A review of cohort sampling designs for Cox's regression model: Potentials in epidemiology

Ørnulf Borgan[1] and Sven Ove Samuelsen[1,2]

[1] *Department of Mathematics, University of Oslo*

[2] *Biostatistics Group, Division of Epidemiology, Norwegian Institute of Public Health*

E-mail: borgan@math.uio.no osamuels@math.uio.no

ABSTRACT

Cox regression is much used in epidemiology to assess the influence of exposure variables and other covariates on mortality or morbidity. Estimation in Cox's model requires ascertainment of covariate values for all individuals in a cohort even when only a small fraction of these actually get diseased or die (fail). For large cohorts this may be very costly. Cohort sampling techniques, where covariate information is collected for all failing individuals (cases), but only for a sample of the non-failing ones (controls), then offer useful alternatives. Such case-control data can contain almost as much statistical information as the full cohort. Two common cohort sampling designs are nested case-control and case-cohort sampling. The paper reviews, discusses and compares the two sampling designs. We also point out the potential benefits of stratified sampling of controls.

Key words: Case-cohort studies; Counter-matching; Cox regression; Matching; Nested case-control studies; Partial likelihood; Pseudo-likelihood; Stratified sampling.

1 Introduction

Cox regression is central to modern survival analysis, and it is the most used method when one wants to assess the influence of risk factors and other covariates on mortality or morbidity. Estimation in Cox's regression model is based on a partial likelihood [see (2) below], which at each observed death or disease occurrence (failure) compares the covariate values of the failing individual to those of all individuals at risk at the time of the failure. In large epidemiologic cohort studies of a rare disease, (standard) use of Cox regression requires collection of covariate information on all individuals in the cohort even though only a small fraction of these actually get diseased or die. This may be very expensive, or even logistically impossible. Cohort sampling techniques, where covariate information is collected for all failing individuals (cases), but only for a sample of the non-failing individuals (controls) then offer useful alternatives that may drastically reduce the resources that need to be allocated to a study. Further, as most of the statistical information is contained in the cases, such studies may still be sufficient to give reliable answers to the questions of interest.

There are two important classes of cohort sampling designs: nested case-control studies[1] and case-cohort studies. For nested case-control sampling, one for each case selects a small number of controls from those at risk at the case's failure time, and a new sample of controls is selected for each case. For the case-cohort design a subcohort is selected from the full cohort, and the individuals in the subcohort are used as controls at all failure times when they are at risk. In their original forms, the nested case-control and case-cohort designs both use simple random sampling without replacement for the selection of controls and subcohort (Thomas 1977, Prentice 1986). Later both designs have been modified to allow for stratified random sampling (Samuelsen 1989, Langholz & Borgan 1995, Borgan *et al.* 2000). Such stratified sampling may be advantageous when a surrogate measure of the covariate of main interest is available for everyone and can be used to classify the individuals into sampling strata.

[1] In epidemiological literature it is common to let the term "nested case-control study" mean any case-control study undertaken in a well-defined cohort or population. Here we will use this term in a more strict sense. Epidemiologists often refer to this design as density or incidence sampling of controls. Another term that is sometimes used is sampling from the risk set.

The purpose of this paper is to describe and discuss both the classical versions of the nested case-control and case-cohort designs using simple random sampling and their modifications with stratified sampling. The outline of the paper is as follows. In Section 2 we review Cox's regression model, describe the type of failure time data we consider for the cohort, and remind the readers about the usual methods of inference for cohort data. The simple nested case-control design is considered in Section 3. We describe how the controls are sampled, and review how statistical inference can be based on a partial likelihood similar to the one for the full cohort. Alternative methods of estimation are also briefly discussed, and the similarity between nested case-control studies and matched case-control studies is pointed out. In Section 4 we consider the simple case-cohort design. Again we describe how the sampling is performed, and discuss alternative methods for estimation of the regression coefficients. In Section 5 we compare the statistical efficiency of the different sampling designs and estimation methods with each other and with a full cohort study. The stratified versions of nested case-control and case-cohort sampling are briefly discussed in Section 6, while some concluding remarks are given in the final Section 7.

2 Model and inference for cohort data

We first review Cox regression for cohort data. Consider a cohort of n individuals, and let $\lambda_i(t)$ be the hazard rate for the ith individual with covariates $x_{i1}, \ldots, x_{ip}$. Here the time-variable t may be age, time since employment, or some other time-scale relevant to the problem at hand. The covariates may be time-fixed (like gender) or time-dependent (like cumulative exposure), but in the latter case we have supressed the time-dependency from the notation. The covariates may be indicators for categorical covariates (like the exposure groups "non-exposed," "low," "medium," and "high") or numeric (as when actual amount of exposure is recorded). We assume that the covariates of individual i are related to its hazard rate by Cox's regression model:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_1 x_{i1} + .. + \beta_p x_{ip}\}. \qquad (1)$$

Here $\beta_1, \ldots, \beta_p$ are regression coefficients describing the effects of the covariates, while the baseline hazard rate $\lambda_0(t)$ corresponds to the hazard rate of an individual with all covariates equal to zero. In particular we interpret the $RR_j = \exp(\beta_j)$ as hazard-rate ratios or more loosely as relative risks. This in the sense that $RR_j = \lambda_{i'}(t)/\lambda_i(t)$ when the covariates of individuals i' and i are equal except for covariate j for which $x_{i'j} - x_{ij} = 1$.

