestimation of AR, if X_1 and X_2 are positively correlated, and negative if their correlation is negative. When considering several factors X_j ($j = 2, \ldots, J$), conditions (*a*) and (*b*) can be extended to a set of $2(J-1)$ analogous sufficient conditions concerning factors X_1 and X_j ($j = 2, \ldots, J$) as shown by Walter.[13]

4. ADJUSTED METHODS OF ESTIMATION BASED ON STRATIFICATION

Walter[8] was the first to propose an adjustment for estimating the AR by decomposing it into the exposure effect and the confounding effect. However, two more general approaches based on stratification are now available, the Mantel–Haenszel approach and the weighted-sum approach.

4.1. The Mantel–Haenszel approach

This approach has been developed by Kuritz and Landis[16,17] and Greenland.[18] It allows adjustment for one or more polychotomous factors forming J levels or strata. For instance, smoking and age can be adjusted for in Problem 1 of the example and they form $J = 3 \times 3 = 9$ levels. It is based on the use of a common adjusted OR for all J levels of adjustment, namely the Mantel–Haenszel OR.[19] From formula (4), one obtains the following estimate:

$$\widehat{AR} = \hat{P}(E|D)\frac{\widehat{OR}_{MH} - 1}{\widehat{OR}_{MH}} \tag{11}$$

where $\hat{P}(E|D) = n_{1.}/n$, the ratio of the number of exposed cases ($n_{1.}$) to the total number of cases (n), is obtained from the cases and $\widehat{OR}_{MH}$ is the Mantel–Haenszel OR estimate. Instead of using OR_{MH}, other quantities have been proposed such as an internally standardized mortality ratio[3] or the MLE from logistic regression,[18] and the approach could be termed instead, the common OR approach. The rationale for the use of OR_{MH} is its lack of (or very small) bias even for sparse data, and its good efficiency except in extreme circumstances.[2,20–22] The crucial assumption of the Mantel–Haenszel approach is that of a common or homogeneous OR which amounts to the absence of interaction between the adjustment factor(s) and the exposure factor. The point estimate in (11) is simple to obtain but variance estimators are more complex. They have been developed by Kuritz and Landis[16,17] and by Greenland[18] and are described in Appendix I.1.

Finite sample properties of these asymptotic point and variance estimators have been studied by simulation. Kuritz and Landis[16] found that the bias in the AR was negligible for simple random sampling and for individual matching under the assumption of homogeneity of the odds ratio. Furthermore, the variance is also consistently estimated and the coverage of the CIs is close to nominal for both sampling schemes. Greenland's limited simulation results concerned stratified random sampling and displayed the same favourable properties as results by Kuritz and Landis.

In the example, smoking and age both strongly influence the risk of oesophageal cancer and could therefore potentially confound the relationship between alcohol consumption and oesophageal cancer. In Problem 1, the Mantel–Haenszel approach allows adjustment for both smoking and age. We formed $J = 9$ strata of smoking $\times$ age and obtained the following results. In Problem 1*a*, the OR for alcohol consumption was estimated at 4·79, a smaller value as a result of adjustment, and a corresponding smaller value of $\widehat{AR}$, namely 0·380, was obtained. Its SD

estimate slightly increased to 0·0441 since more factors were taken into account (Table II). The 95 per cent CI, based on (6), was [0·294, 0·466]. In Problem 1*b*, there was an increase in $\widehat{OR}$ to 6·17 as a result of adjustment and a corresponding slight increase in $\widehat{AR}$ to 0·716 (Table II). A slight decrease in the SD estimate of $\widehat{AR}$ was observed due to the fact that $1 - \widehat{AR}$ decreased but it can be checked that the relative $\widehat{SD}$ given by $\widehat{SD}(1 - \widehat{AR})/(1 - \widehat{AR})$ still increased. The 95 per cent CI based on (6) was [0·616, 0·816].

In Problem 2 we formed $J = 3$ age strata and obtained estimates adjusted for age. $\widehat{OR}$ increased to 11·09 as a result of adjustment and a corresponding small increase in $\widehat{AR}$ to 0·869 was obtained with a slight decrease in $\widehat{SD}$ to 0·0436 as in Problem 1*b* (Table II). The 95 per cent CI based on (6) for the AR for alcohol and smoking was [0·783, 0·954]. These results show the effect of adjustment for age and smoking (Problem 1) and for age alone (Problem 2). In this example, adjustment resulted in a maximum relative change in $\widehat{OR}$ of 15 per cent (Problem 1*a*), while the effect on $\widehat{AR}$ was more modest, with a maximum relative change of only 4 per cent.

The Mantel–Haenszel approach allows for adjustment while avoiding small-sample bias in the AR which can arise with the weighted-sum approach (see below). This is true even for sparse data and it can be used for individual matching, but the critical assumption of a common OR can be misleading if there is an interaction between the exposure factor and the adjustment factor(s), and can in turn cause bias. A general discussion of the AR and interaction is beyond the scope of this paper and elements can be found in Miettinen,[3] Rothman[23] and Walter.[13,14] For our purpose, it must be kept in mind that interaction has to be looked for when using the Mantel–Haenszel approach even though interaction tests have usually little power. If there is evidence of interaction, Greenland[18] suggested using a modified estimator based on the following hybrid OR. If it is possible to define H levels out of the J levels of adjustment ($H < J$) such that they each have a homogeneous OR, then the hybrid OR estimate is given by:

$$\widehat{OR} = \sum_{h=1}^{H} n_{1h} \Big/ \sum_{h=1}^{H} e_h, \tag{12}$$

in which $e_h = n_{0h}m_{1h}/m_{0h}$, and the adjusted hybrid AR estimate is given by:

$$\widehat{AR} = \sum_{h=1}^{H} (n_{1h} - e_h)/n = \sum_{h=1}^{H} \widehat{AR}_h. \tag{13}$$

A variance estimator is given in Appendix I.1

This constitutes a possible solution to take care of interaction in the framework of the Mantel–Haenszel approach. However, the choice of the H levels, which is critical for this approach, is somewhat arbitrary, particularly in view of the low power of tests to detect interaction. Moreover, even though the hybrid estimate is asymptotically unbiased given this correct choice, finite sample properties of the proposed estimators have not been studied. In particular, as H increases, bias might arise as in the weighted-sum approach with which a comparison would be useful.

4.2. The Weighted-sum approach

This approach suggested by Walter[8] and mainly developed by Whittemore[11,15] allows adjustment for one or more polychotomous factors forming J levels or strata. The AR is written as a weighted sum of the ARs over strata, namely:

$$AR = \sum_{j=1}^{J} w_j AR_j, \tag{14}$$

where AR_j is the AR specific to level j, and w_j, the weight corresponding to level j, is the proportion of cases in level j. This choice of weighting is termed the 'case-load method'. The corresponding estimator can be derived as a maximum-likelihood estimator and is thus asymptotically unbiased.[10] An alternative choice of weighting will be discussed below.

