STATISTICS IN MEDICINE, VOL. 10, 1025–1035 (1991)

FLOATING ABSOLUTE RISK: AN ALTERNATIVE TO RELATIVE RISK IN SURVIVAL AND CASE-CONTROL ANALYSIS AVOIDING AN ARBITRARY REFERENCE GROUP

D. F. EASTON, J. PETO AND A. G. A. G. BABIKER

Section of Epidemiology, Institute of Cancer Research, Block D, 15 Cotswold Road, Sutton, Surrey SM2 5NG, U.K.

SUMMARY

We discuss the problem of describing multiple group comparisons in survival analysis using the Cox model, and in matched case-control studies. The standard method of comparing the risk in each group with a baseline group is unsatisfactory because the standard errors and confidence limits relate to correlated parameters, all dependent on precision within the baseline group. We describe the construction of standard errors for the parameters of all groups, without the need to select a baseline group. These standard errors can be regarded as relating to roughly independent parameters, so that groups can be compared efficiently without knowledge of the covariances. The method should assist in graphical presentation of relative risks, and in the combination of results from published studies. Two examples are presented.

INTRODUCTION

Many regresssion models include the effects of a factor which divides individuals into several categories, or levels. If there are s categories $0, 1, \ldots, s-1$ it is usual to fit $s-1$ parameters $\beta_1, \ldots, \beta_{s-1}$, with the zero group forming a baseline or comparison category. This standard parametrization, used for example in the package GLIM,[1] has two disadvantages. All the estimates of the regression parameters will necessarily be correlated by virtue of their dependence on the baseline category, and if the baseline category is small, they will have large standard errors.

In many linear models, this problem can be avoided by dispensing with a parameter for the overall mean and fitting s parameters $\alpha_0, \ldots, \alpha_{s-1}$ relating to category membership. This 'natural' parametrization of the effect of the factor will result in estimates whose only correlations are those induced by other parameters in the model. This manoeuvre is carried out easily in GLIM, and would apply, for example, to logistic regression analysis of case-control studies with large strata.[2]

This multiple comparison problem also occurs in proportional hazards regression[3] and in the analysis of matched case-control studies using logistic regression.[2] Here, however, the problem is more fundamental, because analysis is based on a conditional likelihood which contains no analogue of the overall mean. Some additional manoeuvring, described here, is required to produce uncorrelated parameter estimates.

METHODS

Survival analysis

We consider first a typical survival analysis problem, in which survival is to be related to a factor with several levels, such as stage of disease. Suppose initially that no additional covariates, other

0277–6715/91/071025–11$05.50

Received June 1990
Revised September 1990

than the factor of interest, are considered. The Cox proportional hazards model[3] assumes that the risk of death at time t for individuals in category j, $\lambda_j(t)$, is of the form:

$$\lambda_j(t) = \lambda_0(t)\exp(\beta_j). \tag{1}$$

Suppose there are K risk sets, one for each time at which a death occurs. Let n_{jk} be the number of individuals in category j and risk set k, and let M_j denote the total number of deaths occurring in individuals in category j. The standard method of analysis is to maximize the partial log-likelihood:

$$\log L = \sum_{j=0}^{s-1} M_j\beta_j - \sum_{k=1}^{K}\left[\log\left(\sum_{j=0}^{s-1} n_{jk}\exp\beta_j\right)\right]. \tag{2}$$

This is exact if one death occurs in each risk set; in the case of ties, it is an approximation to the exact likelihood suggested by Peto[4] (in which case the last term in (2) is repeated for each death). Only $s-1$ of the β_j's can be estimated by maximizing this likelihood, because the addition of an arbitrary constant to each β_j will only add a constant to the log-likelihood. The usual practice is to set β_0 to zero, so that $\exp(\beta_j)$ is the relative risk for category j relative to the zero (or baseline) category. The information matrix for $\beta_1, \ldots, \beta_{s-1}$ is then:

$$\frac{\partial^2 \log L}{\partial\beta_j^2} = -\sum_{k=1}^{K}\frac{n_{jk}\exp\beta_j}{\sum_l n_{lk}\exp\beta_l} + \sum_{k=1}^{K}\frac{(n_{jk}\exp\beta_j)^2}{(\sum_l n_{lk}\exp\beta_l)^2}$$

and

$$\frac{\partial^2 \log L}{\partial\beta_i\partial\beta_j} = \sum_{k=1}^{K}\frac{n_{ik}n_{jk}\exp\beta_i\exp\beta_j}{(\sum_l n_{lk}\exp\beta_l)^2}.$$

