



## Exposure Stratified Case-Cohort Designs

ØRNULF BORGAN

borgan@math.uio.no

*Department of Mathematics, University of Oslo, P. O. Box 1053 Blindern, N-0316 Oslo, Norway*

BRYAN LANGHOLZ

*Department of Preventive Medicine, University of Southern California, School of Medicine, 1540 Alcazar Street  
CHP-220, Los Angeles, California 90089-9011, U.S.A.*

SVEN OVE SAMUELSEN

*Department of Mathematics, University of Oslo, P. O. Box 1053 Blindern, N-0316 Oslo, Norway*

LARRY GOLDSTEIN

*Department of Mathematics, University of Southern California, 1042 W. 36th Place, Los Angeles, California  
90089-1113, USA*

JANICE POGODA

*Statology, 10355 Pine Cone Way, Truckee, California 96161, USA*

*Received June 25, 1998; Revised May 6, 1999; Accepted June 10, 1999*

**Abstract.** A variant of the case-cohort design is proposed for the situation in which a correlate of the exposure (or prognostic factor) of interest is available for all cohort members, and exposure information is to be collected for a case-cohort sample. The cohort is stratified according to the correlate, and the subcohort is selected by stratified random sampling. A number of possible methods for the analysis of such exposure stratified case-cohort samples are presented, some of their statistical properties developed, and approximate relative efficiency and optimal allocation to the strata discussed. The methods are compared to each other, and to randomly sampled case-cohort studies, in a limited computer simulation study. We found that all of the proposed analysis methods performed well and were more efficient than a randomly sampled case-cohort study.

**Keywords:** case-cohort studies, Cox regression, optimal allocation, pseudo-likelihood, relative efficiency, score-unbiasedness, stratified sampling, survival analysis

### 1. Introduction

As proposed by Prentice (1986), a case-cohort study for failure time data consists of a random sample from the cohort, the subcohort, and any additional cases not in the subcohort. Covariate information is collected on this sample, rather than the entire cohort. Using a case-cohort design can be very cost-efficient in that a sample much smaller than the full cohort generally results in only a small loss in statistical efficiency. Because the same subcohort may be used as a control group for multiple outcomes, it is particularly well suited for clinical studies in which various clinical outcomes, such as relapse or death, are evaluated with respect to a fixed set of prognostic factors or treatments.

Recently, “two-stage” designs have been proposed in which exposure-related information, available on the entire study group, is used to obtain a sample that is more informative about exposure-related questions than simple random sampling. Here, we are using the term “ex-

posure” loosely to refer to a factor that is of primary interest in the study. This could be some agent that is believed to play a role in causing disease or a treatment to be investigated. Stratified sampling by a correlate of the exposure results in large efficiency gains for unmatched case-control studies (Breslow and Cain, 1988) and in nested case-control samples using the counter-matching method (Langholz and Borgan, 1995). The success of these designs motivated us to explore whether analogous methods would be advantageous for case-cohort sampling. In such a design the subcohort would consist of subjects randomly sampled within two or more exposure-related strata, typically with some strata disproportionately represented. For example, PCR analysis is an accurate, but expensive, way to assess the viral load among HIV infected patients and, thus should be a good predictor of time to AIDS and to death. There are other less accurate, but much less expensive, assays that measure viral load such as the level of P-24 antigen. Thus, a natural study design to investigate the prognostic value of PCR analysis would be to determine P-24 levels in a cohort of HIV infected patients and select a subcohort which over-samples subjects with high P-24 values.

It is not completely clear how one should analyze a stratified case-cohort study. In this paper, therefore, we investigate a number of potential strategies for analyzing case-cohort data where the subcohort is selected by stratified random sampling, and compare their performance relative to each other and relative to the existing methods for analyzing case-cohort data with simple random sampling of the subcohort. The estimators we consider are described in Section 2. All of these are based on a pseudo-likelihood in the spirit of Prentice (1986). In Section 3 we investigate which of the proposed estimators are score-unbiased in the sense that the expectation of the pseudo-score (i.e. the derivative of the log-pseudo-likelihood) is exactly equal to zero at the true parameter value. Asymptotic distribution properties are derived in Section 4, and relative efficiency and optimal allocation of individuals from the strata are considered in Section 5. The performance of the estimators is compared in a small simulation study described in Section 6, while Section 7 offers some additional comments and a discussion of future research directions.

## 2. Pseudo-Likelihood Estimators

We assume throughout that failures in the cohort occur according to Cox’s (1972) proportional hazards model, where the hazard function for a subject with vector of covariates  $\mathbf{Z}(t)$  is given by

$$\alpha(t; \mathbf{Z}) = \alpha_0(t) \exp\{\beta_0' \mathbf{Z}(t)\}. \quad (1)$$

Here the baseline hazard  $\alpha_0(t)$  corresponds to the hazard for an individual with covariate vector identically equal to zero, while the regression coefficients  $\beta_0$  measure the effect of the covariates. We denote by  $t_1 < t_2 < \dots$  the times when failures occur and, assuming no tied failures, we let  $i_j$  be the index of the failure at time  $t_j$ .

The subcohort is selected by stratified random sampling as follows. Based on information which is available for everyone, the cohort is partitioned into  $L$  strata. We then select by random sampling  $m_l$  subcohort members without replacement from the  $n_l$  subjects in stratum  $l$ . The subcohort  $\tilde{\mathcal{C}}$  consists of the  $m = \sum_l m_l$  individuals selected from the  $L$

strata. Covariate information is collected for all failing individuals (cases) as well as for the non-failures in the subcohort. Covariate information for non-failures outside the subcohort is, however, not collected.

The estimators for  $\beta_0$  considered in this paper, are all based on maximizing a pseudo-likelihood function of the form

$$\tilde{\mathcal{L}}(\beta) = \prod_{t_j} \left[ \frac{\exp\{\beta' \mathbf{Z}_{i_j}(t_j)\} w_{i_j}(t_j)}{\sum_{k \in \tilde{\mathcal{R}}(t_j)} Y_k(t_j) \exp\{\beta' \mathbf{Z}_k(t_j)\} w_k(t_j)} \right]. \quad (2)$$

Here  $\tilde{\mathcal{R}}(t_j)$  is a “case-cohort set” which may depend on the failure time  $t_j$  and the case  $i_j$ ,  $Y_k(t_j)$  is an at risk indicator for subject  $k$ , and  $w_k(t_j)$  is a weight for this individual which does not depend on  $\beta$  but may depend on  $t_j$  and  $\tilde{\mathcal{R}}(t_j)$ . The various estimators differ in the choice of  $w_k(t_j)$  and  $\tilde{\mathcal{R}}(t_j)$ , and we define the estimators in terms of these. For the special case of no stratification (i.e.  $L = 1$ ), Prentice’s (1986) original suggestion corresponds to  $w_k(t_j) = 1$  and  $\tilde{\mathcal{R}}(t_j) = \tilde{\mathcal{C}} \cup \{i_j\}$ , the subcohort augmented with the case when it occurs outside the subcohort. Self and Prentice (1988), for the purpose of studying large sample properties of Prentice’s estimator, considered the asymptotically equivalent modification where  $\tilde{\mathcal{R}}(t_j) = \tilde{\mathcal{C}}$  only includes the case when it happens to occur inside the subcohort.

We will consider three different type of estimators for the stratified case-cohort design. The idea underlying the first two is to simply replace the denominator of the full cohort partial likelihood by an unbiased estimator computed from the case-cohort sample. We let  $\mathcal{D}$  be the set of all cases, and write  $n_l^0$  and  $m_l^0$ , respectively, for the total number of non-failures in stratum  $l$  and the number of these which belong to the subcohort. Then, with  $s(k)$  the sampling stratum of individual  $k$ , the first two estimators are given by

*Estimator I:*  $\tilde{\mathcal{R}}(t_j) = \tilde{\mathcal{C}}$  and  $w_k(t_j) = n_{s(k)} / m_{s(k)}$

*Estimator II:*  $\tilde{\mathcal{R}}(t_j) = \tilde{\mathcal{C}} \cup \mathcal{D}$  and  $w_k(t_j) = \begin{cases} n_{s(k)}^0 / m_{s(k)}^0 & \text{if } k \in \tilde{\mathcal{C}} \setminus \mathcal{D} \\ 1 & \text{if } k \in \mathcal{D} \end{cases}$

Estimator I is the natural generalization of Self and Prentice’s (1988) estimator to stratified sampling, and it was considered in an unpublished Ph.D.-thesis by one of the authors (Samuelsen, 1989). In the spirit of Kalbfleisch and Lawless (1988), Estimator II includes all at risk cases in the denominator weighted with one to reflect that they are included in  $\tilde{\mathcal{R}}(t_j)$  with probability one. A similar approach was adopted in a recent paper by Kulic and Lin (1998) on case-cohort methodology for additive hazards regression models. Note that Estimator II can be considered the special case of Estimator I in which the stratum definitions also depend on outcome, thus making  $\mathcal{D}$  a stratum on its own and redefining the strata  $l = 1, \dots, L$  by excluding the cases.

