Biometrika (1986). **73**, 1. *pp.* 1–11
Printed in Great Britain

A case-cohort design for epidemiologic cohort studies and disease prevention trials

BY R. L. PRENTICE
Fred Hutchinson Cancer Research Center, Seattle, Washington 98104, U.S.A.

SUMMARY

Suppose that a cohort of individuals is to be followed in order to relate failure rates to preceding covariate histories. A design is proposed which involves covariate data only for cases experiencing failure and for members of a randomly selected subcohort. Odds ratio and relative risk estimation procedures are presented for such a 'case-cohort' design. A small simulation study compares case-cohort relative risk estimation procedures to full-cohort and synthetic case-control analyses. Relevance to epidemiologic cohort studies and disease prevention trials is discussed.

Some key words: Case-control study; Cohort study; Epidemiology; Failure time data; Odds ratio; Prevention trial; Relative risk.

1. INTRODUCTION

Epidemiologic cohort studies and disease prevention trials typically require the follow-up of several thousand subjects for a number of years before yielding useful results, and hence can be prohibitively expensive. For example, the recently completed multiple risk factor intervention trial (MRFIT Research Group, 1982) randomized 12,866 subjects, reported results after an average follow-up of seven years and reportedly cost in excess of US$100 million.

Much cost and effort in such studies relates to the analysis of raw materials in order to assemble covariate histories for individual cohort members. Such assembly may involve, for example, the biochemical analysis of blood samples or other specimens, or the hand coding of individual diet records. Because of a usual low rate of disease occurrence, for example only 2% of MRFIT men experienced the primary endpoint of coronary heart disease mortality, much of the covariate information on disease free subjects is largely redundant.

A synthetic case-control design can be used to reduce the number of subjects for whom covariate data are required (Mantel, 1973; Liddell, McDonald & Thomas, 1977; Prentice & Breslow, 1978; Breslow & Patton, 1979; Oakes, 1981; Thomas, 1981; Whittemore, 1981; Whittemore & McMillan, 1982; Breslow et al., 1983; Lubin & Gail, 1984). In this approach each subject developing disease is matched to one or more subjects without disease at the same point in 'time', whence relative risks are estimated using a matched case-control analysis. Covariate histories are required only for cases and their matched controls. Prentice & Breslow (1978) considered sampling from a 'large' practically infinite cohort in which the matched sets at distinct failure times could be assumed disjoint, in which circumstance the case-control analysis possesses a conditional likelihood interpretation. Oakes (1981) noted that the case-control analysis possesses a partial likelihood analysis more generally, provided controls are sampled independently and randomly from the risk sets at distinct failure times.

There are several reasons for considering alternatives to such a 'case-control within a cohort' design. Intuitively the alignment of each selected control subject to its matched case seems inefficient, since that subject may also properly serve as a member of the comparison group for cases occurring at a range of other times. Also, a strict application of the time matched case-control approach would involve the selection of a new set of controls for each distinct disease category under study, whereas intuitively a single comparison group should suffice, as in full-cohort analyses. Finally, in the context of a disease prevention trial, it is desirable to have a subset of the trial cohort for whom covariate data are analysed on an ongoing basis in order to monitor intervention effectiveness and compliance. The case-control approach is not well suited to this purpose since covariate histories are only assembled following case occurrence.

A case-cohort design is proposed to address these issues. This design involves the selection of a random sample, or a stratified random sample, of the entire cohort, and the assembly of covariate histories only for this random subcohort and for all cases. The subcohort in a given stratum constitutes the comparison set of cases occurring at a range of failure times. The subcohort also provides a basis for covariate monitoring during the course of cohort follow-up.

Kupper, McMichael & Spirtas (1975) and Miettinen (1982) suggest a very similar approach under the labels 'hybrid retrospective design' and 'case-base' design, respectively. Both restricted their attention to a binary failure indicator and a binary covariate. The proposed estimation procedures of Kupper et al. (1975) are not applicable to inference on population parameters, since they considered only the randomness introduced by subcohort sampling. Miettinen, on the other hand, proposed a simple estimation procedure for the ratio of disease probabilities at the two covariate values; that is, for the population 'risk ratio'.

General methods for asymptotic odds ratio and relative risk estimation under a case-cohort design are given in §§ 2 and 3, respectively. A small simulation study is reported in § 4 and further discussion follows in § 5.

2. THE CASE-COHORT DESIGN: BINARY RESPONSE

Although our main interest centres on relative risk estimation it is instructive to begin with a short discussion of odds ratio estimation, based on the follow-up of a cohort of size n to observe whether, $D = 1$, or not, $D = 0$, failure, e.g. disease, occurs during a specified time period.