In general it is not possible to invert this matrix algebraically and obtain an explicit expression for the variance–covariance matrix of the β_j's. However, there is one important case when the variance–covariance matrix can be written down explicitly. This occurs when the categories are represented in equal proportions in every risk set, so that the n_{jk} are of the form:

$$n_{jk} = N_j\chi_k \tag{3}$$

where $N_j = \sum_k n_{jk}$ and χ_k, $k = 1, \ldots, K$ are constants.

In this case the information matrix takes the simple form:

$$C + DD^{\mathrm{T}}$$

where

$$C = \frac{K}{\sum_l b_l}\operatorname{diag}(b_1, b_2, \ldots, b_{s-1})$$

$$D^{\mathrm{T}} = \frac{\sqrt{K}}{\sum_l b_l}(b_1, b_2, \ldots, b_{s-1})$$

and

$$b_j = N_j\exp(\beta_j), \quad j = 1, \ldots, s-1.$$

The variance–covariance matrix for the β_j's is then:

$$B = \frac{\sum_l b_l}{K}\begin{bmatrix} \left(\frac{1}{b_0}+\frac{1}{b_1}\right) & \frac{1}{b_0} & \frac{1}{b_0} & \cdots & \frac{1}{b_0} \\ \frac{1}{b_0} & \left(\frac{1}{b_0}+\frac{1}{b_2}\right) & & & \\ \vdots & & \ddots & & \\ \frac{1}{b_0} & & & & \left(\frac{1}{b_0}+\frac{1}{b_{s-1}}\right) \end{bmatrix}.$$

The estimates $\hat{\beta}_j$ are of course correlated, but the covariances are all equal to $(\sum_l b_l)/(b_0 K)$, and depends only on the sample size of the 'zero' category. These variances and covariances are precisely those which would be obtained if there were *independent* parameters $\alpha_0, \ldots, \alpha_{s-1}$ associated with each category, such that

$$\beta_j = \alpha_j - \alpha_0$$

with the variance of $\hat{\alpha}_j$ given by:

$$\text{var}(\hat{\alpha}_j) = \frac{\sum_l N_l \exp\beta_l}{K N_j \exp\beta_j} = \frac{\sum_l N_l \exp\alpha_l}{K N_j \exp\alpha_j}.$$

We interpret the parameters $\alpha_0, \ldots, \alpha_{s-1}$ as logarithms of the *absolute* risks associated with each category, so that subtraction of α_0 gives logarithms of the relative risks β_j. This representation of the results has a major advantage. The estimates $\hat{\alpha}_j$, together with their standard errors, provide a complete description of the results, whereas the standard representation is inadequate without the complete covariance matrix of the $\hat{\beta}_j$'s. The only problem is that, although we know what the standard errors of the $\hat{\alpha}_j$'s ought to be, the parameters themselves cannot be estimated from the Cox likelihood, which only allows $s - 1$ relative risks to be estimated. This problem is dealt with in the next section.

Why are s independent parameters obtained in this case? The Cox model (1) allows the absolute risk to vary in an arbitrary manner between different risk sets. The partial likelihood (2) is equivalent to a Poisson likelihood in which a separate risk parameter is estimated for each risk set. In general the categories will not be equally represented in all risk sets, and the parameter estimates for the categories are therefore partially confounded with the parameter estimates for the risk sets, and in turn correlated with each other. However, in the special case specified by condition (3), group membership is orthogonal to the confounding effect of risk set and no correlations are induced. This condition will often hold approximately in survival analysis, as in the example below, provided that the pattern of censoring is similar in each group. The proportions at risk in different groups will change over time due to differences in event rates, but the effect will be large only if (a) the β_j's differ substantially, and (b) the death rates are sufficiently large to eliminate a substantial proportion of individuals.

