BIOMETRICS 42, 293–299
June 1986

More on "Biased Selection of Controls for Case–Control Analyses of Cohort Studies"

James M. Robins

Occupational Health Program, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

and

Mitchell H. Gail and Jay H. Lubin

National Cancer Institute, Bethesda, Maryland 20205, U.S.A.

SUMMARY

The authors consider several aspects of the design and analysis of synthetic case–control studies of cohort data under a proportional hazards model. First, in highly stratified data, consistent estimates of the relative risk are shown to result only if controls are sampled randomly with replacement from the entire risk set or without replacement from the noncases. Second, if previous controls are excluded from consideration as future controls but are included as cases if they fail, then inconsistent estimates of the relative risk can occur if "time" in the proportional hazards model represents an individual's chronological age and age at entry into follow-up is variable. On the other hand, if "time" represents time since the beginning of follow-up, estimates of the relative risk will be consistent, but the usual variance estimator will be inconsistent.

1. Introduction

Lubin and Gail (1984) analyze biases that may result from inappropriate constraints on the selection of controls in a synthetic case–control analysis of cohort data under a proportional hazards model. They were concerned primarily with biases that result when, for example, certain members of the risk set are excluded as potential controls for a current case because it is known that these members subsequently develop disease. This note serves to clarify two issues discussed by Lubin and Gail.

In synthetic case–control studies, it is commonly stated that controls are to be sampled at random from cohort members at risk at the failure time of the case. This prescription might plausibly refer to any of four possible sampling schemes, depending on whether sampling is to be done with or without replacement and on whether the case is or is not allowed to be selected as his own control. Lubin and Gail (1984) suggested that any of the four sampling schemes would be valid. Oakes (1981) and Cox and Oakes (1984) demonstrated that sampling controls without replacement from the noncases (sampling scheme 1) leads to consistent estimates of relative risk. We show that sampling controls with replacement from the entire risk set including the case (sampling scheme 2) is also valid. Two sampling schemes that lead to inconsistent estimates in small subcohorts resulting from extensive stratification (i.e., sparse data) are to select controls without replacement from the entire risk set (sampling scheme 3) or with replacement from the noncases (sampling scheme 4). These ideas are developed in Section 2. They do not affect the results

Dr Robins' research was funded in part by the National Institute of Environmental Health Sciences grants 5 R23 ES03405 and 5 P30 ES00002.

Key words: Case–control studies; Cox proportional hazards model; Relative risk.

of Lubin and Gail (1984), who considered the case in which the number of individuals in each stratum is large.

Second, Lubin and Gail suggest that in order to parallel the prospective approach, controls should be selected from the risk set without regard to the results of the previous control sampling. In this prescription, subjects may be selected as controls for more than one case and cases may be selected as controls prior to their failure times. As an aside, they suggest excluding individuals who have been previously chosen as controls from consideration as future controls, yet including them as cases if and when they subsequently fail.

This latter suggestion leads to consistent estimators of the relative risk in certain special cases but not in general. In order to define when this latter sampling scheme is valid, we require the following definition of a "closed" cohort. Let "time" in the Cox model be represented by t. Then a cohort is defined to be "closed" if and only if within any stratum the set of individuals at risk at t is contained (strictly or nonstrictly) in the set of individuals at risk at t' whenever $t' < t$. Otherwise, the cohort is "open." A cohort may be "closed" under one definition of time and "open" under another. As an example, in the analysis of occupational mortality studies "time" is often chosen to be either an individual's chronological age or the time since the beginning of follow-up for that individual. If "time" is chosen as time since initiation of follow-up, and censored individuals cannot reenter into follow-up, then the cohort is "closed." But if "time" is identified with chronological age, and fine stratification on age at initiation of follow-up is not employed, individuals will have different "times" at the beginning of follow-up. As a result, the cohort will be "open." For instance, if one individual began follow-up at age 15, and another began at age 30, the latter individual would be at risk at $t = 30$ but not at $t' = 15$.