Suppose initially that one is interested in the dependence of failure probability on the presence, $z = 1$, or absence, $z = 0$, of some covariate. Denote $p_{ij} = \text{pr}(D = i, z = j)$ $(i, j = 0, 1)$. Assuming no censoring a conventional cohort approach would involve observation of the number of failures d_0 and d_1, and the number of subjects n_0 and n_1, corresponding to $z = 0$ and $z = 1$, respectively. The maximum likelihood estimate of the odds ratio $\lambda = p_{11}p_{00}(p_{10}p_{01})^{-1}$ is then

$$\hat{\lambda} = d_1(n_0 - d_0)\{d_0(n_1 - d_1)\}^{-1}.$$

Quite generally $\hat{\beta} = \log \hat{\lambda}$ has asymptotic variance consistently estimated by $d_0^{-1} + d_1^{-1} + (n_0 - d_0)^{-1} + (n_1 - d_1)^{-1}$.

Suppose now that the entire cohort is monitored for failure as before, but that covariate values are assembled only for a randomly selected subcohort of size $m \leqslant n$, and for failing subjects. Upon reparameterizing so that $p_{00} = p\alpha$, $p_{01} = p(1-\alpha)$, and noting

that $p = 1-p_{10}-p_{11}$, we may write the likelihood function for such case-cohort data as

$$p_{10}^{d_0} p_{11}^{d_1}(1-p_{10}-p_{11})^{n-d}\alpha^{m_0-k_0}(1-\alpha)^{m_1-k_1},$$

where (m_0, k_0) and (m_1, k_1) are the numbers of subjects and cases, i.e. failures, corresponding to $z = 0$ and $z = 1$, respectively, in the randomly selected subcohort and $d = d_0+d_1$. It follows easily that $\hat{p}_{10} = d_0 n^{-1}$, $\hat{p}_{11} = d_1 n^{-1}$ and $\hat{\alpha} = (m_0-k_0)(m-k)^{-1}$, where $k = k_0+k_1$, so that the maximum likelihood estimate of the odds ratio is $\hat{\lambda} = \{d_1(m_0-k_0)\}\{d_0(m_1-k_1)\}^{-1}$. Application of standard asymptotic likelihood formulae then shows $\hat{\beta} = \log\hat{\lambda}$ to have asymptotic variance consistently estimated by $d_0^{-1}+d_1^{-1}+(m_0-k_0)^{-1}+(m_1-k_1)^{-1}$. It follows that the case-cohort odds ratio estimate will have good efficiency properties whenever m_0-k_0 and m_1-k_1 are large in comparison to the smaller of d_0 and d_1, with high probability.

Suppose now that the covariate z is an arbitrary row p-vector. An odds ratio model

$$\mathrm{pr}(D=1\,|\,z)\,\mathrm{pr}(D=0\,|\,z=0)\{\mathrm{pr}(D=1\,|\,z=0)\,\mathrm{pr}(D=0\,|\,z)\}^{-1} = \exp(z\beta)$$

with $p\times 1$ column vector β is equivalent to a binary logistic failure probability model

$$\mathrm{pr}(D\,|\,z) = \exp\{(\alpha+z\beta)D\}/\{1+\exp(\alpha+z\beta)\} \tag{1}$$

and to a logistic-type covariate density model

$$\mathrm{pr}(z\,|\,D) = C_D\exp\{\gamma(z)+z\beta D\} \tag{2}$$

(Prentice & Pyke, 1979), where α is a scalar parameter, $\gamma(.)$ is a nuisance function and C_D is an integration constant. Suppose now that a subcohort of size m is randomly selected from the full cohort of size n and that z-values are assembled for subcohort members and for all cases. Denote by z_{ij} $(j = 1, \ldots, s_1)$ the covariate vectors for the $s_1 = d$ failing subjects and by z_{0j} $(j = 1, \ldots, s_0)$ the covariate vectors for the $s_0 = m-k$ subcohort members who turn out not to fail. If we use (2) the overall likelihood function can be written

$$\left\{\prod_{i=0}^{1}\prod_{j=1}^{s_i}\mathrm{pr}(z_{ij}\,|\,D=i)\,\mathrm{pr}(D=i)\right\}\mathrm{pr}(D=0)^{n-s_0-s_1}$$

$$= \left\{\prod_{i=0}^{1}\prod_{j=1}^{s_i}\mathrm{pr}(z_{ij}\,|\,D=i)\right\}q^{s_1}(1-q)^{n-s_1} \tag{3}$$

$$= \left[\prod_{i=0}^{1}\prod_{j=1}^{s_i}C_i\exp\{\gamma(z_{ij})+iz_{ij}\beta\}\right]q^{s_1}(1-q)^{n-s_1}, \tag{4}$$