Two alternative definitions of the problem

(i) *An heuristic formulation*

The general situation, in which completely independent parameter estimates cannot be derived, can be approached in two quite different ways. The first is quite general and applies to any

analysis of data in s categories which yields $s-1$ (approximately normal) parameter estimates $\hat{\beta}_j$ ($j = 1, \ldots, s-1$) and the corresponding covariance matrix $[B_{ij}]$, including normal or logistic regression analysis and matched case control analysis, as well as Cox regression. These $s-1$ estimates $\hat{\beta}_j$ from a conventional analysis are then assumed to have been derived from s parameter estimates $\hat{\alpha}_j$ ($j = 0, \ldots, s-1$) with covariance matrix $[A_{ij}]$, such that

$$\hat{\beta}_j = \hat{\alpha}_j - \hat{\alpha}_0. \tag{4}$$

This implies that the covariance matrices A and B must satisfy

$$B_{ij} = A_{ij} - A_{i0} - A_{0j} + A_{00}, \qquad i, j = 1, \ldots, s-1. \tag{5}$$

The problem may thus be restated as follows. We wish to 'invent' an extra parameter estimate $\hat{\alpha}_0$ for the baseline group (hence defining $\hat{\alpha}_j$ ($j = 1, \ldots, s-1$) from equation (4)), and a corresponding matrix $[A_{ij}]$ which satisfies equation (5). Without loss of generality, we can assume that $\hat{\alpha}_0$ happens to equal zero. The objective is thus to select a variance–covariance matrix A which satisfies equation (5) and is 'almost diagonal' in some sense. For the procedure to be well-defined, we also require that the same matrix A should be generated regardless of which baseline group was chosen.

There is then one matrix A that appears particularly 'natural', as it satisfies (and is uniquely defined by) three attractive and equivalent criteria:

1. The sum of covariances in each row is zero.
2. The sum of squares of the covariances is minimized.
3. A_{00}, the variance of the baseline group parameter estimate $\hat{\alpha}_0$, is the average of the covariances B_{ij} ($i \neq j$).

This matrix A can be calculated from B as follows:

$$A_{00} = \frac{1}{(s-1)(s-2)} \sum_{i=1}^{s-1} \sum_{\substack{j=1 \\ j \neq i}}^{s-1} B_{ij}$$

$$A_{0i} = A_{00} - \frac{1}{(s-2)} \sum_{\substack{j=1 \\ j \neq i}}^{s-1} B_{ij} \qquad i = 1, \ldots, s-1$$

$$A_{ij} = B_{ij} + A_{0i} + A_{0j} - A_{00} \qquad i, j = 1, \ldots, s-1.$$

If there are other covariates, then further constraints are required to completely specify the off diagonal terms of the covariance matrix. A reasonable choice is that which again minimizes the sum of squares of the covariances. The remaining terms of A are then given by:

$$A_{0j} = -\frac{1}{s} \sum_{l=1}^{s-1} B_{lj}$$

$$A_{ij} = B_{ij} - \frac{1}{s} \sum_{l=1}^{s-1} B_{lj} \qquad i = 1, \ldots, s-1, \quad j = s, \ldots, t+s-1$$

$$A_{ij} = B_{ij} \qquad i, j = s, \ldots, t+s-1$$

where the parameters $s, \ldots, t+s-1$ refer to the additional covariates. The calculation of covariances between parameter estimates is usually of little practical importance, as they are so rarely published. The approach described above is probably not the most appropriate way to deal

with covariances between estimates for two categorical variables which are strongly inter-related, such as age at first birth and number of children. This question requires further consideration.

