Dept. of Math. University of Oslo
Statistical Research Report No. 1
ISSN 0806–3842 February 2005

Stratified case-cohort analysis of general cohort sampling designs

Sven Ove Samuelsen
Department of Mathematics, University of Oslo &
Division of Epidemiology, Norwegian Institute of Public Health
P.O. Box 1053 Blindern, 0316 Oslo, Norway
email: osamuels@math.uio.no,

Hallvard Ånestad
Ullevål University Hospital, Oslo, Norway

and

Anders Skrondal
Department of Statistics, London School of Economics &
Division of Epidemiology, Norwegian Institute of Public Health

Abstract

We first point out that variance estimates for regression coefficients in exposure stratified case-cohort studies (Borgan et al., 2000) can easily be obtained from influence terms routinely calculated in standard software for Cox regression. We also place the estimators proposed by Chen (2001) for a general class of cohort sampling designs within the Borgan et al. framework by allowing for post-stratification on outcome. This facilitates simple variance estimation for this class of cohort sampling designs. Finally, we extend the approach of Chen to accommodate stratified designs with surrogate variables available for all cohort members, such as stratified case-cohort and counter-matching designs.

[0]*Key words:* Bernoulli sampling, case-cohort studies, counter-matched studies, DFBETAS, generalized case-cohort studies, inverse probability weighting, nested case-control studies, post-stratification, stratified case-cohort studies, variance estimation

1 Introduction

In cohort studies the event under investigation is often rare. In such situations the use of case-control designs can considerably reduce the number of individuals for which covariate information must be gathered without much loss in efficiency. The resulting savings can then potentially be spent on acquiring relevant covariates to reduce omitted variable bias and accurate measurement of covariates to reduce measurement error bias.

The two main variants of cohort sampling designs are the nested case-control design (Thomas, 1977) and the case-cohort design (Prentice, 1986). In a nested case-control design controls are sampled from the risk sets during the follow-up at event times. In a case-cohort design, which provides the background for the approach proposed in this paper, covariates are obtained for individuals who experience the event (cases) and for a subcohort sampled at the outset of the study. Proportional hazards models are typically fitted to case-cohort data using estimating equations that resemble partial likelihoods (Cox, 1972), such as the pseudolikelihood of Prentice(1986). Alternatively one may, similarly to for instance Self & Prentice (1988) or Chen & Lo (1999), use weighted partial likelihoods with inverse probability weighting (e.g. Robins et al., 1994).

Often some covariate information is available for all cohort members, including "surrogate" variables that are predictive of the main exposure variables. For instance, blood type may be known for an entire population whereas DNA-typing must be performed for each individual to determine the alleles of a particular gene. If the gene frequency is known to depend on blood type we may use blood type as a surrogate when sampling controls for which DNA-typing is performed. A more powerful study design can then be constructed by stratified sampling of the subcohort where the surrogate variables define the strata (Samuelsen, 1989, Borgan et al., 2000, Kulich & Lin, 2000). Alternatively, surrogate variables can be used for counter-matched designs (Langholz & Borgan, 1995) with stratified sampling of controls at each event time.

For stratified case-cohort studies Borgan et al. (2000) present large sample results for estimators derived from weighted partial likelihoods where weights are inverse sampling fractions to the subcohort. The large sample covariance matrix for this estimator can be split into two components; the cohort covariance matrix and a covariance matrix due to sampling the subcohort from the full cohort, which depends on the stratum-specific covariance matrix of score influence terms. Borgan et al. also argue that a more efficient estimator can be obtained by redefining the strata and sampling fraction by using the cases as a separate stratum in addition to the original strata (see also Chen & Lo, 1999), which amounts to post-stratification (e.g. Cochran, 1977). However, the efficiency improvement is typically modest in practice.

Self & Prentice (1988) derived the large sample properties of the original case-cohort estimator of Prentice (1986). Their variance estimator was subsequently sim-

plified by Samuelsen (1989) and Lin & Ying (1993). Building on the simplified representation, Therneau & Li (1999) show that the variation due to sampling can be calculated from estimated influence terms ("DFBETAS"). They also give concrete examples of implementation in the software packages SAS and Splus.

The first objective of this paper is to provide similar scripts for the stratified case-cohort estimator suggested by Borgan et al. (2000), which is possible since variance estimators in this case also depend on DFBETAS. This was indicated in Borgan et al. (2000) although implementation was not described.

A perusal of recent applications of stratified case-cohort designs suggests that it is useful to explicitly state the variance estimator of Borgan et al. (2000) in a simpler form. For instance, De Roos et al. (2005) and Li et al. (2006) appear to use robust variance estimates (Barlow, 1994) which we will show can be very conservative for stratified case-cohort studies. Hisada et al. (2005) claims to use an "appropriate" bootstrapping technique but no further details were given. However, even for standard case-cohort data bootstrapping should proceed with caution (Wacholder et al., 1989).

The second objective of this paper is to point out the relation between stratified case-cohort analysis and the approach of Chen (2001). We show that his "local averaging" estimators can be viewed as post-stratified case-cohort estimators where strata are defined by both case-control status and by right-censored time grouped into intervals. Variance estimation for the estimators within Chen's general cohort sampling framework can then be carried out using the script presented for stratified case-cohort analysis.

In addition to case-cohort studies, Chen's framework also includes nested case-control studies and what he calls "traditional case-control studies". Many other sampling designs such as replenishing the subcohort (Prentice, 1986, Barlow, 1994) are also possible within his framework. Chen argued that all these designs may be analyzed by the same approach. This proceeds by dividing the individuals into groups according to whether they experienced the event and into the intervals in which they were censored or became cases. His "local averaging" approach amounts to counting the numbers in these groups in the cohort and the sampled cohort and subsequently weighting by the inverse of the sampling fraction in the groups. In other words: Chen stratifies the cohort after follow-up and implicitly treats the sampled data as post-stratified.

The third objective of this paper is to extend the class of designs discussed by Chen (2001). Although the scope of his approach is quite general, it is confined to designs where no surrogate variables are available for all cohort members. Moreover, we point out that the local averaging or post-stratification technique can also be used for the stratified case-cohort designs of Borgan et al. (2001), counter-matched designs (Langholz & Borgan, 1995) and Bernoulli sampling designs (Kalbfleisch & Lawless, 1988, Robins et al., 1995). The connection to standard stratified case-cohort designs makes variance estimation straightforward.

The outline of the paper is as follows. In the next section we describe the frame-

work of Borgan et al. (2000) for stratified case-cohort analysis, present their regression parameter estimator and variance estimator and show how the latter can be obtained from the DFBETAS. We also present a small simulation study investigating the performance of this variance estimator. In Section 3 we discuss Chen's approach in more detail and point out how it is related to stratified case-cohort analysis. In Section 4 we study an extension of Chen's generalized case-cohort design to allow for surrogate dependent sampling and show how such data may be analyzed with the post-stratification method. In Section 5 we use simulations to investigate the performance of estimators which can be interpreted as poststratified case-cohort estimators. Finally, we close the paper with a brief discussion.

2 Stratified case-cohort studies

2.1 Cohort data

We represent cohort data as survival data in counting process notation (Andersen et al., 1993):

$$\mathcal{F} \;=\; \{i = 1, \ldots, n; 0 \leq t \leq \tau : (N_i(t), Y_i(t), Z_i)\}$$

where, for individual i, $N_i(t)$ is an indicator of event (case) before (or at) time t, $Y_i(t)$ is an indicator of being at risk just before time t and Z_i is a p-dimensional vector of covariates. For notational simplicity we omit possible time-dependency for Z_i.

