



Designs for Synthetic Case-Control Studies in Open Cohorts

James M. Robins; Ross L. Prentice; Donald Blevins

Biometrics, Vol. 45, No. 4 (Dec., 1989), 1103-1116.

Stable URL:

<http://links.jstor.org/sici?sici=0006-341X%28198912%2945%3A4%3C1103%3ADFSCSI%3E2.0.CO%3B2-L>

Biometrics is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/ibs.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

Designs for Synthetic Case–Control Studies in Open Cohorts

James M. Robins

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

Ross L. Prentice

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center,
1224 Columbia Street, Seattle, Washington 98104, U.S.A.

and

Donald Blevins

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

SUMMARY

Several designs are proposed for case–control studies within cohorts when the cohort is open to late entry. These and previously proposed designs are examined with respect to consistency and efficiency of relative risk parameter estimation, and a small simulation study is reported. If study costs increase in proportion to the total number of “at-risk” controls, the most efficient design, Design C, is as follows. For a case failing at time t , controls are selected at random (and without regard to “at-risk” status) from among cohort members who are (i) known not to have failed prior to t and (ii) have not been previously selected as controls. At each t , control sampling proceeds until a prespecified number of controls who are “at risk” at t have been obtained. The efficiency advantage of Design C over that of the standard case–control design proposed by Thomas (in Appendix to Liddell, McDonald, and Thomas, 1977, *Journal of the Royal Statistical Society, Series B* **140**, 469–490) will often be small. If, on the other hand, the costs increase in proportion to the number of distinct “at-risk” controls, Design C is no longer the most efficient design. In this case, several alternative designs are proposed.

1. Introduction

Thomas (1977) proposed a design, hereafter referred to as *Design A*, for sampling within a cohort. In this design, controls for a case occurring at time t are randomly sampled (without replacement) from cohort members known not to have failed by time t . Controls are selected independently at distinct failure times. Time may refer, for example, to chronological time or to study subject age. This sampling procedure has been shown to yield a partial likelihood function (Oakes, 1981; Prentice, 1986a) so that standard asymptotic likelihood procedures will generally be appropriate for the estimation of relative risk parameters. The purposes of such sampling include computational reduction and, importantly, reduction in the number of cohort members for whom covariate data need be assembled.

Robins, Gail, and Lubin (1986) and Prentice (1986a) consider modifications to the above sampling procedure and some aspects of the (asymptotic) biases and efficiency properties of corresponding relative risk estimators. The fact that Design A may involve sampling the same individual as a control for more than one case prompted consideration of designs in which a subject selected as a control at time u would be ineligible for control selection at

any $t > u$. Robins et al. noted that inconsistent estimates of relative risk may arise from cohorts open to late entry if the controls for a case occurring at t are merely a random sample of "at-risk" subjects who have not previously failed and who have not previously (i.e., at any $u < t$) been selected as a control. A cohort member is "at risk" at time t if the subject is under active follow-up for the study disease at that time. Prentice (1986a) argued that this problem could be avoided by selecting controls at time t without regard to their "at-risk" status. This leads to a sampling *Design B* in which the controls for a case occurring at time t are a random sample of cohort members without failure prior to t and not previously selected as controls. If the case had been selected as a control at (at most one) time $u < t$, then the controls are matched to the case with respect to "at-risk" status at time u . This is essentially Design B in Prentice (1986a), though the earlier description was incomplete in not specifying the matching on "at-risk" status at u if the case had been selected as a control at $u < t$.

In this paper further sampling options are considered toward avoiding the rather cumbersome matching in Design B and toward possible improvements in estimating efficiency. Attention is also paid to whether sampling at time t continues until a specified number of controls are obtained or until a specified number of "at-risk" controls are obtained. Possible variance estimators for relative risk regression parameters are given and a simulation study is reported.

We shall see that the preferred design depends on the cost structure of the study. In a case-control study, costs may increase in proportion to the total number of at-risk controls or, alternatively, in proportion to the total number of "distinct" at-risk controls (Langholz and Thomas, unpublished manuscript, 1987). For example, costs would be proportional to the total number of distinct controls if all covariates were time-independent. Alternatively, costs would be proportional to the total number of at-risk controls if, for example, each subject provided a blood specimen at weekly intervals and, for a case failing in week j , the j th blood specimen of the case would be compared to the j th blood specimens of the case's matched controls.

2. Design Options for Synthetic Case-Control Studies

Consider now an additional sampling *Design C* in which the controls for a case occurring at time t are simply a random sample of cohort members without prior failure and not previously selected as controls. Thus, sampling proceeds without regard to whether the case had previously been selected as a control and without regard to the "at-risk" status of the controls.

Note that in describing Designs A-C we have referred to any subject sampled as part of the comparison group for a given case as a "control," without regard to the subject's "at-risk" status at the time of case occurrence. Of course it is only the "at-risk" controls that contribute to the data analysis and for whom covariate data need be assembled. (A more descriptive terminology might have been to refer to our controls as precontrols or potential controls, and to describe the "at-risk" controls as true controls.) Note that in our description of Design A, we have assumed that controls were selected without regard to at-risk status. Thomas, in his description of Design A, assumed controls at t were sampled only from those "at risk" at t . Because, under Design A, control selection proceeds independently at distinct failure times, these two descriptions are equivalent (since the non-"at-risk" controls do not contribute to the data analysis).

Note also that our description of Designs A-C has not provided a prescription for stopping control sampling at a given point in time. In accordance with standard practice for Design A, we will assume for each design that control sampling continues until a specified number of "at-risk" controls have been obtained, or until the available control

pool is exhausted. If covariates are time-independent, the costs associated with ascertaining the covariate status of a particular subject will not depend on the number of cases for which the subject serves as a control. Therefore, in this setting, we might wish to consider a *Design D*, a modification of *Design C*, in which control sampling at time t is from cohort members without prior failure and continues until a specified number of “at-risk” controls have been chosen who have not previously been selected as controls, or until the available control pool is exhausted. In this design, we do not exclude previously selected controls as later controls.

Another approach to control selection at time t would be to continue sampling until a specified number of controls (without prior failure) are obtained, without regard to their “at-risk” status, or until the available control pool is depleted. The above designs with this stopping prescription will be referred to as *Designs A’–D’*, respectively. (Under *Design D’* control sampling proceeds until a specified number of controls have been chosen who have not previously been selected as controls.) These designs may be of lesser practical interest, due to the fact that a random number of “at-risk” controls will be obtained in each risk set, but they are useful in the consideration of statistical properties.

3. Relative Risk Estimation

In the notation of Prentice (1986a), let $Z(t)$ denote a covariate measurement for a study subject at time t and let $\lambda\{t; Z(u), 0 \leq u < t\}$ denote the (instantaneous) failure rate at time t among subjects with covariate history $\{Z(u), 0 \leq u < t\}$ prior to time t . A relative risk regression model can then be written

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

where the modeled regression vector $\mathbf{X}(t) = \{X_1(t), \dots, X_p(t)\}$ consists of functions of $\{Z(u), 0 \leq u < t\}$ or products of such functions with time, β is a corresponding column p -vector of relative risk regression parameters to be estimated, $r(\cdot)$ is a fixed function, usually $r(\cdot) = \exp(\cdot)$ or $r(\cdot) = 1 + (\cdot)$, standardized so that $r(0) = 1$, and $\lambda_0(\cdot)$ is a baseline hazard function corresponding to a standard covariate history for which $\mathbf{X}(t) \equiv 0$ (Cox, 1972).