The standard variance–covariance matrix B can be routinely output by standard conditional logistic regression and Cox regression packages such as PECAN,[5] and this method is therefore trivial to implement. In spite of its simplicity, however, this formulation has the unattractive property of being defined in terms of the estimates and their covariances rather than any explicit model in which the parameters are well-defined. The formulation outlined in the following section is in this sense more satisfactory, as it is based on an explicit (if somewhat arbitrary) likelihood with a natural interpretation.

(ii) *An augmented likelihood*

An alternative method for defining and formally deriving the $\hat{\alpha}_j$'s and their covariances is by the addition of a term to the log-likelihood (2) of the form:

$$-E + M \log E \tag{6}$$

where $M = \sum M_j$ is the total number of deaths and

$$E = \lambda \sum_{j=0}^{s-1} [N_j \exp \alpha_j] \tag{7}$$

with $N_j = \sum_k n_{kj}$ and λ an arbitrary constant. The term (6) can be interpreted as a Poisson likelihood for the total number of deaths M, with expectation E which is given by some underlying rate λ multiplied by the number of individuals at risk in category j weighted by their risks $\exp \alpha_j$, and summed over all categories and risk sets. This is a likelihood familiar from analyses of cohort studies. The total log-likelihood, with the β_j's replaced by α_j's, is then:

$$\log L = \sum_{j=0}^{s-1} M_j \alpha_j - \sum_{k=1}^{K} \log \left\{ \sum_{l=0}^{s-1} [n_{lk} \exp \alpha_l] \right\} - \lambda \sum_{j=0}^{s-1} [N_j \exp \alpha_j] + M \log \left[\sum_{j=0}^{s-1} N_j \exp \alpha_j \right] + M \log \lambda. \tag{8}$$

We provide below a more rational basis for this likelihood. For the moment, we simply note that (8) can be maximized with respect to all s parameters $\alpha_0, \ldots, \alpha_{s-1}$. Moreover, having obtained estimates of the α_j's the corresponding log-relative risk estimates $\hat{\beta}_j$ and their variances and covariances are precisely the same as those obtained from the original Cox likelihood. This is because the first derivatives of (8) are given by:

$$\frac{\partial \log L}{\partial \alpha_j} = M_j - \sum_{k=1}^{K} \frac{n_{jk} \exp \alpha_j}{\sum_l n_{lk} \exp \alpha_l} - \lambda N_j \exp \alpha_j + \frac{M N_j \exp \alpha_j}{\sum_l N_l \exp \alpha_l} \tag{9}$$

and the maximum likelihood solution to (9) must satisfy:

$$\lambda \sum_j N_j \exp \alpha_j = M. \tag{10}$$

At the maximum likelihood estimate, therefore, the final two terms of (9) cancel out and the equations reduce (after substituting $\beta_j = \alpha_j - \alpha_0$) to the first derivatives of the Cox partial

likelihood (2). The second derivatives of (8) are given by:

$$\frac{\partial^2 \log L}{\partial \alpha_j^2} = -\sum_{k=1}^{K} \frac{n_{jk} \exp \alpha_j}{\sum_l n_{lk} \exp \alpha_l} + \sum_{k=1}^{K} \frac{(n_{jk} \exp \alpha_j)^2}{(\sum_l n_{lk} \exp \alpha_l)^2}$$
$$+ N_j \exp \alpha_j - \frac{M N_j \exp \alpha_j}{\sum_l N_l \exp \alpha_l} + \frac{M (N_j \exp \alpha_j)^2}{(\sum_l N_l \exp \alpha_l)^2}$$
$$= -\sum_{k=1}^{K} \frac{n_{jk} \exp \alpha_j}{\sum_l n_{lk} \exp \alpha_l} + (\exp \alpha_j)^2 \sum_{k=1}^{K} \left\{ \frac{n_{jk}^2}{(\sum_l n_{lk} \exp \alpha_l)^2} - \frac{N_j^2}{(\sum_l N_l \exp \alpha_l)^2} \right\} \quad (11)$$

at the maximum likelihood solution, and

$$\frac{\partial^2 \log L}{\partial \alpha_i \partial \alpha_j} = (\exp \alpha_i \exp \alpha_j) \sum_{k=1}^{K} \left\{ \frac{n_{ik} n_{jk}}{(\sum_l n_{lk} \exp \alpha_l)^2} - \frac{N_i N_j}{(\sum_l N_l \exp \alpha_l)^2} \right\}. \quad (12)$$