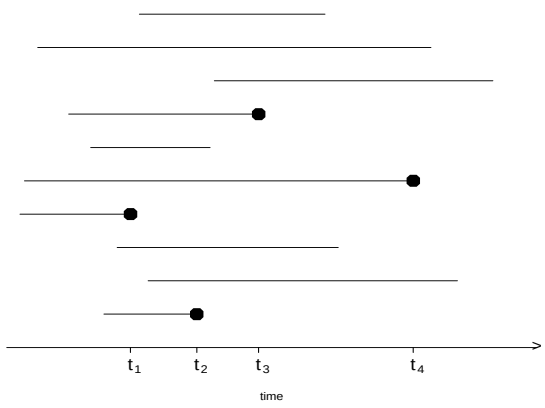

Figure 1: Illustration of data from a hypothetical cohort of ten individuals with four observed failures. Each individual is represented by a line starting at an entry time and ending at an exit time corresponding to censoring or failure. Failure times are indicated by dots (•).

The individuals in the cohort may be followed over different periods of time, from an entry time to an exit time corresponding to failure or censoring. The risk set $\mathcal{R}(t)$ is the collection of all individuals who are under observation just before time t, and $n(t)$ is the number at risk at that time. We will number the individuals so that $i = 1, 2, \ldots, d$ correspond to the failures with ordered failure times $t_1 < t_2 < \cdots < t_d$. (Here we have assumed that there are no ties, that is, no failure times are equal. A few ties may, however, be broken at random.)

Figure 1 illustrates the data for a small hypothetical cohort of 10 individuals. Each individual in the cohort is represented by a horizontal line starting at some entry time and ending at some exit time. If the exit time corresponds to a failure, this is represented by a "•" in the figure. In the hypothetical cohort considered, four individuals are observed to fail.

We assume throughout that late entries and censorings are independent in the sense that the additional knowledge of which individuals have entered the study or have been censored before any time t do not carry information on the risks of failure at t (Kalbfleisch & Prentice 2002, sections 1.3 and 6.2). Then the regression coefficients in (1) are estimated by $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$, the values of $\beta_1, \ldots, \beta_p$ maximizing Cox's partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{\exp\{\beta_1 x_{j1} + .. + \beta_p x_{jp}\}}{\sum_{k \in \mathcal{R}(t_j)} \exp\{\beta_1 x_{k1} + .. + \beta_p x_{kp}\}} \qquad (2)$$

It is a standard result that $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ can be treated

as ordinary maximum likelihood estimators. In particular, as implemented in many statistical packages, standard errors se_j for the $\widehat{\beta}_j$ are obtained automatically from the partial likelihood[2], 95% confidence intervals for $RR_j = \exp(\beta_j)$ are then given as $\exp(\widehat{\beta}_j \pm 1.96 se_j)$. Furthermore nested models may be compared by likelihood ratio tests.

3 Nested case-control sampling

3.1 Design: sampling of controls

The nested case-control design was originally suggested by Thomas (1977). For this design, if an individual fails at time t, one selects $m-1$ controls by simple random sampling (without replacement) from the $n(t)-1$ non-failing individuals in the risk set $\mathcal{R}(t)$. The set $\widetilde{\mathcal{R}}(t)$ consisting of the case and these $m-1$ controls is denoted a sampled risk set. Covariate values are ascertained for the individuals in the sampled risk sets, but are not needed for the remaining individuals in the cohort. Thus the design can be summarized as

- Case occurs at time t
- Sample $m-1$ controls from the risk set $\mathcal{R}(t)$
- Sampled risk set $\widetilde{\mathcal{R}}(t)$ consists of the case and the sampled controls
- Ascertain covariates for the individudals in $\widetilde{\mathcal{R}}(t)$

Figure 2 illustrates the basic features of a nested case-control study for the hypothetical cohort of Figure 1 when one control is selected per case (i.e. when $m = 2$). The potential controls for the four cases are indicated by a "|" in the figure, and are given as all non-failing individuals at risk at the times of the failures. Among the potential controls one is selected at random as indicated by a "$\circ$" in the figure. The four sampled risk sets are then represented by the four $\bullet$, $\circ$ pairs in Figure 2.

Note that the selection of controls is done independently at the different failure times. Thus subjects may serve as controls for multiple cases, and cases may serve as controls for other cases that failed when the case was at risk. For example, the case at time t_4 in the figure had been selected as control at the earlier time t_1.