Therefore, the ARs are estimated separately within each level of the adjustment factors and the adjusted estimate is the weighted-sum of AR estimates across the levels of adjustment. Unlike the Mantel–Haenszel approach, no assumption of a common OR is made. The odds ratio is free to vary across strata which correspond to a saturated model for the odds ratio. Thus, the weighted-sum approach not only allows one to control for confounding but also for interaction or effect-modification. An extension of previous notation to incorporate subscript j for levels of the adjustment factors yields the adjusted AR:[10]

$$\widehat{AR} = 1 - \sum_{j=1}^{J} m_{.j} n_{0j} / n m_{0j}, \tag{15}$$

where $m_{.j}$ is the number of controls in level j of adjustment. This estimate is valid in case of simple random sampling of the controls to adjust for one or several confounders or secondary exposure factors, and in case of stratified random sampling or frequency matching of the controls to adjust for the stratificaton factors. Variance estimators have been developed by Whittemore[11] and are given in Appendix I.2.

Even though the Mantel–Haenszel approach and the weighted-sum approach seem totally different, they can be reconciled. Indeed, the weighted-sum approach could be used with the assumption of a common odds ratio, notwithstanding the fact that it was not originally restricted in this way. In Appendix II.1, we show that, under this assumption, the two approaches yield the same expression for the AR and the same AR estimates as long as the common odds ratio is estimated the same way. Therefore, the Mantel–Haenszel approach can be regarded as a special case of the weighted-sum approach.

Finite sample properties of point and variance estimators and coverage probabilities of CIs have been studied by simulation for frequency-matching and simple random sampling of the controls.[11,16] There is a negative bias in AR which becomes substantial with sparse data and a high probability of exposure. In an example with $n = m = 100$, $J = 10$ and the same 80 per cent probability of exposure for cases and controls in all levels j, a situation with no confounding and $AR = 0$, the mean value of $\widehat{AR}$ was $-0{\cdot}486$ falsely indicating a protective effect of exposure.[11] As noted by Whittemore,[11] this bias is not surprising in view of the inconsistency of the maximum likelihood estimator (MLE) for the common OR in several 2×2 tables using, for example, unconditional logistic regression, as the number of tables increases while each table remains sparse.[20] This feature makes the method inappropriate for individual $1:k$ matching because each level j, corresponding to a matched set, has only one case and k controls, and consequently very severe bias is to be anticipated.[11] Variance estimates tend to be too large, leading to conservative CIs whether a transform is used or not. Therefore, while the weighted-sum approach allows one to control both for confounding and interaction, small-sample bias can be substantial and renders it impractical for individual matching.

Ejigou[24] has suggested using a different weighting given by $w_j = \{\widehat{\text{var}}(\widehat{AR}_j)\}^{-1}$. This 'precision-based method' can be shown to be inconsistent. In simulations the same problems of bias for point and variance estimates as with 'case-loading' occurred, the main difference being that the bias in AR, though very close in absolute value to the bias with 'case-loading', was positive.[16]

In the example, we applied the weighted-sum approach with case-load weighting. In Problem 1, we formed $J = 9$ strata of smoking × age as with the Mantel–Haenszel approach and adjusted for

smoking and age. In Problem 1*a*, we obtained $\widehat{AR} = 0{\cdot}380$ as with the Mantel–Haenszel approach with an $\widehat{SD}$ of 0·0445 (Table II). The corresponding 95 per cent CI, based on (6), for the AR of alcohol consumption was [0·293, 0·467]. The fact that no assumption of a common odds ratio was made did not lead to a change in $\widehat{AR}$ which suggests an absence of detectable interaction between alcohol consumption on the one hand and smoking and age on the other hand. Notice that two values of $\widehat{AR}$ appear in Table II. The first one corresponds to the original data for which no variance can be estimated since division by zero occurs (see Appendix I.2). In order to avoid this problem, values 0·5 were assigned to all zero cells as suggested by Whittemore[11] and the second value of $\widehat{AR}$ and $\widehat{SD}$ correspond to these modified data . Notice that SD is slightly higher than with the Mantel–Haenszel approach which should be expected from the extra-variability in taking interaction into account.

In Problem 1*b*, we obtained $\widehat{AR} = 0{\cdot}700$ with SD = 0·0579 (Table II). The corresponding 95 per cent CI based on (6) was [0·587, 0·813]. The decrease in $\widehat{AR}$ compared to the Mantel–Haenszel approach reflected an interaction between alcohol consumption as defined in Problem 1*b*, and smoking and age. There was a marked increase in SD reflecting the extra-variability due to interaction. The second value of $\widehat{AR}$ based on the modified data as described above was markedly smaller than 0·700 and a modified 95 per cent CI was [0·571, 0·797]. Finally, in Problem 2, the AR estimate for smoking and alcohol consumption was 0·868, SD = 0·0452, almost identical to that with the Mantel–Haenszel approach. The 95 per cent CI based on (6) was [0·779, 0·957] or [0·772, 0·950] using the second value of $\widehat{AR}$. These results show the effect of allowing for interaction of alcohol consumption with smoking and age (Problem 1), and for interaction of age with alcohol consumption and smoking (Problem 2). The effect on $\widehat{AR}$ was modest since the maximum relative change compared to the Mantel–Haenszel approach was 2 per cent (Problem 1*b*).

5. ADJUSTED METHODS OF ESTIMATION BASED ON REGRESSION MODELS

The methods described in the previous section have several limitations. The weighted-sum approach leads to bias in estimating the AR. This bias can be quite substantial in some instances, particularly when the data are sparse, and the method is thus inappropriate for individual matching. The Mantel–Haenszel approach does not apply when the odds ratio is not homogeneous across adjustment levels and the procedure proposed by Greenland[18] to take interaction into account, while being somewhat data-dependent in its implementation, might in turn lead to some bias. Furthermore, when sampling of the controls has been stratified or frequency matched, current results on variance estimators for both approaches only allow one to take into account stratification factors but no other factors such as other exposure or confounding factors. For individual matching, the Mantel–Haenszel approach is the only available approach but does not offer the possibility to take into account other factors besides the matching factors.

Regression methods have become an essential tool for modelling relative risk because of their advantages over stratification methods. Indeed, regression models allow more confounding and effect modifying factors to be taken into account. Regression modelling may only require that confounders enter additively, reducing the dimensionality of the problem. Interactions with exposure can easily be accommodated by introducing interaction terms in the regression model. Finally, regression methods represent a flexible and unified approach to parameter estimation and hypothesis testing, and MLE which arise from regression models have favourable asymptotic properties. In particular, the importance of interaction terms can be evaluated before choosing a final model. All these advantages can potentially apply to AR estimation and methods of estimation of the AR based on regression models have been developed.