Provided "time" in a Cox proportional hazards model is defined so that the cohort is "closed," sampling schemes 1 and 2, modified so that previous controls are excluded from consideration as future controls (but not as future cases), yield consistent estimates of the relative risk parameter. However, the usual variance estimator based on the inverse of the observed information will not be consistent. If time is defined so that the cohort is "open," these modified sampling schemes can also yield inconsistent estimates of the relative risk, as discussed in Section 3.

2. Plans for Selecting Controls from the Risk Set

For simplicity we assume continuous failure time distributions so that at most one case fails at a time. Following the notation of Prentice (1986), we let $R(t)$ be the risk set of all cohort members at risk at time t, including the case who fails at t. The corresponding covariate histories are the functions $\{Z_i(u), 0 < u < t\}$, and the relative hazard for member i of $R(t)$ is $\Theta_i(t) = \exp\{\mathbf{X}_i(t)\boldsymbol{\beta}\}$, where $\mathbf{X}_i(t)$ is a functional on the entire covariate history. Let $\Theta_1(t)$ correspond to the observed case. Suppose $\tilde{R}(t)$ is a subsample of m elements from the r members of $R(t)$; $\tilde{R}(t)$ includes the case plus $m - 1$ additional "controls" selected according to one of the four sampling schemes in Section 1. Finally, let $\tilde{S} = \sum \Theta_i(t)$ be the sum over the m elements in $\tilde{R}(t)$. Oakes (1981) and Prentice (1986) showed that $\exp(L) = \prod \{(\Theta_1(t_j)/\tilde{S}(t_j)\}$ is a partial likelihood under sampling scheme 1, where this product is over distinct death times, t_j. In the Appendix we show that $\exp(L)$ is also equal to a partial likelihood under sampling scheme 2, apart from irrelevant constants, but not under schemes 3 or 4.

We now consider the property of the estimator $\tilde{\boldsymbol{\beta}}_v$ that maximizes e^L with v indexing the four control sampling schemes under two different large-sample limiting models. In limiting model A (the large-stratum model), we suppose that in any stratum and at any fixed time t at which the density function for failure is nonzero, the number of individuals at risk at t increases without bound. In large-sample limiting model B (the sparse-data model), we suppose that both the number of risk sets and number of strata increase without bound as the cohort size increases, but the number of individuals per risk set stays bounded in some nonnegligible fraction of the risk sets. Under the limiting models A and B, $\tilde{\beta}_1$ and $\tilde{\beta}_2$ will be consistent from general results on the consistency of partial maximum likelihood estimators. Under large-stratum limiting model A, if the number of controls selected per case is bounded, $\tilde{\beta}_3$ and $\tilde{\beta}_4$ will also be consistent because the chance of redundancy or of selecting the case as its own control goes to zero as the risk set size increases. However, it is easy to show by example that $\tilde{\beta}_3$ and $\tilde{\beta}_4$ may be inconsistent under the sparse limiting data model B.

The first author encountered a typical example with many small risk sets when analyzing cardiac mortality of copper smelter workers. Survival time was taken to be the age of the worker at death. Strata were defined by joint 10-year intervals of time since hire, calendar time, and years since termination of employment. Fifteen percent of the risk sets had more than 200 members and 30% had fewer than ten members. To save computing costs, it was decided to select five controls (or the maximum number of available controls if less than five) from each risk set. In these circumstances, sampling schemes 3 and 4 may lead to inconsistent estimates and should be avoided. The usual method of selecting controls without replacement from the noncases (sampling scheme 1) is probably to be preferred to sampling scheme 2 on grounds of efficiency. For this reason, we consider only scheme 1 and its modifications in the rest of the paper.

3. Subsequent Exclusion of Previously Selected Controls: A Modification of Scheme 1

We now examine the consequences of the following modification of scheme 1, which was mentioned by Lubin and Gail (1984). Controls are selected without replacement from the noncases in a given risk set. However, once selected as a control, that individual is excluded from serving as a future control. If a previously selected control becomes a subsequent case, he is treated like any other case. Thus, scheme 1 is modified only by excluding previously selected controls from subsequent selection as controls.