As discussed more closely in Section 3, Prentice’s (1986) estimator is score-unbiased, while this is only approximately the case for Self and Prentice’s (1988) suggestion and the Estimators I and II. Further, Prentice’s choice  $\tilde{\mathcal{R}}(t_j) = \tilde{\mathcal{C}} \cup \{i_j\}$  is score-unbiased for stratified sampling only if, when the case occurs outside the subcohort, only the subcohort members in the same sampling stratum as the case are included in the denominator. This is clearly an inefficient estimation method. It turns out that we can obtain score-unbiasedness

as well as an effective use of the information from the subcohort in the following way. Let  $J_l$  be a randomly selected subject among the subcohort members from stratum  $l$ . Then our third estimator is

$$\text{Estimator III: } w_k(t_j) = n_{s(k)}/m_{s(k)} \text{ and } \tilde{\mathcal{R}}(t_j) = \begin{cases} \tilde{\mathcal{C}} & \text{if } i_j \in \tilde{\mathcal{C}} \\ \tilde{\mathcal{C}} \cup \{i_j\} \setminus \{J_{s(i_j)}\} & \text{if } i_j \notin \tilde{\mathcal{C}} \end{cases}$$

Note that in this estimator, if the case occurs outside the subcohort, the subcohort member  $J_{s(i_j)}$  swaps place with the case so that the case  $i_j$  is inside the ‘‘case-cohort set’’  $\tilde{\mathcal{R}}(t_j)$  while the ‘‘swapper’’  $J_{s(i_j)}$  is removed from this set.

In all of Estimators I–III, the weights depend on the number of individuals in the strata and the number of subjects sampled from these at entry to the study, i.e. at  $t = 0$ . However, as time proceeds the number at risk,  $n_l(t)$ , in stratum  $l$  will differ from  $n_l$ , as will the number at risk,  $m_l(t)$ , in the subcohort from this stratum differ from  $m_l$ . This suggests a modification of the above estimators. In Estimator I we replace the weights  $n_{s(k)}/m_{s(k)}$  by the time-dependent ones  $w_k(t_j) = n_{s(k)}(t_j)/m_{s(k)}(t_j)$ . The same modification takes place for Estimator III, but, when the case occurs outside the subcohort, we also replace  $J_{s(i_j)}$  by a time-dependent ‘‘swapper’’  $J_{s(i_j)}(t_j)$  sampled at random among those at risk in the subcohort from the case’s stratum. Finally for Estimator II we replace the weights  $n_{s(k)}^0/m_{s(k)}^0$  for the non-failing individuals by  $w_k(t_j) = n_{s(k)}^0(t_j)/m_{s(k)}^0(t_j)$ , where  $n_l^0(t)$  and  $m_l^0(t)$  are the total number at risk at  $t$ , respectively, among the non-failures in stratum  $l$  and the number of these which belong to the subcohort. As for counter-matching, if the categoric stratification variable is the only covariate in the model, each of these time dependent weight variants yields the full cohort partial likelihood. Thus, in this sense these estimators bring the full cohort marginal information from the exposure-related stratification variable into the sample.

### 3. Unbiasedness Considerations

In order to study score-unbiasedness for estimators based on pseudo-likelihoods of the form (2), we need to define our statistical model more carefully. We first describe the model for the full cohort assumed to consist of  $n = \sum_l n_l$  individuals. For that purpose we fix throughout a time interval  $[0, \tau]$ , and following the counting process formulation of the Cox model as given by Andersen and Gill (1982), we let  $N_i$ ,  $Y_i$ , and  $\mathbf{Z}_i$  be the counting, at risk, and covariate processes for the  $i$ th subject in the cohort. As is usual, we assume that there is a non-decreasing family of  $\sigma$ -algebras  $(\mathcal{H}_t)_{t \in [0, \tau]}$  such that the  $N_i$  are  $(\mathcal{H}_t)$ -adapted and the  $Y_i$  and  $\mathbf{Z}_i$  are predictable with respect to  $(\mathcal{H}_t)$ . Thus  $\mathcal{H}_t$  is the ‘‘cohort history’’ including failure time, censoring, and covariate information up to time  $t$ . The  $(\mathcal{H}_t)$ -intensity process  $\lambda_i$  of  $N_i$  is given heuristically by  $\lambda_i(t) dt = \mathbb{P}\{dN_i(t) = 1 \mid \mathcal{H}_{t-}\}$ , where  $dN_i(t)$  is the increment of  $N_i$  over the small time interval  $[t, t + dt)$ . Assuming censoring to be independent (Andersen *et al.* 1993, Section III.2.2), (1) yields the intensity process

$$\lambda_i(t) = Y_i(t)\alpha_0(t) \exp\{\beta_0' \mathbf{Z}_i(t)\} \quad (3)$$

for  $N_i$ .

Now that a model for the cohort has been given, we describe how the sampling of the subcohort may be superimposed onto this model. To keep our presentation simple, we restrict our attention to the situation where the weights do not depend on time, i.e.  $w_k(t_j) = w_k$  in (2), and assume that the stratification and the weights are based on information available at entry to the study. Thus our formulation covers Estimators I and III, while this is not the case for Estimator II. At the end of this section, we comment upon why Estimator II is not covered by our general set-up, and indicate how our results may be modified to cover the time-dependent modifications of Estimators I and III mentioned in the last paragraph of Section 2.

We introduce  $\mathcal{S}_l$  for the subset of the cohort members who belong to stratum  $l$  so that  $n_l = |\mathcal{S}_l|$ . Since stratification is assumed to depend on information available at time zero, the  $\mathcal{S}_l$  will be  $\mathcal{H}_0$ -measurable. Further we let  $\mathcal{P}$  be the power set of  $\{1, 2, \dots, n\}$ , i.e. the set of all subsets of  $\{1, 2, \dots, n\}$ , and introduce

$$\mathbf{C} = \{\mathbf{c} \in \mathcal{P}: |\mathbf{c} \cap \mathcal{S}_l| = m_l, l = 1, \dots, L\}$$

for the set of possible sampled subcohorts. Finally we let

$$\pi(\mathbf{c}) = \mathbb{P}(\tilde{\mathcal{C}} = \mathbf{c}) = 1 / \prod_{l=1}^L \binom{n_l}{m_l}, \quad (4)$$

for  $\mathbf{c} \in \mathbf{C}$ , be the sampling distribution for the subcohort  $\tilde{\mathcal{C}}$ .

The sampling of the subcohort will induce extra random variation. In order to take care of this, we will now have to work with the enlarged family of  $\sigma$ -algebras  $(\mathcal{F}_t)_{t \in [0, \tau]}$  obtained by augmenting the ‘‘cohort history’’ by the sampling information (at time zero). This may have the consequence that the intensity processes corresponding to the counting processes  $N_i$  may change, i.e. their  $(\mathcal{F}_t)$ -intensity processes may differ from their  $(\mathcal{H}_t)$ -intensity processes (3). To rule out such possibilities we need to assume that the *sampling is independent* in the sense that the additional knowledge of which individuals have been selected to the subcohort does not alter the failure intensities. Thus  $\mathbb{P}\{dN_i(t) = 1 \mid \mathcal{F}_{t-}\} = \mathbb{P}\{dN_i(t) = 1 \mid \mathcal{H}_{t-}\}$  so that the  $(\mathcal{F}_t)$ -intensity processes of the  $N_i$  are also given by (3).

We are then in a position to take a closer look at the pseudo-likelihood (2) and the corresponding pseudo-score. To this end let  $\tilde{\mathcal{R}}_i$  be the ‘‘case-cohort set’’ to be used when/if individual  $i$  fails. Thus for simple random sampling (i.e.,  $L = 1$ ) Prentice (1986) and Self and Prentice (1988) considered the choices  $\tilde{\mathcal{R}}_i = \tilde{\mathcal{C}} \cup \{i\}$  and  $\tilde{\mathcal{R}}_i = \tilde{\mathcal{C}}$ , respectively. The latter is also the one used for Estimator I, while  $\tilde{\mathcal{R}}_i = \tilde{\mathcal{C}} \cup \{i\} \setminus \{J_{s(i)}\}$  for Estimator III. Note that with this notation  $\tilde{\mathcal{R}}(t_j) = \tilde{\mathcal{R}}_{i_j}$ , so that (2) may be reformulated as

$$\tilde{\mathcal{L}}(\beta) = \prod_{i \in [0, \tau]} \prod_{i=1}^n \left[ \frac{\exp\{\beta' \mathbf{Z}_i(t)\} w_i(\tilde{\mathcal{R}}_i)}{\sum_{k \in \tilde{\mathcal{R}}_i} Y_k(t) \exp\{\beta' \mathbf{Z}_k(t)\} w_k(\tilde{\mathcal{R}}_i)} \right]^{\Delta N_i(t)}$$

where we have written  $w_k = w_k(\tilde{\mathcal{R}}_i)$  for the weights in order to emphasize that these may depend on the sets  $\tilde{\mathcal{R}}_i$ . For simple random sampling Prentice (1986) and Self and Prentice (1988) both used the weights  $w_k = 1$ , while in Estimators I and III  $w_k = n_{s(k)}/m_{s(k)}$ . It

should be noted that for all these estimators, the  $\tilde{\mathcal{R}}_i$  and the  $w_k = w_k(\tilde{\mathcal{R}}_i)$  are known at time zero, i.e. they are  $\mathcal{F}_0$ -measurable.