Under the orthogonality given by (3), the off diagonal terms (12) equal zero, and the estimates $\hat{\alpha}_j$ are independent. More generally, provided that the proportion of individuals in each category is approximately the same in all risk sets, the off diagonal terms will be small compared with the diagonal terms (11), and the $\hat{\alpha}_j$'s should therefore be only weakly correlated.

We note that the actual values of the α_j (as opposed to their differences) are still arbitrary since by suitable choice of λ in (7) the α_j can take any value. It will often in fact be convenient to arrange for $\hat{\alpha}_0$ to be zero so that the 'zero' group can be regarded as a baseline group, as in the standard analysis. Note, however, that $\hat{\alpha}_0$ does have a well defined standard error in this formulation.

The augmented likelihood (8) can be interpreted in the following way. Instead of the Cox partial likelihood, one can consider a full Poisson likelihood with one parameter γ_k for each risk set:

$$\log L = \sum_{k=1}^{K} \gamma_k + \sum_{j=1}^{s-1} M_j \beta_j - \sum_{k=1}^{K} \sum_{j=1}^{s-1} n_{jk} \delta_k \exp(\gamma_k + \beta_j). \quad (13)$$

Here δ_k is the time period associated with risk set k. At first sight this does not appear very satisfactory, because the number of parameters may increase without limit with the number of deaths. However, Breslow[6] demonstrated that if this likelihood is maximized with respect to the γ_k's, one obtains (in the absence of ties) precisely the Cox likelihood (2), and therefore the two likelihoods lead to identical inferences. Here we wish to maximize (13) with respect to just $K - 1$ parameters, so that s can be retained, one for each group. The idea, loosely, is to maximize (13) with respect to the relative values of the parameters γ_k, in other words the *shape* of $\lambda_0(t)$, the baseline hazard function, but not the overall level of risk. Thus instead of the usual estimates for the γ_k's:

$$\exp(\hat{\gamma}_k) = 1 \Big/ \left(\delta_k \sum_j n_{kj} \exp \hat{\beta}_j \right)$$

we have

$$\exp(\hat{\gamma}_k) = \eta \Big/ \left(\delta_k \sum_j n_{kj} \exp \hat{\beta}_j \right)$$

where η is an arbitrary constant. When these terms are substituted in (13) they give the Cox likelihood plus the augmented term:

$$K \log \eta - K \eta.$$

There are various ways in which the γ_k's could be constrained, but one simple option is to choose η such that

$$\sum_k \left[\exp(\gamma_k)\delta_k \sum_j n_{jk} \exp \beta_j \right] = \sum_k \sum_j \delta_k n_{jk} \exp \beta_j$$

that is, such that the personyears weighted sum of the γ_k's is 1.

Then

$$\eta = \left[\sum_k \sum_j \delta_k \exp \beta_j \right] \Big/ K$$

and this gives the augmented log-likelihood (8).

Case-control analysis

Consider now the related problem of matched case-control analyses using logistic regression.[2] In a design where a single case is matched to a number of controls, the likelihood for logistic regression is identical to the partial likelihood (2), where each risk set now consists of the case and all its matched controls.[7] If the number of controls per case is large, the same augmentation to the likelihood may be made to provide approximately orthogonal estimates, although it can no longer be interpreted as a Poisson likelihood. Where the number of controls per case is small, it is of course not possible for each group to be equally represented in each risk set. Nevertheless, approximate independence should often still result. Consider the extreme case of a matched pair design, with m_{ij} pairs where the case is in group i and the control is in group j. The first term in the expression for an off diagonal element of the information matrix (12) is then

$$(\exp \alpha_i \exp \alpha_j) \frac{(m_{ij} + m_{ji})}{(\exp \alpha_i + \exp \alpha_j)^2} .$$

