Exposure stratified case-cohort designs

Ørnulf Borgan * Larry Goldstein † Bryan Langholz ‡ Janice Pogoda§
Sven Ove Samuelsen*

June 1998

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 and some of their statistical properties developed. The bias and efficiency of 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 reasonably well and were more efficient than a randomly sampled case-cohort sample. We conclude that these methods are well suited for the "clinical trials setting" in which subjects enter the study at time zero (at diagnosis or treatment) and a correlate of an expensive prognostic factor is collected for all study subjects at the time of entry to the study. In such studies, a correlate stratified subcohort can be much more cost-efficient for investigation of the expensive prognostic factor than a randomly sampled subcohort.

Key words: Case-cohort studies; Cox regression; Pseudo-likelihood; 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

*Institute of Mathematics, University of Oslo, P.O.Box 1053 Blindern, N-0316 Oslo 3, Norway

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

‡Department of Preventive Medicine, University of Southern California, School of Medicine, 1540 Alcazar Street CHP-220, Los Angeles, California 90033, U.S.A.

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

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 "exposure" 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 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 an exposure 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 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 discussed in Section 4, while the performance of the estimators is compared in a small simulation study described in Section 5. Some concluding remarks are given in Section 6.

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

$$\alpha(t;\mathbf{Z}) = \alpha_0(t)\exp\{\boldsymbol{\beta}_0'\mathbf{Z}(t)\}. \tag{1}$$

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 $\boldsymbol{\beta}_0$ measure the effect of the covariates. We denote by $t_1 < t_2 < \cdots$ 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 $\widetilde{\mathcal{C}}$ 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. In the main body of the paper, we assume that the there are no delayed entries, and that the stratified sampling is based on information available at time zero. In Section 6 we briefly discuss the special problems with case-cohort sampling when this is not the case.

The estimators for $\boldsymbol{\beta}_0$ considered in this paper, are all based on maximizing a pseudo-likelihood function of the form

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

Here $\widetilde{\mathcal{R}}(t_j)$ is a "sampled risk 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 $\boldsymbol{\beta}$ but may depend on t_j and $\mathcal{R}(t_j)$. The various estimators differ in the definitions of $w_k(t_j)$ and $\widetilde{\mathcal{R}}(t_j)$, 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 $\widetilde{\mathcal{R}}(t_j) = \widetilde{\mathcal{C}} \cup \{i_j\}$, the subcohort augmented with the case when it occurs outside the subcohort. Self and Prentice (1988), for the purpose of studying large sample properties, considered the (asymptotically equivalent) modification where $\widetilde{\mathcal{R}}(t_j) = \widetilde{\mathcal{C}}$ 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: $\widetilde{\mathcal{R}}(t_j) = \widetilde{\mathcal{C}}$ and $w_k(t_j) = n_{s(k)}/m_{s(k)}$

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

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 $\widetilde{\mathcal{R}}(t_j)$ with probability one. 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 $l = 1, \ldots, L$ 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 $\widetilde{\mathcal{R}}(t_j) = \widetilde{\mathcal{C}} \cup \{i_j\}$ 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

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

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 "sampled risk set" $\widetilde{\mathcal{R}}(t_j)$ 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 $n^0_{s(k)}/m^0_{s(k)}$ for the non-failing individuals by $w_k(t_j) = n^0_{s(k)}(t_j)/m^0_{s(k)}(t_j)$, where $n^0_l(t)$ and $m^0_l(t)$ 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 with 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-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, censoring, and covariate processes for the ith subject in the cohort. As is usual, we assume that there is a non-decreasing family of σ-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 λ_i of N_i is given heuristically by $\lambda_i(t)\mathrm{d}t = \mathrm{pr}\{\mathrm{d}N_i(t) = 1 \,|\, \mathcal{H}_{t-}\}$, where $\mathrm{d}N_i(t)$ is the increment of

N_i over the small time interval $[t, t + \mathrm{d}t)$. Assuming censoring to be independent (Andersen *et al.*, 1993, Section III.2.2), (1) yields the intensity process

$$\lambda_i(t) = Y_i(t)\alpha_0(t)\exp\{\boldsymbol{\beta}_0'\mathbf{Z}_i(t)\} \tag{3}$$