A basic assumption for valid inference is that not only delayed entries and censorings, but also the sampling of controls, are independent in the sense that the additional knowledge of which individuals have entered the study, have been censored or have been selected as controls before any time t do not carry information on the risks of failure at t. This assumption will be violated if, e.g., in a prevention trial, individuals selected as controls change their behavior in such a way that their risk of failure is different from similar individuals who have not been selected as controls.

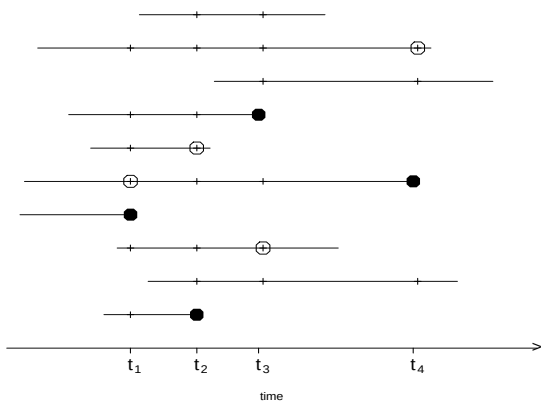

Figure 2: Illustration of nested case-control sampling, with one control per case, from the hypothetical cohort of Figure 1. Each individual is represented by a line starting at an entry time and ending at an exit time corresponding to censoring or failure. Failure times are indicated by dots ($\bullet$), non-failing individuals at risk at the failure times are indicated by bars (|), and the sampled controls are indicated by circles ($\circ$).

3.2 Estimation for nested case-control data

For nested case-control studies, it is common to base estimation of the regression coefficients in (1) on the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{t_j} \frac{\exp\{\beta_1 x_{j1} + .. + \beta_p x_{jp}\}}{\sum_{k \in \widetilde{\mathcal{R}}(t_j)} \exp\{\beta_1 x_{k1} + .. + \beta_p x_{kp}\}} \quad (3)$$

(Thomas 1977, Oakes 1981, Goldstein & Langholz 1992, Borgan *et al.* 1995). Note that the partial likelihood (3) is similar to the full cohort partial likelihood (2), except that the sum in the denominator is only over subjects in the sampled risk set.

Inference concerning the regression coefficients, using usual large sample likelihood methods, can be based on the partial likelihood (3). Thus the maximum partial estimators $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ are approximately normally distributed, and their standard errors may be obtained as for usual maximum likelihood estimators. Further nested models may be compared by the likelihood ratio test.

For computing one may use standard software for Cox regression, formally treating the label of the sampled risk sets as a stratification variable in the

[2] The variance estimates for maximum (partial) likelihood estimators are given as the inverse of the information matrix. This matrix is minus the second derivatives of $\log(L(\beta))$ evaluated at the parameter estimates.

Cox regression. Alternatively the partial likelihood (3) is of the same form as a conditional likelihood for logistic regression. Thus software typically used for analysis of matched case-control studies can also be applied. For both types of software each case and control has to be entered as a separate line in the datafile for every sampled risk set they are members of.

In the partial likelihood (3), cases are included only at their failure times. Samuelsen (1997) and Chen (2001) have developed estimation methods that use information from the cases and all the sampled controls whenever they are at risk. Estimation is then carried out by maximizing pseudo-likelihoods or weighted likelihoods of the form (6), as will be discussed in connection with case-cohort studies in Section 4.2, with weights equal to $1/p_k$ for inclusion probabilities p_k. For instance Samuelsen (1997) suggested using $p_k = 1$ for cases and

$$p_k = 1 - \prod \left[1 - \frac{m-1}{n(t_j) - 1} \right]$$

for controls. Here the product is over the failure times t_j for which individual k is at risk. Estimation of the relative risks is then, as discussed in Section 4.2, fairly straightforward. However, variance estimation requires some more attention. When the disease under study is fairly common, or if follow-up time depends strongly on covariates, these alternative methods of estimation give more precise estimates than the one based on the partial likelihood. However, nested case-control sampling is mainly used for rare diseases, and then the gain of these more complicated methods is often modest.

3.3 Matching in nested case-control studies

In order to keep the presentation simple, we have so far considered the proportional hazards model (1) where the baseline hazard rate is assumed to be the same for all individuals in the cohort. Sometimes this may not be reasonable. To control for the effect of one or more confounding factors one may want to adopt a stratified version of (1) where the baseline hazard differs between population strata generated by the confounders. The regression coefficients are, however, assumed to be the same across population strata. Thus the hazard rate of an individual i from population stratum c is assumed to take the form

$$\lambda_i(t) = \lambda_{0c}(t) \exp\{\beta_1 x_{i1} + .. + \beta_p x_{ip}\}. \qquad (4)$$

When the stratified proportional hazards model (4) applies, the sampling of controls in a nested case-control study need to be restricted to those at risk in the same population stratum as the case. We say that the controls are matched by the stratification variable. In particular if an individual in population stratum c fails at time t, one selects at random $m-1$ controls from the $n_c(t)-1$ non-failing individuals at risk in this population stratum. The partial likelihood (3) then still applies. Thus estimation of the relative risks is carried out as described in Section 3.2.