In a closed cohort, the consistency of estimates of $\boldsymbol{\beta}$ follows, under regularity conditions, from the fact that the score statistic has expectation 0. To see that the expectation of the score statistic is 0, note that the selected controls constitute a random sample of all noncases at risk at t_j. Given $\tilde{R}(t_j)$, the corresponding covariate histories prior to t_j, and the fact that a death occurred at t_j, the conditional probability that the person who died at t_j has relative hazard $\Theta_1(t_j)$ is therefore still $\Theta_1(t_j)/\tilde{S}(t_j)$, which implies that the score has expectation 0. Prentice (1986) has shown that the product $\prod \Theta_1(t_j)/\tilde{S}(t_j)$ is no longer a partial likelihood because relevant sigma-fields are not nested; we follow his terminology and call the product a "pseudolikelihood."

The following artificial example illustrates that contributions to the score statistic from successive risk sets are correlated under modified sampling scheme 1, and that standard variance estimators may be inconsistent. Suppose we have uncensored survival data on two exposed and two unexposed individuals in each of N strata. Suppose the nuisance hazard varies arbitrarily among strata, but the relative risk is constant across strata. In each stratum at the first death time we choose two controls without replacement from the noncases, and at the second death time we sample without replacement all noncases who have not previously been sampled as controls. Let $\dot{L}_i(\beta)$, $\dot{L}_2(\beta)$ represent stratum-specific contributions to the score statistic from the first and second risk sets, respectively, with $\dot{L}_2(\beta) \equiv 0$ if no individuals remain eligible to be selected as controls at the second death time. A calculation described in the following paragraph shows that within each stratum $E[\dot{L}_1(\beta)] = 0$, and $E[\dot{L}_2(\beta)] \equiv 0$, but

$$\operatorname{cov}[\dot{L}_1(\beta), \dot{L}_2(\beta)] = \frac{-4e^{2\beta}}{(2e^{\beta} + 1)^2(e^{\beta} + 1)^2(e^{\beta} + 2)}.$$

Thus, the usual large-sample tests and confidence intervals which assume zero covariance will fail to obtain their nominal significance levels and coverage rates.

Table 1 defines events (i, j) and their probabilities and contains the contributions to the score statistics from the first and second risk set in each stratum. The index i refers to four possible events at the first failure time, and j refers to the exposure status of the second case. As shown in Table 1, the probability that the first case is exposed and two unexposed controls are selected for the first case, $\Pr(i = 1)$, is $[e^{\beta}/(e^{\beta} + 1)](\frac{1}{3})$, since the chance the first case is exposed is $e^{\beta}/(e^{\beta} + 1)$ and, given the first case is exposed, the chance of getting two unexposed controls is $\frac{1}{3}$. Furthermore, the contribution to the pseudolikelihood from the first failure in the risk set, $e^{L_{1,i}}$, is $e^{\beta}/(e^{\beta} + 2)$ when $i = 1$.

The probability that the second case is nonexposed ($j = 2$), given $i = 1$, namely $\Pr[(i, j) \mid i] = \Pr[(1, 2) \mid 1]$ equals $2/(e^{\beta} + 2)$, because conditional on an exposed individual having failed in risk set 1, exactly one exposed and two unexposed individuals were at risk for the second failure. The corresponding contribution to the pseudolikelihood from the second failure is $e^{L_{2,12}} = 1/(e^{\beta} + 1)$ because when an unexposed individual fails in risk set 2, the remaining unexposed noncase will have been previously sampled as a control. As a consequence, only the exposed noncase may be selected as a control. Therefore, $\dot{L}_{2,12} = e^{\beta}/(e^{\beta} + 1)$. The quantity $\dot{L}_{2,ij}$ is 0 for $(i, j) = (1, 1)$ or $(4, 1)$ since both noncases in risk set 2 will have been previously sampled as controls. Also, $\dot{L}_{2,31} = 0$ because if the exposed case in the second risk set was not the exposed control in the first risk set, both noncases in the second risk set have been previous controls. If the exposed case of the second risk set is the exposed control of the first risk set, the only eligible noncase in the second risk set is exposed so that $e^{L_{2,31}} = e^{\beta}/(e^{\beta} + e^{\beta})$ and $\dot{L}_{2,31} = 0$. The rest of the entries in Table 1 can be filled in similarly. The results given in the preceding paragraph follow by noting that $E(\dot{L}_1) = \sum_i \dot{L}_{1,i}\Pr(i)$, $E(\dot{L}_2) = \sum_{ij}\dot{L}_{2,ij}\Pr[(ij)]$, and $\text{cov}(L_1, L_2) = E(\dot{L}_1\dot{L}_2) = \sum_{ij} \dot{L}_{1,i}\dot{L}_{2,ij}\Pr[(i, j)]$, where $\Pr[(i, j)] = \Pr(i)\Pr[(i, j) \mid i]$.