where $q = 1-p = \mathrm{pr}(D=1)$. The first factor in (3) is a conditional likelihood for z-values in the sample, given their corresponding D-values. Under (2) it is precisely the likelihood maximized by Prentice & Pyke (1979). It follows from their work that given D the maximum conditional likelihood estimate of β can be obtained by applying the prospective disease probability model (1) to the s_0+s_1 cases and noncases as if a prospective study had been conducted. It also follows directly, under weak conditions, that a variance estimator for this conditional maximum likelihood estimator is given by the corresponding $p\times p$ submatrix of the inverse observed information matrix in such a prospective application of (1). Finally, since there are no constraints to link the parameter q in (4) to the parameters $\{\gamma(.), \beta\}$, it is clear that the maximum conditional likelihood estimate $\hat{\beta}$ is also the overall maximum likelihood estimate. In summary,

asymptotic inference on the odds ratio in a case-cohort study can be carried out by applying the logistic model (1) directly to the s_0+s_1 subjects for whom covariate data is assembled. The case-cohort data also provide a natural estimator $\hat{q}=s_1/n$ of the marginal disease probability $q=\text{pr}(D=1)$, but information on q is not, in itself, useful for large sample odds ratio estimation. It is straightforward to permit the parameters in (1) and the subcohort selection to be stratified on baseline characteristics.

3. The case-cohort design: Time to response data

A number of generalizations of the above formulation will be simultaneously considered. These include use of the actual times of failure for cases; the replacement of odds ratios by relative risks; the allowance for late entry into the cohort, censorship and even intermittent exclusion from the cohort risk set; and a relaxation to allow nonexponential relative risk forms. Allowance for stratification on baseline covariates will be deferred for notational convenience.

Let $Z(t)$ denote a covariate measurement on a subject at time t. Here time can be thought of as time since the beginning of cohort follow-up, though other specifications, such as subject age, may be more suitable in some applications. Let $\lambda\{t;\, Z(u),\, 0\leqslant u<t\}$ denote the failure rate of interest at time t for a subject with preceding covariate history $\{Z(u);\, 0\leqslant u<t\}$. Consider a relative risk regression model (Cox, 1972)

$$\lambda\{t;\, Z(u),\, 0\leqslant u<t\}=\lambda_0(t)\, r\{X(t)\,\beta\},$$

where $r(x)$ is a fixed function with $r(0)=1$, for example $r(x)=e^x$ or $r(x)=1+x$; $X(t)$ is a row p-vector consisting of functions of $\{Z(u);\, 0\leqslant u<t\}$ and possibly product terms between such functions and t; β is a column p-vector of regression parameters to be estimated; and $\lambda_0(t)$ is a baseline hazard function corresponding to a standard covariate history for which the modelled regression vector is $X(t)\equiv 0$.

Consider now a cohort of size n. Let $\{N_i(u),\, Y_i(u), Z_i(u);\, 0\leqslant u<t\}$ denote counting, censoring and covariate histories for the ith subject prior to time t. Specifically N_i, with right-continuous sample paths, takes value zero prior to an observed failure on the ith subject and value one thereafter, while Y_i, with left continuous sample paths, takes value one at times at which the ith subject is at risk for failure and value zero otherwise. Sample paths for the modelled covariate X_i are required to be left continuous with right-hand limits. With this notation one can specify the time t_i of failure or censorship for the ith subject as $t_i=\min\{t\,|\,Y_i(u)=0;\ \text{all } u>t\}$, while the censoring indicator δ_i takes value one if $N_i(t_i)\neq N_i(t_i^-)$ and value zero otherwise. Under standard independent failure time and independent censorship assumptions and full cohort data, a partial likelihood function (Cox, 1975) can be written

$$L(\beta)=\prod_{i=1}^{n}\left(r_{ii}\Big/\sum_{l=1}^{n}r_{li}\right)^{\delta_i}, \tag{5}$$

where $r_{li}=Y_l(t_i)\, r\{X_l(t_i)\,\beta\}$.

Suppose now that a random subcohort C of size m is selected from the entire cohort. Suppose further that $\{N_i,\, Y_i\}$ processes are available for all cohort members, as above, but that corresponding covariate histories are available only for members of C and for subjects that fail. More specifically, let $K(t)=\{i\,|\,N_i(t)=1\}$ denote the set of subjects failing at or before time t. Covariate histories at time t will be assumed available only for subjects in $M(t)=K(t)\cup C$. Also write $D(t)=\{i\,|\,N_i(t)\neq N_i(t^-)\}$, so that $D(t)$ is empty unless a failure occurs at time t, and let $\tilde{R}(t)=D(t)\cup C$. Finally let $\Delta(t)$ equal one if

$\tilde{R}(t) \neq C$ and value zero otherwise. Therefore $\Delta(t)$ takes value one at a time of failure only if failure occurs outside the subcohort.