Suppose that, in fact, matching was unnecessary, and let η_j be the expected proportion of controls in exposure category j. Then the expectation of m_{ij} is

$$\frac{M\eta_i \eta_j \exp \alpha_i}{\sum_l \eta_l \exp \alpha_l}$$

so that for large numbers of case-control pairs the first term of (12) can be approximated by

$$\frac{M(\eta_i \exp \alpha_i)(\eta_j \exp \alpha_j)}{(\sum_l \eta_l \exp \alpha_l)(\exp \alpha_i + \exp \alpha_j)}$$

or approximately

$$\frac{M(N_i \exp \alpha_i)(N_j \exp \alpha_j)}{2(\sum_l N_l \exp \alpha_l)^2}$$

provided the α_j's are not too large. This is exactly half the second term in (12). We propose, therefore, that to achieve good independence between the α_j's in a 1–1 matched case-control analysis, the augmented term in the likelihood (6) and thus in the information matrix (11) and (12) should be multiplied by 1/2. Note that this is a perfectly legitimate manoeuvre, since the resulting parameter estimates $\hat{\alpha}_j$ and their covariances still give rise to the original $\hat{\beta}_j$'s upon subtracting $\hat{\alpha}_0$. In general the augmented terms should be multiplied by $C_k/(1 + C_k)$ where C_k is the number of controls in risk set k. This adjustment is usually of no consequence in survival analysis where risk sets are large, but can markedly improve independence of the $\hat{\alpha}_j$'s for matched pairs.

For a general stratified design with multiple cases and controls in any given stratum, the exact partial log-likelihoods are more complicated.[2] However, the same adjustment terms should apply, but with $C_k/(C_k + 1)$ replaced by $C_k/(C_k + D_k)$, where D_k is the number of cases in risk set k.

Additional covariates

If t additional covariates with parameters δ_i $(i = 1, \ldots, t)$ are included in the model, the 'expectation' E which augments the Poisson likelihood (6) is replaced by:

$$E = \lambda \sum_l r_l \exp(\alpha_{j(l)} + \sum_i Z_{il}\delta_i),$$

where the sum is over all individuals, r_l is the number of risk sets containing individual l, $j(l)$ is the category to which he belongs, and Z_{il} is his ith covariate measurement. This leads to straightforward corrections to the augmented terms for the information matrix in (11) and (12). It is necessary in this situation to arrange that the covariates have weighted mean zero over all risk sets and individuals, that is

$$\sum_l r_l Z_{il} \exp(\sum_j Z_{jl}\delta_j) = 0 \qquad (i = 1, \ldots, t)$$

by subtracting from the covariates the weighted mean:

$$\bar{Z}_i = \frac{\sum_l r_l Z_{il} \exp(\sum_j Z_{jl}\delta_j)}{\sum_l r_l \exp(\sum_j Z_{jl}\delta_j)}.$$

This ensures that the terms in the information matrix relating only to the δ_j's are not affected by augmenting the likelihood.

EXAMPLES

Small cell lung cancer

Vincent *et al.*[8] describe an analysis of prognostic factors in a series of 333 patients with small cell lung cancer. A number of important prognostic factors were identified including performance status defined by the WHO five point scale. An analysis of performance status is shown in Table I. (The group of 281 patients analysed here differs slightly from that presented by Vincent *et al.*[8]). In a standard analysis assuming proportional hazards there is a highly significant trend in risk with increasing performance status. However, the confidence limits for the relative risk, compared with the baseline category (zero performance status) overlap considerably. This unsatisfactory phenomenon is a consequence of the baseline category being a small group, and inflating the width of all the confidence intervals. The right hand column of Table I gives the standard errors and confidence intervals using the suggested heuristic approach. The covariance matrix shows that these estimates are nearly independent; this is true despite the marked effect of performance status on survival and the high overall death rate leading to substantial changes in the proportions in the different groups over time. The confidence limits for this analysis indicate that patients with performance status 2 and 3 clearly fare worse than those with performance status 1, which was not obvious from the standard analysis. The covariance matrix obtained using the augmented likelihood gives somewhat higher correlations than the heuristic approach for most of the estimates, although the standard errors and confidence limits are very similar.