Under the proportional hazards assumption, the hazard of the event for individual i is given as $\lambda_i(t) \;=\; \exp(\beta' Z_i)\lambda_0(t)$, where β is a vector of regression coefficients and $\lambda_0(t)$ a baseline hazard function. Cox (1972) suggested that β could be estimated by maximizing the (log-)partial likelihood which in counting process notation can be written

$$\log(L(\beta)) \;=\; \sum_{i=1}^{n} \int_0^\tau [\beta' Z_i - \log(S^{(0)}(\beta, s))] dN_i(s)$$

with

$$S^{(0)}(\beta, s) \;=\; \sum_{i=1}^{n} Y_i(s) \exp(\beta' Z_i).$$

2.2 Stratified case-cohort sampling

Assume that the full cohort has been divided into L strata based on covariates or surrogate variables available for all individuals. A subcohort is subsequently sampled from the full cohort using stratified sampling. Supposing that there are n_l^0 individuals in stratum $l = 1, 2, \ldots, L$ and that m_l^0 of these are sampled, the sampling fraction in stratum l is $\pi_{l,n}^0 = m_l^0/n_l^0$. Let V_i^0 be the indicator for individual i being sampled

to the subcohort and p_i^0 the corresponding inclusion probability for individual i such that $p_i^0 = \pi_{l,n}^0$ when individual i belongs to stratum l.

Covariate information is obtained for the entire subcohort and for the cases in the full cohort. From such data one may estimate β by maximizing a weighted partial log-likelihood

$$\tilde{l}_I(\beta) \;=\; \sum_{i=1}^{n} \int_0^\tau [\beta' Z_i - \log(\tilde{S}_I^{(0)}(\beta, s))] dN_i(s)$$

where

$$\tilde{S}_I^{(0)}(\beta, s) \;=\; \sum_{i=1}^{n} Y_i(s) \frac{V_i^0}{p_i^0} \exp(\beta' Z_i).$$

Note that $\tilde{S}_I^{(0)}(\beta, s)$ only depends on the subcohort, so that $\tilde{l}_I(\beta)$ may be computed from case-cohort data. Maximization of $\tilde{l}_I(\beta)$ produces 'Estimator I' of Borgan et al. (2000). For the special case where the whole cohort is a single stratum, this estimator coincides with the Self & Prentice (1988) estimator. The estimator is asymptotically equivalent to the original estimator of Prentice (1986), but is somewhat unstable in small samples. Borgan et al. (2000) also suggested an alternative estimator ('Estimator III') that coincides with the Prentice estimator when there is only one stratum.

We will in the sequel focus on 'Estimator II' of Borgan et al. (2000) where the strata are redefined by excluding all cases. Let n_l be the total number and m_l the sampled number of individuals in stratum l after redefining the strata. The sampling fraction in stratum l among the non-cases is thus $\pi_{l,n} = m_l/n_l$ and the inclusion probability p_i for a non-case i in stratum l is $p_i = \pi_{l,n}$.

For 'Estimator I' the cases that have not been sampled are not represented in $\tilde{S}_I^{(0)}(\beta, s)$. Since covariates are available it seems more efficient to make use of this information. This can be accomplished by weighting the contribution from the cases by one, or equivalently using a inclusion probability $p_i = 1$ (Kalbfleisch & Lawless, 1988, Chen & Lo, 1999). Using the modified inclusion indicator $V_i = \max(V_i^0, N_i(\tau))$, 'Estimator II', $\tilde{\beta}_{II}$, can then be obtained by maximizing

$$\tilde{l}_{II}(\beta) \;=\; \sum_{i=1}^{n} \int_0^\tau [\beta' Z_i - \log(\tilde{S}_{II}^{(0)}(\beta, s))] dN_i(s)$$

where

$$\tilde{S}_{II}^{(0)}(\beta, s) \;=\; \sum_{i=1}^{n} Y_i(s) \frac{V_i}{p_i} \exp(\beta' Z_i).$$

2.3 Large sample covariance matrices

Borgan et al. (2000) present the asymptotic covariance matrix for 'Estimator I', $\tilde{\beta}_I$, as $\sqrt{n}(\tilde{\beta}_I - \beta) \to \mathrm{N}(0, \Sigma^{-1} + \Sigma^{-1}\Delta_I\Sigma^{-1})$. Here Σ is defined as the limit of

$-n^{-1}\partial^2\tilde{l}_{II}(\beta)/\partial\beta^2$ and the variation due to the sampling is given by

$$\Delta_I \;=\; \sum_{l=1}^{L} q_l^0 \frac{1-\pi_l}{\pi_l}\Delta_l^0.$$

Here, q_l^0 is the limit proportion in stratum l in the cohort, i.e. $n_l^0/n \to q_l^0$, π_l is the limit of $\pi_{l,n}^0$ and Δ_l^0 is the limit of the stratum specific covariance matrix of the terms

$$X_i^0 \;=\; \int_0^\tau [Z_i - \frac{\tilde{S}_I^{(1)}(\tilde{\beta}_I, s)}{\tilde{S}_I^{(0)}(\tilde{\beta}_I, s)}]Y_i(t)\exp(\tilde{\beta}_I' Z_i)\frac{dN_\bullet(s)}{\tilde{S}_I^{(0)}(\tilde{\beta}_I, s)},$$

where $N_\bullet(s) \;=\; \sum_{i=1}^n N_i(s)$ and $\tilde{S}_I^{(0)}(\beta, s) = \sum_{i=1}^n Z_i Y_i(s)(V_i^0/p_i^0)\exp(\beta' Z_i)$.

Correspondingly, for 'Estimator II', the large sample distribution of $\sqrt{n}(\tilde{\beta}_{II} - \beta)$ is multivariate normal with covariance matrix $\Sigma^{-1} + \Sigma^{-1}\Delta_{II}\Sigma^{-1}$, where

$$\Delta_{II} \;=\; \sum_{l=1}^{L} q_l \frac{1-\pi_l}{\pi_l}\Delta_l,$$

and q_l is the limit of n_l/n, π_l the limit of $\pi_{l,n}$ and Δ_l the limit over the non-cases in stratum l of the covariance matrix of

$$X_i \;=\; \int_0^\tau [Z_i - \frac{\tilde{S}_{II}^{(1)}(\tilde{\beta}_{II}, s)}{\tilde{S}_{II}^{(0)}(\tilde{\beta}_{II}, s)}]Y_i(t)\exp(\tilde{\beta}_{II}' Z_i)\frac{dN_\bullet(s)}{\tilde{S}_{II}^{(0)}(\tilde{\beta}_{II}, s)}.$$

Note that since the original sampling did not depend on the outcome, the limit sampling fraction π_l in stratum l remains the same among the non-cases as before redefining the strata.

'Estimator II' is asymptotically more efficient than 'Estimator I' for two reasons: (1) the variation of X_i among the non-cases in stratum l, i.e. the Δ_l, is smaller than the overall variation in stratum l, i.e. Δ_l^0, and (2) because $q_l < q_l^0$. However, since the proportion of cases is typically small in case-cohort studies the efficiency improvement is usually modest.

2.4 Variance estimation

We will only give details of the variance estimation for 'Estimator II', since it is more efficient than 'Estimator I' and both model fitting and variance estimation is easier.

The natural estimator of Σ is $n^{-1}\tilde{I}$, where $\tilde{I} = -\partial^2\tilde{l}_{II}(\tilde{\beta}_{II})/\partial\beta^2$ is the observed information matrix evaluated at $\tilde{\beta}_{II}$. The covariance matrix of $\tilde{\beta}_{II}$ can be estimated by

$$\tilde{I}^{-1} + \sum_{l=1}^{L} m_l \frac{1-\pi_{l,n}}{\pi_{l,n}^2}\tilde{I}^{-1}\tilde{\Delta}_l\tilde{I}^{-1}$$

where $\tilde{\Delta}_l$ is the covariance matrix of the X_i among the sampled non-cases in stratum l. Note that, with $D_i = -\tilde{I}^{-1}X_i/p_i$ and $\mathcal{S}_l$ denoting the set of individuals sampled in stratum l after removal of cases, we can write the elements of the above sum as

$$m_l \frac{1-\pi_{l,n}}{\pi_{l,n}^2}\tilde{I}^{-1}\tilde{\Delta}_l\tilde{I}^{-1} \;=\; \frac{m_l(1-\pi_{l,n})}{m_l-1}\sum_{i\in\mathcal{S}_l}(D_i-\bar{D}_l)(D_i-\bar{D}_l)^\top,$$

where $\bar{D}_l$ is the average of D_i in $\mathcal{S}_l$. Thus, the left hand side of the equation is proportional to the stratum specific covariance matrix of the D_i.