Introduce for  $\mathbf{r} \in \mathcal{P}$  the notation

$$S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta}, t) = \sum_{k \in \mathbf{r}} Y_k(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_k(t)\} w_k(\mathbf{r}), \quad (5)$$

$$\mathbf{S}_{\mathbf{r}}^{(1)}(\boldsymbol{\beta}, t) = \sum_{k \in \mathbf{r}} Y_k(t) \mathbf{Z}_k(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_k(t)\} w_k(\mathbf{r}), \quad (6)$$

$$\mathbf{E}_{\mathbf{r}}(\boldsymbol{\beta}, t) = \mathbf{S}_{\mathbf{r}}^{(1)}(\boldsymbol{\beta}, t) / S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta}, t). \quad (7)$$

Then the pseudo-score becomes

$$\tilde{\mathbf{U}}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log \tilde{\mathcal{L}}(\boldsymbol{\beta}) = \int_0^\tau \sum_{i=1}^n \left\{ \mathbf{Z}_i(t) - \mathbf{E}_{\tilde{\mathcal{R}}_i}(\boldsymbol{\beta}, t) \right\} dN_i(t), \quad (8)$$

and the maximum pseudo-likelihood estimator  $\tilde{\boldsymbol{\beta}}$  is the value of  $\boldsymbol{\beta}$  which maximizes (2) or solves  $\tilde{\mathbf{U}}(\boldsymbol{\beta}) = \mathbf{0}$ . Let us then evaluate the expected value of the pseudo-score at the true parameter vector  $\boldsymbol{\beta}_0$  and consider conditions for score-unbiasedness. Using (3), it is seen that the pseudo-score at  $\boldsymbol{\beta}_0$  may be written as

$$\begin{aligned} \tilde{\mathbf{U}}(\boldsymbol{\beta}_0) &= \int_0^\tau \sum_{i=1}^n \left\{ \mathbf{Z}_i(t) - \mathbf{E}_{\tilde{\mathcal{R}}_i}(\boldsymbol{\beta}_0, t) \right\} dM_i(t) \\ &\quad + \int_0^\tau \sum_{i=1}^n \left\{ \mathbf{Z}_i(t) - \mathbf{E}_{\tilde{\mathcal{R}}_i}(\boldsymbol{\beta}_0, t) \right\} Y_i(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_i(t)\} \alpha_0(t) dt, \end{aligned} \quad (9)$$

where by standard counting process theory (e.g. Andersen *et al.* 1993, Section II.4.1) the  $M_i(t) = N_i(t) - \int_0^t \lambda_i(u) du$  are orthogonal local square integrable ( $\mathcal{F}_t$ )-martingales. Since  $\mathbf{Z}_i(\cdot)$  and  $\mathbf{E}_{\tilde{\mathcal{R}}_i}(\boldsymbol{\beta}_0, \cdot)$  are ( $\mathcal{F}_t$ )-predictable processes, the first term on the right hand side of (9) is a vector valued stochastic integral, and therefore a local square integrable ( $\mathcal{F}_t$ )-martingale. In particular, the expected value of this term is zero. (Here and below we tacitly assume that all expectations considered actually do exist.)

To investigate the expected value of the second term on the right hand side of (9), note that by taking expectation over the sampling, conditional on the entire cohort history, we get for each  $t \in [0, \tau]$

$$\begin{aligned} &\sum_{i=1}^n \mathbf{E}\{\mathbf{E}_{\tilde{\mathcal{R}}_i}(\boldsymbol{\beta}_0, t) | \mathcal{H}_\tau\} Y_i(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_i(t)\} \\ &= \sum_{i=1}^n \left\{ \sum_{\mathbf{r} \in \mathcal{P}} \mathbf{E}_{\mathbf{r}}(\boldsymbol{\beta}_0, t) \mathbf{P}(\tilde{\mathcal{R}}_i = \mathbf{r}) \right\} Y_i(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_i(t)\} \\ &= \sum_{\mathbf{r} \in \mathcal{P}} \frac{\sum_{k \in \mathbf{r}} Y_k(t) \mathbf{Z}_k(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_k(t)\} w_k(\mathbf{r})}{\sum_{k \in \mathbf{r}} Y_k(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_k(t)\} w_k(\mathbf{r})} \sum_{i \in \mathbf{r}} Y_i(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_i(t)\} \mathbf{P}(\tilde{\mathcal{R}}_i = \mathbf{r}) \\ &\quad + \sum_{\mathbf{r} \in \mathcal{P}} \frac{\sum_{k \in \mathbf{r}} Y_k(t) \mathbf{Z}_k(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_k(t)\} w_k(\mathbf{r})}{\sum_{k \in \mathbf{r}} Y_k(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_k(t)\} w_k(\mathbf{r})} \sum_{i \notin \mathbf{r}} Y_i(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_i(t)\} \mathbf{P}(\tilde{\mathcal{R}}_i = \mathbf{r}). \end{aligned} \quad (10)$$

Here  $P(\tilde{\mathcal{R}}_i = \mathbf{r})$  may be derived from the subcohort distribution (4) and the relation between the subcohort  $\tilde{\mathcal{C}}$  and the sets  $\tilde{\mathcal{R}}_i$ , cf. below. In general, it does not seem possible to give a simple expression for (10). However, for an important special case this is possible, namely when the following two conditions are fulfilled:

For all  $k$  and  $\mathbf{r} \in \mathcal{P}$  we have:

$$(A) \quad P(\tilde{\mathcal{R}}_k = \mathbf{r}) = 0 \text{ for } k \notin \mathbf{r}.$$

$$(B) \quad P(\tilde{\mathcal{R}}_k = \mathbf{r}) = \text{const}(\mathbf{r})w_k(\mathbf{r}) \text{ for } k \in \mathbf{r}.$$

Note that Condition A requires the cases to be included in the ‘‘case-cohort sets,’’ while Condition B assumes the weights  $w_k = w_k(\tilde{\mathcal{R}}_i)$  to be proportional to the probability of selecting  $\tilde{\mathcal{R}}_i$  as the ‘‘case-cohort set’’ had  $k$  been the failure.

When Condition A is fulfilled, the second term at the right hand side of (10) vanishes. Moreover, introducing  $\mathcal{P}_k = \{\mathbf{r} \in \mathcal{P}: k \in \mathbf{r}\}$  and using Condition B, the first term equals

$$\begin{aligned} & \sum_{\mathbf{r} \in \mathcal{P}} \sum_{k \in \mathbf{r}} Y_k(t) \mathbf{Z}_k(t) \exp\{\beta'_0 \mathbf{Z}_k(t)\} P(\tilde{\mathcal{R}}_k = \mathbf{r}) \\ &= \sum_{k=1}^n Y_k(t) \mathbf{Z}_k(t) \exp\{\beta'_0 \mathbf{Z}_k(t)\} \sum_{\mathbf{r} \in \mathcal{P}_k} P(\tilde{\mathcal{R}}_k = \mathbf{r}) \\ &= \sum_{k=1}^n Y_k(t) \mathbf{Z}_k(t) \exp\{\beta'_0 \mathbf{Z}_k(t)\} \end{aligned}$$

since  $P(\tilde{\mathcal{R}}_k = \mathbf{r})$  is a probability distribution over sets  $\mathbf{r}$  in  $\mathcal{P}_k$ . Thus if Conditions A and B are fulfilled,

$$\sum_{i=1}^n E\{E_{\tilde{\mathcal{R}}_i}(\beta_0, t) \mid \mathcal{H}_\tau\} Y_i(t) \exp\{\beta'_0 \mathbf{Z}_i(t)\} = \sum_{k=1}^n Y_k(t) \mathbf{Z}_k(t) \exp\{\beta'_0 \mathbf{Z}_k(t)\} \quad (11)$$

so that the expected value over the sampling of the second term at the right hand side of (9) is zero. Therefore Conditions A and B are sufficient for the pseudo-score to have expected value zero. We conjecture that they are necessary as well.