3.4 Matched and nested case-control studies

As mentioned Section 3.2 nested case-control studies may be analyzed as matched case-control studies. In fact one may think of nested studies as matched studies where time is a matching factor, and there are many examples of nested case-control studies that have been described as matched studies. The distinction, however, is that in the nested studies non-cases can be sampled controls for several cases, and cases still at risk can be chosen as controls. But under a rare disease assumption the chance of sampling cases or controls repeatedly is small.

4 Case-cohort sampling

4.1 Design: sampling of the subcohort

The case-cohort design was originally suggested by Prentice (1986). For this design one selects by simple random sampling (without replacement) a subcohort $\widetilde{\mathcal{C}}$ of size $\widetilde{m}$ from the full cohort, and the individuals in the subcohort are then used as controls at all the failure times when they are at risk. Covariate values are ascertained for the individuals in $\widetilde{\mathcal{C}}$ as well as for the cases occurring outside the subcohort, but they are not needed for the non-failures outside the subcohort. The design can be summarized as

- Sample subcohort $\widetilde{\mathcal{C}}$ from full cohort
- Case occurs at time t
- Sampled risk set $\mathcal{S}(t)$ at time t consists of the case and the individuals in $\widetilde{\mathcal{C}}$ that are still at risk
- Ascertain covariates for individudals in the subcohort $\widetilde{\mathcal{C}}$ and for cases not in $\widetilde{\mathcal{C}}$

Figure 3 illustrates a case-cohort study for the hypothetical cohort of Figure 1 with a subcohort size of four (i.e. with $\widetilde{m} = 4$). The individuals selected to the subcohort are indicated by thick lines.

As for nested case-control sampling, it is also for case-cohort sampling an assumption for valid inference that individuals sampled to the subcohort do not change their behavior in such a way that their risk of failure is different from similar individuals who have not been selected to the subcohort.

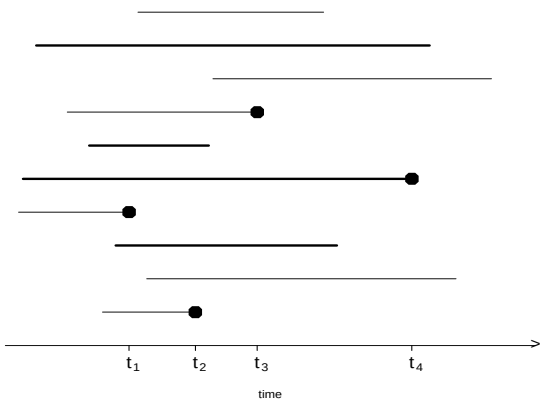

Figure 3: Illustration of case-cohort sampling, with a subcohort of size $\widetilde{m} = 4$, from the hypothetical cohort of Figure 1. Each individual is represented by a line starting at an entry time and ending at an exit time corresponding to censoring or failure. Failure times are indicated by dots (•), and the individuals in the subcohort are indicated by thick lines.

4.2 Estimation for case-cohort data

Different methods have been suggested for estimation of the regression coefficients in (1) from case-cohort data. The original suggestion of Prentice (1986) consist of maximizing what is referred to as a pseudo-likelihood

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{\exp\{\beta_1 x_{j1} + .. + \beta_p x_{jp}\}}{\sum_{k \in \mathcal{S}(t_j)} \exp\{\beta_1 x_{k1} + .. + \beta_p x_{kp}\}} \quad (5)$$

Here the summation in the denominator is over the set $\mathcal{S}(t_j)$ consisting of subcohort individuals at risk at time t_j with the case added whenever it occurs outside the subcohort.

Each term in the product in (5) is of the same form as a term in the product in (3). However controls from the subcohort are used over again for each case. For this reason (5) is not a partial likelihood (Langholz & Thomas 1991). This has the drawback that estimation of standard errors becomes more complicated and that likelihood ratio tests are not valid. Nevertheless one may show that the maximum pseudo-likelihood estimators $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ are approximately normally distributed (Self & Prentice 1988, Borgan *et al.* 2000).

Fitting of the pseudo-likelihood (5) was early on implemented in the program Epicure. The standard error estimates in this implementation uses the procedure originally proposed by Prentice (1986) and required substantial computational power (around 1990). Self & Prentice (1988) derived alternative estimates for standard errors, but presented them in a form that was difficult to implement. This earned case-cohort studies the reputation of being hard to analyze.

Actually the standard error estimates of Self & Prentice (1988) can be rewritten in a much simpler form as shown by Samuelsen (1989) and Therneau & Li (1999). The latter paper also provides computer code in the programs SAS, S-PLUS and R for fitting case-cohort studies with correct standard errors.