A fair amount of programming may appear to be required to obtain X_i and D_i, but the D_i are fortunately calculated by many software packages (Therneau & Li, 1999). Specifically, the D_i's are the so-called "DFBETAS" for the controls, and approximate the influence on parameter estimates from removing individual i. The score of $\tilde{l}_{II}(\beta)$ can be written as

$$\begin{aligned}\tilde{U}_{II}(\beta) = \tfrac{\partial \tilde{l}_{II}(\beta)}{\partial\beta} &= \textstyle\sum_{i=1}^n \int_0^\tau [Z_i - \frac{\tilde{S}_{II}^{(1)}(\beta)}{\tilde{S}_{II}^{(0)}(\beta)}]dN_i(s)\\ &= \textstyle\sum_{i=1}^n \int_0^\tau [Z_i - \frac{\tilde{S}_{II}^{(1)}(\beta)}{\tilde{S}_{II}^{(0)}(\beta)}][dN_i(s) - \frac{V_i}{p_i}Y_i(t)\exp(\beta' Z_i)\frac{dN_\bullet(t)}{\tilde{S}_{II}^{(0)}(\beta)}]\end{aligned}$$

where $\tilde{U}_{II}(\tilde{\beta}_{II}) = 0$. Hence, the score contribution for a non-case simplifies to $-X_iV_i/p_i$.

Software which handles either weights or "offset"-terms is required to perform stratified case-cohort analysis. The weights are the inverses of the inclusion probabilities $1/p_i$ and the offsets are $\log(1/p_i)$. After fitting the Cox-model, the D_i are calculated and the sum of their stratum-specific covariances weighted by $m_l(1-\pi_l)$ is calculated, giving the covariance matrix due to sampling.

An example script for implementation in S-Plus and R is given below:

```
stratcox<-coxph(Surv(time,d)~z1+z2,weights=1/p)
dfb<-resid(stratcox,type='dfbeta')

gamma<-numeric(0)
for (str in 1:no.strata){
indst<-(1:length(time))[stratum==str]
if (m[str]>1) gamma<-gamma+(1-m[str]/n[str])*m[str]*var(dfb[indst,])
}
adjvar<-stratcox$var+gamma
```

Here, `time, d, z1, z2, p` and `stratum` are, respectively, the individual follow-up times, the case-indicators, two covariates, the individual inclusion probabilities and the stratum variable in the case-cohort study. The inclusion probabilities have been redefined such that cases have inclusion probability 1. The variable `stratum` has

levels $1, 2, \ldots, L$ for the non-cases and some other value for the cases. The number of strata L is denoted `no.str`. To obtain the variance estimates we also need the number sampled in each stratum `m` and the total number in each stratum `n` (after redefining the strata) as vectors of length L. The covariance matrix due to the sampling is stored in the variable `gamma` and the estimated covariance matrix for the regression coefficient estimators `adjvar` is given by adding the "naive" covariance matrix estimates $\tilde{I}^{-1}$ from `stratcox$var`.

For 'Estimator I' and 'Estimator III' of Borgan et al. (2000) fitting requires that data are set up in a more complicated way because the cases outside the subcohort do not contribute to $\tilde{S}_I^{(0)}(\beta, s) = \sum_{i=1}^n Y_i(s)\frac{V_i^0}{p_i^0}\exp(\beta' Z_i)$. Therneau & Li (1999) give details for standard case-cohort data, but since 'Estimator II' is more efficient than the other estimators we do not pursue this.

2.5 A small simulation study

We conducted a small simulation study in order to investigate the performance of the variance estimator. Survival times T_i were drawn from a proportional hazards model with one covariate $Z_i \sim U[0,1]$, regression parameter $\beta = 1$ and a Weibull baseline $\lambda_0(t) = 2t$. Censoring times were uniformly distributed on the interval $[0, 0.5]$ and independent of T_i. This resulted in a proportion of cases of about 12.5%. The strata were defined by a surrogate indicating whether Z_i was smaller or greater than 0.5 and a sampling fraction of 13% was chosen for both strata.

This simulation was replicated 5000 times with sample sizes $n = 1000$ and $n = 10000$. In each replication we estimated $\tilde{\beta}_{II}$ and its variance estimator se^2 according to the method described in Section 2.4. In addition, we recorded the robust variance estimate (Barlow, 1994, Therneau & Grambsch, 2000). Below we report the average of the parameter estimates, the averages variance estimates, the empirical variance, the proportion of confidence intervals $\tilde{\beta}_{II} \pm 1.96\, se$ covering the true value $\beta = 1$ and the average of the robust variance estimates.

Table 1: Result from simulations for stratified case-cohort designs.

	Average $\tilde{\beta}$	Mean variance estimator	Empirical variance	Coverage probability	Mean robust variance
$n = 1000$	1.023	0.198	0.210	0.944	0.250
$n = 10000$	1.003	0.0192	0.0186	0.952	0.0244

The estimator of the regression parameter was practically unbiased, the average variance estimates corresponded well to the empirical variances and the coverage corresponded well to the nominal value of 95%. There appeared to be a tendency of

under-estimated variances with a sample size of $n = 1000$, but coverage should be considered satisfactory based on only around 125 cases. The robust variance estimates were clearly larger than the estimated variances and 95% confidence intervals based on the robust variances had coverage of 0.97 for $n = 1000$ and 0.976 for $n = 10000$, thus this procedure was clearly conservative.

3 Generalized case-cohort designs and post-stratification

Chen (2001) discusses a general design for sampling controls – and cases – within a cohort study. In this section we present his framework and discuss how it is related to stratified case-cohort studies. Importantly, the "local averaging" approach proposed by Chen can be represented as post-stratification on censoring times grouped into strata. This enables us to use the variance estimation method described in the previous section.

Generalized case-cohort designs are defined as follows by Chen (2001, p. 793): (a) the design consists of a number of sampling steps, (b) each step takes a random sample of a certain size without replacement from a certain subset of the cohort and (c) the design of the sample size and subset at each step and of the total number of steps must not use information about the observed covariates.

In a standard case-cohort study the sampling is carried out in one step at the outset. The subcohort sampling is carried out by simple random sampling from the total cohort and does not depend on covariates. Thus, a standard case-cohort study clearly falls within this generalized case-cohort design.

In a nested case-control study (Thomas, 1977, Langholz & Goldstein, 1993) controls are sampled from the risk sets at event times with simple random sampling and without knowledge of covariates. The sampling steps are thus given by the event times and do not depend on covariates. Chen (2001) and Chen & Lo (1999) also discuss a traditional case-control design in which controls are sampled after observing the cases. For this design there is only one sampling step and the sampling does not depend on the covariates of the sampled individuals. Another design captured by the framework of Chen is studies in which new subcohorts are sampled at specified times (Prentice, 1986).

For generalized case-cohort designs Chen (2001) suggested a weighting technique termed "local averaging". This involves choosing partitions, separately for cases and controls, of the time axis and calculating weights that are assigned specifically to individuals with exit times in the intervals defined by the partition. In contrast to Chen we assume that covariate information is obtained on all cases and need only consider a partition $0 = s_0 < s_1 < \cdots < s_L = \tau$ for the controls. The weights are then given by

$$w_{(s_{j-1},s_j]} = \frac{\sum_{i=1}^{n} I(Y_i(s_{j-1} = 1, Y_i(s_j) = 0, N_i(\tau) = 0)}{\sum_{i=1}^{n} I(Y_i(s_{j-1} = 1, Y_i(s_j) = 0, N_i(\tau) = 0, V_i = 1)},$$

where V_i is the indicator that individual i was selected by the sampling design and $I(\cdot)$ is the indicator function. Thus, the numerator of $w_{(s_{j-1},s_j]}$ counts the number of individuals censored in $(s_{j-1}, s_j]$ and the denominator the number of these that were sampled. Individual i is then assigned weight $w_i = w_{(s_{j-1},s_j]}$ if censored within interval $(s_{j-1}, s_j]$ and $w_i = 1$ if the individual is a case.