Consider a cohort of size n and let $\{N_i(u), Y_i(u), 0 \leq u < t\}$ denote counting and censoring processes up to time t for the i th subject, so that N_i takes value 0 up to an observed failure on subject i and value 1 thereafter, while Y_i takes value 1 at times at which the i th subject is “at risk,” and value 0 otherwise.

For the purposes of this paper, we say a cohort is *closed* if a subject “at risk” at time t , $t > u$, was also at risk at u (if any subject was “at risk” at u). Otherwise, the cohort is *open*.

A standard independent censorship condition requires

$$\lambda\{t; \{Z_i(u), Y_i(u), 0 \leq u < t\}\} = Y_i(t)\lambda\{t; Z_i(u), 0 \leq u < t\},$$

where $Y_i(t)$ is the limit as u increases to t of $Y_i(u)$. This implies that given $\{Z(u); u < t\}$, previous at-risk history is not a predictor of failure among those at risk at t .

For each of the above designs we propose that the relative risk parameter be estimated as the value, $\hat{\beta}$, that maximizes

$$L(\beta) = \prod_{i=1}^n \left(\frac{r_{ii}}{\sum_{l \in \tilde{R}(t_i)} r_{li}} \right)^{\delta_i}, \quad (1)$$

where $\tilde{R}(t)$ is the matched case-control set for a failure occurring at t , and where the observed failure or terminal right-censoring time for the i th subject is given by $t_i = \inf\{t \mid Y_i(u) = 0, \text{ all } u > t\}$, and where the corresponding censoring indicator is defined by $\delta_i = 1$ if $N_i(t_i) \neq N_i(t_i^-)$ and $\delta_i = 0$ otherwise, while $r_{ii} = Y_i(t_i)r\{X_i(t_i)\beta\}$.

Note that $\hat{\beta}$ is a solution to the “score” equation $U(\beta) = \partial \log[L(\beta)]/\partial \beta = 0$. Unbiasedness of the score statistic $U(\beta)$ at the “true” value of β is a key step in the development of a consistency (asymptotic unbiasedness) result for $\hat{\beta}$. At a failure time t , simple random sampling of controls from cohort members without prior failure, that is independent of control sampling at other failure times, gives, as the conditional probability that the i th subject is the case,

$$\Pr\{N_i(t) \neq N_i(t^-) \mid \tilde{F}(t), \tilde{N}(t) \neq \tilde{N}(t^-)\} = \frac{Y_i(t)r\{\mathbf{X}_i(t)\beta\}}{\sum_{l \in \tilde{R}(t)} Y_l(t)r\{\mathbf{X}_l(t)\beta\}}, \tag{2}$$

where $\tilde{F}(t) = [\tilde{R}(t), \{N_l(u), Y_l(u), Z_l(u), 0 \leq u < t, l \in \tilde{R}(t)\}]$ and $\tilde{N}(t) \neq \tilde{N}(t^-)$ means that $N_l(t) \neq N_l(t^-)$ for some $l \in \tilde{R}(t)$. The fact that (2) is a probability distribution implies that the conditional, and hence the unconditional, expectation of the logarithm of (2) with respect to β is 0. Therefore, the score statistic under Design A', in which a prespecified number of controls are selected independently at each case occurrence, has mean zero. The same is true for the standard Design A since (2) continues to hold if one adds to the conditioning event $F(t)$, the requirement that exactly $m(t)$ controls are “at risk” at time t . Specifically, the information that exactly $m(t) + 1$ of the $Y_l(t)$ values in $\tilde{R}(t)$ are unity provides no differential information concerning which of these $m(t) + 1$ subjects is the case at time t . In fact, a stronger result holds for Designs A and A' in that (2) continues to hold if one conditions on $\{\tilde{F}(u), u \leq t\}$, so that (1) is a partial likelihood function for each of these designs.

The conditional probabilities (2) also obtain for Design B' so that the score statistic also has a mean of zero for these designs. As in Prentice (1986a) let $\Delta(u, t) = 1$ if the case occurring at time t was selected as a control at time $u < t$, while $\Delta(u, t) = 0$ otherwise, and add $\{\Delta(u, t), 0 \leq u < t\}$ to the conditioning event in (2). Under Design B', $\tilde{R}(t)$ consists of a prespecified number of subjects that are matched with regard to the value of $\{\Delta(u, t), 0 \leq u < t\}$, and also with respect to “at-risk” status at time u , if the case occurring at t had been selected as a control at $u < t$. This matching, along with the independent censorship assumption mentioned above, implies that the sampling procedure conveys no differential information concerning which subject in $\tilde{R}(t)$ fails, so that (2) continues to hold with this additional conditioning. Therefore, the score statistic has conditional, and hence unconditional, mean zero under Design B'.

For Designs C' and D' the matched case-control set $\tilde{R}(t)$ consists of any case occurring at time t along with a random sample of prespecified size (if available), selected from cohort members without prior failure. The fact that none of these controls have been selected as controls at earlier time points (Design C'), or that a prespecified number had not been selected at any earlier time point (Design D'), provides no differential information concerning the failing individual, so that (2) obtains and the score estimating equation is unbiased under these designs.

This leaves Designs B, C, and D. Since with these designs, control selection at time t depends on sampling at earlier time points as well as on the “at-risk” histories $\{Y_l(t), 0 \leq u < t\}$ of selected individuals, there is potential for (2) not to hold. The following example makes it clear that score statistic bias can occur with these designs. Suppose a cohort consists of four subjects {a, b, c, d} and that subjects {a, b, c} with modeled covariate $X \equiv 0$ are followed until a first failure at time u . Immediately thereafter subject d with $X \equiv 1$ becomes “at risk” and follow-up terminates immediately following the second failure at time t . Suppose Design C or D is applied with two “at-risk” not-previously-selected controls selected at time u and one such “at-risk” control selected at time t . The expectation of the score statistic at time u is zero since $\tilde{R}(u)$ is either (a, b, c, d) or (a, b, c), and hence always includes all cohort members “at risk” at u . Without loss of generality,

suppose subject a fails at time u . If $\tilde{R}(u) = (a, b, c, d)$ then the possible values for $\tilde{R}(t)$ under Design C are (b), (c), or (d), there being no controls that have not been previously selected. On the other hand, if $\tilde{R}(u) = (a, b, c)$, possible values for $\tilde{R}(t)$ are (b, d), (c, d), or (d), according to whether b, c, or d fails at time t . It follows that whenever $\tilde{R}(t) = (b, d)$ or (c, d) then b or c is the case, respectively, in violation of (2). For example, at $\beta = \mathbf{0}$ the conditional expectation of the score statistic given $\tilde{R}(t) = (b, d)$ is $-\frac{1}{2}$ and the (marginal) expectation of the score statistic is $-\frac{1}{2}$ times the probability of matched sets (a, d), (b, d), or (c, d) arising at time t , that is, $(-\frac{1}{2})(\frac{1}{9}) = -\frac{1}{18}$.