Another option for obtaining correct inference was suggested by Barlow (1994). Details for obtaining his corrected standard error estimates in SAS are given in Barlow *et al.* (1999). In S-PLUS this method is accomplished simply by using the robust variance estimator.

Another proposal for case-cohort studies, in the spirit of Kalbfleisch & Lawless (1988), is to maximize a weighted pseudo-likelihood

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{\exp\{\beta_1 x_{j1} + .. + \beta_p x_{jp}\}}{\sum_{k \in \widetilde{\mathcal{S}}(t_j)} \exp\{\beta_1 x_{k1} + .. + \beta_p x_{kp}\} w_k} \quad (6)$$

where now $\widetilde{\mathcal{S}}(t_j)$ is the joint set of the subcohort individuals along with all cases that are at risk. The weights are $w_k = 1$ for the cases (whether in the subcohort or not), and $w_k = 1/p_k$ for an inclusion probability p_k for the non-failures belonging to the subcohort. Borgan *et al.* (2000) suggested using p_k equal to the proportion non-cases in the subcohort compared to all non-cases. When the disease under study is fairly common, this alternative method of estimation will perform better than the one originally suggested by Prentice (1986). However, for rare diseases the difference is of less importance. Chen & Lo (1999) and Chen (2001) argued that other weights will reduce the variability in the estimates even further. However with a rare disease and with length of follow-up not strongly dependent on covariates the gain will again be modest.

Variance estimation under (6) can be carried out similarly to the procedure of Therneau & Li (1999). Correction of the variances can also be obtained by specifying robust variances in S-PLUS or by probability weighting in Stata. These corrections can be conservative, that is produce standard errors that are too large and confidence intervals that are to wide. However for this effect to be pronounced the effect of covariates needs to be quite strong.

5 Relative efficiency

5.1 Relative efficiency of simple random sampling

The relative efficiency of a cohort sampling method compared to a full cohort analysis, is the ratio of the variance of the estimator for full cohort data to the variance of the estimator based on the sampled

data. Thus, e.g., a relative efficiency of 1/2 means that the variances of the sampled data estimator is twice as large as the variance of the full cohort estimator (and that its standard error is $\sqrt{2} = 1.41$ as large). With a relative efficiency of 1/2 it would also require two such case-control studies to obtain the same statistical power as a cohort study. We may thus say that such a case-control study is only half as informative as a cohort study.

If there is only one covariate in the model, and its regression coefficient equals zero, the (large sample) relative efficiency of the simple nested case-control design compared to a full cohort study is $(m-1)/m$, independent of censoring and covariate distributions (Goldstein & Langholz 1992). Thus for testing a simple association the relative efficiency is 1/2 with one control per case and 2/3 when three controls are used. The relative efficiencies by the $(m-1)/m$ rule for several values of m are given in the following table.

Table 1: Relative efficiency of a nested case-control study with $m-1$ controls compared to a cohort study when $\beta = 0$.

m	2	3	5	10
Efficiency	0.5	0.67	0.8	0.9

When the regression coefficient departs from zero, and when more than one regression coefficient has to be estimated, the efficiency of the nested case-control design may be much lower than given by the "$(m-1)/m$ efficiency rule". E.g., with one binary covariate for exposure with relative risk $e^{\beta} = 4$, the relative efficiency of the nested case-control design with one control per case is about 1/4 when 10% of the cohort is exposed rather than 1/2 as the rule suggests.

For the simple case-cohort design, it does not seem possible to derive a similar general and simple result as the "$(m-1)/m$ efficiency rule" (Self & Prentice 1988). Although published results are somewhat conflicting (Langholz & Thomas 1991, Barlow *et al.* 1999), the relative efficiencies of simple nested case-control and case-cohort studies seem to be about the same when they involve the same number of individuals for whom covariate values have to be ascertained.

5.2 Simulation

To illustrate the relative merits of nested case-control and case-cohort studies we have carried out a Monte-Carlo simulation experiment. In this experiment the cohort size was set to $n = 1000$ and the average number of cases to 125. The model for the cohort was a proportional hazard model $\lambda_i(t) = \lambda_0(t)\exp(x_i)$, that is the log-relative risk $\beta = 1$. The covariates x_i come from a uniform distribution over the interval $[0, 1]$. The baseline hazard is given as $\lambda_0(t) = t$, thus survival times are drawn from Weibull distributions.

We will consider consider three different censoring patterns in the simulations. First we assume that every individual has the same potential follow-up time, thus censoring is at a fixed time. This will mimic studies in which all individuals are recruited at the same time and where there is no loss of follow-up. In the second set of simulations we let half of the individual have censoring according to a uniform censoring distribution and the other half have censoring time at the maximum value of this uniform distribution. This may correspond studies where all individuals are recruited at the same time, but with considerable loss of follow-up. In a third set of simulation the censoring times all come from a uniform distribution. This will correspond to studies in which individuals are recruited over time and follow-up is ended at the same date for all individuals.