Table I. Relative risk of death for small cell lung cancer patients by performance status (based on Vincent *et al.*[8])

Performance status	Number of patients	Number of deaths	Relative risk	Standard presentation 95% confidence limits	Suggested presentation* Standard error	Suggested presentation* 95% confidence limits
0	36	18	1·00		0·24	(0·63–1·60)
1	144	117	1·93	(1·17–3·17)	0·10	(1·59–2·34)
2	63	54	4·15	(2·41–7·13)	0·14	(3·17–5·43)
3	31	29	5·31	(2·91–9·67)	0·19	(3·67–7·67)
4	7	6	5·06	(2·00–12·83)	0·41	(2·27–11·29)

* Using the heuristic approach. Confidence limits obtained by the augmented likelihood approach are very similar.

Covariance matrix for standard analysis:

	1	2	3	4
1	0·0643			
2	0·0558	0·0763		
3	0·0560	0·0584	0·0940	
4	0·0558	0·0578	0·0586	0·2253

Covariance matrix using heuristic approach:

	0	1	2	3	4
0	0·0571				
1	0·0012	0·0096			
2	− 0·0003	− 0·0003	0·0187		
3	− 0·0006	− 0·0005	0·0005	0·0353	
4	− 0·0003	− 0·0004	0·0001	0·0006	0·1676

Covariance matrix using augmented likelihood approach:

	0	1	2	3	4
0	0·0582				
1	0·0022	0·0105			
2	− 0·0003	− 0·0004	0·0176		
3	− 0·0013	− 0·0013	− 0·0014	0·0331	
4	− 0·0001	− 0·0002	− 0·0008	− 0·0011	0·1668

Oral contraceptives and breast cancer

The U.K. National Case-Control Study Group[9] conducted a matched case-control study of oral contraceptive use and breast cancer risk in women aged under 36 years. 755 cases and individually matched population based controls were studied. Table II shows the effect of duration of oral contraceptive use on breast cancer risk, after allowing for five possible confounding variables in a conditional logistic regression analysis. There is a highly significant ($p < 0·001$) trend in risk with increasing duration of use of oral contraceptives, but again the 95 per cent confidence limits for the relative risks compared with the baseline category (never used) overlap considerably. The confidence limits give the impression that the three categories of pill use do not differ significantly; they all appear consistent with a relative risk of 1·3 compared with non-users, but this is an artefact of the small size of the baseline category. The right hand column

Table II. Relative risk of breast cancer by duration of oral contraceptive use (from U.K. National Case Control Study Group[9])

Duration of use (months)	Number of cases	Number of controls	Odds ratio*	Standard presentation 95% confidence limits	Suggested presentation† Standard error	Suggested presentation† 95% confidence limits
Never	67	80	1·00		0·18	(0·71–1·41)
1–48	218	285	0·95	(0·64–1·41)	0·10	(0·79–1·15)
49–96	272	247	1·43	(0·97–2·12)	0·09	(1·20–1·71)
97+	198	143	1·74	(1·15–2·62)	0·10	(1·39–2·18)

* Adjusted for age, age at menarche, nulliparity, age at first full-term pregnancy, breast feeding and family history of breast cancer.
† Using the heuristic approach. Confidence limits obtained by the augmented likelihood approach are very similar.