Walter[8] was the first to suggest using regression models to estimate the AR while Sturmans *et al.*[25] noticed that the logistic model was the most natural one to use in the context of case-control studies. Greenland[18] proposed a modification of the Mantel–Haenszel approach consisting of replacing $\widehat{OR}_{MH}$ by the more efficient MLE from conditional logistic regression in formula (11). His approach to variance estimation described in Appendix I.1 remains valid and the limitations remain the same as with the Mantel–Haenszel approach. However, the full generality and flexibility of the regression approach has best been exploited by Bruzzi *et al.*[26] The principle is not to model the AR directly, but to incorporate adjusted RR estimates based on regression models in the AR estimate. Bruzzi *et al.*[26] started by noticing, as did Miettinen,[3] that for one exposure factor X with $I + 1$ levels ($X = x_0, x_1, \ldots, x_I$), the AR can be written as:

$$AR = 1 - \sum_{i=0}^{I} \rho_i RR_i^{-1}, \tag{16}$$

where $\rho_i = P(X = x_i|D)$ and RR_i is the RR associated with level i of exposure. If X is binary, this is equivalent to formulae (2) and (4). Furthermore, one obtains the estimate (3) by treating X as a binary factor. Although an informal proof of (16) is given by Bruzzi *et al.*,[26] a formal proof can be obtained from conditional probability arguments (Appendix II.2). If one or several other factors Y forming J levels ($Y = y_1, \ldots, y_J$) are considered, then the adjusted AR is given by:

$$AR = 1 - \sum_{j=1}^{J} \sum_{i=0}^{I} \rho_{ij} RR_{i|j}^{-1}, \tag{17}$$

where now $\rho_{ij} = P(X = x_i, Y = y_j|D)$ and $RR_{i|j}$ is the RR associated with level i of exposure for level j of Y, that is

$$RR_{i|j} = P(D|X = x_i, Y = y_i)/P(D|X = x_0, Y = y_j).$$

A proof is given in Appendix II.2. In the most general form, sampling of the controls may be stratified or frequency matched on one or several factors Z forming K levels. The AR adjusted for the stratification factor(s) Z and other secondary exposure or confounding factors Y is then given by:[27]

$$AR = 1 - \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=0}^{I} \rho_{ijk} RR_{i|j,k}^{-1}, \tag{18}$$

where now

$$RR_{i|j,k} = P(D|X = x_i, Y = y_j, Z = z_k)/P(D|X = x_0, Y = y_j, Z = z_k)$$

and

$$\rho_{ijk} = P(X = x_i, Y = y_j, Z = z_k|D).$$

Formulations (17) and (18) allow one to take into account one or several stratification factors Z and one or several other confounding or secondary exposure factors Y. The quantities ρ_{ij} or ρ_{ijk} can be estimated from the observed distribution of the cases. The RR can be estimated by using the OR estimated through logistic regression. Therefore, an adjusted AR estimate can be obtained from case-control data. From formula (18), it is given in its most general form by:

$$\widehat{AR} = 1 - \sum_{i=1}^{K} \sum_{j=1}^{J} \sum_{i=0}^{I} \hat{\rho}_{ijk} RR_{i|j,k}^{-1}(\hat{\theta}), \tag{19}$$

where $\hat{\rho}_{ijk} = n_{ijk}/n$ and θ is the vector of parameters from logistic regression. Therefore a logistic model can be fitted with factors X, Y and Z in order to estimate θ as is usually done in case-

Table III. Comparison of models for the attributable risk of alcohol consumption in oesophageal cancer with adjustment for smoking and age (Problem 1*a*)

Model	log (OR)	$-2l$	$\widehat{AR}(\widehat{SD})$
1	$\alpha + \beta Al$	893·06	0·395 (0·0420)
2	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al$	756·64	0·382 (0·4440)
3	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot Ag$	756·36	0·380 (0·0442)
4	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot S$	755·43	0·381 (0·0444)
5	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot (S \cdot Ag)$	750·28	0·380 (0·0440)

In all five models, alcohol consumption (Al) was considered a binary (0–79; 80 + g/day) factor, and nine parameters were used for the main effects of age (Ag), smoking (S) and their interaction (Ag · S) in models 2–5. In model 5, eight parameters were used to model the interactions of alcohol consumption with smoking and age

control studies. Then, an AR estimate is obtained from formula (19). As mentioned before, this allows great generality and flexibility since hypotheses can be tested and a model can be selected accordingly. In particular, interaction terms can be introduced in the model, tested and retained or not, depending on the result of the test. If there is no interaction between X and factors Y and Z, the triple sum in (19) collapses into a single sum in i, since then: $RR_{i|j,k}(\hat{\theta}) = RR_i(\hat{\theta})$. For simple random sampling, stratified random sampling or frequency matching of the controls, unconditional logistic regression can be used, but if the data get too sparse, conditional logistic regression[28] should be used. For individual matching, conditional regression should be used and the AR estimate can be obtained from formula (17). The matching factors appear in the AR formula only if they have an interaction with X or Y.

Variance estimates have been worked out by Benichou and Gail.[27] They are complex because they involve not only the variance of quantities $\hat{p}_{ijk}$ and of OR estimates from logistic regression, but also covariance terms between the two. A brief outline is given in Appendix I.3.

Simulations have been performed to study the finite sample properties of the AR estimators based on logistic regression.[27] They show little or no bias in the AR for most situations. However, when the data get sparse and when unconditional logistic regression is used to estimate the OR, bias toward zero appears. In one example with frequency-matching on a three-level stratification factor and with 100 cases and 100 controls, the mean AR estimate for a four-level exposure factor was 0·178, whereas the true value of the AR was 0·213.[27] Even though the bias in the AR is small, this indicates that in such situations it would be preferable to use either simpler models for the OR in order to decrease the number of estimated parameters or conditional logistic regression.[28] However, with conditional logistic regression, the approach outlined in Appendix I.3 to obtain variance estimates carries through only for individual matching, and extensions are needed for other sampling schemes.[27] Variance estimates for $\widehat{AR}$ (Appendix I.1) are consistent even for small sample sizes in the four situations studied by simulations,[27] namely for simple random sampling, stratified random sampling and frequency matching of the controls with unconditional logistic regression and for individual matching with conditional logistic regression. The resulting 95 per cent CIs have coverages very close to the nominal 95 per cent level even for small sample sizes in the same four situations,[27] whether they are based on formulae (6), (8) or (10).