Similarly, under Design D, if $\tilde{R}(u) = (a, b, c, d)$, with subject a failing at time u , then $\tilde{R}(t) = (b, c, d)$, there being no controls that have not previously been selected. If, however, $\tilde{R}(u) = (a, b, c)$ then $\tilde{R}(t)$ must be (b, c, d), (b, d), or (c, d). If $\tilde{R}(t) = (b, d)$ or $\tilde{R}(t) = (c, d)$ then failure at time t must have occurred in subject b or c, respectively, while if $\tilde{R}(t) = (b, c, d)$ there is a probability in excess of (2) that the failure is subject d. Hence, (2) is violated at time t and it is straightforward to show that the score statistic generally has a nonzero (marginal) mean. For example, at $\beta = \mathbf{0}$ the score statistic has mean $-\frac{1}{54}$. A similar example shows the score may have nonzero expectation under Design B in the presence of interval censoring, where a subject is said to be interval-censored if there is some time t at which the subject is not at risk, but the subject was at risk at times prior to and subsequent to t .

Even though $U(\beta)$ is not unbiased under Designs B, C, and D, the expected value of $n^{-1/2}U(\beta)$ will approach zero as the cohort size $n \rightarrow \infty$ under fairly weak conditions. In particular, such will generally be the case if the fraction of the cohort that is "at risk" and has not previously been selected as a control is bounded away from 0, uniformly in t , and the modeled covariate is bounded. A sketch of the proof of this result is given in Appendix 1. More generally, one expects $n^{-1/2}U(\beta)$ to converge in distribution to a normal variate with mean zero and finite variance for each of the Designs A to D and A' to D' under appropriate stability and regularity conditions, along with the conditions just alluded to for Designs B, C, and D.

Variance estimation A Taylor expansion of $U(\beta) = \partial \log[L(\beta)]/\partial \beta$ about the "true" β -value, evaluated at $\hat{\beta}$, gives $n^{1/2}(\hat{\beta} - \beta) = \{n^{-1}I(\beta_*)\}^{-1}n^{-1/2}U(\beta)$ for β_* between $\hat{\beta}$ and the "true" β , where $I(\beta) = -\partial^2 \log[L(\beta)]/\partial \beta^2$. Hence, $n^{1/2}(\hat{\beta} - \beta)$ will rather generally have an asymptotic normal distribution with mean zero and variance the limit of $n\Sigma(\beta)^{-1}V(\beta)\Sigma(\beta)^{-1}$ where $n^{-1}\Sigma(\beta)$ is the expected value of $n^{-1}I(\beta)$, and $n^{-1}V(\beta)$ is the variance matrix for $n^{-1/2}U(\beta)$. Now $n^{-1}\hat{\Sigma}(\hat{\beta})$ quite generally provides a consistent estimator of $n^{-1}\Sigma(\beta)$, where

$$\begin{aligned}\hat{\Sigma}(\hat{\beta}) &= \sum_{j=1}^n \delta_j \mathbf{v}_{jj}, \\ \mathbf{v}_{jj} &= \text{var}\{U_j(\beta) | \hat{F}(t_j), \hat{N}(t_j) \neq \hat{N}(t_j^-)\} \\ &= \sum_{l \in \hat{R}(t_j)} \mathbf{c}_{lj} \mathbf{c}_{lj}^T r_{lj} R_j^{-1} - \mathbf{B}_j \mathbf{B}_j^T R_j^{-2},\end{aligned}$$

and where $\mathbf{b}_{lj} = Y_l(t_j)\mathbf{X}_l(t_j)r' \{ \mathbf{X}_l(t_j)\beta \}$, $\mathbf{c}_{lj} = \mathbf{b}_{lj}/r \{ \mathbf{X}_l(t_j)\beta \}$, $\mathbf{B}_j = \sum_{l \in \hat{R}(t_j)} \mathbf{b}_{lj}$, and $R_j = \sum_{l \in \hat{R}(t_j)} r_{lj}$, and $U_j(\beta)$ is the score statistic contribution from the j th factor in equation (1).

Because $L(\beta)$ is a partial likelihood function under Designs A and A', $n^{-1}\hat{\Sigma}(\hat{\beta})$ also provides a consistent estimator of $n^{-1}V(\beta)$ under weak conditions. It is also a natural candidate estimator of $n^{-1}V(\beta)$ under Designs D and D', as $L(\beta)$ is "nearly" a partial likelihood function for these designs (see Appendix 2). For Designs B and B' the argument

of Prentice (1986a) gives an estimator $n^{-1}\hat{V}_1(\hat{\beta})$ for $n^{-1}V(\beta)$, where $\hat{V}_1(\beta) = \hat{\Sigma}(\beta) + \hat{C}_1(\beta)$, where

$$\hat{C}_1(\beta) = \sum_{j=1}^n \delta_j \sum_{\{i | t_i < t_j\}} \delta_i \Delta(t_i, t_j)(\mathbf{v}_{ji} + \mathbf{v}_{ji}^T),$$

and where, for $t_i < t_j$,

$$\mathbf{v}_{ji} = - \sum_{k \in \tilde{R}(t_j)} \left[\frac{\mathbf{B}_i + \mathbf{b}_{ki} - \mathbf{b}_{ji}}{R_i + r_{ki} - r_{ji}} \right] (\mathbf{c}_{kj} - \mathbf{B}_j R_j^{-1})^T r_{kj} R_j^{-1}.$$

This variance formula derives from the fact that score statistic contributions at t_i and t_j are uncorrelated unless the case occurring at t_j was selected as a control at $t_i < t_j$ (so that $\Delta(t_i, t_j) = 1$), in which situation $\tilde{R}(t_i)$ and $\tilde{R}(t_j)$ have exactly one member in common and equation (2) gives the probability that the common subject is $i \in \tilde{R}(t_j)$. These same circumstances prevail for Design C' so that $n^{-1}\hat{V}_1(\hat{\beta})$ provides a variance estimator for this design. As discussed in Appendix 2, we have been unable to determine whether $n^{-1}\hat{V}_1(\hat{\beta})$ provides a consistent variance estimator under Design C. Simulation results, reported in Section 4, suggest that it may be consistent.

Because calculation of $\hat{V}_1(\beta)$ is somewhat demanding computationally, a more "empirical" estimator for $n^{-1}V(\beta)$ has also been considered, namely, $n^{-1}\hat{V}_2(\hat{\beta})$, where

$$\hat{V}_2(\beta) = \hat{\Sigma}(\beta) + \hat{C}_2(\beta)$$

and

$$\hat{C}_2(\beta) = \sum_{j=1}^n \delta_j \sum_{\{i | t_i < t_j\}} \delta_i \Delta(t_i, t_j)(\mathbf{U}_i \mathbf{U}_j^T + \mathbf{U}_j \mathbf{U}_i^T),$$

where, as before, $\mathbf{U}_i = \mathbf{U}_i(\beta)$ is the score contribution from the i th term in $L(\beta)$. Appendix 2 indicates that $n^{-1}\hat{V}_2(\hat{\beta})$ will rather generally provide a consistent estimator of $n^{-1}E\{\mathbf{U}(\beta)\mathbf{U}(\beta)^T\}$ under regularity conditions for Designs B, B', C, and C'. In fact, as shown in Appendix 2, in defining $\hat{C}_2(\hat{\beta})$ under Design B or B', we can replace $\Delta(t_i, t_j)$ by $\Delta_B(t_i, t_j) = \Delta(t_i, t_j)Y_j(t_i)$, where $Y_j(t_i)$ is the at-risk status of the subject j (the failure at t_j) at time t_i .