Covariance matrix for standard analysis:

	1–48	49–96	97+
1–48	0·04081		
49–96	0·03184	0·04003	
97+	0·03087	0·03119	0·04390

Covariance matrix using heuristic approach:

	Never	1–48	49–96	97+
Never	0·0313			
1–48	−0·0001	0·0094		
49–96	−0·0002	0·0003	0·0083	
97+	0·0003	0·0002	−0·0001	0·0131

Covariance matrix using augmented likelihood approach:

	Never	1–48	49–96	97+
Never	0·0321			
1–48	0·0009	0·0105		
49–96	0.0000	0·0006	0·0079	
97+	−0·0004	−0·0007	−0·0013	0·0110

of Table II shows the confidence limits given by the suggested method. The risk associated with 1–48 months can be seen to be significantly less than that for 49–96 or 97 + months use. Note again the covariance matrix for the new analysis, indicating the near independence of the parameter estimates, in marked contrast to the standard analysis. The augmented likelihood method gives very similar confidence limits, but the majority of the correlations are somewhat larger than those given by the heuristic approach.

CONCLUSIONS

Our parameter estimates with their associated confidence intervals, which we suggest should be referred to as 'floating absolute risk' (FAR) estimates, are superficially similar to conventional relative risks. Indeed, the parameter estimates are identical, and the variance–covariance matrices can be regarded as transformations of one another. To avoid confusion, we recommend that FAR

confidence limits should be quoted in addition to conventional relative risk confidence intervals, as in Tables I and II.

The above results have all been described assuming a log-linear relative risk function, but in principle similar methods should apply to general relative risk functions.[10] We have not formally examined the effect on the covariances of variation between risk sets or strata in the proportion of individuals in each category; but the correlations were small in the data sets we have analysed, even when the proportions varied markedly. This approximate independence of the estimates is an important property, since it enables standard errors and confidence limits to describe their uncertainty without the need for covariances which are, in practice, not given.

For practical purposes the 'heuristic' approach seems clearly preferable to the augmented likelihood method of analysis. The results are virtually identical, but the heuristic method, by its definition, gives lower average covariances between estimates, and is algebraically trivial.

A fundamental statistical principle, that a satisfactory summary of the data should include the sufficient statistics, underlies the inadequacy of the common practice of describing data on s categories by $s - 1$ relative risks and their standard errors, a total of $2s - 2$ parameters. Subject to the usual asymptotic normal approximation, in cases where the analysis recommended here yields s exactly independent relative risk estimates, there are $2s - 1$ sufficient statistics ($s - 1$ relative risks, their standard errors and the standard error of the baseline group). In most situations, the recommended estimates are almost independent and these $2s - 1$ statistics can perhaps be described as 'virtually sufficient'.

ACKNOWLEDGEMENTS

We thank members of the U.K. National Case Control Study Group and Dr. Ian Smith for allowing us to use their data as examples, and Sandra McVeigh and Rosemary Booth for typing this manuscript. The Institute of Cancer Research receives support from the Cancer Research Campaign and the Medical Research Council.

REFERENCES

1. Baker, R. J. and Nelder, J. A. *The GLIM system: Release 3*, Numerical Algorithms Group, Oxford, 1978.
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, Lyon, 1980.
3. Cox, D. R. 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
4. Peto, R. 'Contribution to discussion of paper by D. R. Cox', *Journal of the Royal Statistical Society, Series B*, **34**, 205–207 (1972).
5. Storer, B. E., Wacholder, S. and Breslow, N. E. 'The maximum likelihood fitting of general risk models to stratified data', *Applied Statistics*, **32**, 172–181 (1983).
6. Breslow, N. E. 'Covariance analysis of censored survival data', *Biometrics*, **39**, 173–184 (1974).
7. Prentice, R. E. and Breslow, N. E. 'Retrospective studies and failure time models', *Biometrics*, **65**, 153–158 (1978).
8. Vincent, M. D., Ashley, S. E. and Smith, I. E. 'Prognostic factors in small cell lung cancer: A simple prognostic index is better than conventional staging', *European Journal of Cancer and Clinical Oncology*, **23**, 1598–1599 (1987).
9. U.K. National Case-Control Study Group. 'Oral contraceptive use and breast cancer risk in young women', *Lancet*, **i**, 973–982 (1989).
10. Thomas, D. C. 'General relative risk models for survival time and matched case-control analysis', *Biometrics*, **37**, 673–686 (1981).