Let us then investigate the implications of this result for the estimators mentioned earlier. Note first that for simple random sampling, i.e.  $L = 1$ , Self and Prentice’s estimator does not include the case in the ‘‘case-cohort set’’ and hence is not score-unbiased. Prentice’s estimator, however, does include the case, and for  $\mathbf{r} \in \mathcal{P}_k$  we find

$$\begin{aligned} P(\tilde{\mathcal{R}}_k = \mathbf{r}) &= P(\tilde{\mathcal{C}} = \mathbf{r}) + P(\tilde{\mathcal{C}} = \mathbf{r} \setminus \{k\}) \\ &= \binom{n}{m}^{-1} I(|\mathbf{r}| = m) + \binom{n}{m}^{-1} I(|\mathbf{r}| = m + 1). \end{aligned}$$

Here the first term on the right hand side corresponds to the situation where  $k$  is a member of the subcohort, while the second corresponds to the situation where  $k$  is not a member. It is seen that Prentice’s estimator fulfills Conditions A and B and hence, as noted by Prentice (1986), is score-unbiased.

Now, consider stratified sampling, where the subcohort  $\tilde{\mathcal{C}}$  is selected according to the sampling probability (4). First note that Estimator I does not include the non-subcohort cases in their “case-cohort sets”, so this estimator is not score-unbiased. Next, consider the situation where  $\tilde{\mathcal{R}}_i = \tilde{\mathcal{C}} \cup \{i\}$ . By a similar reasoning as just given for Prentice’s estimator, we get for  $\mathbf{r} \in \mathcal{P}_k$ ,

$$P(\tilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r}) + \pi(\mathbf{r} \setminus \{k\}).$$

To see the implications of Conditions A and B, let  $i$  correspond to the case so that  $\mathbf{r} \in \mathcal{P}_i$ . If then  $\mathbf{r} \in \mathbf{C}$ , i.e. the case occurs within the subcohort,  $P(\tilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r})$  and Conditions A and B are fulfilled for the weights  $w_k = 1$ . However, if  $\mathbf{r} \notin \mathbf{C}$ , i.e. the case occurs outside the subcohort,  $P(\tilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r} \setminus \{k\})$  which is zero except when  $i$  and  $k$  belong to the same stratum. Thus score-unbiasedness can only be obtained if all non-zero weights are just one, but, if the case is not in the subcohort, only subcohort members in the same sampling stratum as the case are included in the denominator. It is seen that the reason why score-unbiasedness leads to this clearly inefficient estimator when  $\tilde{\mathcal{R}}_i = \tilde{\mathcal{C}} \cup \{i\}$ , is that the structure of the case-cohort sets gives too much information about the case when it occurs outside the subcohort (Pogoda, 1993).

This motivated the construction of our Estimator III, which may be given as follows. Conditional on the chosen subcohort  $\tilde{\mathcal{C}}$ , we select at random (in principle, at time zero), a “swapper”  $J_l \in \tilde{\mathcal{C}} \cap \mathcal{S}_l$  for each stratum  $l$ . Thus  $P(J_l = j \mid \tilde{\mathcal{C}} = \mathbf{c}) = 1/m_l$  for each  $j \in \mathbf{c} \cap \mathcal{S}_l$ . Then, for  $\mathbf{r} \in \mathbf{C} \cap \mathcal{P}_k$ ,

$$P(\tilde{\mathcal{R}}_k = \mathbf{r}) = P(\tilde{\mathcal{C}} = \mathbf{r}) + \sum_j P(\tilde{\mathcal{C}} = \mathbf{r} \cup \{j\} \setminus \{k\}, J_{s(k)} = j),$$

where the sum is over all  $j \notin \mathbf{r}$  with  $s(j) = s(k)$ . Now for all such  $j$

$$P(\tilde{\mathcal{C}} = \mathbf{r} \cup \{j\} \setminus \{k\}, J_{s(k)} = j) = \pi(\mathbf{r} \cup \{j\} \setminus \{k\})/m_{s(k)} = \pi(\mathbf{r})/m_{s(k)},$$

with  $\pi(\mathbf{r})$  given by (4). Since the sum over  $j$  has  $n_{s(k)} - m_{s(k)}$  terms this gives

$$P(\tilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r}) \frac{n_{s(k)}}{m_{s(k)}}.$$

Thus we have proved that Conditions A and B hold for the “swapper approach,” so Estimator III is score-unbiased.

To simplify the presentation, we have in this section assumed the weights not to depend on time. The results extend immediately, however, to the modifications of Estimators I and III mentioned in the last paragraph of Section 2. There are two reasons for this. Firstly, the time-dependent weights of Estimators I and III are predictable, so that the leading term on the right hand side of (9) has expected value zero also for the modified estimators. Secondly, conditional on the cohort history, those at risk in the subcohort at a given time constitute a stratified random sample from everyone at risk, so the unbiasedness arguments based on Conditions A and B also continue to hold for the modified Estimators I and III. Hence the version of Estimator III using time-dependent weights is score-unbiased, while this is not the case for the modified Estimator I.

As mentioned earlier, Estimator II is not covered by the framework considered above. The main reason for this is that the stratification and the weights used for this estimator depend on the complete cohort history  $\mathcal{H}_\tau$ , making the integrand in the leading term on the right hand side of (9) non-predictable. Thus this term no longer has expected value zero and, as a consequence, Estimator II is not score-unbiased.

#### 4. Asymptotic Distribution and Variance Estimation

In their study of the asymptotic properties of the case-cohort estimator for simple random sampling, Self and Prentice (1988) concentrated on the situation where the subcohort  $\tilde{\mathcal{C}}$  is used as the “case-cohort sets” in (2) for all  $t_j$ . Since stratified random sampling is simple random sampling independently between strata, the asymptotic properties of the corresponding Estimator I may be derived as a simple extension of their results. Following Samuelsen (1989, 1997), we will here sketch the main steps in this derivation. At the end of the section we discuss the extent to which similar results hold for the asymptotic distribution of the other estimators considered in this paper.

So for the time being, we restrict our attention to Estimator I with time-fixed weights  $w_k = n_{s(k)}/m_{s(k)}$ . We assume that  $n_l/n \rightarrow v_l > 0$  and  $m_l/n_l \rightarrow \pi_l > 0$  as  $n \rightarrow \infty$ , and that, within each stratum, the regularity conditions of Self and Prentice (1988) hold when simplified to the situation with exponential relative risk function  $r(x) = \exp(x)$ . Write  $\tilde{\mathcal{C}}_l = \tilde{\mathcal{C}} \cap \mathcal{S}_l$  for the subset of the subcohort that belongs to stratum  $l$ , and introduce, for  $\gamma = 0, 1$  (later we avoid boldfacing for  $\gamma = 0$ ),

$$\mathbf{S}_{\tilde{\mathcal{C}}_l}^{(\gamma)}(\boldsymbol{\beta}, t) = \sum_{k \in \tilde{\mathcal{C}}_l} Y_k(t) \mathbf{Z}_k(t)^\gamma \exp\{\boldsymbol{\beta}' \mathbf{Z}_k(t)\} (n_l/m_l), \quad (12)$$

as well as the corresponding cohort quantities  $\mathbf{S}_{(l)}^{(\gamma)}(\boldsymbol{\beta}, t)$  obtained from (12) by omitting the weights and summing over  $k \in \mathcal{S}_l$  instead of  $k \in \tilde{\mathcal{C}}_l$ . Then, in particular, we assume that  $1/n_l$  times (12) and  $n_l^{-1} \mathbf{S}_{(l)}^{(\gamma)}(\boldsymbol{\beta}, t)$  both converge (uniformly over  $\boldsymbol{\beta}$  and  $t$ ) in probability to the same limit as  $n \rightarrow \infty$ . Further, for  $\gamma = 0, 1$ , introduce

$$\mathbf{S}^{(\gamma)}(\boldsymbol{\beta}, t) = \sum_{k=1}^n Y_k(t) \mathbf{Z}_k(t)^\gamma \exp\{\boldsymbol{\beta}' \mathbf{Z}_k(t)\} = \sum_{l=1}^L \mathbf{S}_{(l)}^{(\gamma)}(\boldsymbol{\beta}, t)$$

and

$$\mathbf{S}_{\tilde{\mathcal{C}}}^{(\gamma)}(\boldsymbol{\beta}, t) = \sum_{l=1}^L \mathbf{S}_{\tilde{\mathcal{C}}_l}^{(\gamma)}(\boldsymbol{\beta}, t),$$

and from these define  $\mathbf{E}(\boldsymbol{\beta}, t)$  and  $\mathbf{E}_{\tilde{\mathcal{C}}}(\boldsymbol{\beta}, t)$  as in (7).