For a fixed number of "at-risk" controls per case, the results of Prentice (1986a) cause one to expect Design B typically to give rise to slightly more efficient parameter estimation than does Design A. Furthermore, one might conjecture that Design C is more efficient than Design B since, in Design B, the non-risk factor "at-risk status at a previous time" is often matched on. Simulation results in Section 4 suggest that this is indeed the case. One would expect Design D to have some efficiency advantage relative to Designs A, B, and C, because of the greater number of at-risk controls per case.

4. Simulation Study

A simulation study was conducted to compare the performance of Designs A–D. We were also concerned to determine whether the failure of equation (2) to hold for Designs C and D may affect the performance of the corresponding relative risk estimators in cohorts of moderate size. Hence, Designs C' and D' are also included in the simulation study. A cohort was considered in which 1,200 exposed $\{X(t) \equiv 1\}$ and 400 unexposed $\{X(t) \equiv 0\}$ subjects entered into follow-up at time 0. At time 1, an additional 400 exposed and 1,200 unexposed subjects entered into follow-up. Follow-up ended at time 2. No subjects were lost to follow-up prior to time 2 except due to failure. An exponential failure rate was chosen to give a 10% expected cumulative probability of failure in the time interval (0, 1)

for both the exposed and unexposed (i.e., $\beta = 0$). In the time interval (1, 2) the exponential failure rate was increased to give an expected cumulative probability of failure of 20% over this interval.

Case-control sampling was conducted such that, for each case occurring in the interval (0, 1), four “at-risk” controls who had not previously been selected as controls were included in the control group for Designs B, C, and D; in Design A, also, four “at-risk” controls were selected for each case. For Designs C’ and D’, eight controls, who had not been previously selected as controls, were sampled. For cases failing in the interval (1, 2), a single not-previously-selected “at-risk” control was sampled under Designs B, C, D, C’, and D’, and a single “at-risk” control was sampled under Design A. Note that in the interval (1, 2) all controls are “at-risk” controls.

We maximized (1) under an exponential relative risk model (i.e., $r(\cdot) = \exp(\cdot)$).

Notable features of this Monte Carlo experiment include: Under Designs C and C’, over 30% of the cases failing in the interval (1, 2) are expected to have been selected as controls. Also, there is a marked association between exposure status and “at-risk” history. Finally, under all relevant designs, at each failure time, a large number of “at-risk” subjects remain who have not been previously selected as controls.

We performed 550 trials in 55 batches of 10 trials so that standard errors of empirical variances could be estimated. Five hundred fifty trials were chosen so that the actual coverage rates of nominal 90% and 95% confidence intervals for β could be determined to within several percentage points. The results are reported in Table 1. For each design we estimated $\text{avg}(\hat{\beta})$, the average of the 550 $\hat{\beta}$ ’s, and the sample variance of $\hat{\beta}$ over the 550 trials. (To the third significant digit, the sample variance equaled the average of the 55 batch-specific empirical estimates of the variance of $\hat{\beta}$.) Also, the Monte Carlo standard error of the sample variance of $\hat{\beta}$ was computed from the empirical between-batch variability of the within-batch empirical variance of $\hat{\beta}$.

Table 1

Simulation summary statistics for synthetic case-control studies under various sampling designs. All calculations arise from $\beta = 0$, where β is the logarithm of the relative risk.

Design	Avg($\hat{\beta}$) ($\times 10^2$)	Sample variance $\hat{\beta}$ ($\times 10^2$)	Average $\hat{\Sigma}(\hat{\beta})^{-1}$ ($\times 10^2$)	Average $\hat{\Sigma}(\hat{\beta})^{-1} \hat{V}_1(\hat{\beta}) \hat{\Sigma}(\hat{\beta})^{-1}$ ($\times 10^2$)	Average $\hat{\Sigma}(\hat{\beta})^{-1} \hat{V}_2(\hat{\beta}) \hat{\Sigma}(\hat{\beta})^{-1}$ ($\times 10^2$)	Empirical coverage rates for nominal confidence intervals	
						90%	95%
A	1.2 (.48) ^a	1.27 (.07)	1.339 (.003)			90.7	95.8
B	.7 (.46)	1.18 (.08)	1.431 (.004)	1.218 (.004)	1.218 (.005)	90.9	94.6
C	1.3 (.46)	1.20 (.07)	1.341 (.003)	1.160 (.003)	1.140 (.005)	87.8	94.2
D	-1.2 (.44)	1.08 (.08)	1.127 (.002)			89.8	95.6
C’	-1.3 (.45)	1.14 (.10)	1.346 (.003)	1.161 (.003)	1.144 (.005)	90.0	95.3
D’	2.0 (.44)	1.09 (.08)	1.132 (.002)			89.5	93.5

^a Estimated standard errors given in parentheses.

For each design we report the average, $\text{avg}(\hat{\Sigma}(\hat{\beta})^{-1})$, of the 550 inverses of the "expected" information, as well as an empirical standard error estimate for this average. In addition, for Designs C and D, we give the average of variance estimators for $\hat{\beta}$ using $\hat{V}_1(\hat{\beta})$ and $\hat{V}_2(\hat{\beta})$, respectively, as score statistic variance estimators, along with empirical standard error estimates for these averages.

Table 1 also gives empirical coverage rates of nominal 90% confidence intervals for β . These nominal confidence intervals are computed as $\hat{\beta} \pm 1.64\{\hat{\Sigma}(\hat{\beta})^{-1}\}$ for Designs A, D, and D', and as $\hat{\beta} \pm 1.64\{\hat{\Sigma}(\hat{\beta})^{-1}\hat{V}_2(\hat{\beta})\hat{\Sigma}(\hat{\beta})^{-1}\}$ for Designs B, C, and C'. Empirical coverage rates for nominal 95% confidence intervals, constructed by replacing 1.64 by 1.96 in the above expressions, are also listed.

For Design B, $\hat{\Sigma}(\hat{\beta})^{-1}$ overestimated the corresponding covariance-corrected variance estimators by about 16%–18%. Similar results were found for Designs C and C'. Furthermore, the sample variance for these designs agreed closely with the average of the "covariance-corrected" variance estimators, $\hat{\Sigma}(\hat{\beta})^{-1}\hat{V}(\hat{\beta})\hat{\Sigma}(\hat{\beta})^{-1}$, although the variability in the sample variance of $\hat{\beta}$ is considerable. The two covariance-corrected variance estimators appear to agree closely under these simulation conditions. The usual (partial likelihood) variance estimator, $\hat{\Sigma}(\hat{\beta})^{-1}$, corresponds reasonably to the sample variance for Designs A, D, and D'.

The actual coverage rates of the 90% and 95% covariance-corrected confidence intervals were at or near their nominal level under all designs, except that the coverage of the 90% interval under Design C (87.8%) was about 2 standard errors less than the nominal value of 90%. The coverage rates of the confidence intervals without covariance correction (not shown) somewhat exceeded the nominal rates of 90% and 95% for Designs B, C, and C'. For example, the empirical coverage rates were 93.4% and 96.4% for Design B. This is about as expected since the actual coverage rates of nominal 90% and 95% normal confidence intervals when using a variance estimator biased 18% upward are 92.8% and 96.5%, respectively.