For estimation of the relative risk parameter β using such case-cohort data consider maximizing the function

$$\tilde{L}(\beta) = \prod_{i=1}^{n} (r_{ii} / \sum_{l \in \tilde{R}(t_i)} r_{li})^{\delta_i}, \tag{6}$$

which differs from (5) only in that the ith denominator factor is a sum over subjects at risk in $\tilde{R}(t_i)$ rather than over subjects at risk in the entire cohort. Expression (6) does not generally possess a partial likelihood interpretation, and hence will be termed a pseudolikelihood; see Besag (1977) for a similar usage.

The maximum pseudolikelihood estimate $\hat{\beta}$ is defined by $U(\hat{\beta}) = 0$, where

$$U(\beta) = \partial \log \tilde{L}(\beta)/\partial\beta = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} \delta_i (c_{ii} - \sum_{l \in \tilde{R}(t_i)} b_{li} / \sum_{l \in \tilde{R}(t_i)} r_{li}),$$

and where $b_{li} = Y_l(t_i)\, X_l(t_i)\, r'\{X_l(t_i)\,\beta\}$, $c_{li} = b_{li}\, r^{-1}\{X_l(t_i)\,\beta\}$ and $r'(u) = dr(u)/du$. Asymptotic properties of $\hat{\beta}$ will derive from those of the score statistic $U(\beta)$.

Write $\mathscr{F}(t)$ for $\tilde{R}(t)$, $\Delta(t)$ and all available counting, censoring and covariate information up to and including time t, with the exception of the values $N(t) = \{N_1(t), \ldots, N_n(t)\}$ and $\{Z_i(t),\, i \in M(t)\}$. Note that $\{\mathscr{F}(t), t \geqslant 0\}$ does not form a nested sequence of σ-algebras since $\mathscr{F}(t)$ does not specify $\{\tilde{R}(u), \Delta(u), u < t\}$. At a failure time t one can calculate, using standard independent failure and censorship assumptions, that

$$\mathrm{pr}\{N_i(t) \neq N_i(t^-) \mid \mathscr{F}(t), N(t) \neq N(t^-)\} = Y_i(t)\, r\{X_i(t)\,\beta\} / \sum_{l \in \tilde{R}(t)} Y_l(t)\, r\{X_l(t)\,\beta\} \tag{7}$$

for any $i \in \tilde{R}(t)$. At any uncensored failure time t_j expression (7) immediately gives

$$\begin{aligned} E\{U_j(\beta) \mid \mathscr{F}(t_j), N(t_j) \neq N(t_j^-)\} &= (\sum_{i \in \tilde{R}(t_j)} c_{ij} r_{ij} - \sum_{l \in \tilde{R}(t_j)} b_{lj})\, R_j^{-1} = 0, \\ \mathrm{var}\{U_j(\beta) \mid \mathscr{F}(t_j), N(t_j) \neq N(t_j^-)\} &= (\sum_{i \in \tilde{R}(t_j)} c'_{ij} c_{ij}\, R_j^{-1} - B'_j B_j\, R_j^{-2}) = v_{jj}, \end{aligned} \tag{8}$$

where

$$R_j = \sum_{l \in \tilde{R}(t_j)} r_{lj}, \quad B_j = \sum_{l \in \tilde{R}(t_j)} b_{lj}.$$

Therefore each $U_i(\beta)$ $(i = 1, \ldots, n)$ and hence $U(\beta)$ also, has mean zero. Also write

$$\mathrm{var}\{U(\beta)\} = \sum_{j=1}^{n} [\mathrm{var}\{U_j(\beta)\} + 2 \sum_{\{k \mid t_k < t_j\}} \mathrm{cov}\{U_k(\beta), U_j(\beta)\}],$$

and consider $\mathrm{cov}\{U_k(\beta), U_j(\beta) \mid \mathscr{F}(t_j), N(t_j) \neq N(t_j^-)\}$ for (k, j) such that $\delta_k = \delta_j = 1$ and $t_k < t_j$. Suppose first that $\Delta(t_j) = 0$. This implies $C = \tilde{R}(t_j)$, so that

$$\tilde{R}(t_k) = \{l \mid N_l(t_k) \neq N_l(t_k^-)\} \cup \tilde{R}(t_j).$$

Hence both $\tilde{R}(t_k)$ and $U_k(\beta)$ are fixed and

$$E\{U_k(\beta)\, U_j(\beta) \mid \mathscr{F}(t_j), N(t_j) \neq N(t_j^-)\} = U_k(\beta)\, E\{U_j(\beta) \mid \mathscr{F}(t_j), N(t_j) \neq N(t_j^-)\} = 0.$$