Chen (2001) suggests estimating a proportional hazard model by solving the weighted estimating equation

$$\tilde{U}_h(\beta) \;=\; \sum_{i=1}^{n} \int_0^\tau [h_i(t) - \frac{\sum_{i=1}^n w_i V_i h_i(t) Y_i(t) \exp(\beta' Z_i)}{\sum_{i=1}^n w_i V_i Y_i(t) \exp(\beta' Z_i)}] dN_i(t) \;=\; 0$$

where the $h_i(t)$ are some functions of the covariates. In particular, with $h_i(t) = Z_i$ this becomes the score equation of a weighted partial likelihood. Chen (2001) argues that a properly chosen $h_i(t)$ can give an efficiency improvement as compared to the conventional $h_i(t) = Z_i$. We will, however, only consider the standard $h_i(t) = Z_i$ here.

Defining $p_i = 1/w_i$, we see that p_i can be interpreted as the proportion of individuals sampled among those who were censored in the same interval $(t_{j-1}, t_j]$ as individual i. Thus the weights can be interpreted as inverse sampling fractions. Also, for the cases $p_i = 1$ which corresponds to sampling all cases. Using this notation and $h_i(t) = Z_i$, the estimating equation becomes

$$\tilde{U}(\beta) \;=\; \sum_{i=1}^{n} \int_0^\tau [Z_i - \frac{\sum_{i=1}^n \frac{V_i}{p_i} Z_i Y_i(t) \exp(\beta' Z_i)}{\sum_{i=1}^n \frac{V_i}{p_i} Y_i(t) \exp(\beta' Z_i)}] dN_i(t) \;=\; 0,$$

which is formally identical to the estimating equation for 'Estimator II' within the stratified case-cohort design. However, the strata are in this setting determined by the length of follow-up instead of a surrogate variable for the covariates.

The method of Chen (2001) can be described as first carrying out the sampling by any sampling scheme within the class of generalized case-cohort studies, then dividing the cohort and the sampled data into strata according to event status and to length of follow-up and finally fitting a model to the data as if they were obtained by stratified case-cohort sampling. It is thus evident that the method amounts to post-stratification (see e.g. Cochran, 1977). Indeed, redefining the strata after observing whether the individuals are cases or non-cases, as was done for estimator II in the stratified case-cohort study, is just a more moderate form of post-stratification.

Due to the post-stratification argument, the large sample covariance matrix of the score of the weighted partial likelihood will be the same as if the data had originally been obtained by stratified sampling. It follows that the large sample properties of the estimator will also be the same as if data were originally collected by stratified sampling. The variances can hence be expressed and calculated as for the stratified case-cohort design. Specifically, this is so when the original sampling is simple random sampling from the full cohort as in the standard case-cohort design, or by

stratified sampling based on case-status in the traditional case-control design. The usual variance result with post-stratification relies on an original simple random or stratified sampling (Cochran, 1977).

The argument is somewhat more convoluted with for instance nested case-control sampling. Although the control sets at the different event times are all sampled by simple random sampling this does not imply that the set of controls are sampled in this way. Indeed, Samuelsen (1997) pointed out that the probability of ever being sampled as a control increases with length of follow-up. Within a post-stratum defined as a follow-up time in the interval $< s_{j-1}, s_j]$ the sampling fraction can vary considerably. However, when making the interval lengths $s_j - s_{j-1}$ all go to zero as sample size increases, the sampling fraction will become approximately equal for individuals censored in $< s_{j-1}, s_j]$. The sampling scheme will then correspond to stratified sampling.

For large sample results Chen (2001) assumed that the maximum number of individuals sampled in a censoring interval grows at a smaller rate than $n^{1/2}$, i.e. as $o_P(n^{1/2})$. For practical purposes this implies that $\max(s_j - s_{j-1}) \to 0$. However, the above post-stratification argument shows that this is not a necessary condition for asymptotic normality and consistency of estimators based on standard case-cohort and traditional case-control designs. However, for nested case-control and other sampling designs with several sampling steps the requirement of Chen is necessary since sampling fractions are usually not constant over censoring intervals. The choice of partitions may hence require some care to avoid biased estimates.

Although it is not always necessary for consistency and asymptotic normality to let the censoring intervals become small, there may be efficiency gains by decreasing their length. However, since the large sample results of Borgan et al. (2001) require that stratum sizes become large, a large number of strata can be a difficulty in particular study. The main efficiency gain might be obtained by using only a moderate number of censoring intervals.

4 Post-stratification for other sampling designs

The results of Chen (2001) required that sampling does not depend on covariates and that simple random sampling is used at each sampling step. Here, we argue that post-stratification (or local averaging) can be used in more general settings. Three sampling designs will be considered in detail, but application may also be possible for other designs. The main idea is that the sampling fractions within the strata should be approximately equal after post-stratification.

4.1 Stratified case-cohort studies

In stratified case-cohort studies the sampling fractions may depend on surrogate variables available for the complete cohort. Within a stratum, sampling of a subcohort

is carried out with simple random sampling. For estimator II of Borgan et al. (2000) which was discussed in Section 2, the strata and sampling fractions were redefined after observing which individuals became cases. It is then a fairly straightforward extension to redefine the strata for censoring, grouped into intervals, as well.

Borgan et al. (2001) discuss time-dependent weights defined as the number at risk in the cohort at a specific time divided by the sampled number at risk at that time, separately for each stratum. Time-dependent weighting has been shown to have good efficiency properties (Nan, 2004, Kulich & Lin 2004), but may be cumbersome to implement. Furthermore, a variance estimator is yet to be developed for this method.

Post-stratification on censoring (or local averaging) is a related way of improving the correspondence between the sampled data and the cohort data throughout the study period and may have similar efficiency gains. Furthermore, the weights are not time-dependent, which makes estimation easier. The variance estimator developed for estimator II of Borgan et al. (2001), modified by the censoring strata, can be used.

4.2 Counter-matched studies

Counter-matched studies (Langholz & Borgan, 1995) are similar to stratified case-cohort studies in the sense that the sampling depends on a surrogate variable known for all individuals in the cohort. On the other hand, the design is an extension of nested case-control studies since controls are sampled from the risk set of the cases. In particular, with L levels of the surrogate variable, m_l controls are sampled from strata l except for the stratum of the case at a time t_j. From stratum l' of the case $m_{l'} - 1$ controls are sampled. In this way the sampled risk set $\widetilde{\mathcal{R}}(t_j)$ at t_j, consisting the case and the controls sampled at that time, at all event times contains exactly m_l individuals from stratum l. With $n_l(t_j)$ individuals at risk right before time t_j in stratum l this risk set gives a likelihood contribution

$$L_j = \frac{\exp(\beta' Z_j)}{\sum_{k \in \widetilde{\mathcal{R}}(t_j)} w_{jk} \exp(\beta' Z_k)}$$

where $w_{jk} = m_l / n_l(t_j)$ when individual k has level l on the surrogate variable. The counter-matching estimator under the proportional hazards assumption proposed by Langholz & Borgan (1995) is obtained by maximizing the product of the L_j over the event times t_j as a function of β. This product possesses a partial likelihood property and large sample inference follows from this (Langholz & Borgan, 1995).

The post-stratification approach can be applied immediately also to counter-matched studies. We then define new strata according to event (case or non-case), censoring interval and the surrogate variable. Weights are again given as inverse sampling fractions within strata defined as the number of sampled individuals divided by the number of individuals in the cohort.

As for nested case-control designs the probability of being sampled will not be constant within a censoring interval, but with a fair number of censoring intervals it will not vary much. Large sample inference thus requires that the lengths of all intervals tend to zero as the sample size increases. However, the situation is otherwise similar to the nested case-control design and the post-stratification argument for variance estimation is valid.

Similarly to post-stratification for stratified case-cohort studies we may end up with a large number of strata. The theory of Borgan et al. (2000) also requires that the number sampled in each stratum is large, a requirement which may be difficult to satisfy for a given sample size. Consequently, the censoring intervals should be chosen with care.