Although, under all designs, the sample mean of $\hat{\beta}$ was at most .02, under Design D' this sample mean was over 4 standard errors removed from zero. Under Designs A, C, and D this mean was more than 2 standard errors from zero. It is unclear whether these discrepancies represent actual biases under these cohort conditions, or are simulation artifacts.

A second simulation (not shown) with $\beta = \log(2)$ gave quite similar results.

Designs A, B, and C have the same number of at-risk controls in each risk set. Therefore, if costs are in proportion to the total number of at-risk controls, the efficiency of these three designs can be meaningfully compared. Because of the large residual sampling variability of the sample variance of $\hat{\beta}$, better efficiency comparisons may result from comparing $\hat{\Sigma}(\hat{\beta})^{-1}$ for Designs A, D, and D' and $\hat{\Sigma}(\hat{\beta})^{-1}\hat{V}(\hat{\beta})\hat{\Sigma}(\hat{\beta})^{-1}$ for Designs C, B, and B'. As expected, Design C is more efficient than Design B. Nonetheless, the difference ($1.22 \times 10^{-2} - 1.14 \times 10^{-2} = .08 \times 10^{-2}$) is surprisingly small (although statistically significant) and entirely due to the differences in $\hat{\Sigma}(\hat{\beta})^{-1}$. This is surprising in view of the sizable correlation between exposure and "at-risk status." As expected, Design C is more efficient than Design A. Furthermore, as expected, Design D is more efficient than Design C, because the number of controls per risk set is much larger.

5. Sampling with Stratified Models

To allow for stratification the above regression model can be generalized to

$$\lambda\{t; Z(u), 0 \leq u < t\} = \lambda_{0s}(t)r\{\mathbf{X}(t)\beta\},$$

where the stratification variable $s = s(t)$ is defined in terms of functions of $\{Z(u), 0 \leq u < t\}$ and may be time-dependent. A cohort will be defined to be stratum-closed if all

subjects at risk in stratum s at any time $t > u$ are also at risk in stratum s at time u , if any subject is at risk at u . Otherwise, the cohort is stratum-open. A closed cohort with time-independent stratification will be stratum-closed.

A natural generalization of each of the above sampling designs arises by regarding all subjects ever at risk in stratum s as constituting a cohort for each $s = 1, 2, \dots$. Relative risk estimation is then based on the product of terms in equation (1) across strata. For example, sampling in stratum s at time t would take place for Design D by randomly selecting subjects in the stratum s cohort without failure prior to time t , until a specified number of controls "at risk" and in stratum s at t have been obtained who have not previously been selected as a control in stratum s . The fact that the same study subject may serve as a control in more than one stratum does not introduce biases in parameter estimation since estimation is based wholly on within-stratum comparisons.

The previous discussion of estimation will apply directly to this stratified model under a "large-stratum limiting model" in which the stratum size $n_s \rightarrow \infty$ for each $s = 1, 2, \dots$. Under a "sparse-stratum limiting model" in which the stratum sizes are bounded for a nonnegligible fraction of the overall cohort, the bias in $n^{-1/2}U(\beta)$ generally will not approach zero in stratum-open cohorts under Designs B, C, and D. Hence $\hat{\beta}$ can be inconsistent in these circumstances. ($\hat{\beta}$ will be consistent under Design B in the absence of interval censoring.)

6. Discussion

The standard case-control-within-cohort sampling procedure, Design A, has the benefit of attractive asymptotic properties and simple variance estimation under a range of limiting models. It is, however, a drawback that the effective number of "at-risk" controls may be less than a prespecified number for some cases owing to a sharing of common controls among cases. Therefore, if costs increase in proportion to the total number of controls, Design C may be slightly preferred to Design A for reasons of efficiency.

If, on the other hand, costs are proportional to the total number of "distinct" at-risk controls, Design D would be preferable to Designs A and C. When costs are proportional to the number of distinct controls, it is not quite fair to compare the efficiency of Design A with that of Design D since the number of distinct controls on whom covariate data must be assembled can be somewhat greater under Design D than under Design A. Specifically, if under Design A, we specify r "at-risk" controls per risk set and, under Design D, we specify r "at-risk" controls that have not been previously selected as controls, the total number of distinct "at-risk" controls will, in open cohorts, be greater under Design D. This reflects the fact that the number of "newly-at-risk" controls at time t under Design D may exceed r since some of the "newly-at-risk" controls at t may have been previously selected as controls before becoming at risk. Nonetheless, we can modify Design D to produce a design that has the same number of distinct controls as Design A and yet is guaranteed to be more efficient than Design A. Specifically, modify Design D so that control selection proceeds at each failure time until we obtain r "at-risk" controls who have not been at-risk controls for a previous case (without regard to whether these r new "at-risk" controls were previous controls). Under this modified Design D, β maximizing equation (1) will still be consistent if, at each failure time, the pool of "at-risk" subjects who have not been previously selected as controls remains large. In fact, if costs are proportional to the number of distinct controls, we could profitably further expand upon this modification of Design D by replacing $\hat{R}(t)$ in equation (1) by the union of $\hat{R}(u)$, $u \leq t$, thereby using every control selected up to time t who was at risk at t , and allowing control selection to proceed at each failure time t until r "at-risk" controls have been obtained who were not at-risk controls for a previous case. This sampling procedure can be viewed as a variation of case-cohort sampling (Prentice, 1986b) in which the subcohort

is selected to be rich in "at-risk" controls at each failure time. Further study of this design is indicated.

ACKNOWLEDGEMENT

This work was partially supported by grant GM-24472 from the National Institutes of General Medical Sciences, and grants KO4-ES00180, 5-P30-ES00002, and RO1-ES03405 from NIEHS.

RÉSUMÉ

Plusieurs plans d'échantillonnage ont été proposés pour la réalisation d'études cas-témoin au sein de cohortes ouvertes. La consistance et l'efficacité de l'estimation du risque relatif est présentée pour ces différents plans, et une courte étude de simulation est rapportée. Si le coût de l'étude augmente proportionnellement au nombre total de témoins "à risque", le plan C, décrit ci-dessous, est le plus efficace: Pour un échec survenant au temps t , les témoins sont tirés au sort (indépendamment de leur exposition au risque), parmi une cohorte telle que: (i) il n'y ait pas eu d'échec avant le temps t , et, (ii) les sujets n'aient pas été choisis auparavant comme témoins. A chaque t , l'échantillonnage des témoins se poursuit jusqu'à ce qu'un nombre préétabli de témoins "à risque" au temps t soit atteint. L'avantage en terme d'efficacité du schéma C sur le schéma habituel proposé par Thomas (dans Appendix to Liddell, McDonald, and Thomas, 1977, *Journal of the Royal Statistical Society, Series B* **140**, 469-490) est souvent faible. Par ailleurs, si le coût augmente proportionnellement au nombre de témoins "à risque" différents, le schéma C n'est alors plus le plan le plus efficace. Dans ce cas, plusieurs alternatives sont proposées.