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 $\{1, 2, \ldots, n\}$, i.e. the set of all subsets of $\{1, 2, \ldots, n\}$, and introduce

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

for the set of possible sampled subcohorts. Finally we let

$$\pi(\mathbf{c}) = \mathrm{pr}(\widetilde{\mathcal{C}} = \mathbf{c}) = 1 \Big/ \prod_{l=1}^{L} \binom{n_l}{m_l}, \tag{4}$$

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

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 σ-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 $\mathrm{pr}\{\mathrm{d}N_i(t) = 1 \,|\, \mathcal{F}_{t-}\} = \mathrm{pr}\{\mathrm{d}N_i(t) = 1 \,|\, \mathcal{H}_{t-}\}$ 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 $\widetilde{\mathcal{R}}_i$ be the "sampled risk 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 $\widetilde{\mathcal{R}}_i = \widetilde{\mathcal{C}} \cup \{i\}$ and $\widetilde{\mathcal{R}}_i = \widetilde{\mathcal{C}}$, respectively. The latter is also the one used for Estimator I, while $\widetilde{\mathcal{R}}_i = \widetilde{\mathcal{C}} \cup \{i\} \setminus J_{s(i)}$ for Estimator III. Note that with this notation $\widetilde{\mathcal{R}}(t_j) = \widetilde{\mathcal{R}}_{i_j}$, so that (2) may be reformulated as

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

where we have written $w_k = w_k(\widetilde{\mathcal{R}}_i)$ for the weights in order to emphasize that these may depend on the sets $\widetilde{\mathcal{R}}_i$. 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 $\widetilde{\mathcal{R}}_i$ and the $w_k = w_k(\widetilde{\mathcal{R}}_i)$ are known at time zero, i.e. they are $\mathcal{F}_0$-measurable.

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

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

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

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

Then the pseudo-score becomes

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

and the maximum pseudo-likelihood estimator $\widetilde{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ which maximizes (2) or solves $\widetilde{\mathbf{U}}(\boldsymbol{\beta}) = 0$. 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) and (5) – (8), it is seen that the pseudo-score at $\boldsymbol{\beta}_0$ may be written as

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

where by standard counting process theory (e.g. Andersen *et al.*, 1993, Section II.4.1) the $M_i(t) = N_i(t) - \int_0^t \lambda_i(u) \mathrm{d}u$ are orthogonal local square integrable $(\mathcal{F}_t)$-martingales. Since $\mathbf{Z}_i(\cdot)$ and $\mathbf{E}_{\widetilde{\mathcal{R}}_i}(\boldsymbol{\beta}_0, \cdot)$ 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 $t \in [0, \tau]$

$$\begin{aligned} & \sum_{i=1}^n \mathrm{E}\{\mathbf{E}_{\widetilde{\mathcal{R}}_i}(\boldsymbol{\beta}_0, t) | \mathcal{H}_\tau\} Y_i(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_i(t)\} \\ & = \sum_{i=1}^n \left\{ \sum_{\mathbf{r} \in \mathcal{P}} \mathbf{E}_{\mathbf{r}}(\boldsymbol{\beta}_0, t) \mathrm{pr}(\widetilde{\mathcal{R}}_i = \mathbf{r}) \right\} Y_i(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_i(t)\} \end{aligned} \tag{10}$$

$$
\begin{aligned}
= & \sum_{\mathbf{r}\in\mathcal{P}} \frac{\sum_{k\in\mathbf{r}} Y_k(t)\mathbf{Z}_k(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_k(t)\}w_k(\mathbf{r})}{\sum_{k\in\mathbf{r}} Y_k(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_k(t)\}w_k(\mathbf{r})} \sum_{i\in\mathbf{r}} Y_i(t)\exp\{\boldsymbol{\beta}_0'\mathbf{Z}_i(t)\}\mathrm{pr}(\widetilde{\mathcal{R}}_i=\mathbf{r}) \\
& + \sum_{\mathbf{r}\in\mathcal{P}} \frac{\sum_{k\in\mathbf{r}} Y_k(t)\mathbf{Z}_k(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_k(t)\}w_k(\mathbf{r})}{\sum_{k\in\mathbf{r}} Y_k(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_k(t)\}w_k(\mathbf{r})} \sum_{i\notin\mathbf{r}} Y_i(t)\exp\{\boldsymbol{\beta}_0'\mathbf{Z}_i(t)\}\mathrm{pr}(\widetilde{\mathcal{R}}_i=\mathbf{r}).
\end{aligned}
$$