For nested case-control samples we use one control for every case, i.e. $m = 2$. The size of the subcohorts in the corresponding case-cohort studies were then set such that the total number of individuals in both types of case-control studies are equal on average.

For each censoring scheme the simulations were repeated 2000 times. Results from the simulations are given in Table 2. For each censoring scheme we present in the first column averages of the estimates, in the second column the empirical variances of the estimates, and in the third column the averages of the variance estimates. These results are given for the complete cohort data Cox-estimates, for nested case-control data both with the usual partial likelihood estimates and the weighted (pseudo-likelihood) estimates of Samuelsen (1997), and for case-cohort data using the Prentice estimates based on (5) and the Kalbfleisch & Lawless (K & L) estimates based on (6).

The simulations shows that the log-relative risk β is estimated without any noticeable bias for all designs and estimation methods. Similarly all variances seems to be estimated without important bias. The variances are for most of the case-control estimators about twice the variance of the cohort estimator and thus in accordance with the relative efficiency rule in Table 1 with $m = 2$.

In general we see that with all censorings at the same time the variances for nested case-control and case-cohort estimates are approximately equal, perhaps with a slight advantage for case-cohort (Table 2a). The same holds true for the second censoring scheme, but now with a slight advantage for nested

Table 2: Results from the simulations

a) The censorings all at a fixed time.

	Ave.est.	Emp.var.	Ave.var.
Cohort	1.00	0.10	0.10
Nested C-C			
partial	1.01	0.22	0.22
pseudo	1.01	0.19	0.20
Case-cohort,			
Prentice	1.01	0.20	0.20
K & L	1.02	0.20	0.20

b) Half of the censorings at fixed time, the rest drawn from a uniform distribution up to this time

	Ave.est.	Emp.var.	Ave.var.
Cohort	1.00	0.10	0.10
Nested C-C			
partial	1.01	0.21	0.21
pseudo	1.00	0.18	0.18
Case-cohort,			
Prentice	0.99	0.24	0.23
K & P	1.00	0.23	0.23

c) All censorings from a uniform distribution

	Ave.est.	Emp.var.	Ave.var.
Cohort	1.02	0.10	0.10
Nested C-C			
partial	1.03	0.22	0.22
pseudo	1.04	0.19	0.18
Case-cohort,			
Prentice	1.05	0.30	0.28
K & P	1.02	0.31	0.28

case-control (Table 2b). This difference can easily be made up for by choosing a slightly larger subcohort. However as the proportion of censorings increases the variance of the case-cohort estimators increases while the variance of the nested case-control estimators seems to remain the same.

Thus for the third censoring scheme the case-cohort design fares rather bad compared to the nested case-control design (Table 2c). With this censoring scheme it may happen that (almost) everyone in the subcohort was censored before the last few cases. Then the case-cohort estimators we have presented are not able to pull much information out of these last cases. The nested case-control design on the other hand ensures that a minimum of controls are available for all cases. In a situation like this a simple case-cohort design should be avoided. Assuming that it is possible to determine the censoring time in advance one should instead carry out a stratified case-cohort study (see Section 6) with censoring time as a stratification variable.

Regarding the different estimation methods we see a consistent improvement for the pseudo-likelihood over the partial likelihood estimator for nested case-control. This is mainly due to the relatively high incidence (12.5 % cases). For case-cohort there is here no improvement of the Kalbfleisch & Lawless estimator (6) over the Prentice estimator (5) partly due to a relatively small effect of the covariate.

Some comments regarding variance estimation are in place. For case-cohort we used both the estimation routine of Therneau & Li (1999) and the robust variance of Barlow (1994). On average these gave the same result. However the variation of the robust variances is higher than that of the model based variances. This may indicate that for very small case-control studies the model based variances should be preferred. The variance estimates for the pseudo-likelihood estimator under the nested case-control design were carried out both by specifying robust variances in the weighted Cox-regression and by a similar routine as that of Therneau & Li (1999). Again these agreed on average. However they are both conservative relatively to the variance estimator in Samuelsen (1997), that is the variances are somewhat too large. In the present situation the conservatism was negligible. In fact the variance estimators of Therneau & Li (1999) and Barlow (1994) for the case-cohort design are also conservative for the estimates based on (6) although no indication of this was visible from these simulations.

6 Stratified Designs

Our presentation of nested case-control studies and case-cohort studies so far has assumed that all covariate information is obtained only on the case-control sample. However, since such studies are performed within well-defined cohorts there will generally be additional background data that are available for all cohort members. For instance a surrogate measure of exposure, like type of work or duration of employment, may be available for everyone. Based on such information it is possible to construct more efficient case-control designs. The main idea is that the cohort can be stratified according to the available

background data, so that predetermined number of controls can be sampled from each stratum. This will, e.g., ensure that we sample control individuals with long and short employment time, and thus likely also individuals with low and high cumulative exposure. In this section we describe variants of the nested case-control and the case-cohort design that effectively takes the background data into account.