We applied this approach to the example. All AR estimates were based on formula (17) since there was no stratification in sampling the controls. We used unconditional logistic regression to obtain OR estimates. Table III compares five models addressing Problem 1*a*. In all five models, alcohol consumption was considered a binary factor with baseline level defined as before in

Problem 1*a*, namely 0–79 g/day. Therefore, only one parameter β was used to model the main effect of alcohol consumption. In model 1, only alcohol consumption was taken into account and it had a significant effect on the risk of oesophageal cancer ($P < 0{\cdot}001$, likelihood-ratio (LR) test). The unadjusted MLE, $\widehat{OR} = \exp \hat{\beta}$, obtained from logistic regression, is equal to $n_1 m_0 / n_0 m_1$, using previous notation, and therefore was identical to the unadjusted $\widehat{OR} = 5{\cdot}64$ in Section 2. Furthermore, as mentioned above and since alcohol consumption is binary, formula (16) is equivalent to formula (2) and (4) and yields the same estimate as in formula (3), namely 0·395, as in Section 2. The $\widehat{SD}$ also remains unchanged and so do CIs. Saturated age and smoking effects were introduced in logistic model 2, eight binary variables were used since age and smoking form nine joint levels, and it resulted in a significant increase in the log-likelihood ($P < 0{\cdot}001$, LR test). Since no interaction terms were used, a common OR for all nine levels of adjustment was assumed, as in the Mantel–Haenszel approach. The adjusted MLE, $\widehat{OR} = \exp \hat{\beta}$, was equal to 4·88, a slightly different value from the Mantel–Haenszel $\widehat{OR}$, which is not an MLE. This slight increase in the $\widehat{OR}$ resulted in a slight increase in the adjusted $\widehat{AR}$ to 0·382 compared to 0·380 for the Mantel–Haenszel $\widehat{AR}$. This difference arose solely from the use of different OR estimates, namely the unconditional logistic maximum likelihood on the one hand and the Mantel–Haenszel estimate on the other hand. The $\widehat{SD}$ was virtually unchanged compared to the Mantel–Haenszel approach and was higher than in the unadjusted case since more parameters were estimated.

The next step was to introduce interaction terms between alcohol consumption and smoking and age in the model. Model 3 has two interaction terms between alcohol consumption and age, while model 4 has two interaction terms between alcohol consumption and smoking. There was no significant increase in the log-likelihood from model 2 to model 3 ($P > 0{\cdot}90$, LR test) nor from model 2 to model 4 ($P > 0{\cdot}50$, LR test) and consequently the changes in $\widehat{AR}$ were marginal. Model 5 has eight interaction terms between alcohol and smoking and age and is therefore a fully saturated model in alcohol consumption, age and smoking. There is no significant increase in the log-likelihood from model 2 to model 5 and, as a result, the change in $\widehat{AR}$ is minor. In Problem 1*a*, there seems to be little interest in going beyond model 2 since there is no detectable interaction between alcohol consumption and either age or smoking. An interesting feature of model 5 is that it allows the OR to vary freely across the nine levels of adjustment like the weighted-sum approach. For levels of adjustment with $n_{\cdot j} \neq 0$, the contribution to $1 - \widehat{AR}$ is, from formula (17), $n_{0j}/n + n_{1j}/n\widehat{RR}_{1|j}^{-1}$, which is equal to $m_{\cdot j} n_{0j} / n m_{0j}$ as in formula (15). This is because, model 5 being saturated, $\widehat{RR}_{1|j}^{-1}$ is equal to $n_{0j} m_{1j} / n_{1j} m_{0j}$ provided $n_{1j} m_{0j} \neq 0$. As a result, model 5 yields the same value of $\widehat{AR}$ as the weighted-sum approach, namely 0·380. The $\widehat{SD}$ is slightly different because, for the weighted-sum approach, it is based on modified data as described in Section 4, while, for the regression approach, it is based on the original data.

Table IV gives results concerning Problem 1*b*. The first five models (6–10) are analogous to models 1–5 in Table III, the only difference being the definition of baseline of alcohol consumption. Therefore, the unadjusted $\widehat{OR}$ from model 6 was equal to 0·709, as in Section 2, and the values of the unadjusted $\widehat{AR}$ and its $\widehat{SD}$ were also unchanged compared to Section 2. Alcohol consumption had a significant effect on the risk of oesophageal cancer ($P < 0{\cdot}001$, LR test). Model 7 resulted in a significant increase in the log-likelihood ($P < 0{\cdot}001$, LR test) and yielded $\widehat{OR} = 6{\cdot}30$, slightly higher than the Mantel–Haenszel $\widehat{OR}$ (6·17). The corresponding adjusted $\widehat{AR}$ (0·719) was also slightly higher. Even though more parameters were estimated than in model 6, the value of $\widehat{SD}$ was slightly lower. As discussed in Section 4 for the Mantel–Haenszel approach this was due to the fact that $1 - \widehat{AR}$ decreased but it can be checked that the relative $\widehat{SD}$ given by $\widehat{SD}(1 - \widehat{AR})/(1 - \widehat{AR})$ still increased. Models 8, 9 and 10 were analogous to models 3, 4 and 5, respectively. Interaction between alcohol consumption and age (model 8), smoking (model 9) and both smoking and age (model 10) were significant, respectively, at the 5 per cent, 10 per cent and

Table IV. Comparison of models for the attributable risk of alcohol consumption in oesophageal cancer with adjustment for smoking and age (Problem 1*b*)

Model	log(OR)	$-2l$	$\widehat{AR}(\widehat{SD})$
6	$\alpha + \beta Al$	899·44	0·709 (0·0511)
7	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al$	743·39	0·719 (0·0504)
8	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot Ag$	737·33	0·723 (0·0502)
9	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot S$	738·48	0·703 (0·0544)
10	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot (S \cdot Ag)$	727·53	0·700 (0·0559)
11	$\alpha + \beta Al$	842·99	0·709 (0·0511)
12	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al$	702·98	0·721 (0·0500)
13	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot Ag$	689·08	0·726 (0·0496)
14	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot S$	697·58	0·703 (0·0545)
15	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma Al \cdot (S \cdot Ag)$	665·22	0·701

In models 6–10, alcohol consumption (Al) was considered a binary (0–39 g, 40 + g/day) factor, and nine parameters were used for the main effects of age (Ag), smoking (S) and their interaction (Ag · S) in models 7–10. In model 10, eight parameters were used to model the interactions of alcohol consumption with smoking and age.
In models 11–15, alcohol consumption (Al) was considered a polychotomous factor with four levels (0–39, 40–79, 80–119, 120 + g/day), and nine parameters were used to model the main effects of age, smoking and their interaction in models 12–15. In model 15, 24 parameters were used to model the interaction of alcohol consumption with smoking and age

5 per cent levels. While interaction with age alone brought $\widehat{AR}$ up to 0·723, interaction with smoking brought it down to 0·703 and interaction with both age and smoking brought it down to 0·700. Again, the value of $\widehat{AR}$ with model 10 was identical to that with the weighted-sum approach while the $\widehat{SD}$ was different for reasons already discussed. In Problem 1*b*, it seems worthwhile to go to the fully saturated model 10 since both age and smoking seem to act like effect modifiers.