REFERENCES

- Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **74**, 187-220.
- Oakes, D. (1981). Survival times: Aspects of partial likelihood. *International Statistical Review* **49**, 235-264.
- Prentice, R. L. (1986a). On the design of synthetic case-control studies. *Biometrics* **42**, 301-310.
- Prentice, R. L. (1986b). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1-11.
- Prentice, R. L. and Self, S. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Annals of Statistics* **11**, 804-813.
- Robins, J. M., Gail, M. H., and Lubin, J. H. (1986). More on "Biased selection of controls for case-control analyses of cohort studies." *Biometrics* **42**, 273-299.
- Thomas, D. C. (1977). In Appendix to Liddell, F. D. K., McDonald, J. C., and Thomas, D. C. *Journal of the Royal Statistical Society, Series B* **140**, 469-490.

Received October 1987; revised January 1989.

APPENDIX 1

Asymptotic Unbiasedness of the Score Statistic Under Designs B, C, and D in the Large-Stratum Limiting Model

We shall consider only Design C in detail since the argument for Designs B and D is similar. We assume that standard stability and regularity conditions hold for asymptotic normality with mean zero of the score statistic under full cohort sampling (e.g., Prentice and Self, 1983). Define $L(t)$ to be those subjects who are "at risk" and without failure at t who have not been selected as controls at any $u < t$ under Design C and let $l(t)$ be the number of such subjects. Let $G(t)$ be the distribution function of observed failure time in the cohort (without regard to covariates).

Theorem A1 Suppose under Design C for some fixed constant α , $\lim_{n \rightarrow \infty} l(t)/n > \alpha$ for all $t < G^{-1}(1)$. Then, for bounded $X(t)$, $E[U(\beta)/n] = 0 + o(n^{-1/2})$. The proof proceeds by a series of lemmas. We require some definitions.

Let $R(t)$ be the set of subjects at risk at t . In a slight abuse of notation, in Appendix 1 only, let $\tilde{R}(t)$ denote the subset of "at-risk" individuals in $\tilde{R}(t)$ as defined above. The quantities $r(t)$ and $\tilde{r}(t)$

are, respectively, the number of subjects in $R(t)$ and $\tilde{R}(t)$. $F(t) = \{[N_i(u), Y_i(u), Z_i(u), 0 \leq u < t, i \in R(t)], R(t)\}$. Let $\tilde{F}(t)$ be $F(t)$ but with $\tilde{R}(t)$ replacing $R(t)$. Let \tilde{r} be the finite maximum of $\tilde{r}(t)$ over all t . It is fixed by design under Design C. Let $S(t)$ be a random subset of $R(t)$ of size $s(t) < \tilde{r}(t)$ chosen without replacement.

Lemma A1 If $\lim_{n \rightarrow \infty} l(t')/n > \alpha$ for $t' < t$ then, under Design C, for any subjects i and k ,

$$\lim_{n \rightarrow \infty} n^{1/2} \left| \frac{\Pr[i \in L(t) - S(t) | i \in R(t) - S(t), t_i > t, F(t)]}{\Pr[k \in L(t) - S(t) | k \in R(t) - S(t), t_k > t, F(t)]} - 1 \right| = 0.$$

Proof A long and difficult proof has been obtained by David Freedman of the University of California at Berkeley in collaboration with the first author and will be published elsewhere. Here we provide a nonrigorous argument that contains the kernel of that proof. Consider a subject alive and at risk at failure time t_j , $t_j < t$, who has not been selected as a control prior to t_j . Then given $l(t_j)$, $l(t_j) > \tilde{r}(t_j)$, the probability that the subject will not be selected as a control at t_j is the probability that $\tilde{r}(t_j)$ of the $l(t_j) - 1$ other subjects "at risk" at t_j who have not been previously selected as controls will be chosen before the subject. This probability is $[l(t_j) - \tilde{r}(t_j)]/l(t_j)$. Consider now a subject who is alive but not at risk at t_j who has not been previously selected as a control. The probability that this subject will not be selected as a control at t_j is the probability that $\tilde{r}(t_j)$ of the $l(t_j)$ subjects at risk at t_j who have not been previously selected as controls will be chosen before the subject. This probability is

$$[l(t_j) - \tilde{r}(t_j) + 1]/[l(t_j) + 1],$$

which is strictly greater than $[l(t_j) - \tilde{r}(t_j)]/l(t_j)$. Define

$$P_{\min}(t) = \prod \frac{l(t_j) - \tilde{r}(t_j)}{l(t_j)} \quad \text{and} \quad P_{\max}(t) = \prod \frac{l(t_j) - \tilde{r}(t_j) + 1}{l(t_j) + 1},$$

where the products are over all death times t_j prior to t . Since, by assumption, with probability approaching 1, $l(t_j) > \tilde{r}(t_j)$ at all death times $t_j < t$, it follows that if the $l(t_j)$ were fixed rather than random, $P_{\min}(t)$ ($P_{\max}(t)$) would be the probability that a subject in $R(t)$ who was at risk at all (no) t_j prior to t had not been selected as a control prior to t . $P_{\min}(t)$ and $P_{\max}(t)$ would therefore bound $\Pr[i \in L(t) | i \in R(t), t_i > t, F(t)]$. More generally, by essentially the same argument, $\Pr[i \in L(t) - S(t) | i \in R(t) - S(t), t_i > t, F(t)]$ would be bounded by $P_{\max}(t)$ and $P_{\min,s}(t) = \prod [l(t_j) - \tilde{r}(t_j) - s]/[l(t_j) - s]$. Of course, the $l(t_j)$ are random, but the formal proof of Freedman and Robins shows that to $o(n^{-1/2})$ we may ignore this randomness. Now,

$$\begin{aligned} 1 > \frac{P_{\min,s}(t)}{P_{\max}(t)} &= \prod \left[1 - \frac{\tilde{r}(t_j)(s + 1)}{[l^2(t_j) - \tilde{r}(t_j) + 1][l(t_j) - s]} \right] \\ &> \prod \left[1 - \frac{\tilde{r}^2}{[l(t_j) - \tilde{r}]^2} \right] > \left[1 - \frac{\tilde{r}^2}{n^2[l(t_j)/n - \tilde{r}/n]^2} \right]^n > 1 - \frac{\tilde{r}}{\alpha^2 n} + o\left(\frac{1}{n}\right), \end{aligned}$$

which proves the lemma. The following is a direct consequence of Lemma A1.

Lemma A2 Under Design C, for any set $M(t) \subset R(t)$ with $\tilde{r}(t) - 1$ elements, under the conditions of Lemma A1,

$$\begin{aligned} \Pr[M(t) \subset L(t) | N(t) \neq N(t^-), M(t) \subset R(t), F(t), t_j > t \text{ for } j \in M(t)] \\ = \left\{ \prod_{j=1}^{\tilde{r}(t)-1} \left[\frac{l(t) + 1 - j}{(r(t) - 1) + 1 - j} \right] [1 + o(n^{-1/2})] \right\}, \end{aligned}$$

where $N(t) \neq N(t^-)$ means that $N_l(t) \neq N_l(t^-)$ for some $l \in \tilde{R}(t)$.