On the other hand, if $\Delta(t_j) = 1$ then $C = \tilde{R}(t_j) - \{i\}$ with conditional probability (7) for any, $i \in \tilde{R}(t_j)$. Hence,

$$\tilde{R}(t_k) = \{l \mid N_l(t_k) \neq N_l(t_k^-)\} \cup [\tilde{R}(t_j) - \{i\}]$$

with conditional probability (7), for any $i \in \tilde{R}(t_j)$. The desired covariance can now be written using realized quantities $\{c_{kk}, B_k, R_k\}$ at t_k as

$$E\{U_k(\beta)\, U_j(\beta) \mid \mathscr{F}(t_j), N(t_j) \neq N(t_j^-)\} = \Sigma\left(c_{kk} - \frac{B_k + b_{jk} - b_{ik}}{R_k + r_{jk} - r_{ik}}\right)'\left(c_{ij} - \frac{B_j}{R_j}\right) r_{ij}\, R_j^{-1}$$

$$= -\Sigma\left(\frac{B_k + b_{jk} - b_{ik}}{R_k + r_{jk} - r_{ik}}\right)'\left(c_{ij} - \frac{B_j}{R_j}\right) r_{ij}\, R_j^{-1}$$

$$= v_{kj}, \tag{9}$$

where the sums are over $i \in \tilde{R}(t_j)$.

Hence, combining (8) and (9), one obtains as variance estimator for $U(\beta)$

$$\tilde{V}(\beta) = \sum_{j=1}^{n} \delta_j \{v_{jj} + 2\Delta(t_j) \sum_{\{k \mid t_k < t_j\}} \delta_k\, v_{kj}\}. \tag{10}$$

One can now argue informally to assert an asymptotic normal distribution for $n^{\frac{1}{2}}(\hat{\beta} - \beta)$. Specifically, a Taylor expansion about the true β evaluated at $\hat{\beta}$ gives

$$n^{-\frac{1}{2}}\, U(\beta) = n^{-1}\, I(\beta_*)\, n^{\frac{1}{2}}(\hat{\beta} - \beta)$$

for β_* between β and $\hat{\beta}$. Sufficient conditions then need to be introduced to require $n^{-\frac{1}{2}}\, U(\beta)$ to converge weakly to a normal variate with mean zero and variance matrix A, and to require $n^{-1}\, I(\beta_*) = -n^{-1}\, \partial^2 \log \tilde{L}(\beta_*)/\partial \beta_*^2$ to consistently estimate a positive-definite matrix Ω for any random β_* between β and $\hat{\beta}$. It will then follow that $n^{\frac{1}{2}}(\hat{\beta} - \beta)$ converges in distribution to a normal variate with mean zero and variance matrix $S = \Omega^{-1} A \Omega^{-1}$, whence it is only necessary to show that $n^{-1}\, \tilde{V}(\hat{\beta})$ consistently estimates A in order to ensure that $nI(\hat{\beta})^{-1}\, \tilde{V}(\hat{\beta})\, I(\hat{\beta})^{-1}$ is a consistent estimator of S.

The martingale convergence results used by Andersen & Gill (1982) and Prentice & Self (1983) for the estimator that maximizes (5) do not immediately apply here in view of the lack of nesting of the $\{\mathscr{F}(t), t \geqslant 0\}$ and the corresponding slight correlations among score statistic contributions at distinct failure times. My colleague Dr Steven Self has, however, noted that the score statistic can quite generally be represented as the sum of a martingale and an asymptotically uncorrelated term to which finite population asymptotic results apply. This representation leads to flexible sufficient conditions for the desired asymptotic normality, but the arguments are rather detailed and will be presented elsewhere. Note, however, that such asymptotic results naturally require the ratio mn^{-1}, of subcohort size to full cohort size, to converge to a positive constant.

A natural estimator of the cumulative baseline failure rate

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$$

can be written

$$\hat{\Lambda}_0(t) = mn^{-1} \int_0^t [\sum_{l \in C} Y_i(w)\, r\{X_l(w)\hat{\beta}\}]^{-1}\, d\bar{N}(w),$$

where $\bar{N} = N_1 + \ldots + N_n$. Note that this estimator, in contrast to the above relative risk estimation procedure, explicitly involves the full cohort size n.

With a limited degree of failure time grouping one may generalize (6) in the approximate manner of Peto (1972) and Breslow (1974), merely by defining $\tilde{R}(t)$ to

consist of the union of C with the set of all subjects with failure time equal to t. Expression (10) corresponds naturally to such a tied data approximation with a distinct $\Delta(t)$, defined to take value one if the subject is outside the subcohort, and value zero otherwise, for each subject failing at t.