Here $\mathrm{pr}(\widetilde{\mathcal{R}}_i=\mathbf{r})$ may be derived from the subcohort distribution (4) and the relation between the subcohort $\widetilde{\mathcal{C}}$ and the sets $\widetilde{\mathcal{R}}_i$, 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 $\mathbf{r}\subset\mathcal{P}$ we have:
A) $\mathrm{pr}(\widetilde{\mathcal{R}}_k=\mathbf{r})=0$ *for* $k\notin\mathbf{r}$.
B) $\mathrm{pr}(\widetilde{\mathcal{R}}_k=\mathbf{r})=\mathrm{const}(\mathbf{r})w_k(\mathbf{r})$ *for* $k\in\mathbf{r}$.

Note that Condition A requires the cases to be included in the "sampled risk sets," while Condition B assumes the weights $w_k = w_k(\widetilde{\mathcal{R}}_i)$ to be proportional to the probability of selecting $\widetilde{\mathcal{R}}_i$ as the "sampled risk 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

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

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

$$
\sum_{i=1}^{n}\mathrm{E}\{\mathbf{E}_{\widetilde{R}_i}(\boldsymbol{\beta}_0,t)|\mathcal{H}_\tau\}Y_i(t)\exp\{\boldsymbol{\beta}_0'\mathbf{Z}_i(t)\} = \sum_{k=1}^{n} Y_k(t)\mathbf{Z}_k(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_k(t)\} \tag{11}
$$

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 investigated 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 "sampled risk 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

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

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 $\widetilde{\mathcal{C}}$ is selected according to the sampling probability (4). First note that Estimator I does not include the case in the "sampled risk sets", so this estimator is not score-unbiased. Next, consider the situation where $\widetilde{\mathcal{R}}_i = \widetilde{\mathcal{C}} \cup \{i\}$. By a similar reasoning as just given for Prentice's estimator, we get for $\mathbf{r} \in \mathcal{P}_k$,

$$\mathrm{pr}(\widetilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r}) + \pi(\mathbf{r} \setminus \{k\}).$$

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, $\mathrm{pr}(\widetilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r})$ 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, $\mathrm{pr}(\widetilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r} \setminus \{k\})$ 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 $\widetilde{\mathcal{R}}_i = \widetilde{\mathcal{C}} \cup \{i\}$, is that the structure of the sampled risk 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 $\widetilde{\mathcal{C}}$, we select at random (in principle, at time zero), for each $i \notin \widetilde{\mathcal{C}}$, a "swapper" $J_i \in \widetilde{\mathcal{C}}$ from individual i's stratum $s(i)$. Thus $\mathrm{pr}(J_i = j \,|\, \widetilde{\mathcal{C}} = \mathbf{c}) = 1/m_{s(i)}$ for each $i \notin \mathbf{c}$ and $j \in \mathbf{c} \cap \mathcal{S}_{s(i)}$. Then, for $\mathbf{r} \in \mathbf{C} \cap \mathcal{P}_k$,

$$\mathrm{pr}(\widetilde{\mathcal{R}}_k = \mathbf{r}) = \mathrm{pr}(\widetilde{\mathcal{C}} = \mathbf{r}) + \sum_j \mathrm{pr}(\widetilde{\mathcal{C}} = \mathbf{r} \cup \{j\} \setminus \{k\}, J_k = j),$$

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

$$\mathrm{pr}(\widetilde{\mathcal{C}} = \mathbf{r} \cup \{j\} \setminus \{k\}, J_k = j) = \pi(\mathbf{r} \cup \{j\} \setminus \{k\})/m_{s(k)} = \pi(\mathbf{r})/m_{s(k)},$$

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

$$\mathrm{pr}(\widetilde{\mathcal{R}}_k = \mathbf{r}) = \pi(\mathbf{r}) \frac{n_{s(k)}}{m_{s(k)}}.$$