4.3 Bernoulli sampling designs

Kalbfleisch & Lawless (1988) and Robins et al. (1994) discuss Bernoulli sampling where individuals are sampled independently. This design allows for inclusion probabilities that depend on covariates and surrogate variables. A variance formula for the estimated regression parameters was developed by Kalbfleisch & Lawless (1988). This formula can be written on a similar form as the one in Section 2.4 by replacing the central 2. order moment $(1/(m_l-1))\sum_{i\in\mathcal{S}_l}(D_i-\bar{D}_l)(D_i-\bar{D}_l)^\top$ by the non-central 2. order moment $(1/m_l)\sum_{i\in\mathcal{S}_l} D_i D_i^\top$, if the same sampling fraction is used for all individuals in stratum $\mathcal{S}_l$.

Formally this design does not belong to the class of Chen (2001) since Bernoulli sampling is not sampling with replacement. However, after conditioning on the number actually sampled in the strata with Bernoulli sampling and using the same sampling probability in each stratum, the sampling frame amounts to stratified random sampling. Thus, after correcting the inclusion probabilities to the number actually sampled divided by the number that could have been sampled within each stratum and weighting by these corrected inclusion probabilities, we obtain the same large sample results as if stratified sampling had been carried out.

Furthermore, this approach may also be extended to post-stratification on censoring intervals. As for stratified case-cohort studies, we then count the total and the sampled number of individuals in each interval and in each stratum among the non-cases and weight by inverse sampling fractions in each group.

5 Simulation studies

In this section we investigate the behavior of the post-stratification (or local averaging) method using simulations. One purpose is to discover when and to what extent this approach produces unbiased estimates and efficiency improvements. Another purpose is to explore the behavior of the variance estimators for stratified case-cohort studies.

We will use the simulation model from Section 2.5, although sometimes with modifications. Standard case-cohort studies, nested case-control studies, stratified case-cohort studies, counter-matched studies and Bernoulli-sampling strategies are considered.

5.1 Case-cohort design

In our first simulation we use the same cohort model as in Section 2.5, with covariate Z uniformly distributed on $[0, 1]$, regression coefficient $\beta = 1$, Weibull baseline hazard $\lambda_0(t) = 2t$ and uniform censoring on the interval $[0, 0.5]$, producing roughly 12.5% cases. This model is simulated 5000 times with a total sample size of 1000 individuals. In each replication of the simulation model a subcohort of size $m^0 = 130$ is sampled from the complete cohort with simple random sampling.

For each replication we obtain the Cox-estimator from the cohort data, the estimator with post-stratification only on case-status and two estimators which are also post-stratified on censoring. For the first of these the censoring interval is stratified into 5 intervals of equal length and for the second into 10 intervals of equal length. For all estimators the variance is estimated. In addition, we estimate the robust variance (Barlow, 1994, Therneau & Grambsch, 2000) for the case-cohort estimators. In Panel A of Table 2 we present the average of regression parameter estimates, the average variance and robust variance estimates and the empirical variances of the estimates. We also calculate the relative efficiency between the case-cohort estimators and the cohort estimator, defined as the ratio of their empirical variances.

There was a very slight bias for the case-cohort estimates, but the magnitude was the same for all three estimators. The variances were also very similar for all estimators, although it appears that post-stratification slightly increased the variances. This is in contrast to large sample results (Chen, 2001), but in a new round of simulations with the sample size increased to $n = 10000$ there was no difference between the variances (results not shown). There was good correspondence between the variance estimates and the empirical variances at least for Estimator II and with 5 interval post-stratification. The robust variance estimator also seemed to be valid for all three estimators.

These results are in clear contrast to the efficiency gains presented by Chen (2001). However, in that paper the censoring times depended on the covariates. To study this effect we will, following Samuelsen (1997), assume that the censoring time is exactly proportional to the covariate. With known censoring times for all individuals in the cohort we have complete cohort information and there is no need to carry out subcohort sampling. This model is still interesting to investigate because we get an idea of how large the efficiency gains can be and how far the weighted likelihood is from the efficient estimator.

Thus 5000 new simulations with the same model for the time to event, but with the censoring time exactly proportional to the covariate, were performed. The censoring

time was uniformly distributed over an interval from 0 to a value chosen to get about 12.5% cases. Subcohorts of size $m^0 = 130$ individuals were then sampled and the same estimators used as in the previous simulation. The results from these simulations are presented in Panel B of Table 2.

The variances were larger for these simulations due to smaller variation in covariate values for late risk sets. There was no evidence of bias of the regression parameter estimates. The correspondence between variance estimates and empirical variances were good, but the robust variance only worked properly for post-stratification only on case-status. Post-stratification also on censoring interval gave estimators that were markedly efficient compared to post-stratification only on case-status and that were not far from efficient compared to the cohort estimator. Using 10 intervals gives some efficiency improvement compared to 5 intervals, but the main gain was achieved with post-stratification on 5 intervals.

The relative efficiencies are somewhat better than those reported by Chen (2001) who used a censoring variable that was not exactly proportional to the covariate. To be able to compare with Chen we performed a third simulation with a correlation between covariate and censoring time approximately equal to 0.9 and with a similar model to that used above. The differences were that the hazard was $\lambda_0(t) = 2.22t$ and that the censoring time was given by $c = \min([3.2z]/6.4 + z'/6.4, 0.5)$, where $[x]$ denotes the largest integer smaller than x and where z' denotes another draw from a $U[0,1]$ independent of z. The results for 5000 replications are shown in Panel C of Table 2.

In these simulations the efficiency gain is still clear, but much more modest. This contrasts with the results of Chen (2001), but may be due to higher incidences in his simulations. There was some bias in the estimates of the regression parameter, variance estimates corresponded well to empirical variances, but the robust variances were clearly conservative for the estimators with post-stratification on censoring.

5.2 Nested case-control design

In nested case-control studies, controls are sampled from the risk sets of the failure times of the cases. Traditionally such studies are fitted using the Thomas (1977) estimator which is obtained by maximizing a Cox-type likelihood given as a product over event times, where the sum in the denominator at an event time is over the case and the controls sampled at that particular time. Goldstein & Langholz (1992) showed that this likelihood is a partial likelihood under the proportional hazards model (See also Oakes, 1981, Borgan et al., 1995).

Samuelsen (1997) instead suggested maximizing a weighted likelihood in which the sum in the denominator at an event time is over all sampled controls and all cases at risk at that time. The weights for the controls were given as the inverses of the estimated inclusion probabilities $p_i = 1 - \prod_s [1 - Y_i(s) m \frac{dN(s)}{Y(s)-1}]$, where m is the number of controls sampled for each case, $Y(s)$ is the number at risk at time $t-$ and

Table 2: Results from simulations of case-cohort studies with censoring time independent of the covariate (Panel A), proportional to the covariate (Panel B), and correlated with the covariate (Panel C).

Panel A: *Censoring independent of the covariate*

	Cohort	Case-cohort with post-stratification on		
	(Cox)	case-status only	5 intervals	10 intervals
Mean estimate	1.006	1.029	1.030	1.032
Mean variance	0.101	0.249	0.259	0.264
Mean robust variance	–	0.251	0.256	0.268
Empirical variance	0.100	0.257	0.263	0.279
Relative efficiency	–	0.39	0.38	0.36

Panel B: *Censoring time proportional to the covariate*

	Cohort	Case-cohort with post-stratification on		
	(Cox)	case-status only	5 intervals	10 intervals
Mean estimate	1.001	1.022	0.999	1.001
Mean variance	0.338	0.582	0.378	0.360
Mean robust variance	–	0.587	0.605	0.607
Empirical variance	0.330	0.600	0.377	0.355
Relative efficiency	–	0.55	0.88	0.93

Panel C: *Censoring time correlated with the covariate*

	Cohort	Case-cohort with post-stratification on		
	(Cox)	case-status only	5 intervals	10 intervals
Mean estimate	1.015	1.037	1.031	1.028
Mean variance	0.269	0.518	0.465	0.443
Mean robust variance	–	0.514	0.528	0.544
Empirical variance	0.268	0.523	0.467	0.438
Relative efficiency	–	0.51	0.57	0.61

$N(t)$ represents the total number of cases in $[0, t]$. It follows that p_i can be interpreted as the probability of ever being sampled as a control. Note that p_i will increase with the length of follow-up and that the weights equal 1 for the cases.