6.1 Counter-matched sampling of controls

Langholz & Borgan (1995) have developed a stratified version of the simple nested case-control design which makes it possible to incorporate additional information into the sampling process in order to obtain a more informative sample of controls. For this design, called counter-matching, one applies the additional information on the cohort subjects to classify each individual at risk into one of say, S, strata. We denote by $\mathcal{R}_s(t)$ the subset of the risk set that belongs to stratum s, and let $n_s(t)$ be the number at risk in this stratum just before time t. If a failure occurs at t, we want to sample our controls such that the sampled risk set will contain a specified number m_s of individuals from each stratum $s = 1, \ldots, S$. This is obtained as follows. Assume that an individual who belongs to stratum r fails at t. Then for $s \neq r$ one samples randomly without replacement m_s controls from $\mathcal{R}_s(t)$. From the case's stratum r only $m_r - 1$ controls are sampled. The failing individual is, however, included in the sampled risk set $\widetilde{\mathcal{R}}(t)$, so this contains a total of m_s from each stratum. The design can be summarized as

- Case occurs at time t from stratum r
- Sample $m_r - 1$ controls from the those at risk in the stratum of the case
- Sample m_s controls from the other strata
- Sampled risk set $\widetilde{\mathcal{R}}(t)$ consists of the case and the sampled controls
- Ascertain covariates for individuals in $\widetilde{\mathcal{R}}(t)$

Even though it is not made explicit in the notation, we note that the classification into strata may be time-dependent; e.g., one may stratify according to the quartiles of a time-dependent surrogate measure of the covariate of main interest. A crucial assumption, however, is that the information on which the stratification is based has to be known just before time t. This assumption is similar to the requirement that time-dependent covariates need to be known prior to events.

By counter-matching, one may be able to increase the variation in the value of the covariate of main interest within each sampled risk set, and this will increase the statistical efficiency for estimating the corresponding regression coefficient. In particular, if this covariate is binary, and we select one control per case, concordant pairs (i.e., the case and its control have the same value of the covariate) do not give any information in estimating the effect of the covariate. For a counter-matched design with $S = 2$ and $m_1 = m_2 = 1$, and where stratification is based on a surrogate measure of the covariate of main interest, the single control is selected from the opposite stratum of the case. This will reduce the number of concordant pairs, and thereby increase the information contained in the pairs of cases and controls. The situation with two strata and one control per case also gives a motivation for the name counter-matching. As the name suggests, it is essentially the opposite of matching where the case and its control are from the same stratum (cf. Section 3.3).

Inference for counter-matched nested case-control studies may be based on a partial likelihood similar to (3). However, weights have to be inserted in the denominator of the partial likelihood in order to reflect the different sampling probabilities in the various strata. Specifically, if individual k in sampled risk set $\widetilde{\mathcal{R}}(t_j)$ belongs to stratum s, its contribution to the partial likelihood is multiplied by $w_k = n_s(t_j)/m_s$. Note that the weight is the same whether individual k is a case or a control.

Inference concerning the regression coefficients, using usual large sample likelihood methods, can be based on the weighted partial likelihood. Moreover, software for Cox regression can be used to fit the model provided the software allows us to specify the logarithm of the weights as "offsets". For further details on counter-matched nested case-control studies, the reader is referred to the review by Langholz & Goldstein (1996).

6.2 Stratified case-cohort studies

Above we indicated how counter-matched sampling in nested case-control studies may increase the variation in the value of the covariate of main interest in the sampled risk sets, and thereby increase the statistical efficiency for estimating the corresponding regression coefficient. In a similar manner, stratified sampling of the subcohort may be advantageous in case-cohort studies when a surrogate measure for the covariate of main interest is available for everyone and may be used to classify the individuals in the cohort into a number of distinct strata. With n_s individuals in stratum s, one then selects a random sample of $\widetilde{m}_s$ individuals to the subcohort $\widetilde{\mathcal{C}}$ from each stratum s (Samuelsen 1989, Borgan *et al.* 2000).

The design can be summarized as

- Sample subcohort $\widetilde{\mathcal{C}}$ from full cohort by stratified sampling

- Case occurs at time t
- Sampled risk set $\mathcal{S}(t)$ at t consists of the case and the individuals $\widetilde{\mathcal{C}}$ that are still at risk
- Ascertain covariates for individuals in the subcohort $\widetilde{\mathcal{C}}$ and for cases not in $\widetilde{\mathcal{C}}$

Note, however, that while the strata may depend on time for nested case-control studies, they need to be fixed over time for the case-cohort design. It is on the other hand possible to define strata according to the length of follow-up.