Suppose now that baseline data available for the entire cohort are used to partition the cohort into q strata, and that a relative risk regression model

$$\lambda_s\{t; Z(u), 0 \leqslant u < t\} = \lambda_{0s}(t)\, r\{X(t)\, \beta_s\}$$

is specified for the disease incidence rate in each stratum. A case-cohort approach to the estimation of $\beta = (\beta_1', \ldots, \beta_q')$ would involve the selection of a subcohort from each stratum and the assembly of covariate histories for cases and subcohort members. The subcohort sampling rates can be allowed to vary among strata. A pseudolikelihood function for β can be written as a product of terms (6) over strata. The corresponding score statistic has mean zero and variance estimated by the sum over strata of matrices (10).

4. Simulation

A small simulation study was conducted to examine the performance of the estimation procedure of § 3 with a moderate number of failures, and in order to compare such performance with that of full-cohort and synthetic case-control estimation procedures. A cohort size of $n = 500$ was selected, and exponentially-distributed failure times, censored at unity, were generated corresponding to a binary covariate which took each of values zero and one for 250 cohort members. The exponential failure rate parameter was selected to give 50 expected failures. One hundred such samples were generated at each of relative risks one and two for the binary covariate. Several analyses were carried out for each sample: the maximum partial likelihood estimate and related quantities were calculated using (5); the maximum pseudolikelihood estimate and related quantities were obtained using (6) and (10), both for randomly selected subcohorts of size 55 and 275; and maximum partial likelihood estimates and related quantities were obtained for synthetic case-control analyses based on either one or five time-matched controls per case. In fact, two synthetic case-control estimators were calculated from each sample. The standard case-control estimator, denoted case-control I, involves the selection of controls randomly from the entire risk set at each failure time, with control selection at a specific failure time independent of that at any other failure time. The second estimator, denoted case-control II, excludes future cases from the control selection and allows a given censored subject to serve as control at at most one failure time. This latter estimator is known to be biased for $\beta \neq 0$ (Lubin & Gail, 1984) but may be expected to possess slightly better properties than the former at $\beta = 0$.

Table 1 gives summary statistics from those simulations. Convergence was achieved in all cases. The upper half of Table 1 gives results at a relative risk of unity ($\beta = 0$). There is no evidence of bias in any of the log relative risk estimates $\hat{\beta}$. Only the 1–1 matched case-control I estimator had a sample mean that differed by more than one sample standard error from zero. Likewise significance levels for testing $\beta = 0$, based on a standard normal distribution for $\hat{\beta}$ divided by its estimated standard error, are all within sampling variation of their nominal values, with the possible exception the 1–1 matched case-control I test for which the estimated 5% significance level exceeds its nominal value by about two standard errors. This point merits further study. With only 100

Table 1. *Simulation summary statistics for various methods of estimating the logarithm of the relative risk* (β).

Estimation procedure	Full-cohort	Case-cohort	Case-control I	Case-control II	Case-cohort	Case-control I	Case-control II
Expected subjects requiring covariate data	500	100	100	100	300	300	300
	(a) Relative risk = 1 ($\beta = 0$)						
Sample mean ($\hat{\beta}$)	−0·031	−0·006	−0·064	−0·034	−0·021	−0·034	−0·365
Sample standard error ($\hat{\beta}$)	0·335	0·412 (23)	0·484 (44)	0·416 (24)	0·348 (4)	0·364 (9)	0·371 (11)
Mean of standard error estimates	0·286	0·389 (36)	0·420 (47)	0·401 (40)	0·304 (7)	0·312 (9)	0·312 (9)
Estimated significance level* ($\alpha = 0{\cdot}10$)	0·10	0·10	0·12	0·08	0·13	0·12	0·13
Estimated significance level* ($\alpha = 0{\cdot}05$)	0·08	0·06	0·10	0·04	0·08	0·09	0·08
	(b) Relative risk = 2 ($\beta = 0{\cdot}693$)						
Sample mean ($\hat{\beta}$)	0·731	0·676	0·748	0·798	0·728	0·729	0·759
Sample standard error ($\hat{\beta}$)	0·258	0·337 (31)	0·417 (62)	0·418 (62)	0·269 (4)	0·282 (9)	0·282 (9)
Mean of standard error estimates	0·303	0·399 (32)	0·444 (47)	0·445 (47)	0·318 (5)	0·330 (9)	0·329 (9)
Estimated rejection probability* ($\alpha = 0{\cdot}10$)	0·84	0·56	0·50	—	0·78	0·78	—
Estimated rejection probability* ($\alpha = 0{\cdot}05$)	0·72	0·33	0·39	—	0·63	0·63	—

Percentage increase over full-cohort value given in parentheses.