In the bottom part of Table IV, five additional model (11–15) were considered to address Problem 1*b*. They were analogous to models 6–10 except that the four levels of alcohol consumption were retained with an unchanged baseline level (0–39 g/day). Three terms were therefore used to model the main effect of alcohol consumption. From model 11, the unadjusted $\widehat{AR}$ (and its $\widehat{SD}$) was identical to that in model 6. As shown in Appendix II.4, this is a general result for a saturated model. In models 12–14 which are not saturated models, we obtained values of $\widehat{AR}$ very close to those from models 7–10. Finally, in model 15, which is a saturated model, we obtained $\widehat{AR} = 0{\cdot}701$ as compared to $\widehat{AR} = 0{\cdot}700$ with model 10. Arguments identical to those in Appendix II.4 show that the two values of $\widehat{AR}$ should be identical since models 10 and 15 are both saturated in alcohol consumption, smoking and age. The difference comes from the fact that the AR estimate from model 15 is less precise than that from model 10. Indeed, 36 parameters were estimated in model 15, which made estimates unstable, as further attested by the fact that the observed information matrix at the MLE was impossible to invert, so that no estimate of SD could be produced. Therefore, theoretical results (Appendix II.4) show and the example illustrates that there is no need to use more than two exposure levels to estimate the AR and, if saturated models are used, the AR estimates are identical whether two or more exposure levels are used. This could also be illustrated for Problem 1*a* in which the three levels of alcohol consumption can be reduced to two levels, and for Problem 2 in which the 12 levels of alcohol consumption × smoking can be reduced to two levels.

Table V. Comparison of models for the attributable risk of alcohol consumption in oesophageal cancer with adjustment for age (Problem 2)

Model	log (OR)	$-2l$	$\widehat{AR}$ ($\widehat{SD}$)
16	$\alpha + \beta X$	907·63	0·862 (0·0457)
17	$\alpha Ag + \beta X$	793·81	0·866 (0·0445)
18	$\alpha Ag + \beta X + \gamma Ag \cdot X$	792·58	0·868 (0·0440)

In all three models the main effect of alcohol consumption and smoking was modelled with one parameter corresponding to one binary variable X with baseline defined by 0–39 g/day and 0–9 g/day of alcohol consumption and smoking, respectively. Three parameters were used for the main effect of age (Ag) in models 17–18. In model 18, two parameters were used to model the interaction between X and Ag

Finally, Table V gives the results concerning Problem 2. The aim was to estimate the AR for alcohol consumption and smoking. One parameter was used to model the main effect of the exposure of interest, namely exposure to alcohol or smoking. The baseline was defined as 0–39 g/day of alcohol consumption and 0–9 g/day of smoking. The effect of alcohol consumption and smoking was significant ($P < 0{\cdot}001$, LR test). The unadjusted $\widehat{OR}$ was 10·23 and the unadjusted $\widehat{AR}$ (model 16) was 0·862, $\widehat{SD} = 0{\cdot}0457$, as in Section 2. When adjusting for age with three levels, the increase in the log-likelihood was significant ($P < 0{\cdot}001$, LR test). The $\widehat{OR}$ increased to 10·73 and the adjusted $\widehat{AR}$ (model 17) was 0·866, $\widehat{SD} = 0{\cdot}0445$, very close to the Mantel–Haenszel $\widehat{AR}$ in Section 4. No intermediate models for interaction were considered since the main effects of alcohol consumption and smoking did not appear as such, and model 18, including two interaction terms between exposure and age, was directly considered. The increase in the log-likelihood was only marginal ($P \simeq 0{\cdot}6$, LR test) and $\widehat{AR}$ slightly increased to 0·868, $\widehat{SD} = 0{\cdot}0440$, as with the weighted-sum approach.

In applying the regression approach to the estimation of the AR in oesophageal cancer, it has been illustrated that the regression approach is more general than the stratification approaches since it includes the unadjusted approach and both the Mantel–Haenszel and weighted-sum approach as special cases. The unadjusted approach corresponds to models with exposure only (models 1, 6, 11 and 16). The Mantel–Haenszel approach corresponds to models with exposure and confounding, but no interaction (models 2, 7, 12, 17) and only the OR estimate slightly differs. The weighted-sum approach corresponds to fully saturated models with all possible interaction terms (models 5, 10, 15 and 18). The regression approach allows for intermediate models such as models with only partial interaction (models 3–4, 8–9 and 13–14). Also, models with fewer parameters for confounding could be used, for instance, with two binary variables for smoking and two for age in Problem 1, instead of eight binary variables.

This flexibility in model selection allows one to balance parsimony against goodness of fit. One strives for parsimonious models, (1) to reduce bias that can arise with unconditional logistic regression when the data get sparse[16] even though the use of conditional logistic regression could alleviate this problem;[27] and (2) to reduce random error. This latter feature was not too apparent in the example but could be more pronounced with more factors and sparser data. However, the desire for parsimony must be balanced against biases that result from model misspecification, such as that resulting from the assumption of a homogeneous OR when there is indeed interaction.

6. DISCUSSION

In this paper, methods of adjustment have been reviewed. The Mantel–Haenszel approach, based on stratification, allows one to control for confounding but not for effect-modification. The weighted-sum approach, also based on stratification, allows one to control both for confounding and effect modification, but bias in estimating the AR occurs when the data get sparse. The approach based on regression is more flexible and general. It includes the two stratification approaches as special cases and provides a unified framework for estimation and hypothesis testing. Parallel to what is done for relative risk estimation, more generality can be gradually built into the models until the increase in the log-likelihood becomes marginal. Following this strategy we would chose the most general model 10 in Problem 1*b*, while we would chose models 2 and 17 with no interaction terms, respectively, in Problems 1*a* and 2.

The discussion of the example has illustrated several practical points. It showed the major impact of the definition of the baseline (or unexposed) level of the exposure of interest. It showed that there is no need to consider more than two levels for the exposure of interest. It illustrated the change in AR estimates incurred by successively controlling for confounding and interaction. That impact was moderate in the example but deserves to be studied on other examples.

Current research on AR estimation in case-control studies concerns the following problems. Improvements of the weighted-sum approach are under study in order to reduce bias.[29] A simple method of estimation of the SD of $\widehat{AR}$ in the regression approach has been proposed for unconditional logistic regression.[30] Since the regression approach is not tied to the (multiplicative) logistic model, using an additive model for the relative risk has been advocated and applied to a case-control study.[31] Finally, as noted before, extensions in the regression approach are still needed in order to obtain SD estimates when conditional logistic regression is used for large strata. Resampling techniques might prove useful in that case.