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 $\widetilde{\mathcal{C}}$ is used as the "sampled risk 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 $n_l/n \to \nu_l > 0$ and $m_l/n_l \to p_l > 0$ as $n \to \infty$, 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 $\widetilde{\mathcal{C}}_l = \widetilde{\mathcal{C}} \cap \mathcal{S}_l$ for the subset of the subcohort that belongs to stratum l, and introduce, for $\gamma = 0, 1$ (later we avoid boldfacing for $\gamma = 0$),

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

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

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

and

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

and from these define $\mathbf{E}(\boldsymbol{\beta}, t)$ and $\mathbf{E}_{\widetilde{\mathcal{C}}}(\boldsymbol{\beta}, t)$ 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 $\mathbf{E}_{\widetilde{\mathcal{R}}_i}(\boldsymbol{\beta}, t)$ by $\mathbf{E}(\boldsymbol{\beta}, t)$ in (8). The pseudo-score for Estimator I, evaluated at $\boldsymbol{\beta}_0$, may be decomposed as

$$\widetilde{\mathbf{U}}(\boldsymbol{\beta}_0) = \mathbf{U}(\boldsymbol{\beta}_0) + \sum_{k=1}^{n} \{1 - (n_{s(k)}/m_{s(k)})I_k\}\mathbf{X}_k \tag{13}$$

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

$$\mathbf{X}_k = \int_0^\tau \{\mathbf{Z}_k(t) - \mathbf{E}(\boldsymbol{\beta}_0, t)\}\, Y_k(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_k(t)\} S^{(0)}_{\widetilde{\mathcal{C}}}(\boldsymbol{\beta}_0, t)^{-1} \mathrm{d}N_\cdot(t)$$

with $N_\cdot = \sum_{i=1}^n N_i$. By approximating $S^{(0)}_{\widetilde{\mathcal{C}}}(\boldsymbol{\beta}_0, t)$ with $S^{(0)}(\boldsymbol{\beta}_0, t)$, it then follows that the normalized pseudo-score $n^{-1/2}\widetilde{\mathbf{U}}(\boldsymbol{\beta}_0)$ is asymptotically equivalent to

$$n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0) + n^{-1/2} \sum_{k=1}^{n} \{1 - (n_{s(k)}/m_{s(k)})I_k\}\mathbf{X}_k^0 \tag{14}$$

with

$$\mathbf{X}_k^0 = \int_0^\tau \{\mathbf{Z}_k(t) - \mathbf{E}(\boldsymbol{\beta}_0, t)\}\, Y_k(t) \exp\{\boldsymbol{\beta}_0' \mathbf{Z}_k(t)\} S^{(0)}(\boldsymbol{\beta}_0, t)^{-1} \mathrm{d}N_\cdot(t). \tag{15}$$

The leading term of (14) is the normalized score for the full cohort, and it is known to converge weakly to a mean zero multivariate normal variate with covariance matrix $\boldsymbol{\Sigma}$, say (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

$$\boldsymbol{\Delta} = \sum_{l=1}^{L} \nu_l \frac{1 - p_l}{p_l} \boldsymbol{\Delta}_l. \tag{16}$$

Here $\boldsymbol{\Delta}_l$ is the limit in probability of the finite-population covariance matrix of the X_k^0 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 $\widetilde{\mathcal{I}}(\boldsymbol{\beta})$ be the observed pseudo-information for Estimator I. Then $n^{-1}\widetilde{\mathcal{I}}(\boldsymbol{\beta}^*)$ converges in probability to the asymptotic cohort information matrix $\boldsymbol{\Sigma}$ for any $\boldsymbol{\beta}^*$ which is consistent for $\boldsymbol{\beta}_0$. The usual Taylor series expansions argument gives that $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges weakly to a mean zero multivariate normal variate with covariance matrix $\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}$ as $n \to \infty$.