Prentice (1986) has extended these results by deriving a consistent estimate of the variance of the pseudolikelihood score. As in the previous example, he found that correlations among contributions to the score were negative near $\boldsymbol{\beta} = \mathbf{0}$, suggesting that the variance estimates based on the "partial likelihood" will be too large. Thus, hypothesis tests based on the usual partial likelihood methods will tend to have subnominal size.

If the cohort is "open," the selection of controls with the modified scheme 1 can result in inconsistent estimates of the relative risk. As an example, suppose "time" is identified as chronological age and stratification on age at the beginning of follow-up is not employed. Assume that, conditional on exposure, age at beginning of follow-up is not a risk factor for death so that it may be ignored in the analysis. Furthermore, assume the null hypothesis of no exposure effect is true. Suppose that the probability of exposure is greater in recent entries into the risk set (immigrants), than in the risk set as a whole. Then, under the above sampling plan, when compared to all individuals at risk, recent immigrants will have an increased chance of selection as control, but not as a case. Thus, controls will have a higher probability of being exposed than cases and an apparent protective exposure effect will be found. To be concrete, assume at age t than 90% of immigrants but only 10% of the "survivors" who have been on test since age t_0 (the youngest age in the study) are exposed. Furthermore, assume that 50% of individuals at risk at t are immigrants, and that 80% of "survivors" and 0% of immigrants have been previously sampled as controls. Then if no exposure effect is present and age at initial risk is not a risk factor, the probability that the case will be exposed is $.5 = (.5)(.9) + (.5)(.1)$ but the probability that an individual at risk at t who has never been previously chosen as a control will be exposed is

$$\frac{(1 - .8)(.5)(.10) + (.5)(.90)}{(1 - .8).5 + .5} = .77.$$

Table 1

Contributions to the score statistic when controls that are selected at failure 1 are excluded from selection as controls at failure 2

Outcome i describing selection of the first risk set	$i = 1$ case exposed 2 controls unexposed		$i = 2$ case exposed 1 control exposed, 1 control unexposed		$i = 3$ case unexposed 1 control exposed, 1 control unexposed		$i = 4$ case unexposed 2 controls exposed	
Outcome (i, j) describing the joint event i and the exposure status j of the second case ($j = 1$ is exposed, $j = 2$ unexposed)	(1, 1) Exposed	(1, 2) Unexposed	(2, 1) Exposed	(2, 2) Unexposed	(3, 1) Exposed	(3, 2) Unexposed	(4, 1) Exposed	(4, 2) Unexposed
$e^{L_{1,i}}$	$\frac{e^{\beta}}{e^{\beta}+2}$		$\frac{e^{\beta}}{2e^{\beta}+1}$		$\frac{1}{e^{\beta}+2}$		$\frac{1}{2e^{\beta}+1}$	
$\dot{L}_{1,i}$	$\frac{2}{e^{\beta}+2}$		$\frac{1}{2e^{\beta}+1}$		$\frac{-e^{\beta}}{e^{\beta}+2}$		$\frac{-2e^{\beta}}{2e^{\beta}+1}$	
$\Pr(i)$, probability of outcome i	$\frac{e^{\beta}}{e^{\beta}+1}\cdot\left(\frac{1}{3}\right)$		$\frac{e^{\beta}}{e^{\beta}+1}\cdot\left(\frac{2}{3}\right)$		$\frac{1}{e^{\beta}+1}\cdot\left(\frac{2}{3}\right)$		$\frac{1}{e^{\beta}+1}\cdot\left(\frac{1}{3}\right)$	
$\Pr[(i, j) \mid i]$	$e^{\beta}/(e^{\beta}+2)$	$2/(e^{\beta}+2)$	$e^{\beta}/(e^{\beta}+2)$	$2/(e^{\beta}+2)$	$2e^{\beta}/(2e^{\beta}+1)$	$1/(2e^{\beta}+1)$	$2e^{\beta}/(2e^{\beta}+1)$	$1/(2e^{\beta}+1)$
$\dot{L}_{2,ij}$	0	$-e^{\beta}/(e^{\beta}+1)$	$1/(e^{\beta}+1)$	0	0	$-e^{\beta}/(e^{\beta}+1)$	$1/(e^{\beta}+1)$	0