We are then in position to take a look at the pseudo-score for Estimator I. To this end we introduce  $\mathbf{U}(\boldsymbol{\beta})$ , the score for the full cohort, obtained by replacing  $\mathbf{E}_{\tilde{\mathcal{C}}}(\boldsymbol{\beta}, t)$  by  $\mathbf{E}(\boldsymbol{\beta}, t)$

in (8). The pseudo-score for Estimator I, evaluated at  $\beta_0$ , may be decomposed as

$$\tilde{\mathbf{U}}(\beta_0) = \mathbf{U}(\beta_0) + \sum_{k=1}^n \{1 - (n_{s(k)}/m_{s(k)})I_k\} \mathbf{X}_k \quad (13)$$

where  $I_k = I(k \in \tilde{\mathcal{C}})$ , and

$$\mathbf{X}_k = \int_0^\tau \{\mathbf{Z}_k(t) - \mathbf{E}(\beta_0, t)\} Y_k(t) \exp\{\beta_0' \mathbf{Z}_k(t)\} S_{\tilde{\mathcal{C}}}^{(0)}(\beta_0, t)^{-1} dN.(t)$$

with  $N. = \sum_{i=1}^n N_i$ . By approximating  $S_{\tilde{\mathcal{C}}}^{(0)}(\beta_0, t)^{-1} dN.(t)$  with  $\alpha_0(t) dt$ , it then follows that the normalized pseudo-score  $n^{-1/2} \tilde{\mathbf{U}}(\beta_0)$  is asymptotically equivalent to

$$n^{-1/2} \mathbf{U}(\beta_0) + n^{-1/2} \sum_{k=1}^n \{1 - (n_{s(k)}/m_{s(k)})I_k\} \mathbf{X}_k^* \quad (14)$$

with

$$\mathbf{X}_k^* = \int_0^\tau \{\mathbf{Z}_k(t) - \mathbf{E}(\beta_0, t)\} Y_k(t) \exp\{\beta_0' \mathbf{Z}_k(t)\} \alpha_0(t) dt. \quad (15)$$

The leading term of (14) is the normalized score for the full cohort, and converges weakly to a mean zero multivariate normal variate with a covariance matrix  $\Sigma$  (Andersen and Gill 1982; see also Andersen *et al.* 1993, Section VII.2). For the second term, we may, conditional on the complete cohort history, apply the finite population large-sample result of Lehmann (1975, pp. 39–40) separately within each stratum. Combining the results over the  $L$  strata, we then get that, conditional on  $\mathcal{H}_\tau$ , the second term of (14) converges weakly to a mean zero multivariate normal variate with covariance matrix

$$\Delta = \sum_{l=1}^L v_l \frac{1 - \pi_l}{\pi_l} \Delta_l. \quad (16)$$

Here  $\Delta_l$  is the limit in probability of the finite-population covariance matrix of the  $X_k^*$  within stratum  $l$  (which exists by our Self-Prentice type conditions). Then, by (3.9) in Samuelsen (1997), it follows that the two terms in (14) are asymptotically independent, and that the unconditional asymptotic distribution of the latter is the same as the conditional one just mentioned. Finally, let  $\tilde{\mathcal{I}}(\beta)$  be the observed pseudo-information for Estimator I. Then  $n^{-1} \tilde{\mathcal{I}}(\beta^*)$  converges in probability to the asymptotic cohort information matrix  $\Sigma$  for any  $\beta^*$  which is consistent for  $\beta_0$ . The usual Taylor series expansions argument gives that  $\sqrt{n}(\tilde{\beta} - \beta_0)$  converges weakly to a mean zero multivariate normal variate with covariance matrix  $\Sigma^{-1} + \Sigma^{-1} \Delta \Sigma^{-1}$  as  $n \rightarrow \infty$ .

The asymptotic covariance matrix of  $\sqrt{n}(\tilde{\beta} - \beta_0)$  may be estimated consistently by  $n$  times  $\tilde{\Sigma}^{-1} + \tilde{\Sigma}^{-1} \tilde{\Delta} \tilde{\Sigma}^{-1}$ . Here  $\tilde{\Sigma} = \tilde{\mathcal{I}}(\tilde{\beta})$  is the observed pseudo-information at  $\tilde{\beta}$ , while

$$\tilde{\Delta} = \sum_{l=1}^L \frac{n_l(n_l - m_l)}{m_l} \Delta_l \quad (17)$$

with  $\tilde{\Delta}_l$  the empirical covariance matrix of the

$$\tilde{\mathbf{X}}_k = \int_0^\tau \{ \mathbf{Z}_k(t) - \mathbf{E}_{\tilde{\mathcal{C}}}(\tilde{\beta}, t) \} Y_k(t) \exp\{ \tilde{\beta}' \mathbf{Z}_k(t) \} S_{\tilde{\mathcal{C}}}^{(0)}(\tilde{\beta}, t)^{-1} dN.(t) \quad (18)$$

based on the sample from stratum  $l$ . Thus

$$\tilde{\Delta}_l = \frac{1}{m_l} \sum_{k \in \tilde{\mathcal{C}}_l} (\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_{(l)})^{\otimes 2} \quad (19)$$

where  $\tilde{\mathbf{X}}_{(l)} = m_l^{-1} \sum_{k \in \tilde{\mathcal{C}}_l} \tilde{\mathbf{X}}_k$ , and  $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$  for any vector  $\mathbf{a}$ . By inserting (18) into (19) and simplifying, one may alternatively write (19) as a double-integral. For simple random sampling ( $L = 1$ ) we have  $\tilde{\mathbf{X}}_{(l)} = \mathbf{0}$ , and it can be shown that this way of reformulating the estimator will specialize to the expressions given by Self and Prentice (1988, p. 74) when they are simplified to the situation with exponential relative risk function. As (17)–(19) are the most convenient expressions for numerical purposes, we chose not to present this alternative formulation.

So far we have in this section considered Estimator I with time-fixed weights. The only difference between the time-fixed weights versions of Estimators I and III is that, when the case occurs outside the subcohort, Estimator III swaps it with an individual in the subcohort from the same stratum. This difference is asymptotically negligible as the cohort and subcohort sizes increase. Thus the two estimators have the same asymptotic distribution, and the variance estimator just mentioned applies for the time-fixed weights version of Estimator III. Estimator II is not asymptotically distributed as the other two. Under suitable regularity conditions, however, the above arguments go through with only small modifications for the time-fixed weights version of this estimator. In particular, for variance estimation, one should replace  $n_l$ ,  $m_l$ ,  $\tilde{\mathcal{C}}$  and  $\tilde{\mathcal{C}}_l$  by, respectively,  $n_l^0$ ,  $m_l^0$ ,  $\tilde{\mathcal{C}} \cup \mathcal{D}$  and  $\tilde{\mathcal{C}}_l \setminus \mathcal{D}$  in (17)–(19).

When the weights depend on time, the score for Estimator I can no longer be decomposed as in (13). We have not been able to derive the asymptotic distribution of the time-dependent weights version of any of our estimators. But by inspecting the simple case where a dichotomous stratification variable is the only covariate in the model, it can be shown that the difference between the two versions of Estimator I is not asymptotically negligible. As discussed in the last paragraph of Section 2, the time-dependent weights version of Estimator I yields the full cohort likelihood when the stratification variable is the only variable in the model and is thus fully efficient in this situation. However, with time-fixed weights (16) does not vanish so the time-fixed weighted Estimator I will be less than fully efficient. Thus one may expect to get an efficiency gain by using time-dependent weights. The study of the asymptotic distribution of the versions of the estimators with time-dependent weights is a topic for further research.

## 5. Asymptotic Relative Efficiency and Optimal Allocation

In this section we present some asymptotic efficiency results and discuss how one should allocate the sampled individuals to the strata. Our calculations are based on the asymptotic

results for Estimator I with time-fixed weights, but should be of relevance also for the other estimators.

We assume that the covariate-vectors  $\mathbf{Z}_i$  are time-fixed, that the population is partitioned into strata according to the value of the stratification variables  $\tilde{Z}_i \in \{1, \dots, L\}$ , and adopt an i.i.d. model where  $(Y_i(t), \mathbf{Z}_i, \tilde{Z}_i)$  are independent copies of  $(Y(t), \mathbf{Z}, \tilde{Z})$ . Then, by the law of large numbers,  $v_l = P(\tilde{Z} = l)$ . Further we write  $p(t) = P(Y(t) = 1)$  for the probability that an individual is at risk at time  $t$ -, and introduce  $p(s, t) = P(Y(s) = 1, Y(t) = 1)$ .