The methods described in this paper may be extended to estimate a more general measure than the AR, namely the generalized impact fraction.[13,32] This is defined as the fractional reduction of the disease that would result from changing the current distribution of the exposure factor(s) to some modified distribution, namely

$$\frac{P\{(\text{disease}|\text{modified distribution of exposure}) - P(\text{disease}|\text{current distribution of exposure})\}}{P(\text{disease}|\text{current distribution of exposure})}.$$

The AR corresponds to the special case in which the modified distribution puts unit mass on the lowest risk configuration. In order to estimate this measure, it is critical to take into account polychotomous or continuous risk factor(s) as such. This points to a further advantage of the regression approach which is readily adapted to polychotomous exposure factors, as already discussed, but also continuous exposure factors.[26,27]

APPENDIX I: VARIANCE ESTIMATES OF $\widehat{AR}$

I.1. Mantel–Haenszel approach

The notation in Section 2 is extended to denote by n_{ij} and m_{ij} the respective numbers of cases and controls with level i of exposure ($i = 0, 1$) and j of adjustment. Moreover we denote by $n_{.j}$ and $m_{.j}$ the respective numbers of cases and controls with level j of adjustment. Variance estimators have been developed, by Kuritz and Landis,[16,17] and by Greenland.[18] Kuritz and Landis[16] noted that

the estimate in (11) can be expressed as a function of the elementary quantities n_{ij} and m_{ij} since:

$$\hat{P}(E|D) = \sum_j n_{1j}/n \quad \text{and} \quad \widehat{\text{OR}}_{\text{MH}} = \frac{\sum_{j=1}^{J} n_{1j}m_{0j}/(n_{\cdot j} + m_{\cdot j})}{\sum_{j=1}^{J} n_{0j}m_{1j}/(n_{\cdot j} + m_{\cdot j})}.$$

For simple random sampling of the controls, the two $2 \times J$ vectors (n_{ij}) $i = 0, 1, j = 1, \ldots, J$ and (m_{ij}) $i = 0, 1, j = 1, \ldots, J$ have independent multinomial distributions and use of the delta-method yields a variance estimate. Though not explicitly worked out by Kuritz and Landis, essentially the same approach could be used for stratified random sampling and frequency matching by considering the appropriate following distributions described for instance by Whittemore.[11] For frequency matching of the controls, the number of controls for each level j of stratification is a multiple of the number of cases for level j, that is $m_{\cdot j} = r_j n_{\cdot j}$, where r_j is fixed and typically common to all levels j (for example, $r_j = r = 1$ for $j = 1, \ldots, J$). The $2 \times J$ vector (n_{ij}) still has a multinomial distribution while the numbers of exposed controls m_{1j} have independent binomial distributions conditionally on the numbers of controls $m_{\cdot j}$. For stratified random sampling, the number of controls $m_{\cdot j}$ for each level of stratification is fixed and the numbers of exposed controls m_{1j} have independent binomial distributions unconditionally. Finally, for individual $1{:}k$ matching, the argument is more subtle and Kuritz and Landis[17] consider the distribution of the quantities L_{xyz}, which denote the number of matched sets (that is one case and k controls), with x exposed cases ($x = 0, 1$), and z controls ($z = 1, \ldots, k$) out of whom y are exposed. Notice that z can be less than k if data are missing. The point estimate in (11) can be expressed as a function of the vector (L_{xyz}) with elements L_{xyz}, since:

$$\widehat{\text{OR}}_{\text{MH}} = \frac{\sum_{z=1}^{k}\sum_{y=0}^{z}(z-y)L_{1yz}/(z+1)}{\sum_{z=1}^{k}\sum_{y=0}^{z}yL_{0yz}/(z+1)} \quad \text{and} \quad \hat{P}(E|D) = \sum_{z=1}^{k}\sum_{y=0}^{z}\frac{L_{xyz}}{n}.$$

Given the multinomial distribution of the vector (L_{xyz}), Kuritz and Landis[17] obtained a variance estimator by applying the delta-method.

Greenland,[18] instead, used the variance estimator of $\widehat{\text{OR}}_{\text{MH}}$ which is known to be valid even for sparse data, for example even in the case of individual matching (when j indexes the matched sets), and is thus termed a dually consistent estimator.[33] He wrote:

$$\begin{aligned}\text{var}\{\log(\widehat{\text{AR}})\} = {} & \text{var}\{\log \hat{P}(E|D)\} + \text{var}\{\log((\widehat{\text{OR}}_{\text{MH}} - 1)/\widehat{\text{OR}}_{\text{MH}})\} \\ & + 2\,\text{cov}\{\log \hat{P}(E|D), \log((\widehat{\text{OR}}_{\text{MH}} - 1)/\widehat{\text{OR}}_{\text{MH}})\}.\end{aligned}$$

The first term can be estimated from the binomial distribution of $n_{1\cdot}$, the total number of exposed cases. The second term is a function of $\text{var}(\widehat{\text{OR}}_{\text{MH}})$. By using the delta-method, it can be seen that the covariance term is asymptotically equivalent to:

$$\frac{1 - P(E|D)}{\text{OR}_{\text{MH}} - 1}\{\text{cov}(\log \widehat{\text{OR}}_{\text{MH}}, \log n_{1\cdot}) - \text{cov}(\log \widehat{\text{OR}}_{\text{MH}}, \log n_{0\cdot})\}.$$

Since $\widehat{\text{OR}}_{\text{MH}}$ is an asymptotically unbiased estimator of OR, the true odds ratio, then, from Cox and Hinkley,[34] asymptotically:

$$\text{cov}\{\log(\widehat{\text{OR}}_{\text{MH}}), U(\text{OR})\} = 1,$$

in which $U(\text{OR}) = n_{1\cdot} - E(n_{1\cdot}|\text{OR})$ is the score calculated in OR. Thus, from the delta-method:

$$\text{cov}(\log \widehat{\text{OR}}_{\text{MH}}, \log n_{1\cdot}) = 1/E(n_{1\cdot}),$$

and

$$\text{cov}(\log \widehat{\text{OR}}_{\text{MH}}, \log n_{0\cdot}) = 1/E(n_{0\cdot}).$$

Hence, Greenland[18] obtained the following variance estimate:

$$\widehat{\text{var}}\{\log(\widehat{\text{AR}})\} = \widehat{\text{var}}(\log \widehat{\text{OR}}_{\text{MH}})/(\widehat{\text{OR}}_{\text{MH}} - 1)^2 + n_{0.}/nn_{1.} + 2/n_{1.}(1 - \widehat{\text{OR}}_{\text{MH}}).$$

Note that this estimator is valid even for sparse data and can therefore be used for all four kinds of sampling of the controls discussed in this paper, namely, simple random sampling, frequency matching, stratified random sampling and individual matching.