Prentice (1986) corrects this deficiency of modified scheme 1 sampling for open cohorts by proposing a novel structured design that excludes previous controls from consideration as future controls but not as future cases, and yet leads to consistent estimates of the relative risk.

RÉSUMÉ

Les auteurs considèrent quelques aspects de la planification et de l'analyse d'études synthétiques cas–contrôle de données organisées en cohortes, en utilisant le modèle de chances proportionnelles. Premièrement, dans les données très stratifiées, on montre que des estimateurs convergents du risque relatif sont obtenus seulement quand les contrôles sont échantillonnés au hasard simple, avec replacement ou sans replacement pour les non-cas. Deuxièmement, si des contrôles préalables ne sont pas considérés comme des contrôles futurs, mais comme des cas, s'ils sont atteints, alors on peut obtenir des estimateurs non convergents du risque relatif si le "temps" dans le modèle des chances proportionnelles représente l'âge chronologique de l'individu. Par ailleurs, si le "temps" représente le temps depuis le début jusqu'à la suite, les estimateurs du risque relatif seront convergents, mais l'estimateur habituel de la variance ne le sera pas.

REFERENCES

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* New York: Chapman and Hall.

Lubin, J. H. and Gail, M. H. (1984). Biased selection of controls for case–control analyses of cohort studies. *Biometrics* **40,** 63–75.

Oakes, D. (1981). Survival times: Aspects of partial likelihood. *International Statistical Review* **49,** 235–264.

Prentice, R. L. (1986). On the design of synthetic case–control studies. *Biometrics* **42,** 301–310.

Received March 1985; *revised February* 1986.

APPENDIX

Oakes (1981) and Prentice (1986) proved $\exp(L)$ was a partial likelihood under sampling scheme 1. We now show that $\exp(L)$ is also equal to a partial likelihood, apart from irrelevant constants, under sampling scheme 2. Let $R(t)$ denote the set of all $r(t)$ persons at risk at t and $\tilde{R}(t)$ the selected risk set consisting of the case and his $m - 1$ controls. $\tilde{R}(t)$ is empty if no individual fails at t. The set $\tilde{R}(t)$ is completely characterized by $\mathbf{m}(t) = (m_1(t), m_2(t), \ldots, m_r(t))$, where $m_i(t)$ is the number of times that individual $i \in R(t)$ is included in $\tilde{R}(t)$, and $r = r(t)$. Note that $m = \sum_{i=1}^{r(t)} m_i(t)$. Let $K(t) = \cup R(u)$ for $0 \le u \le t$ and $\tilde{K}(t) = \cup \tilde{R}(u)$ for $0 \le u \le t$. As in Prentice (1986), for the ith individual, let $N_i(u)$ be a right-continuous death indicator with value 0 while alive and 1 while dead, $Y_i(u)$ be a left-continuous indicator with value 1 when at risk and 0 otherwise, and $Z_i(u)$ be the covariate history through time u. Let

$$F(t) = [\{N_i(u), Y_i(u), Z_i(u), 0 \le u < t, i \in K(t)\}, R(u), 0 \le u \le t].$$