Lemma A3 Under the conditions of Lemma A1, under Design C, for $k \in \tilde{R}(t)$,

$$\Pr[N_k(t) \neq N_k(t^-) | \tilde{F}(t), N(t) \neq N(t^-)] = \frac{r(\mathbf{X}_k(t)\boldsymbol{\beta})}{\sum_{l \in R(t)} r(\mathbf{X}_l(t)\boldsymbol{\beta})} [1 + o(n^{-1/2})]. \tag{A1.1}$$

Proof

$$\begin{aligned} \Pr[N_k(t) \neq N_k(t^-), \tilde{R}(t) | F(t), N(t) \neq N(t^-)] &= \frac{r(\mathbf{X}_k(t)\beta)}{\sum_{l \in R(t)} r(\mathbf{X}_l(t)\beta)} \\ &\times \Pr[\tilde{R}(t) - \{k\} | F(t), N(t_k) \neq N(t_k^-)] \end{aligned} \tag{A1.2}$$

But

$$\begin{aligned} &\Pr[\tilde{R}(t) - \{k\} | F(t), N(t_k) \neq N(t_k^-)] \\ &= \{\Pr[\tilde{R}(t) - \{k\} \subset L(t) | F(t), N(t_k) \neq N(t_k^-)]\} \Pr[\tilde{R}(t) - \{k\} | F(t), \tilde{R}(t) - \{k\} \subset L(t), N(t_k) \neq N(t_k^-)]. \end{aligned}$$

By Lemma A2 and standard combinatorial arguments, this product equals

$$\begin{aligned} &\left\{ \prod_{j=1}^{\tilde{r}(t)-1} \left[\frac{l(t) + 1 - j}{(r(t) - 1) + 1 - j} \right] [1 + o(n^{-1/2})] \right\} \frac{(\tilde{r}(t) - 1)! [l(t) - (\tilde{r}(t) - 1)]!}{(l(t))!} \\ &= [1 + o(n^{-1/2})] \frac{(\tilde{r}(t) - 1)! (r(t) - 1 - [\tilde{r}(t) - 1])!}{(r(t) - 1)!}. \end{aligned}$$

Therefore, dividing equation (A1.2) by $\sum_{k \in \tilde{R}(t)} \Pr[N_k(t) \neq N_k(t^-), \tilde{R}(t) | F(t), N(t) \neq N(t^-)]$ yields equation (A1.1). Finally, equation (A1.1) implies that for bounded $X(t)$, $E[\mathbf{U}(\beta)] = o(1/n^{1/2})$, proving Theorem A1.

APPENDIX 2

In this appendix we shall assume that conditions sufficient for the pseudolikelihood score to be asymptotically unbiased over Designs B, C, and D hold (see, for example, Theorem A1 in Appendix 1). Furthermore, for these three designs, any statements of equality given below are assumed to hold only to the appropriate order. For notational convenience, assume that β is one-dimensional.

Consistency of the empirical covariance estimators We argue that, for $g \in \{B, C\}$, $\tilde{C}_2(\beta)/n$ will, under regularity conditions, be consistent for its expectation $E[\tilde{C}_2(\beta)/n]$ under Design g or g' . If so, under regularity conditions, the empirical covariance estimators will be consistent.

Define $\Delta_C(t_i, t_j) = \Delta(t_i, t_j)$ and $\Delta_B(t_i, t_j) = Y_j(t_i)\Delta(t_i, t_j)$. Let $H_g(t)$ be the event that records both (a) the observed number of deaths in the cohort through time t and (b) the set

$$\{(i, j); (i, j) \in S_g \text{ and } t_j \leq t\} \equiv S_g(t) \text{ where } (i, j) \in S_g \Leftrightarrow \Delta_B(t_i, t_j) = 1.$$

If we can find a random variable $T_{ji}^g(\beta)$ such that under Design g or g' , for $t_i < t_j$,

$$E[T_{ji}^g(\beta) | H_g(t_j)] = \text{cov}[U_i(\beta), U_j(\beta) | H_g(t_j)], \tag{A2.1}$$

then $\sum_{j=i}^n \delta_j \sum_{|i| t_i < t_j} \delta_i T_{ji}^g(\beta)/n \equiv T^*$ will converge, under suitable regularity conditions, to its expectation provided

$$\text{var}[T^*] = 0 + o(1). \tag{A2.2}$$

Lemma If $T_{ji}^g(\beta) = U_i(\beta)U_j(\beta)$ for $(i, j) \in S_g$ and $T_{ji}^g = 0$ otherwise, then equation (A2.1) holds for Design g or g' . (Note that $T^* \equiv \frac{1}{2}\tilde{C}_2(\beta)/n$.)

Proof It is sufficient to show that if $(i, j) \notin S_g$ then

$$\text{cov}[U_i(\beta), U_j(\beta) | H_g(t_j)] = 0 \text{ under Design } g \text{ or } g'.$$

To show this for the case $g = C$, we calculate that under Design C or C' , for $k \in \tilde{R}(t)$,

$$\Pr\{N_k(t) \neq N_k(t^-) | F_g(t), N(t) \neq N(t^-)\} = \frac{Y_k(t)r\{X_k(t)\beta\}}{\sum_{l \in R(t)} Y_l(t)r\{X_l(t)\beta\}}, \tag{A2.3}$$

where, for $g \equiv C$,

$$F_C(t) = \{[N_i(u), Y_i(u), Z_i(u), 0 \leq u < t, i \in K(t)], \tilde{R}(t), A(t, u), \Delta_C(u, t) = 0, 0 \leq u < t\}. \tag{A2.4}$$

Here $\tilde{R}(t)$ is, as previously, the set consisting of the case failing at t , and all controls selected at t ; $A(t, u) = \tilde{R}(t) \cup \tilde{R}(u)$; $K(t)$ is the union of the $\tilde{R}(u)$ for all $u < t$; and $\Delta_C(u, t) \equiv 1$ if the case failing at t has been selected as a control at u , and is 0 otherwise. Now, by the argument given in Prentice (1986a, bottom of p. 304), it follows from equation (A2.3) that if $\Delta_C(t_i, t_j) = 0$, $t_i < t_j$, then

$$\text{cov}\{U_i(\beta), U_j(\beta) | F_g(t_j), N(t_j) \neq N(t_j^-)\} = 0. \tag{A2.5}$$

But $\Delta_C(t_i, t_j) = 0 \Leftrightarrow (i, j) \notin S_C$.

For $g = B$ we use the fact that under Design B or B', equation (A2.3) holds when we define $\tilde{F}_B(t)$ by equation (A2.4) except that we (1) replace all C's by B's and (2) define $\Delta_B(u, t) \equiv 1$ if the case failing at t has been selected as a control at u and was at risk at u , and $\Delta_B(u, t) \equiv 0$ otherwise. Equation (A2.1) again follows since $\Delta_B(t_i, t_j) = 0 \Leftrightarrow (i, j) \notin S_B$.