The asymptotic covariance matrix of $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ may be estimated consistently by n times $\widetilde{\boldsymbol{\Sigma}}^{-1} + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\Delta}}\widetilde{\boldsymbol{\Sigma}}^{-1}$. Here $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\mathcal{I}}(\widetilde{\boldsymbol{\beta}})$ is the observed pseudo-information, while

$$\widetilde{\boldsymbol{\Delta}} = \sum_{l=1}^{L} \frac{n_l(n_l - m_l)}{m_l} \widetilde{\boldsymbol{\Delta}}_l \tag{17}$$

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

$$\widetilde{\mathbf{X}}_k = \int_0^\tau \left\{ \mathbf{Z}_k(t) - \mathbf{E}_{\widetilde{\mathcal{C}}}(\widetilde{\boldsymbol{\beta}}, t) \right\} Y_k(t) \exp\{\widetilde{\boldsymbol{\beta}}' \mathbf{Z}_k(t)\} S^{(0)}_{\widetilde{\mathcal{C}}}(\widetilde{\boldsymbol{\beta}}, t)^{-1} \mathrm{d}N_\cdot(t) \tag{18}$$

based on the sample from stratum l. Thus

$$\widetilde{\boldsymbol{\Delta}}_l = \frac{1}{m_l} \sum_{k \in \widetilde{\mathcal{C}}_l} \left(\widetilde{\mathbf{X}}_k - \widetilde{\mathbf{X}}_{(l)} \right) \left(\widetilde{\mathbf{X}}_k - \widetilde{\mathbf{X}}_{(l)} \right)' \tag{19}$$

with $\widetilde{\mathbf{X}}_{(l)} = m_l^{-1} \sum_{k \in \widetilde{\mathcal{C}}_l} \widetilde{\mathbf{X}}_k$.

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, so the two estimators have the same asymptotic distribution. Estimator II is not asymptotically distributed as the other two. Under suitable regularity conditions, however, the above arguments go through with only small modifications. In particular, for variance estimation, one should replace n_l, m_l, $\widetilde{\mathcal{C}}$ and $\widetilde{\mathcal{C}}_l$ by, respectively, n_l^0, m_l^0, $\widetilde{\mathcal{C}} \cup \mathcal{D}$ and $\widetilde{\mathcal{C}}_l \setminus \mathcal{D}$ in (17) - (19).

When the weights depend on time, as described in the last paragraph of Section 2, 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. (The time-dependent version is fully efficient in this situation, but (16) does not vanish for the time-fixed version.) Thus one may expect to get an efficiency gain by using time-dependent weights. However, the study of the asymptotic distribution of the versions of the estimators with time-dependent weights is a topic for further research.

5 A simulation study

To get some understanding for the potentials of stratification in case-cohort studies and the behavior of our estimators, we performed a small simulation study for model (1). To mimic an instance where stratification may be useful, we considered a situation where a dichotomous surrogate measure $\widetilde{Z}$ is available for everyone, and is used to partition the cohort into two strata. The surrogate measure is related to the covariate of interest Z in such a way that surrogate positive individuals ($\widetilde{Z} = 1$) are more likely to have a high value of Z than surrogate negative individuals ($\widetilde{Z} = 0$). Note that in an actual case-cohort study, the value of Z will only be obtained for the failures and the subjects in the subcohort.

More precisely, we simulate the censored survival times for the individuals in the cohort as follows. First we generate a Bernoulli variable $\widetilde{Z}$ with success probability $p = 0.10$, corresponding to a situation with 10% surrogate positive individuals. Then the covariate Z is selected from the standard normal distribution when $\widetilde{Z} = 0$, and from a normal distribution with mean μ and standard deviation σ when $\widetilde{Z} = 1$. Conditional on $Z = z$, the failure time

Table 1: Average estimate of β ("aver") with empirical standard deviation ("sd") based on repeated sampling of 1000 cohorts each with 1000 individuals [a]. For the case-cohort studies, the subcohort size is $m = 100$ subjects in the first four panels and $m = 30$ in the latter. For the case-control designs, one control is selected per case. The number of cases are approximately equal to the size of the subcohorts [b].