As for the simple case-cohort design, there are different options for the analysis of stratified case-cohort studies. One possibility is to modify the pseudo-likelihood (5) by including weights in the denominator to reflect that the sampling fractions vary between strata. Specifically, if individual k belongs to stratum s its contribution to the pseudo likelihood (5) is weighted by $w_k = n_s/\widetilde{m}_s$. Alternatively, one may modify the pseudo-likelihood (6) and use information from the cases at all failure times when they are at risk. Then the proper weights are $w_k = 1$ for the cases and $w_k = n_s^0/\widetilde{m}_s^0$ for the non-failing subcohort members from stratum s. Here n_s^0 and $\widetilde{m}_s^0$ are the number of non-failures in the cohort and the subcohort, respectively, belonging to stratum s. It should be noted, however, that the gain by including the cases at all failure times when they are at risk, seems to be of less importance for stratified sampling than is the case when the subcohort is selected by simple random sampling (Borgan *et al.* 2000).

6.3 Efficiency gain by stratified sampling

Stratified sampling is a useful option when there is one covariate that is of particular interest in the study. Then one may use a surrogate measure for this covariate (available for everyone) to increase the variation in this covariate within the sampled risk sets or the subcohort, and thereby obtain an efficiency gain for estimating the regression coefficient corresponding to this covariate. However, one should be aware that there is "no free lunch," so stratified sampling may result in a loss in efficiency for other covariates compared to simple random sampling.

For the nested case-control design, the efficiency gain has been documented both by large sample relative efficiency calculations (Langholz & Borgan 1995, Langholz & Goldstein 1996) and by Steenland and Deedens' (1997) study of a cohort of gold miners. For the latter, a counter-matched design (with stratification based on duration of exposure) with three controls per case had the same statistical efficiency for estimating the effect of exposure to crystalline silica as a simple nested case-control study using ten controls. The efficiency gain by using stratified sampling for the case-cohort design has been less studied. But the simulation study of Borgan *et al.* (2000) seems to indicate that the gain by using stratified sampling is of comparable size for nested case-control and case-cohort studies.

7 Discussion

Cohort studies are usually considered to be the most reliable study design in epidemiology, while "classical" case-control studies are easier and quicker to implement, but usually also less reliable. The cohort sampling methods considered in this paper have been developed in response to the need to have available study designs which, like cohort studies, take the time aspect in the development of a disease into account, and at the same time combine the cost-effectiveness of a "classical" case-control study with the presumably greater validity of a cohort study.

If one in a study wants to use a cohort sampling method to reduce the workload of data collection and error checking, a choice between a nested case-control and a case-cohort study has to be made. As the two designs generally are comparable as far as statistical efficiency is concerned, the choice between the two has to be based on other considerations.

The statistical analysis of nested case-control data may be performed using partial likelihood methods and standard software for Cox regression. Since the usual large sample likelihood methods do *not* apply for case-cohort data, the analysis of data from case-cohort studies is more cumbersome.

Control sampling in a nested case-control study are from those at risk at the cases' failure times, while in a case-cohort study the subcohort is selected without consideration of at risk status. This difference in the way sampling is performed, creates two limitations for nested case-control studies which are avoided for case-cohort studies (Barlow *et al.* 1999):

- ascertainment of covariate values for the controls has to wait until failures occur
- choice of time-scale for the analysis has to be decided before the controls are selected

The relevance of these limitations depends on the situation at hand, but they are most likely to be of importance for prospective studies like disease prevention trials.

Often there will be some background information available for all members of a cohort. If some of this information is correlated with the covariate of main interest, it may be advantageous to adopt a stratified study design. However, one should keep in mind that one sometimes has to pay for the increased statistical efficiency for estimating the effect of the covariate of main interest by a lower statistical efficiency for estimating the effect of other covariates.

References

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

Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of case-cohort designs. *J. Clin. Epidemiol.*, **52**, 1165-72.

Borgan, Ø., Goldstein L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Stat.*, **23**, 1749-78.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.*, **6**, 39-58.

Chen, K. (2001). Generalized Case-Cohort Estimation *J. Roy. Statist. Soc. B*, **63**, 791-809.

Chen, K. and Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika*, **86**, 755-64.

Goldstein L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Stat.*, **20**, 1903-28.

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

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data. 2nd edition.* Wiley

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

Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science*, **11**, 35-53.

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

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

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**, 379-94.

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

Steenland, K. and Deddens, J. A. (1997). Estimating exposure-response trends in nested case-control studies: control selection via counter-matching versus random sampling. *Epidemiology*, **8**, 232-242.

Therneau, T. M. and Li, H. (1999). Computing the Cox model for case-cohort designs. *Lifetime Data Anal.*, **5**, 99-112.

Thomas, D. C. (1977). Addendum to: "Methods of cohort analysis: appraisal by application to asbestos mining," by F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *J. Roy. Stat. Soc. A*, **140**, 469-91.