Thus it only remains to show

$$\begin{aligned} & \lim_{n \rightarrow \infty} n \text{var}[\frac{1}{2}\tilde{C}_2(\beta)/n] \\ &= \lim_{n \rightarrow \infty} \left[\sum_{(i,j) \in S_g} \frac{\text{var}[U_i(\beta)U_j(\beta)]}{n} + \frac{2}{n} \sum_{\substack{(i,j),(k,l) \in S_g \\ (i,j) \neq (k,l)}} \text{cov}[U_i(\beta)U_j(\beta), U_k(\beta)U_l(\beta)] \right] \\ &= O(1). \end{aligned} \tag{A2.6}$$

To show that equation (A2.6) holds it is sufficient to show that each of the two terms in the sum is $O(1)$. Since $\text{var}[U_i(\beta)U_j(\beta) | H_g(t_j)]$ is $O(1)$, the first term in the sum will be $O(1)$ if the number of elements (ordered pairs) in S_g is $O(n)$. For Designs C, C', B, and B', it immediately follows that the number of elements in S_g is $O(n)$ since the number of elements in S_g is less than the number of failures.

We have been unable to characterize regularity conditions under which the second term in the sum in equation (A2.6) will be $O(1)$, although we expect the conditions will be fairly mild. Hence, simply assume that the regularity conditions necessary for this term to be $O(1)$ hold.

Consistency of Prentice covariance estimators under Designs B, B', C' It follows from the above discussion that we would expect $\tilde{C}_i(\hat{\beta})/n$ to also be consistent for $E[C_2(\beta)/n]$ under Design g or g' , $g \in (B, C)$, provided, for $t_i < t_j$,

$$E[v_{ji} | H_g(t_j)] = \text{cov}[U_i(\beta), U_j(\beta) | H_g(t_j)] \quad \text{for } (i, j) \in S_g. \tag{A2.7}$$

Now with $g = B$, equation (A2.3) holds under Designs B and B', with $\Delta_B(u, t) = 1$ substituted for $\Delta_B(u, t) = 0$ in the previous definition of $F_B(t)$. But if $\Delta_B(t_i, t_j) = 1$ (i.e., $(i, j) \in S_B$), then the controls at t_i include exactly one member of the set $\tilde{R}(t_j)$. In fact,

$$\tilde{R}(t_i) = A(t_j, t_i) - \tilde{R}(t_i) + \{k\}$$

with probability proportional to (A2.3) for $k \in \tilde{R}(t_j)$. Therefore,

$$\text{cov}[U_j(\beta), U_i(\beta) | F_g(t), N(t_j) \neq N(t_j^-)] = v_{ji}, \tag{A2.8}$$

which implies that equation (A2.7) holds (Prentice, 1986a).

Under Design C', equation (A2.3) holds when we replace $\Delta_C(u, t) = 0$ with $\Delta_C(u, t) = 1$ in the definition of $F_C(t)$. Thus, equation (A2.8) and, therefore, equation (A2.7) hold under Design C' with $g = C$.

On the other hand, under Design C, equation (A2.3), with $\Delta_C(u, t) = 1$ replacing $\Delta_C(u, t) = 0$ in the definition of $F_C(t)$, does not hold. To see why, suppose $\Delta_C(u, t) = 1$, four of the six members of $\tilde{R}(t)$ were at risk at u , $\tilde{R}(u)$ contains, by design, six subjects, and $A(t, u) - \tilde{R}(t)$ contained only five subjects. Therefore, the case failing at t (who was a control at u) must have been at risk at u (in order that $\tilde{R}(u)$ contain six members). Therefore, the two members of $\tilde{R}(t)$ who were not at risk at u cannot have been the case at t given the information in $F_C(t)$. Therefore, equation (A2.3), and thus (A2.8), would not hold. Thus, we are unable to prove that equation (A2.7) holds.

Note the fact that we are unable to prove that $\tilde{C}_i(\hat{\beta})/n$ is consistent under Design C does not prove that $\tilde{C}_i(\hat{\beta})/n$ is inconsistent. In fact, our simulation results suggest that it may be consistent.

Covariances under Designs D and D' In this subsection, we show that $\text{cov}[U_i(\beta), U_j(\beta)]$ can be nonzero under Design D' or D. Nonetheless, we conjecture, but have been unable to prove that, as

suggested by our simulation study, the nonzero $\text{cov}[U_i(\beta), U_j(\beta)]$ are asymptotically negligible under a limiting model in which the number of subjects in each stratum increases without bound.

To show that $\text{cov}[U_i(\beta), U_j(\beta)]$ need not be zero, define the event $F_D(t)$ to be the intersection of the event $F_C(t)$ as defined in equation (A2.4) with the event $S(t)$ where $S(t) = \{s(t, u), 0 \leq u < t\}$ and $s(t, u)$ is the number of "at-risk" controls selected at t who were selected as controls at u . Under Design D or D', we calculate that equation (A2.3) still holds provided $s(t, u) = 0$ and thus $\text{cov}[U_i(\beta), U_j(\beta)] = 0$ if $s(t_j, t_i) = 0$.

Next suppose $s(t_j, t_i) \neq 0$. Redefine $F_D(t)$ as the intersection of $F_C(t)$ with $S(t)$ except that, in the definition of $F_C(t)$, we replace $\Delta_C(u, t) = 0$ by $\Delta_C(u, t)$. Then, under Design D', for any time $u < t$ and for $k \in \tilde{R}(t)$,

$$\begin{aligned} \Pr[N_k(t) \neq N_k(t^-), IS(t, u, k) = \{a_1, a_2, \dots, a_{s(t,u)}\} | F_D(t), N(t) \neq N(t^-)] \\ = \left[\frac{Y_k(t)r\{X_k(t)\beta\}}{\sum_{l \in \tilde{R}(t)} Y_l(t)r\{X_l(t)\beta\}} \right] \binom{m-1}{s(t, u)}^{-1}, \end{aligned}$$

where conditional on subject k being the case, $IS(t, u, k)$ is a list of $s(t, u)$ members of $\tilde{R}(t_j) - \{k\}$ who were the controls at u , m is the number of subjects in $\tilde{R}(t)$, and $\binom{a}{b} = a!/[b!(a-b)!]$. It then follows that when $\Delta_C(t_i, t_j) = 0, t_i < t_j$,

$$\begin{aligned} \text{cov}[U_i(\beta), U_j(\beta) | F_D(t_j), N(t_j) \neq N(t_j^-)] \\ = \binom{m-1}{s(t_j, t_i)}^{-1} \sum_{k \in \tilde{R}(t_j)} r_{kj} R_j^{-1} (c_{kj} - \beta_j R_j^{-1}) \left[\sum_{IS(t_j, t_i, k)=1}^{\binom{m-1}{s(t_j, t_i)}} \left[c_{ii} - \frac{\sum_l b_{li}}{\sum_l r_{li}} \right] \right] \equiv v_{ji}, \quad (\text{A2.9}) \end{aligned}$$

where $l \in A(t_j, t_i) - \tilde{R}(t_j) + IS(t_j, t_i, k)$ and we have arbitrarily numbered the sets $IS(t_j, t_i, k)$ from 1 to $\binom{m-1}{s(t_j, t_i)}$. If $\Delta_C(t_i, t_j) = 1$, equation (A2.9) is unchanged except that $l \in A(t_j, t_i) - \tilde{R}(t_j) + IS(t_j, t_i, k) + \{k\}$.

Using equation (A2.9), it is easy to construct a simple example to show that unconditionally $\text{cov}[U_i(\beta), U_j(\beta)]$ need not equal 0 if $s(t_j, t_i) \neq 0$.