	$m = 100$ $\mu = 2$, $\sigma = 1$ $\beta = 0.30$		$m = 100$ $\mu = 4$, $\sigma = 1$ $\beta = 0.20$		$m = 100$ $\mu = 4$, $\sigma = 4$ $\beta = 0.20$		$m = 100$ $\mu = 4$, $\sigma = 4$ $\beta = 0.30$		$m = 30$ $\mu = 4$, $\sigma = 4$ $\beta = 0.30$	
Method	aver	sd	aver	sd	aver	sd	aver	sd	aver	sd
Full cohort	0.298	0.088	0.196	0.060	0.196	0.052	0.299	0.044	0.294	0.069
Self & Prentice	0.314	0.148	0.210	0.105	0.221	0.114	0.337	0.133	0.557	0.691
Est II unstratified	0.311	0.140	0.206	0.097	0.212	0.092	0.314	0.085	0.377	0.182
Prentice	0.309	0.143	0.205	0.101	0.213	0.103	0.320	0.109	0.374	0.229
Est I time-fixed	0.302	0.124	0.196	0.071	0.197	0.062	0.302	0.053	0.306	0.131
Est I time varying	0.302	0.123	0.196	0.070	0.198	0.061	0.302	0.052	0.309	0.125
Est II time-fixed	0.301	0.122	0.195	0.070	0.197	0.061	0.300	0.051	0.300	0.111
Est II time varying	0.301	0.121	0.196	0.069	0.197	0.060	0.301	0.050	0.304	0.106
Est III time-fixed	0.298	0.123	0.195	0.071	0.197	0.062	0.301	0.053	0.302	0.109
Est III time varying	0.299	0.121	0.195	0.069	0.197	0.060	0.301	0.052	0.301	0.102
Nested case-control	0.309	0.144	0.204	0.102	0.208	0.096	0.311	0.091	0.328	0.174
Counter-matched	0.286	0.124	0.193	0.067	0.196	0.058	0.299	0.050	0.295	0.080

a) In the first four panels $\alpha = 0.10$ and $c = 3$, in the latter $\alpha = 0.05$ and $c = 1.25$.

b) The average number of failures are 89 - 91 for the first three panels, 101 for the fourth and 39 for the last.

T is generated from the exponential distribution with parameter $\alpha e^{\beta z}$. The censored survival time is $\widetilde{T} = \min(T, U)$, where the censoring variable $U = \min(1, V)$ with V independent of T and uniformly distributed over $[0, c]$. Various choices of the parameters μ, σ and β were used to illustrate situations with different covariate distributions and effects of the covariate, while the parameters α and c were adjusted to obtain the desired number of failures and censorings before time $t = 1$.

For all simulated data sets we compute the full cohort estimator as well as estimators based on simple and stratified case-cohort sampling. The simple and counter-matched nested case-control designs are also included for comparison (cf. below). For the unstratified case-cohort study, we consider 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 are considered both with time-fixed and time-varying weights. 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 stratified design, an equal number of individuals are selected to the subcohort among those with $\widetilde{Z} = 0$ and $\widetilde{Z} = 1$.

The first four panels of Table 1 give results for a situation where more than half of the cohort subjects are still at risk at the end of the study (i.e. at $t = 1$), and where the number of cases is approximately the same as the subcohort size of 100 individuals. In the last panel, the subcohort consists of 30 individuals, the number of failures are about the same, while only a fifth of the individuals are still at risk at the end of the study. The choices of μ, σ and β in the three first panels correspond to situations where an increase in the value of the covariate equal to four standard deviations of the (unconditional) distribution of Z, corresponds to

a relative risk of 3.5 - 4. For the fourth and fifth panel, a similar increase corresponds to a relative risk which is about twice as large. The main conclusion from Table 1, 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.

For a subcohort size of 100 subjects, i.e. the first four panels of Table 1, all the stratified estimators give similar results, and the differences observed are of little practical importance. Nevertheless, it is worth noting, that for all three methods, the version with time-dependent weights performs slightly better than the version where the weights are time-fixed. Further Method III performs consistently better than Method I, while Methods II and III give almost identical results. For the three unstratified methods, the Self-Prentice estimator has the poorest performance. It is biased upwards and has consistently the largest standard deviation. Among the other two, Prentice's estimator tends to have slightly smaller bias than Estimator II, while the latter has the smallest standard deviation. These qualitative conclusions for a subcohort size of one hundred, are confirmed by additional simulations (not shown) with more censorings and/or less failures. A tendency is observed, however, of slightly improved performance of the Prentice estimator relative to Estimator II as the number of failures are reduced. Subcohorts as small as 30 individuals are not used in practice, but may mimic situations where a time-dependent effect of a covariate is to be estimated from data in the latter part of a case-cohort study. As illustrated in the last panel of Table 1, somewhat larger differences between the methods can show up in such extreme situations.