* Fraction of samples in which $|\hat{\beta}/(\text{est. std. err. } \hat{\beta})|$ exceeds 1·65 ($\alpha = 0{\cdot}10$) or 1·96 ($\alpha = 0{\cdot}05$).

samples there is considerable random variation in the sample standard errors for $\hat{\beta}$. Upon accounting for such randomness one can note that sample standard errors are in good agreement with the corresponding mean of the standard error estimates, again with the possible exception of the 1–1 matched case-control I procedure where the sample standard error, 0·484, exceeds the corresponding mean, 0·420, by about two estimated standard errors. The sample standard errors are in fairly close agreement between the case-cohort and case-control procedures at a specified expected number, 100 or 300, of subjects, though the 1–1 matched case-control I estimator has a somewhat larger sample standard error than does its competitors. The sample standard errors for analyses involving 300 subjects are within 11% of the full cohort sample standard errors for all methods.

The situation is rather different at a relative risk of two, $\beta = 0{\cdot}693$. First the bias induced by excluding future cases in the choice of matched controls is evident in that the sample mean of $\hat{\beta}$ for the 1–1 matched case-control II procedure is 2·5 estimated standard errors greater than β, while that for the 5–1 matched case-control II procedure is 2·3 estimated standard errors greater than β. More importantly, the case-cohort sample standard errors are about midway between the full-cohort and corresponding case-control I or II sample standard errors, both at 100 and 300 expected subjects. This reduction in standard error is practically important at 100 expected subjects. With 300 expected subjects, however, both case-cohort and case-control sample standard errors are within 10% of the full-cohort sample standard errors. The final two rows of Table 1 indicate that $\hat{\beta}$ divided by its standard error gives a test for $\beta = 0$ for which the empirical power based on a case-cohort or a case-control I sample of expected size 300 approaches that of the full-cohort analysis. Power estimates are not given for the case-control II estimator in view of the bias in $\hat{\beta}$. The close agreement between the case-cohort and case-control I empirical powers seems surprising in view of the rather different sample standard errors for $\hat{\beta}$. Such agreement may be a simulation artifact corresponding to the fact that samples with a large number of failures involve a larger comparison group for the case-control than for the case-cohort procedure.

A few other observations on the simulation are as follows: the standard error of the difference between $\hat{\beta}$ and the corresponding full-cohort β estimator was approximately 0·25 and 0·10 for the case-cohort estimator at 100 and 300 expected subjects, respectively. The corresponding differences had similar estimated standard errors for the case-control estimators, except for the 1–1 matched case-control estimators at a relative risk of 2, for which the estimated standard errors were somewhat larger, about 0·32. The standard error estimates for $\hat{\beta}$ were quite stable for all five estimators. Specifically the sample standard errors of these standard error estimates were somewhat larger for the case-control than for the case-cohort or full-cohort estimators. For a given sample, the estimated information matrices $-\partial^2 \log L(\hat{\beta})/\partial\hat{\beta}^2$ were consistently virtually identical under (5) and (6).

Whittemore & McMillan (1982), drawing on results reported by Whittemore (1981) and by Breslow et al. (1983), noted that the efficiency of a time-matched case-control design agreed with that of an unmatched case-control design at a relative risk of unity, but that the efficiency of the time-matched design compared poorly to that for the unmatched design if relative risk departed substantially from zero. In fact, a comparison with unmatched case-control results, or equivalently, with unmatched case-cohort results, § 2, gives additional insight into Table 1. Specifically, at a relative risk of unity asymptotic standard errors for the maximum likelihood estimate of the log odds ratio are 0·298, 0·400 and 0·310 under full-cohort, case-cohort 100 expected, and case-cohort 300

expected, respectively. These numbers agree well with the corresponding entries in the upper part of Table 1. The corresponding asymptotic standard errors for the log odds ratio estimator at a relative risk of two are 0·313, 0·411 and 0·324, respectively. Note that the percentage increase of the latter two compared to the first are 31% and 5%, respectively, virtually identical to the corresponding case-cohort to full-cohort standard error increments in Table 1. It then seems sensible to interpret Table 1 as indicating that the poor efficiency properties of the synthetic case-control design for the estimation of relative risks different from one, as compared to the corresponding unmatched case-control design for odds ratio estimation, are attributable to the somewhat artificial alignment of a control subject with only a single case. It appears that this loss can be obviated by employing a case-cohort design in which a selected control contributes to the comparison group of all cases in the same stratum in that control's risk period. One might speculate further that efficiency results for case-cohort designs will parallel closely those for a corresponding unmatched case-control design in situations in which the latter design applies.