$\tilde{F}(t)$ is defined similarly, except that $\tilde{K}(t)$ and $\tilde{R}(u)$ replace $K(t)$ and $R(u)$. Let

$$\tilde{F}(t^-) = \left[\left\{N_i(u), Y_i(u), Z_i(u), 0 \le u < t, i \in \bigcup_{0 \le v < t} \tilde{R}(v)\right\}, \tilde{R}(u), 0 \le u < t\right]$$

and let

$$B(t) = [\{N_i(u), Y_i(u), Z_i(u), 0 \le u < t, i \in \tilde{R}(t)\}, \tilde{R}(t)],$$

so that $\tilde{F}(t) = \tilde{F}(t^-) \cap B(t)$.

Theorem. Under sampling scheme 2, e^L is proportional to a partial likelihood.

Proof. Since the σ-algebras $(\tilde{F}(t); t \ge 0)$ are nested, the theorem would follow if we could show that

$$\Pr[N_i(t) \ne N_i(t^-) \mid \tilde{F}(t), N(t) \ne N(t^-)] = m_i(t)[\Theta_i(t)/\tilde{S}(t)] \tag{A.1}$$

where $N(t) \neq N(t^-)$ means that $N_l(t) \neq N_l(t^-)$ for some $l \in R(t)$, and $\Theta_i(t)$ and $\tilde{S}(t)$ are defined in Section 2. Under independent failure and censoring mechanisms,

$$\Pr[N_i(t) \neq N_i(t^-), \tilde{B}(t) \mid \tilde{F}(t^-), F(t), N(t) \neq N(t^-)] = [\Theta_i(t)/S(t)]P_i(t),$$

where

$$S(t) = \textstyle\sum_{i=1}^{r(t)} \Theta_i(t) \quad \text{and} \quad P_i = r^{-(m-1)}\begin{pmatrix} m-1 \\ m_1, m_2, \ldots, m_i - 1, \ldots, m_r \end{pmatrix}$$

is the probability of obtaining $\tilde{R}(t)$ given that individual i failed at t and the controls are selected with replacement from the entire risk set $R(t)$. The probability

$$\Pr\{B(t) \mid \tilde{F}(t^-), F(t), N(t) \neq N(t^-)\} = \textstyle\sum_1^r \{\Theta_i/S\}P_i = \begin{pmatrix} m-1 \\ m_1, m_2, \ldots, m_i, \ldots, m_r \end{pmatrix} \sum_{i=1}^r \{m_i\Theta_i/S\},$$

where the dependence on t has been suppressed. Dividing, we obtain the conditional probability

$$\begin{aligned}\Pr\{N_i(t) \neq N_i(t^-); \mid B(t), \tilde{F}(t^-), F(t), N(t) \neq N(t^-)\} &= \Pr\{N_i(t) \neq N_i(t^-) \mid \tilde{F}(t), F(t), N(t) \neq N(t^-)\} \\ &= m_i\Theta_i/\tilde{S},\end{aligned}$$

which is conditionally independent of $F(t)$, given $\tilde{F}(t)$, $N(t) \neq N(t^-)$. Thus, (A.1) is proved.

We now show that sampling schemes 3 and 4 do not produce partial likelihoods. Suppose $R(t)$ includes individuals {1, 2, 3, 4} and $\tilde{R}(t)$ = {1, 2, 2}. Then, if one has selected two controls with replacement from the noncases, the left side of A.1 has probability 1 [rather than being proportional to $\Theta_i(t)/\tilde{S}(t)$], since the repetition of individual 2 implies that 2 must have been a control. Likewise, if $\tilde{R}(t)$ = {1, 1, 2}, the left side of A.1 has probability 1 if controls have been selected without replacement from $R(t)$, since, by repetition, the case must be individual 1.