The two last lines of Table 1 give the results for the simple and counter-matched case-control designs with one control selected per case. The results are not directly comparable, however, since the number of distinct subjects used are not exactly the same for the case-cohort and the case-control designs. The case-control designs require information from somewhat fewer subjects for the situations considered in the first three panels of Table1, and from slightly more for the latter two. Although this caveat makes it difficult to make precise statements about the comparative efficiency, the results indicate that the case-control designs have a performance which is approximately comparable to that of the corresponding case-cohort studies.

6 Discussion

Our simulation 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 three stratified estimators, it is also important to note that the data requirements are not the same for all three. Estimator II requires the full covariate histories for the cases, while Estimators I and III only need the cases' covariate values at their failure times.

One rather strong assumption we have made in our development, is that the stratified sampling of the subcohort depends only on exposure information known at time zero. This poses no difficulties with time-fixed covariates and for clinical trial type situations such as the HIV study example mentioned in Section 1 in which the P-24 levels used in the sampling would be collected at entry to the study. Further research is needed to assess the magnitude of the bias that may occur in other situations, such as in occupational cohort settings. Such studies would often require stratifying the subcohort based on a time-varying covariate evaluated at a time other than zero or on a covariate at the time of entry when there is late entry into the cohort. This cautionary note only applies to the sampling of the subcohort. If the assumption holds, then estimation of parameters associated with time-dependent factors included appropriately in the model will be valid.

We have discussed variance estimation for the stratified case-cohort estimators with time-fixed weights in Section 4. For simple random sampling, the $\tilde{\mathbf{X}}_k$ given by (18) sum to zero, and it can be shown that our variance estimator (17) equals the one given by Self and Prentice (1988, p. 74) when simplified to the situation with exponential relative risk function. A similar reformulation of the Self-Prentice covariance estimator is given by Therneau and Li (1998), who show how their version of the estimator can be easily computed using standard computer packages like Splus and SAS. 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.

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. In principle, once one has split the cohort into strata, question *(ii)* has a precise answer. If the covariate Z_1 is the one of main interest, it is optimal to chose the m_l proportional to $n_l\delta_{11l}$, where $\Delta_l = \{\delta_{jkl}\}$ is the matrix defined just below (16). Thus one should sample more than the proportional share from strata where there is a large variation among the X_k^0; cf. (15). 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 acknowledges the Department of Statistics, Stanford University, for providing the best working facilities during his sabbatical year 1997/98. He additionally acknowledges support from The Norwegian Research Council, Johan and Mimi Wessmann's foundation and the Norwegian Association of Professional Writers and Translators. The authors are grateful to Ola Hestnes and Svein Børre Solvang for programming assistance.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models based on Counting Processes*. Springer Verlag, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.*, **10**, 1100–20.

Barlow, W. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, **50**, 1064–72.

Breslow, N. and Cain, K. (1988). Logistic regression for two stage case-control data. *Biometrika*, **75**, 11–20.

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

Kalbfleisch, J. and Lawless, J. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. in Med.*, **7**, 149–60.

Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika*, **82**, 69–79.

Langholz, B. and Thomas, D. C. (1991). Efficiency of cohort sampling designs: Some surprising results. *Biometrics*, **47**, 1563–71.

Lehmann, E. (1975) *Nonparametrics* Holden-Day, San Francisco.

Lin, D. Y. and Ying, Z . (1993) Cox regression with incomplete covariate measurements *J. Amer. Statist. Assoc.*, **88**, 1341-1349.

Pogoda, J. (1993). *Variance Estimation in Complex Cohort Problems*. PhD thesis, University of Southern California.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **73**, 1–11.

Samuelsen, S. O. (1989). *Two Incomplete Data Problems in Life-History Analysis: Double Censoring and the Case-Cohort Design*. PhD thesis, University of Oslo.

Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies *Biometrika*, **84**, 379-394.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.*, **16**, 64–81.

Therneau, T. and Li, H. (1998). Computing for case-cohort designs Manuscript, Mayo Clinic.