As an improvement Chen (2001) suggested using local averaging weights, which we have argued amounts to post-stratification on censoring times grouped into interval strata. The inclusion probability for this approach will be constant over the time interval, but it may well decrease from one interval to the next. This may reflect the actual sampling better than the monotonous (in length of followup) p_i of Samuelsen (1997).

However, the actual choice of intervals is somewhat arbitrary and there could be problems both with intervals that are too short and too long. An alternative inclusion probability can be obtained by using some smooth function over time that properly describes the proportion of sampled controls. Several ways of implementing this idea are possible, for instance smoothing indicators of being sampled against censoring times with generalized additive models (GAM, Hastie & Tibshirani, 1990).

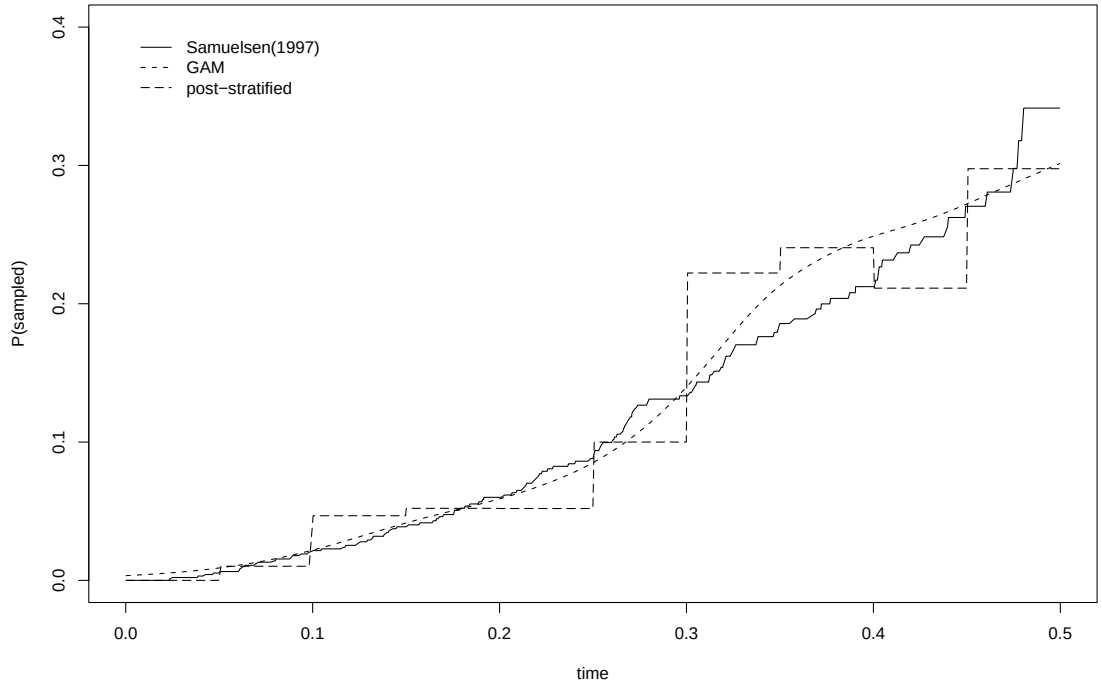
As an example we simulated the model in Section 2.5 once and sampled $m = 1$ control per case. Estimates of the probability of being sampled as a control are displayed in Figure 1. For post-stratification we only show the estimates based on 10 equal length intervals. A potential problem with the post-stratified estimate with 10 intervals is that some intervals do not contain controls, corresponding to a zero sampling fraction.

We replicated simulations with nested case-control sampling of $m = 1$ control per case 5000 times. In each simulation we used (1) the cohort (Cox) estimator, (2) the Thomas (1977) estimator and (3) weighted partial likelihood estimators. We used weights from (3a) the inclusion probabilities of Samuelsen (1997), (3b) GAM, (3c) post-stratification with 5 equal length intervals and (3d) post-stratification with 10 equal length intervals. Variance estimates were obtained for the Cox-estimator and the traditional nested case-control estimator (as the inverse information) and for the post-stratified estimators. Samuelsen (1997) developed a variance estimator for his estimator, but this was not used in these simulations. For the GAM-weighting no variance estimator is developed.

Results from the simulations are reported in Panel A of Table 3, apart from the results for the Cox-estimator and the post-stratified estimator with 5 intervals. For the Cox-estimator the results were very close to those in Panel A of Table 2. The reported efficiency is relative to the Cox-estimator.

The estimators all had a slight bias. Variance estimation worked well for the traditional nested case-control method and the post-stratified method. Both for the inclusion probability of Samuelsen (1997) and the GAM approach the robust variances performed well. The traditional nested case-control estimator was somewhat inefficient compared to the weighted estimators, but the inclusion probability and the GAM approaches produced as precise estimates as the post-stratified estimator. Post-stratification with 5 intervals gave similar results as with 10 intervals.

Figure 1: Estimated probability of being sampled as a control as function of censoring time based on inclusion probabilities from 1) the method of Samuelsen (1997), 2) generalized additive models and 3) post-stratification with 10 intervals.



To evaluate the effect of a (relatively) small sample size we performed 5000 new simulations, increasing the sample size to $n = 10000$. The slight bias of the regression parameter estimators vanished. Average variance estimates, average robust variances estimates and empirical variances were in good agreement. The traditional nested case-control estimator was still somewhat inefficient compared to the weighted estimators.

Similar to the case-cohort study we also conducted 5000 simulations of nested case-control studies with the censoring times proportional to the covariates. Thus, full information about the covariate is available and the simulations were performed only to study the behavior in this extreme case. Results are given in Panel B of Table 3.

In this case the weights from the generalized additive models produced a practically efficient estimate. Post-stratification with 10 intervals also gave an estimator with small variation, although having a clear bias. Post-stratification with 5 intervals was less biased, but had a somewhat larger variance (results not shown). Variance estimation for the post-stratified estimators appeared to work well with 10 intervals,

Table 3: Results from simulations of nested case-control studies with censoring time independent of the covariate (Panel A), proportional to the covariate (Panel B), and correlated with covariate (Panel C).

Panel A: *Censoring time independent of the covariate*

		Inclusion probabilities		
	Traditional			Post-stratified
	Thomas (1977)	Samuelsen (1997)	GAM	10 intervals
Mean estimate	1.021	1.017	1.017	1.019
Mean variance	0.218	–	–	0.190
Mean robust variance	–	0.185	0.187	0.192
Empirical variance	0.221	0.190	0.192	0.198
Relative efficiency	0.46	0.54	0.53	0.52

Panel B: *Censoring time proportional to the covariate*

		Inclusion probabilities		
	Traditional			Post-stratified
	Thomas (1977)	Samuelsen (1997)	GAM	10 intervals
Mean estimate	1.030	0.983	0.999	0.929
Mean variance	0.711	–	–	0.350
Mean robust variance	–	0.525	0.521	0.552
Empirical variance	0.725	0.479	0.354	0.377
Relative efficiency	0.48	0.72	0.98	0.92

Panel C: *Censoring time correlated with the covariate*

		Inclusion probabilities		
	Traditional			Post-stratified
	Thomas (1977)	Samuelsen (1997)	GAM	10 intervals
Mean estimate	1.022	0.984	0.998	0.911
Mean variance	0.558	–	–	0.346
Mean robust variance	–	0.437	0.437	0.457
Empirical variance	0.581	0.416	0.348	0.363
Relative efficiency	0.47	0.65	0.78	0.75

but these estimates were somewhat too small with 5 intervals.

The traditional nested case-control estimator is quite inefficient in this situation with efficiency comparable to Panel A. This is not surprising as no information about the relation between covariate and censoring is used for this estimator. The estimator based on the inclusion probability of Samuelsen (1997) provides a great improvement from the traditional nested case-control estimator, but is still far from efficient. However, it had little bias. The robust variance estimator exceeded the empirical variance somewhat, which is in accordance with the variance expression in Samuelsen (1997).