5. DISCUSSION

Epidemiologic cohort studies or disease prevention trials in which raw materials for covariate data ascertainment have been stored on all cohort members, provide a primary application area for the case-cohort design. Such raw materials may include, for example, blood serum samples, tissue specimens, or occupational exposure records. The prediagnostic storage of such raw materials will in most circumstances assure the comparability of case and subcohort covariate histories. In the prevention trial context a case-cohort approach can eliminate much of the costly analysis of raw materials in the assembly of covariate histories, while still allowing the monitoring of such histories on an ongoing basis by means of the subcohort.

The case-cohort approach may also be considered as an alternative to a case-control design in the presence of a population-based disease registry, since the above relative risk estimation procedures do not require a cohort roster. Efficiency gains relative to time-matched, i.e. age-matched, case-control analyses can be anticipated, though recall bias would be as much an issue for the case-cohort design as for the case-control design in such a context.

The topic of efficiency relative to time-matched case-control studies merits additional study. Certainly one would expect better efficiency properties for the case-cohort approach in a broad range of circumstances. One can, however, imagine situations with detailed stratification and short individual risk periods in which the subcohort risk set may be small at some failure times or in which a small stratum-specific subcohort constitutes the comparison group for a large number of cases. In such circumstances reduced efficiency may arise relative to a case-control approach which requires the size of the reference group to be specified for each case. Even here, however, one can presumably define a subcohort, retrospectively, that is appropriately related to the strata and failure times of the failing subjects in order to ensure an efficiency improvement relative to a matched case-control analysis. Specifically, one could choose to augment the subcohort at preselected follow-up times with minimal change in the procedures of §3.

ACKNOWLEDGEMENT

The author would like to thank Mark Mason for considerable computational assistance. Helpful comments from Drs Neil Dubin, Sander Greenland, Bernard Pasternack, Steven Self and Noel Weiss are also acknowledged. This work was supported by grants from the National Institute of General Medical Sciences and the National Cancer Institute.

REFERENCES

ANDERSEN, P. K. & GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–20.

BESAG, J. E. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64**, 616–8.

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

BRESLOW, N. E. & PATTON, J. (1979). Case-control analysis of cohort studies. In *Energy and Health*, Ed. N. E. Breslow and A. S. Whittemore, pp. 226–42. Philadelphia: SIAM.

BRESLOW, N. E., LUBIN, J. H., MAREK, P. & LANGHOLTZ, B. (1983). Multiplicative models and cohort analysis. *J. Am. Statist. Assoc.* **78**, 1–12.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

KUPPER, L. L., MCMICHAEL, A. J. & SPIRTAS, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *J. Am. Statist. Assoc.* **70**, 524–8.

LIDDELL, F. D. K., MCDONALD, J. C. & THOMAS, D. C. (1977). Methods for cohort analysis: appraisal by application to asbestos mining (with discussion). *J. R. Statist. Soc.* A **140**, 469–90.

LUBIN, J. H. & GAIL, M. H. (1984). Biased selection of controls for case-control analyses of cohort studies. *Biometrics* **40**, 63–75.

MANTEL, N. (1973). Synthetic retrospective studies and related topics. *Biometrics* **29**, 479–86.

MIETTINEN, O. S. (1982). Design options in epidemiologic research: an update. *Scand. J. Work Environ. Health* **8**, Suppl 1, 7–14.

MRFIT RESEARCH GROUP (1982). Multiple risk factor intervention trial: risk factor changes and mortality results. *J. Am. Med. Assoc.* **248**, 1465–77.

OAKES, D. (1981). Survival times: Aspects of partial likelihood (with discussion). *Int. Statist. Rev.* **49**, 235–64.

PETO, R. (1972). Contribution to discussion of paper by D. R. Cox. *J. R. Statist. Soc.* B **34**, 205–7.

PRENTICE, R. L. & BRESLOW, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–8.

PRENTICE, R. L. & PYKE, R. L. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

PRENTICE, R. L. & SELF, S. G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Ann. Statist.* **11**, 804–13.

THOMAS, D. C. (1981). General relative risk models for survival time and matched case-control analysis. *Biometrics* **37**, 673–86.

WHITTEMORE, A. S. (1981). The efficiency of synthetic retrospective studies. *Biom. J.* **23**, 73–8.

WHITTEMORE, A. S. & MCMILLAN, A. (1982). Analyzing occupational cohort data: application to U.S. uranium miners. In *Environmental Epidemiology: Risk Assessment*, Ed. R. L. Prentice and A. S. Whittemore, pp. 65–81. Philadelphia: SIAM.

[*Received May* 1984. *Revised July* 1985]