As an exact evaluation of the quantities from Section 4 is quite involved, we assume, as an approximation, that the distribution of the covariates among those at risk in a given sampling stratum is constant over time. This is a reasonable assumption when the disease is rare and there is ‘‘administrative censoring.’’ Then the usual asymptotic cohort information (e.g., Andersen *et al.* 1993, Condition VII.2.1.e) simplifies to

$$\Sigma = \left( E\{\mathbf{Z}^{\otimes 2} e^{\beta'_0 \mathbf{Z}}\} - \mathbf{e}(\beta_0)^{\otimes 2} E\{e^{\beta'_0 \mathbf{Z}}\} \right) \int_0^\tau p(t) \alpha_0(t) dt, \quad (20)$$

where  $\mathbf{e}(\beta_0) = E\{\mathbf{Z} \exp(\beta'_0 \mathbf{Z})\} / E\{\exp(\beta'_0 \mathbf{Z})\}$ . From (15) it further follows that  $\Delta_l = E_l\{\mathbf{X}_0^{\otimes 2}\} - (E_l\{\mathbf{X}_0\})^{\otimes 2}$ , where  $E_l(\cdot) = E(\cdot \mid \tilde{Z} = l)$  denotes the expectation within the  $l$ th stratum, and

$$\mathbf{X}_0 = \int_0^\tau \{\mathbf{Z} - \mathbf{e}(\beta_0)\} Y(t) \exp\{\beta'_0 \mathbf{Z}\} \alpha_0(t) dt.$$

Therefore, under our assumptions,

$$\begin{aligned} \Delta_l &= E_l \left\{ (\mathbf{Z} e^{\beta'_0 \mathbf{Z}} - \mathbf{e}(\beta_0) e^{\beta'_0 \mathbf{Z}})^{\otimes 2} \right\} \int_0^\tau \int_0^\tau p(s, t) \alpha_0(s) \alpha_0(t) ds dt \\ &\quad - \left( E_l \{\mathbf{Z} e^{\beta'_0 \mathbf{Z}}\} - \mathbf{e}(\beta_0) E_l \{e^{\beta'_0 \mathbf{Z}}\} \right)^{\otimes 2} \left( \int_0^\tau p(t) \alpha_0(t) dt \right)^2. \end{aligned} \quad (21)$$

For right-censored data  $p(s, t) = p(s \vee t)$ , which simplifies the evaluation of the double integral on the right-hand side of (21).

Before we illustrate how the approximations (20) and (21) may be used to compute asymptotic efficiencies and optimal allocation rules for specific models, we need to define a criterion to optimize. We will assume that one covariate,  $Z_1$ , is of prime importance in the study and that we wish to minimize the asymptotic variance of  $\tilde{\beta}_1$  over the proportions sampled from each stratum,  $\pi_l$ , given a fixed overall subcohort sampling fraction  $\pi$ ,

$$\pi = \sum_{l=1}^L v_l \pi_l, \quad (22)$$

from the cohort. (Remember from Section 4 that  $n_l/n \rightarrow v_l$  and  $m_l/n_l \rightarrow \pi_l$  as  $n \rightarrow \infty$ .) Thus we want to minimize  $(\Sigma^{-1} \Delta \Sigma^{-1})_{11}$  subject to the constraint (22). (We adopt the notation  $\mathbf{A}_{ij}$  for the  $(i, j)$ th element of a matrix  $\mathbf{A}$ .) Using (16) it is then fairly simple to

show that the optimal sampling fractions are

$$\pi_l = \pi \frac{\{(\Sigma^{-1} \Delta_l \Sigma^{-1})_{11}\}^{1/2}}{\sum_{j=1}^L v_j \{(\Sigma^{-1} \Delta_j \Sigma^{-1})_{11}\}^{1/2}}, \quad (23)$$

i.e. it is optimal to choose the  $m_l$  proportional to  $n_l \{(\Sigma^{-1} \Delta_l \Sigma^{-1})_{11}\}^{1/2}$ . This allocation rule is similar to the well known Neyman allocation (e.g. Cochran 1977, p. 99) in classical sampling theory.

In practice one may combine the optimal allocation rule (23) with the approximations (20) and (21) to derive an allocation rule which is approximately optimal. To this end one either needs a reasonable model for the covariates in order to evaluate the expected values in (20) and (21) (analytically or by stochastic simulation), or one needs a data set with a structure similar to the data set to be collected from which these expected values may be estimated. A simplification is possible, however, when  $(Z_1, \tilde{Z})$  is independent of  $\mathbf{Z}_2 = (Z_2, \dots, Z_p)'$ . For then the entries  $(1, j)$  and  $(i, 1)$  of (20) vanish for all  $i, j \geq 2$ , so that the contributions from  $\Sigma^{-1}$  cancel in (23). It is therefore approximately optimal to choose the  $m_l$  proportional to  $n_l \{\Delta_{l11}\}^{1/2}$ . Further if we write  $\beta_0 = (\beta_{01}, \beta'_{02})'$ , we have under our assumptions

$$\begin{aligned} \Delta_{l11} &= E_l \left\{ \left( Z_1 e^{\beta_{01} Z_1} - \frac{E(Z_1 e^{\beta_{01} Z_1})}{E(e^{\beta_{01} Z_1})} e^{\beta_{01} Z_1} \right)^2 \right\} \\ &\quad \times E \left( e^{2\beta'_{02} Z_2} \right) \int_0^\tau \int_0^\tau p(s, t) \alpha_0(s) \alpha_0(t) ds dt \\ &\quad - \left\{ E_l \left( Z_1 e^{\beta_{01} Z_1} \right) - \frac{E(Z_1 e^{\beta_{01} Z_1})}{E(e^{\beta_{01} Z_1})} E_l \left( e^{\beta_{01} Z_1} \right) \right\}^2 \left\{ E \left( e^{\beta'_{02} Z_2} \right) \right\}^2 \left( \int_0^\tau p(t) \alpha_0(t) dt \right)^2. \end{aligned}$$

This simplification may be useful, since it allows the expected values to be computed without knowledge of the joint distribution of all the covariates.

We now consider two example situations that apply these results.

*Example 1—One Binary Covariate.* Let  $Z$  be a binary covariate, e.g., measuring the absence ( $Z = 0$ ) or presence ( $Z = 1$ ) of exposure. We write  $r_l = P(Z = 1 \mid \tilde{Z} = l)$  for the fraction exposed in stratum  $l$ , and let  $r = \sum_l v_l r_l$  be the proportion exposed in the population. Furthermore, in order to get nice analytic results, we approximate  $p(t)$  and  $p(s, t)$  by 1 for all  $s, t$ , corresponding to a situation with a rare event and no censoring. Then  $E(Z^2 e^{\beta_0 Z}) = E(Z e^{\beta_0 Z}) = r e^{\beta_0}$  and  $E(e^{\beta_0 Z}) = 1 - r + r e^{\beta_0}$ , and the (scalar) cohort information (20) takes the form

$$\Sigma = \frac{r(1-r)e^{\beta_0}}{1-r+re^{\beta_0}} \int_0^\tau \alpha_0(t) dt. \quad (24)$$

When the subcohort is sampled by simple random sampling, we get in a similar manner from (16) and (21),

$$\Delta_{\text{srs}} = \frac{(1-\pi)r(1-r)e^{2\beta_0}}{\pi(1-r+re^{\beta_0})^2} \left( \int_0^\tau \alpha_0(t) dt \right)^2. \quad (25)$$

Further, from (21),

$$\Delta_l = \frac{r_l(1-r_l)e^{2\beta_0}}{(1-r+re^{\beta_0})^2} \left( \int_0^\tau \alpha_0(t) dt \right)^2, \quad (26)$$

so that  $\Sigma^{-1}\Delta_l\Sigma^{-1} = r_l(1-r_l)/[r(1-r)]^2$ . Thus the optimal sampling fractions (23) for stratified sampling of the subcohort becomes

$$\pi_{l,\text{opt}} = \pi \frac{[r_l(1-r_l)]^{1/2}}{\sum_j v_j [r_j(1-r_j)]^{1/2}},$$

i.e., they are proportional to  $\sqrt{r_l(1-r_l)}$ , independent of the value of the regression parameter  $\beta_0$ .

Now, assume that the overall sampling fraction is

$$\pi = M(1-r+re^{\beta_0}) \int_0^\tau \alpha_0(t) dt, \quad (27)$$

corresponding (approximately) to a subcohort size of  $M$  times the expected number of cases. We further adopt the approximation  $(1-\pi)/\pi \approx 1/\pi$ , valid when the subcohort is a small fraction of the cohort. Then, from (16) and (24)–(27), the relative efficiencies of all the case-cohort designs, relative to the full cohort take the form

$$\text{ARE} = \frac{\Sigma^{-1}}{\Sigma^{-1} + \Sigma^{-1}\Delta\Sigma^{-1}} \approx \frac{M}{M + Q\{e^{\beta_0}/(1-r+re^{\beta_0})^2\}}, \quad (28)$$

the differences between the designs being reflected in the value of  $Q$ . For the simple case-cohort study  $Q_{\text{srs}} = 1$ , while for the stratified designs with proportional and optimal allocation to the strata  $Q_{\text{prop}} = \sum_l v_l [r_l(1-r_l)/r(1-r)]$  and  $Q_{\text{opt}} = \{\sum_l v_l [r_l(1-r_l)/r(1-r)]^{1/2}\}^2$ , respectively. By Schwarz' inequality we have  $\text{ARE}_{\text{srs}} \leq \text{ARE}_{\text{prop}} \leq \text{ARE}_{\text{opt}}$  with equality only when the fractions exposed are the same in all strata.