Through similar arguments, Greenland[18] obtained the following variance estimators for the hybrid AR estimate in (13):

$$\widehat{\text{var}}(\widehat{\text{AR}}) = \left(\sum_{h=1}^{H} \widehat{\text{AR}}_h^2 \hat{V}_h\right) - \widehat{\text{AR}}^2/n,$$

with:

$$\hat{V}_h = \widehat{\text{var}}(\log \widehat{\text{OR}}_h)/(1 - \widehat{\text{OR}}_h)^2 + n_{0h}/n_{1h}n_{.h} + 2/n_{1h}(1 - \widehat{\text{OR}}_h) + 1/n_{.h}.$$

I.2. Weighted-sum approach

In case of simple random sampling of the controls, the two $2 \times J$ vectors (n_{ij}) and (m_{ij}) have independent multinomial distributions, as described in Appendix I.1, and Whittemore[11] obtained the following variance estimate of $\widehat{\text{AR}}$ using the delta-method:

$$\widehat{\text{var}}(\widehat{\text{AR}}) = \left\{\sum_{j=1}^{J} (m_{.j}n_{0j}/m_{0j})^2 (1/n_{0j} + m_{1j}/m_{.j}m_{0j}) - n(1 - \widehat{\text{AR}})^2\right\} \Big/ n^2.$$

In case of frequency matching of the controls, given the distributions described in Appendix I.1, Whittemore[11] obtained the exact same variance estimate as for simple random sampling. Finally, Whittemore did not deal with stratified random sampling but it is straightfoward to show that the variance estimate is again the same.

I.3. Regression Approach

From formula (19) and the delta-method, the variance of $\widehat{\text{AR}}$ is given by:

$$\text{var}(\widehat{\text{AR}}) = \sum_{i,j,k} \sum_{i',j',k'} [\text{RR}^{-1}_{i|j,k}(\hat{\theta})\text{RR}^{-1}_{i'|j',k'}(\hat{\theta})\,\text{cov}(\hat{p}_{ijk}, \hat{p}_{i'j'k'}) + \hat{p}_{ijk}\hat{p}_{i'j'k'}\,\text{cov}\{\text{RR}^{-1}_{i|j,k}(\hat{\theta}), \text{RR}^{-1}_{i'|j',k'}(\hat{\theta})\} + 2\text{RR}^{-1}_{i|j,k}(\hat{\theta})\hat{p}_{i'j'k'}\,\text{cov}\{\hat{p}_{ijk}, \text{RR}^{-1}_{i'|j',k'}(\hat{\theta})\}].$$

The covariance in the first term can be estimated from the multinomial distribution of the quantities $n_{ijk} = n\hat{p}_{ijk}$, and in the second term by using the delta-method and the observed information matrix from the logistic model. The third covariance term between the vector $(\hat{p}_{ijk})$ and the estimated relative risks arises from implicit functional relationships among these quantities as described in Benichou and Gail.[27] Briefly, the log-likelihood of the fitted logistic model involves the quantities $\hat{p}_{ijk}$ as well as the vector of parameters θ and the score equations define relationships between the quantities $\hat{p}_{ijk}$ and the MLE $\hat{\theta}$. The score equations cannot in general be solved explicitly. However these equations still define implicit relationships between $\hat{p}_{ijk}$ and $\hat{\theta}$ and an extension of the delta-method to implicitly related random variables[35] can be used to obtain the required covariance estimates.

APPENDIX II. COMPLEMENTS

II.1. Identity of the Mantel–Haenszel AR and the weighted-sum AR with 'case-loading' for a homogeneous relative risk

The weighted-sum approach allows adjustment for one (or several) factor(s) that we will denote by Y and which form J levels $Y = y_1, \ldots, Y = y_J$. By definition, we have:

$$1 - \mathrm{AR} = \sum_{j=1}^{J} w_j (1 - \mathrm{AR}_j)$$
$$= \sum_{j=1}^{J} P(Y = y_j | D) \frac{P(D | \bar{E}, Y = y_j)}{P(D | Y = y_j)}$$

by definition of the 'case-load' weighting and of AR_j.

Therefore

$$1 - \mathrm{AR} = \frac{1}{P(D)} \sum_{j=1}^{J} P(Y = y_j) P(D | \bar{E}, Y = y_j)$$

from Bayes' theorem,

$$= \frac{1}{P(D)} \sum_{j=1}^{J} P(Y = y_j) P(D | \bar{E}, Y = y_j) \{P(\bar{E} | Y = y_j) + P(E | Y = y_j)\}$$

$$= \frac{1}{P(D)} \sum_{j=1}^{J} P(D | \bar{E}, Y = y_j) \{P(\bar{E}, Y = y_j) + P(E, Y = y_j)\}$$

by definition of conditional probability,

$$= \frac{1}{P(D)} \sum_{j=1}^{J} \left\{ P(D, \bar{E}, Y = y_j) + \frac{1}{\mathrm{RR}} P(D, E, Y = y_j) \right\}$$

by definition of conditional probability and of the common relative risk RR,

$$= \frac{1}{P(D)} \left\{ P(D, \bar{E}) + \frac{1}{\mathrm{RR}} P(D, E) \right\}$$

$$= P(\bar{E} | D) + \frac{1}{\mathrm{RR}} P(E | D)$$

$$= 1 - P(E | D) \frac{\mathrm{RR} - 1}{\mathrm{RR}}.$$

Therefore, if the relative risk is assumed to be common to all levels J, the weighted-sum AR has the same expression as the Mantel–Haenszel AR and their corresponding estimates are the same as long as the relative risk is estimated the same way (for instance by the Mantel–Haenszel odds ratio).

II.2. Proof of formula (16)

$$\sum_{i=0}^{I} p_i \mathrm{RR}_i^{-1} = \sum_{i=0}^{I} P(X = x_i | D) \frac{P(D | X = x_0)}{P(D | X = x_i)}$$

by definition,

$$= \sum_{i=0}^{I} P(X = x_i|D) \frac{P(D|X = x_0)P(X = x_i)}{P(X = x_i|D)P(D)}$$

from Bayes' theorem,

$$= \frac{P(D|X = x_0)}{P(D)}$$

$$= 1 - \mathrm{AR}.$$

II.3. Proof of formula (17)

$$\sum_{j=1}^{J} \sum_{i=0}^{I} \rho_{ij} \mathrm{RR}_{i|j}^{-1} = \sum_{j=1}^{J} \sum_{i=0}^{I} P(X = x_i, Y = y_j|D) \frac{P(D|X = x_0, Y = y_j)}{P(D|X = x_i, Y = y_j)}$$

by definition,

$$= \sum_{j=1}^{J} \sum_{i=0}^{I} P(X = x_i, Y = y_j|D) \frac{P(D|X = x_0, Y = y_j)P(X = x_i, Y = y_j)}{P(X = x_i, Y = y_j|D)P(D)}$$

from Bayes' theorem,

$$= \frac{1}{P(D)} \sum_{j=1}^{J} P(D|X = x_0, Y = y_j)P(Y = y_j)$$

$$= \frac{P(D|X = x_0)}{P(D)}$$

$$= 1 - \mathrm{AR}.$$

II.4. Identity of AR estimates when a single polychotomous exposure factor with $I + 1$ levels is taken into account in the logistic model

This shows that AR estimates are identical whether the factor is considered polychotomous (I binary variables in the model) or binary (one binary variable in the model) for unconditional logistic regression.