The interval lengths for the post-stratified estimators were of fixed and equal length which does not preclude that no controls could be sampled in one or more strata. With 5 strata this happened only in 5 out of 5000 times, but with 10 strata it occurred in more than 20% of the replications. To avoid this we also, similar to Chen (2001), considered estimators where the intervals were determined by having an equal number of controls sampled in each. This, however, produced very biased results with for instance an average estimate of 1.72 with 5 intervals. The bias decreased with more intervals, but the average estimate was as high as 1.21 even with 40 intervals. Thus, choice of intervals may be a problem for the post-stratification approach at least in some extreme situations. It is noteworthy that use of both the GAM-weights and the inclusion probability weights of Samuelsen (1997) produced almost unbiased estimates.

Additional simulations with $n = 10000$ were conducted. In this case the post-stratified estimator with 10 intervals was practically unbiased as were the traditional nested case-control estimator and the estimator based on the method of Samuelsen (1997). The estimator based on GAM-weights showed a very slight bias (mean estimate 1.023) as did the post-stratified estimator with 5 intervals (mean estimate 1.039). The post-stratified estimators with intervals determined by having equal number of controls, however, were as biased as for $n = 1000$.

Variance estimates and empirical variances corresponded well for the post-stratified estimators with both 5 and 10 intervals and had relative empirical efficiencies of 0.95 and 0.98, respectively, compared to the cohort estimator. The GAM-estimator, had an efficiency of 1.00 compared to the cohort estimator. For the inclusion probability estimator (Samuelsen, 1997) and the traditional nested case-control estimator the efficiencies were as in Panel B of Table 3.

As for the case-cohort design we also performed simulations where the censoring time had a correlation of 0.9 with the covariate, using the same model as used for Panel C of Table 2. The results are presented in Panel C of Table 3.

Weights with post-stratification again produced clearly biased estimates and the variance estimates were generally too small compared to the empirical variances. The estimates based on the GAM-weights and the inclusion probability weights were much less biased. The GAM-weights produced a smaller variance than the other estimators with an efficiency of 0.79 compared to the cohort estimator.

5.3 Stratified case-cohort design

We have extended the post-stratification or local averaging method proposed by Chen (2001) to designs where sampling can depend on covariates. To study the potential benefits of our extension we simulated the same model and sampling scheme as in Section 2.5 with $n = 1000$. Thus, the surrogate variable was again an indicator for the U$[0, 1]$ covariate taking a value above 0.5. In addition to the surrogate we then post-stratified the data into 5 equal length censoring intervals, giving 10 strata in total, and to 10 intervals, giving 20 strata. Results are given in Panel A of Table 4.

The estimates appeared to be somewhat biased. However, in the similar simulation reported in Section 2.5, the bias for the original stratified case-cohort estimator was much smaller. The average estimated variance was in good agreement with the empirical variance for 5 intervals, but perhaps a bit too small for 10 intervals. The robust variance estimator was again markedly conservative.

The main observation from these simulations is that the variances are considerably reduced after post-stratification on censoring intervals when censoring and covariates are independent. This is in contrast to the effect of post-stratification in the usual case-cohort studies discussed in Section 5.1 where a rather strong dependence was required to demonstrate an efficiency improvement. The efficiency improvement for stratified case-cohort studies may be explained by inspecting the DFBETAS or the

$$X_i \;=\; \int_0^\tau [Z_i - \frac{\tilde{S}_{II}^{(1)}(\tilde{\beta}_{II}, s)}{\tilde{S}_{II}^{(0)}(\tilde{\beta}_{II}, s)}] Y_i(t) \exp(\tilde{\beta}'_{II} Z_i) \frac{dN_\bullet(s)}{\tilde{S}_{II}^{(0)}(\tilde{\beta}_{II}, s)}.$$

Within standard case-cohort studies these have an average over the controls close to zero unless covariates are strongly predictive of case-status. Taking averages within post-strata defined by length of follow-up then typically also produces values close to zero. In contrast, for a stratified case-cohort study the average of the X_i will differ from zero in the different strata, but the X_i will also depend on the length of follow-up. Taking averages over post-strata defined by both the original stratification variable and the length of follow-up will produce systematically different X_i in the post-strata and variation within these post-strata may be smaller than the variation within the original strata.

5.4 Counter-matched design

Counter-matching was described in Section 4.2 where the original estimator of Langholz & Borgan (1995) was presented. It was argued that we could alternatively use a post-stratification method with strata defined as censoring intervals for each level of the surrogate. Furthermore, it is possible to calculate an inclusion probability, similar to that of Samuelsen (1997), of individual i ever being sampled as a control. This is given by $p_i = 1 - \prod_s [1 - Y_i(s) m_l(s) \frac{dN(s)}{n_l(s) - 1}]$ when individual i belongs to stratum l, where $m_l(s) = m_l - 1$ if the case at time s comes from stratum l and where

$m_l(s) = m_l$ otherwise. An alternative estimator could be obtained by maximizing a weighted partial likelihood where cases are weighted by 1 and controls by $1/p_i$. Another option could be to smooth indicators of being sampled as controls against censoring times separately for each stratum using for instance GAMs.

To investigate the performance of such methods we performed a simulation study with the same model as used for Panel A of Table 3 with censoring independent of the covariate. In addition we used an indicator for the uniform [0,1] covariate taking a value above 0.5 as stratum variable. We obtained the cohort Cox-estimator, the traditional counter-matching estimator of Langholz & Borgan (1995), an estimator with inclusion probabilities similar to that of Samuelsen (1997) for both strata, an estimator with inclusion probabilities based on GAM and a post-stratified estimator with 5 equal length censoring intervals and 2 levels of surrogate variable (10 strata in total). Variance estimates were obtained for the traditional counter-matched estimator as the inverse of the information and for the post-stratified method using the correction method described in Section 2.4.

Contrary to all other simulation results reported in this paper, the traditional method clearly outperformed the post-stratified and all other estimators in this case. It should be noted that the traditional counter-matched estimator attained a very high efficiency of 0.85, higher than any other estimator based on simulations from such a model (see Table 1 and Panel A of Tables 2, 3 and 4). The estimator thus worked exceptionally well in this situation. The relative efficiencies for the other methods were roughly in accord with those for stratified case-cohort studies shown in Panel A of Table 3. However, although there was no efficiency improvement the estimators were only slightly biased and variance estimation seemed to work well.

In order to investigate whether the observed results were due to small samples the sample size was increased to $n = 10000$. The original estimator for counter-matching was clearly superior, still having a relative efficiency of 0.85. However, the other methods came somewhat closer with efficiencies of 0.71 for the inclusion probabilities of Samuelsen (1997), 0.77 for GAM probabilities and 0.74 for post-stratified inclusion probabilities. The bias practically vanished for all estimators, except for the scenario with 5 intervals where the average estimate was 0.83.

5.5 Bernoulli sampling design

In Section 4.3 we argued that we could also invoke the post-stratification method if the subcohort was sampled with Bernoulli sampling. The standard approach to analyzing such data would be to weight by the inverse of the sampling fractions. Alternatively, the redefined weights after observing how many were sampled in each interval should give a closer correspondence to the cohort data and might thus produce more precise estimates.

To demonstrate this we performed a simulation similar to the one in Section 2.5, but with Bernoulli sampling in both strata. As previously, the strata were determined

Table 4: Results from simulations of stratified case-cohort studies (Panel A), counter-matched studies studies (Panel B), and Bernoulli-sampling (Panel C). In all cases the censoring time is independent of the covariate.

Panel A: *Stratified case-cohort studies*

	Original stratification scheme	Post-stratified 5 intervals	Post-stratified 10 intervals
Mean estimate	1.040	1.038	1.043
Mean variance	0.198	0.159	0.157
Mean robust variance	0.251	0.263	0.282
Empirical variance	0.208	0.164	0.176
Relative efficiency	0.50	0.63	0.59

Panel B: *Counter-matched studies*

		Inclusion probabilities		
	Traditional counter-matching	Similar to Samuelsen (1997)	GAM	Post-stratified 10 intervals
Mean estimate	1.010	1.027	1.036	1.029
Mean variance	0.122	–	–	0.145
Mean robust variance	–	0.195	0.198	0.206
Empirical variance	0.123	0.157	0.147	0.151
Relative efficiency	0.85	0.67	0.71	0.69

Panel C: *Bernoulli sampling*

	Original sampling fraction	Corrected sampling fraction	Post-stratified scheme
Mean estimate	1.040	1.035	1.030
Mean variance	0.249	0.200	0.160
Mean robust variance	0.251	0.253	0.266
Empirical variance	0.277	0.212	0.167
Relative efficiency	0.37	0.48	0.61

by whether the uniform $[0, 1]$ covariate Z was above or below 0.5. The sampling fraction for the Bernoulli sampling was 0.13 which was also the fixed sampling fraction for the stratified case-cohort studies.