In Table 1 we present the approximate relative efficiencies (28) of the stratified designs with two sampling strata (i.e.  $L = 2$ ) when  $e^{\beta_0} = 2$ . As in Langholz and Borgan (1995) these efficiencies are given as functions of the fraction exposed  $r$ , the sensitivity  $1 - \alpha = P(\tilde{Z} = 2 \mid Z = 1) = r_2 v_2 / r$ , and the specificity  $1 - \beta = P(\tilde{Z} = 1 \mid Z = 0) = (1 - r_1) v_1 / (1 - r)$ . When sensitivity and specificity are both 50%, we have  $r_1 = r_2$ , so the efficiencies for these values of  $1 - \alpha$  and  $1 - \beta$  correspond to the ones for the simple case-cohort design. The efficiencies relative to the full cohort are higher when 50% of the individuals are exposed than when this is only the case for 5% of the cohort. From the efficiencies for the optimal design we see, however, that the potential gain, by using a stratified case-cohort design rather than the simple one, is largest when few individuals are exposed. The efficiencies increase with increasing sensitivity and specificity, and this increase is symmetric in  $1 - \alpha$  and  $1 - \beta$  for the optimal design and for the proportional design when  $r = 0.50$ . As a final point, we note that when 50% of the cohort is exposed, the proportional and optimal design perform about equally well, while the latter is clearly superior when only 5% of the individuals are exposed.

Table 1. Approximate relative efficiencies in per cent for the optimal and proportional stratified case-cohort designs versus a full cohort study. The results are for one binary covariate  $Z$  with  $r = P(Z = 1)$  when stratification is based on a binary surrogate  $\tilde{Z}$  with specificity  $1 - \beta$  and sensitivity  $1 - \alpha$  and the subcohort size equals the expected number of cases<sup>a</sup>. (The relative efficiency of the case-cohort design with simple random sampling is equal to those of the stratified designs when sensitivity and specificity both equal 50%.)

(a) $r = 0.05$						
$1 - \alpha$	Proportional			Optimal		
	$1 - \beta$	$1 - \beta$	$1 - \beta$	$1 - \beta$	$1 - \beta$	$1 - \beta$
0.50	0.50	0.70	0.90	0.50	0.70	0.90
0.50	35.5	35.7	37.3	35.5	36.5	40.8
0.70	35.7	36.4	39.4	36.5	39.6	47.3
0.90	36.2	37.4	42.4	40.8	47.3	60.5

  

(b) $r = 0.50$						
$1 - \alpha$	Proportional			Optimal		
	$1 - \beta$	$1 - \beta$	$1 - \beta$	$1 - \beta$	$1 - \beta$	$1 - \beta$
0.50	0.50	0.70	0.90	0.50	0.70	0.90
0.50	52.9	54.0	58.2	52.9	54.0	58.4
0.70	54.0	57.3	64.3	54.0	57.3	64.7
0.90	58.2	64.3	75.8	58.4	64.7	75.8

<sup>a</sup>) Computed according to (28) with  $M = 1$  assuming  $e^{\beta_0} = 2$ , and with  $r_1$  and  $r_2$  adjusted to give the stated sensitivities and specificities.

*Example 2—One Normal Covariate.* In this example we assume that the covariate  $Z$ , e.g., measuring some exposure, is normally distributed within each of the two strata defined by the surrogate  $\tilde{Z} \in \{1, 2\}$ . More precisely  $Z$  is standard normally distributed for surrogate negative individuals ( $\tilde{Z} = 1$ ), while it is normally distributed with mean  $\mu$  and standard deviation  $\sigma$  for surrogate positive individuals ( $\tilde{Z} = 2$ ). Conditional on  $Z = z$ , the failure time  $T$  is assumed to have an exponential distribution with parameter  $\alpha_0 e^{\beta_0 z}$ , while the censoring variable is  $U = \min(1, V)$  with  $V$  independent of  $T$  and uniformly distributed over  $[0, c]$  with  $c \geq 1$ . In Table 2 we give approximate efficiencies for this situation for the simple case-cohort design as well as for the stratified designs with proportional and optimal allocation to the strata. The optimal sampling fractions are also given. The results are based on (20) and (21) substituting  $P(U \geq t)$  for  $p(t)$  and using the moment generating function of the normal distribution (and its derivatives) to evaluate the relevant expected values. For all situations considered, the overall sampling fraction is  $\pi = 0.10$ , the true value of the regression parameter is  $\beta_0 = 0.20$ , and the fraction surrogate positive is  $\nu_2 = 0.10$ . Different values of  $\mu$  and  $\sigma$  are presented to illustrate the importance of the distribution of the covariate among the surrogate negative and surrogate positive individuals. The parameters  $\alpha_0$  and  $c$  are adjusted to obtain a failure probability of 10%

Table 2. Approximate relative efficiencies in per cent for the simple case-cohort design and for the stratified case-cohort designs with proportional and optimal allocation versus the full cohort. The results are for a single covariate normally distributed within each stratum and for an overall sampling fraction and failure rate of 10% (see text for details).

$\mu$	$\sigma$	Approximate efficiencies			Optimal sampling fractions	
		Simple	Proportional	Optimal	$\pi_1$	$\pi_2$
(a) 20% censored						
2.0	1.0	44.1	54.2	56.9	0.089	0.197
4.0	1.0	35.8	63.4	75.3	0.073	0.340
4.0	2.0	27.2	40.5	73.7	0.051	0.539
(b) 60% censored						
2.0	1.0	40.3	48.7	52.1	0.088	0.209
4.0	1.0	32.3	53.5	69.2	0.071	0.365
4.0	2.0	24.2	34.3	67.8	0.052	0.534

and a probability of censoring of 20% or 60% before “the closure of the study” at time  $t = 1$ .

The main conclusion from the table, is that a substantial improvement of a case-cohort study may be achieved using a stratified design. Not surprisingly, the gain increases with the difference between the distribution of the covariate among surrogate negative and surrogate positive individuals. When there is little variation in the covariate values within each stratum ( $\sigma = 1$ ), the stratified case-cohort design with proportional allocation to the strata is quite efficient. However, when the covariate values among the surrogate positive individuals are highly variable ( $\sigma = 2$ ), the design with optimal allocation is clearly superior.

## 6. A Simulation Study

The results of the previous section show that much can be gained by using a stratified case-cohort design compared to the standard one where the subcohort is selected by simple random sampling. However, these results are derived for Estimator I with time-fixed weights and are based on an approximation to the asymptotic distributions. Therefore they are of no use in comparing the asymptotic equivalent Estimators I and III, to compare these to Estimator II, or to investigate whether the versions of the estimators with time-varying weights may be better than their fixed-weights counterparts. In order to get some understanding for such problems, and to see whether the asymptotic results are representative for what one gets in finite samples, we performed a small simulation study.

To this end we generated censored survival data from the model described in Example 2 in the previous section with 20% censoring. For all simulated data sets we computed the full cohort estimator as well as estimators based on simple and stratified case-cohort

*Table 3.* Average estimates of  $\beta$  (“ave”), empirical standard deviations (“sd”) and empirical efficiencies relative to the full cohort (“ere”) based on repeated sampling of 1000 cohorts each with 1000 individuals. The true value of  $\beta$  is 0.20 for all situations, while the baseline hazard and the censoring distribution are adjusted to get a failure probability of 10% and a probability of censoring of 20% before  $t = 1$ . The subcohort size is  $m = 100$ , and only optimal sampling fractions are considered for the stratified case-cohort designs. For the case-control designs, one control is selected per case.

Method	$\mu = 2, \sigma = 1$			$\mu = 4, \sigma = 1$			$\mu = 4, \sigma = 2$		
	ave	sd	ere	ave	sd	ere	ave	sd	ere
Full cohort	0.198	0.082	100	0.200	0.053	100	0.200	0.045	100
Self & Prentice	0.213	0.132	37.7	0.217	0.100	27.3	0.229	0.112	15.0
Est II unstratified	0.213	0.129	39.3	0.215	0.093	32.0	0.219	0.087	24.9
Prentice	0.210	0.129	39.4	0.213	0.097	29.8	0.222	0.100	19.0
Est I time-fixed	0.202	0.115	50.1	0.200	0.064	70.3	0.200	0.053	72.0
Est I time varying	0.202	0.114	50.7	0.200	0.063	71.5	0.201	0.053	72.0
Est II time-fixed	0.202	0.115	50.5	0.200	0.063	71.3	0.200	0.053	72.3
Est II time varying	0.202	0.114	50.8	0.200	0.062	72.6	0.201	0.053	72.3
Est III time-fixed	0.201	0.114	51.1	0.200	0.063	70.9	0.200	0.053	72.1
Est III time varying	0.201	0.113	51.8	0.200	0.063	72.2	0.200	0.052	72.8
Nested case-control	0.205	0.129	40.1	0.208	0.093	32.2	0.209	0.085	27.2
Counter-matched	0.189	0.120	45.6	0.199	0.061	76.4	0.201	0.052	75.1

sampling. The simple and counter-matched nested case-control designs were also included for comparison (Langholz and Borgan, 1995). For the unstratified case-cohort study, we considered Self and Prentice’s estimator, Estimator II specialized to the situation with only one stratum, and Prentice’s original estimator. For the stratified design, Estimators I–III were considered both with time-fixed and time-varying weights. Table 3 gives the average estimated regression parameter, the empirical standard deviation of these estimates, and the empirical relative efficiencies for the three combinations of the parameters considered in panel (a) of Table 2. The latter are computed as ratios between the empirical variance of the cohort estimator and the sum of the empirical variances and the squares of the average difference between the cohort estimator and the sampling based estimators. All results are based on repeated sampling of 1000 cohorts, each with censored survival times for 1000 individuals (the same for all estimators). For the case-cohort designs the subcohort size is 100, equal to the expected number of failures. Only results for the stratified case-cohort designs with optimal allocation to the strata (as given in the two rightmost columns of Table 2) are presented.