When the factor is considered as binary, the MLE of the OR from unconditional logistic regression is given by:

$$\widehat{\mathrm{OR}} = \frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} m_i} \frac{m_0}{n_0}$$

and from formula (16), one obtains:

$$\widehat{\mathrm{AR}} = 1 - \frac{n_0}{n} - \frac{\sum_{i=1}^{I} n_i}{n} \widehat{\mathrm{OR}}^{-1}$$

$$= 1 - \frac{n_0}{n} \frac{m}{m_0}$$

which is the same estimate as given in formula (3).

When I binary variables are used in the logistic model, the unconditional MLE of OR_i, the OR associated with level i of exposure is given by:

$$\widehat{OR}_i = \frac{n_i m_0}{m_i n_0}$$

and from formula (16), one obtains:

$$\begin{aligned}\widehat{AR} &= 1 - \frac{n_0}{n} - \sum_{i=1}^{I} \left(\frac{n_i}{n}\right) \widehat{OR}_i^{-1} \\ &= 1 - \frac{n_0}{n}\left(1 + \sum_{i=1}^{I} \frac{m_i}{m_0}\right) \\ &= 1 - \frac{n_0}{n}\frac{m}{m_0}\end{aligned}$$

which is again the same estimate as given in formula (3).

ACKNOWLEDGEMENTS

I wish to thank the editor, reviewers and Dr. Mitchell Gail for helpful comments. I also wish to thank Jennifer Donaldson and Stéphanie Henry for their great care and skill in typing this paper.

REFERENCES

1. Levin, M. L. 'The occurrence of lung cancer in man', *Acta Unio Internationalis contra Cancrum*, **9**, 531–541 (1953).
2. Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research Vol I: The Analysis of Case-control Studies*, International Agency for Research on Cancer Scientific Publications, No. 32, Lyon, 1980.
3. Miettinen, O. S. 'Proportion of disease caused or prevented by a given exposure, trait or intervention', *American Journal of Epidemiology*, **99**, 325–332 (1974).
4. MacMahon, B. and Pugh, T. F. *Epidemiology: Principles and Methods*, Little, Brown and Co, Boston, 1970.
5. Cole, P. and MacMahon, B. 'Attributable risk percent in case-control studies', *British Journal of Preventive and Social Medicine*, **25**, 242–244 (1971).
6. Ouellet, B. L., Romeder, J. M. and Lance, J. M. 'Premature mortality attributable to smoking and hazardous drinking in Canada', *American Journal of Epidemiology*, **109**, 451–463 (1979).
7. Tuyns, A. J., Pequignot, 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', *Bulletin of Cancer*, **64**, 45–60 (1977).
8. Walter, S. D. 'The estimation and interpretation of attributable risk in health research', *Biometrics*, **32**, 829–849 (1976).
9. Rao, C. R. *Linear Statistical Inference and its Applications*, 2nd edn, Wiley, New York, 1973.
10. Walter, S. D. 'The distribution of Levin's measure of attributable risk', *Biometrika*, **62**, 371–374 (1975).
11. Whittemore, A. S. 'Statistical methods for estimating attributable risk from retrospective data', *Statistics in Medicine*, **1**, 229–243 (1982).
12. Leung, H. M. and Kupper, L. L. 'Comparisons of confidence intervals for attributable risk', *Biometrics*, **37**, 293–302 (1981).
13. Walter, S. D. 'Prevention for multifactorial diseases', *American Journal of Epidemiology*, **112**, 409–416 (1980).
14. Walter, S. D. 'Effects of interaction, confounding and observational error on attributable risk estimation', *American Journal of Epidemiology*, **117**, 598–604 (1983).
15. Whittemore, A. S. 'Estimating attributable risk from case-control studies', *American Journal of Epidemiology*, **117**, 76–85 (1983).
16. Kuritz, S. J. and Landis, J. R. 'Summary attributable risk estimation from unmatched case-control data', *Statistics in Medicine*, **7**, 507–517 (1988).

17. Kuritz, S. J. and Landis, J. R. 'Attributable risk estimation from matched case-control data', *Biometrics*, **44**, 355–367 (1988).
18. Greenland, S. 'Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data', *Statistics in Medicine*, **6**, 701–708 (1987).
19. Mantel, N. and Haenszel, W. 'Statistical aspects of the analysis of data from retrospective studies of disease', *Journal of the National Cancer Institute*, **22**, 719–748 (1959).
20. Breslow, N. E. 'Odds ratio estimators when the data are sparse', *Biometrika*, **68**, 73–84 (1981).
21. Birch, M. W. 'The detection of partial association, I: the 2×2 case', *Journal of the Royal Statistical Society, Series B*, **27**, 313–324 (1964).
22. Landis, J. R., Heyman, E. R. and Koch, G. G. 'Average partial association in three-way contingency tables: a review and discussion of alternative tests', *International Statistical Review*, **46**, 237–254 (1978).
23. Rothman, K. J. 'Synergy and antagonism in cause-effect relationships', *American Journal of Epidemiology*, **99**, 385–388 (1974).
24. Ejigou, A. 'Estimation of attributable risk in the presence of confounding', *Biometric Journal*, **21**, 155–165 (1979).
25. Sturmans, F., Mulder, P. G. H. and Walkenburg, H. A. 'Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage', *American Journal of Epidemiology*, **105**, 281–289 (1977).
26. Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A. and Schairer, C. 'Estimating the population attributable risk for multiple risk factors using case-control data', *American Journal of Epidemiology*, **122**, 904–914 (1985).
27. Benichou, J. and Gail, M. H. 'Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models', *Biometrics*, **46**, 991–1003 (1990).
28. Prentice, R. L. and Breslow, N. E. 'Retrospective studies and failure time models', *Biometrika*, **65**, 153–168 (1978).
29. Gart, J. J. and Thomas, D. G. 'Point and interval estimation of the attributable risk in stratified case-control studies', Personal communication.
30. Drescher, K. and Schill, W. 'Attributable risk estimation from case-control data via logistic regression', Personal communication.
31. Coughlin, S. S., Nass, C. C., Pickle, L. W., Trock, B. and Bunin, G. 'Regression methods for estimating attributable risk in population-based case-control studies: a comparison of additive and multiplicative models', *American Journal of Epidemiology*, **133**, 305–313 (1991).
32. Morgenstern, H. and Bursic, E. S. '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 (1982).
33. Robins, J. M., Breslow, N. E. and Greenland, S. 'Estimators of the Mantel–Haenszel variance consistent in both sparse-data and large-strata limiting models', *Biometrics*, **42**, 311–323 (1986).
34. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1974.
35. Benichou, J. and Gail, M. H. 'A delta-method for implicitly defined random variables', *The American Statistician*, **43**, 41–44 (1989).