All the stratified case-cohort estimators give almost identical results, and the small differences observed are of no practical importance. Nevertheless, it is worth noting that, for all three methods, the version with time-dependent weights performs slightly better than the one where the weights are time-fixed. Further Estimator III performs consistently better than Estimator I, while there is no clear ordering between Estimators II and III. For the three unstratified methods, the Self-Prentice estimator has the poorest performance. It is biased

upwards and has consistently the largest standard deviation. Note also that the empirical relative efficiencies are similar to, although somewhat lower than, the approximate ones reported in Table 2.

In our simulations we also computed the variance estimator given in connection with (17)–(19). The average estimated standard errors obtained by this procedure were, respectively, 0.111, 0.064, and 0.055 for the three situations considered in Table 3. This is in good agreement with the empirical standard deviations reported in the table.

The two last lines of Table 3 give the results for the simple and counter-matched case-control designs with one control selected per case. As these estimators automatically adjust the number of controls according to the number of failures in the cohort, while the subcohort size only on the average equals the number of failures, the results are not completely comparable. Nevertheless, the results indicate that the case-control designs have a performance which is approximately equal to that of the corresponding case-cohort studies.

## 7. Discussion

Our results show that if a correlate of exposure is available for all cohort members, it can be advantageous to stratify the sampling of the subcohort to over-represent more highly exposed subjects. We indicated why the natural generalization of Prentice's (1986) pseudo-likelihood for simple random sampling is clearly inefficient for estimation of rate ratio parameters. Estimator III solves this problem while retaining score-unbiasedness. We found little bias, however, in the other stratified estimators in our simulations. Based on the observation that the time-dependent weight variants of the estimators bring the full cohort marginal information about the stratification variable into the sample, we conjectured that these estimators would be superior to the corresponding time-fixed weight versions. In fact, our simulations showed a slight improvement in efficiency using time-dependent weights. Further simulation studies of more complex situations and analyses of real data sets are needed before definitive conclusions can be drawn on the importance of score-unbiasedness and whether the use of time-dependent weights warrants the additional complexity in the analysis. When comparing the stratified estimators, it is important to note that the data requirements are not the same for all of these. Estimator II requires full covariate histories for the cases, while Estimators I and III only need the cases' covariate values at their failure times. Furthermore, the time-dependent weights variants of the estimators require knowledge of the at-risk-status of the full cohort, while this knowledge is only needed for the subcohort for the estimators with time-fixed weights.

A few comments on our assumptions in Sections 3 and 4 are in order. In our study of score-unbiasedness in Section 3 we assume that the stratified sampling of the subcohort depends only on exposure information known at time zero and that the weights are predictable. This is crucial in order to be able to write the pseudoscore (9) for Estimator III as a sum of a martingale and an additional term with mean zero. These assumptions are not needed in Section 4, however, where the pseudoscore (13) is decomposed as a sum of the score for the full cohort and an additional term depending on the sampling. Therefore it causes no difficulties, in terms of large sample properties, if the stratification and

weights depend on the full cohort history as is the case for Estimator II. However, a more involved use of the full cohort history than for Estimator II can only be achieved for retrospective case-cohort studies where all events have already occurred when the sampling is performed.

We have discussed variance estimation for the stratified case-cohort estimators with time-fixed weights in Section 4. For simple random sampling the variance estimator given in connection with (17) is a reformulation of the one given by Self and Prentice (1988). A similar reformulation of the Self-Prentice covariance estimator is given by Therneau and Li (1999), who show how their version of the estimator can be computed using standard computer packages like Splus and SAS. Their approach may easily be modified to allow computation of our variance estimators for the stratified estimators with time-fixed weights. For the unstratified case, Lin and Ying (1993) and Barlow (1994) suggested an alternative to the Self-Prentice covariance estimator. We have not investigated how their approach may be adopted to stratified sampling, and whether it may be modified to handle the situation with time-varying weights.

Zhou and Pepe (1995) present a procedure for missing covariate data in Cox regression with auxiliary information which is similar to the case-cohort methods described here. Their validation set plays the role of the subcohort in our setting, while their auxiliary covariate corresponds to our surrogate. Some algebra shows that the denominators in their “estimated partial likelihood” [(6) in their paper] equal the ones we use in our pseudo-likelihood for the time-dependent weights version of Estimator I. The numerators differ, however, since the covariates are known for cases occurring outside the subcohort in Estimator I, while covariates for these cases are estimated from the validation set in the Zhou and Pepe estimator. Another important difference is that Zhou and Pepe (1995) assume the validation set to be a simple random sample from the full cohort, while our subcohort is selected by stratified random sampling.

Important design questions for stratified case-cohort studies are (i) how one should divide the cohort into strata, and (ii) how many subjects one should choose from each stratum. We discussed (ii) at length in Section 5, and showed how the rule (23) gives an optimal allocation of the individuals to the strata. A strict implementation of this allocation rule is problematic, however, since it depends on the covariate and censoring distributions, none of which are known at the design stage. But, as intuition would predict, the rule indicates that one should over-sample high risk strata where the variation of the exposure tends to be large. Design question (i) is a difficult one. To answer this, one may, for example, have to decide on the number of strata and how to categorize a continuous variable used for creating strata. Further research is needed to get a better understanding of such problems.

### Acknowledgments

This work was primarily supported by grant CA42949 from the United States National Cancer Institute. Ørnulf Borgan and Sven Ove Samuelsen acknowledge support from The Norwegian Research Council and Johan and Mimi Wesmann’s foundation. The authors are grateful to Ola Hestnes and Svein Børre Solvang for programming assistance.

## References

- P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, *Statistical Models based on Counting Processes*, Springer Verlag: New York, 1993.
- P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: A large sample study," *Ann. Statist.* vol. 10 pp. 1100–20, 1982.
- W. Barlow, "Robust variance estimation for the case-cohort design," *Biometrics* vol. 50 pp. 1064–72, 1994.
- N. E. Breslow and K. Cain, "Logistic regression for two stage case-control data," *Biometrika* vol. 75 pp. 11–20, 1988.
- W. G. Cochran, *Sampling Techniques. 3rd edition*. Wiley: New York, 1977.
- D. R. Cox, "Regression models and life-tables (with discussion)," *J. Roy. Statist. Soc. B* vol. 34 pp. 187–220, 1972.
- J. Kalbfleisch and J. Lawless, "Likelihood analysis of multi-state models for disease incidence and mortality," *Statist. in Med.* vol. 7 pp. 149–60, 1988.
- M. Kulic and D. Y. Lin, "Additive hazards regression for case-cohort studies," Manuscript, Charles University, Prague, 1998.
- B. Langholz and Ø. Borgan, "Counter-matching: A stratified nested case-control sampling method," *Biometrika* vol. 82 pp. 69–79, 1995.
- B. Langholz and D. C. Thomas, "Efficiency of cohort sampling designs: Some surprising results," *Biometrics* vol. 47 pp. 1563–71, 1991.
- E. Lehmann, *Nonparametrics*, Holden-Day: San Francisco, 1975.
- D. Y. Lin and Z. Ying, "Cox regression with incomplete covariate measurements," *J. Amer. Statist. Assoc.* vol. 88 pp. 1341–1349, 1993.
- J. Pogoda, *Variance Estimation in Complex Cohort Problems*, PhD thesis, University of Southern California, 1993.
- R. L. Prentice, "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika* vol. 73 pp. 1–11, 1986.
- S. O. Samuelsen, *Two Incomplete Data Problems in Life-History Analysis: Double Censoring and the Case-Cohort Design*, PhD thesis, University of Oslo, 1989.
- S. O. Samuelsen, "A pseudolikelihood approach to analysis of nested case-control studies," *Biometrika* vol. 84 pp. 379–394, 1997.
- S. G. Self and R. L. Prentice, "Asymptotic distribution theory and efficiency results for case-cohort studies," *Ann. Statist.* vol. 16 pp. 64–81, 1988.
- T. M. Therneau and H. Li, "Computing the Cox model for case cohort designs," *Lifetime Data Analysis* vol. 5 pp. 99–112, 1999.
- H. Zhou and M. S. Pepe, "Auxillary covariate data in failure time regression," *Biometrika* vol. 82 pp. 139–149, 1995.