Based on 5000 replicated datasets generated from this model we obtained the cohort Cox-estimator, the weighted Cox-estimators with the original weights of 0.13, the modified weights after observing how many censored individuals were actually sampled in each stratum and also estimators post-stratified both on stratum and censoring interval. Robust and adjusted variances were also recorded for all estimators. The results are given in Panel C of Table 4.

The variances based on the original sampling fractions were clearly larger than for the sampling weights corrected for stratum. An additional efficiency improvement was obtained after post-stratifying also on censoring interval. Indeed, the behavior of the post-stratified estimators with Bernoulli sampling in Panel C of Table 4 is in very good agreement with the post-stratified estimators for stratified case-cohort sampling in Panel A of the same table. The robust variances in Panel C were only valid when using the original sampling fractions.

6 Discussion

We have shown that proportional hazards models can easily be fitted for stratified case-cohort data by using standard Cox regression software accommodating inverse probability weighting. In particular, estimation of the covariance matrix for the regression coefficients can proceed based on the DFBETAS. Simulation studies indicated that such variance estimation performs well. In contrast, robust variance estimates can be markedly conservative for stratified case-cohort studies.

We have also pointed out a relation between post-stratification on censoring intervals and the local averaging weights of Chen (2001) for a general class of sampling designs. The use of stratified case-cohort methods to adjust variance estimates was investigated and such methods appeared to work well. However, for nested case-control studies the estimates were sometimes clearly biased. Some care should thus be exercised in choosing the censoring intervals that constitute the strata for post-stratification. It is interesting to note that the inclusion probabilities of Samuelsen (1997) or smoothed inclusion probabilities based on generalized additive models produced practically unbiased estimates.

Chen (2001) showed that his local averaging estimator was large sample efficient compared to other estimators. In our simulations we found clear efficiency improvements when censoring depended strongly on a covariate, but with independence there was little improvement. In small samples there may even be an efficiency reduction compared to traditional methods.

We have also shown that local averaging can be used for covariate (or surrogate variable) dependent sampling such as stratified case-cohort and counter-matched designs. For stratified case-cohort designs our simulations were very promising since we

obtained efficiency gains even when censoring and covariates were independent. For counter-matching, on the other hand, the method of Langholz & Borgan (1995) was found to be more efficient than post-stratification in our simulation. The generality of this result should be investigated.

Most of the methods discussed in this paper are based on maximizing weighted partial likelihoods. A merit of probability weighting is that it is a very general approach that can be used for a multitude of models, including parametric survival models (Kalbfleisch & Lawless, 1988, Samuelsen, 1997) and semi-parametric additive hazard models (Kulich & Lin, 2000). Furthermore, for competing risk models with nested case-control and counter-matched designs, controls sampled to cases of one type of event can only be used in relation to this type of event when using the traditional estimation techniques. In contrast, weighting makes it straightforward to use all sampled controls for all types of events just as for case-cohort studies. Also, for time-matched designs, the traditional methods do not allow the time-scale to be changed from the original scale (for instance age) to another scale (such as calendar time or time in study). With weighting techniques this does not pose a problem. Another advantage of weighting methods for nested case-control and counter-matched studies is that the efficiency loss due to missing covariates can be reduced. This is so because the traditional methods require that matched sets (both case and controls) with a missing covariate value must be excluded from the analysis. In contrast, weighting by inverse inclusion probabilities enables us to make use of all individuals with complete covariate information. Thus, although our simulations of counter-matched studies did not demonstrate an efficiency improvement from using weighted methods, this approach may still be useful in practice.

Recently semi-parametric maximum partial likelihood estimators for case-cohort studies (Scheike & Martinussen, 2004), for stratified case-cohort studies (Kulich & Lin, 2004) and nested case-control studies (Scheike, & Juul , 2004) have been developed. These methods may sometimes perform better than our approach based on inverse probability weighting. However, the estimators suggested in this paper generally perform very well and model fitting and variance estimation is very easy to carry out using standard software.

Acknowledgement: Sven Ove Samuelsen would like to thank the Center for Advanced Study, Oslo for providing excellent research facilities during his participation in the Research Group on Statistical Analysis of Complex Event History Analysis in the autumn of 2005.

References

[1] Andersen P.K., Borgan Ø., Gill R.D. and Keiding N. (1993). *Statistical models based on counting processes*. Springer Verlag, New York.

[2] Andersen P.K. and Gill R.D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100-1120.

[3] Barlow W.E. (1994). Robust variance estimation for the case-cohort design. *Biometrics* **50**, 1064-1072.

[4] Borgan Ø., Langholz B. and Goldstein L. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.* **23**, 1749-1778.

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

[6] Chen K.N. and Lo S.H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika* **86**, 755-764.

[7] Chen K.N. (2001). Generalized case-cohort sampling. *J. Roy. Statist. Soc. Ser. B* **63**, 791-809.

[8] Cochran W.G. (1977). *Sampling techniques (3rd edition)*. Wiley, New York.

[9] Cox D.R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **74**, 187-200.

[10] De Roos A.J., Ray R.M., Gao D.L., Wernli K.J., Fitzgibbons E.D., Ziding F., Astrakianakis G., Thomas D.B. and Checkoway H. (2005). Colorectal cancer incidence among female textile workers in Shanghai, China: a case-cohort analysis of occupational exposures. *Cancer, Causes and Control* **16**, 1177-1188.

[11] Goldstein L. and Langholz B. (1993). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* **20**, 1903-1928.

[12] Hastie T.J. and Tibshirani R.J. (1990). *Generalized additive models.* Chapman & Hall, London.

[13] Hisada M., Chatterjee N., Kalaylioglu Z., Battjes R.J. and Goedert J.J. (2005). Hepatitis C virus load and survival among injecting drug users in the United States. *Hepatology* **42**, 1446-1452.

[14] Kalbfleisch J.D. and Lawless J.F. (1988). Likelihood analysis of multistate models for disease incidence and mortality. *Statist. Med.* **7**, 149-160.

[15] Kulich M. and Lin D.Y. (2000). Additive hazards regression with covariate measurement error. *J. Amer. Statist. Assoc.* **95**, 238-248.

[16] Kulich M. and Lin D.Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Amer. Statist. Assoc.* **99**, 832-844.

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

[18] Li W., Ray R.M., Gao D.L., Fitzgibbons E.D., Seixas N.S., Camp J.E., Wernli K.J., Astrakianakis G., Feng Z., Thomas D.B. and Checkoway H. (2006). Occupational risk factors for nasopharyngeal cancer among female textile workers in Shanghai, China. *Occup. and Environ. Med.* **63**, 39-44.

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

[20] Nan B. (2004). Efficient estimation for case-cohort studies. *Canad. J. Statist.* **32**, 403-419.

[21] Oakes D. (1981). Survival analysis: aspects of partial likelihood (with discussion). *Int. Statist. Rev.* **49**, 235-64.

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

[23] Robins J.M., Rotnitzky A. and Zhao L.P. (1995). Estimation of regression-coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

[24] Samuelsen S.O. (1989). *Two incomplete data problems in event history analysis: Double censoring and the case-cohort design.* PhD-Dissertation, University of Oslo.

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

[26] Scheike T.H. and Juul A. (2004). Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics* **5**, 193-206.

[27] Scheike T.H. and Martinussen T. (2004). Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scand. J. Statist.* **31**, 283-293.

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

[29] Therneau T.M. and Grambsch P.M. (2000). *Modeling survival data. Extending the Cox model*. Springer Verlag, New York.

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

[31] 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. R. Statist. Soc. Ser. A* **140**, 469-491.

[32] Wacholder S., Gail M.H., Pee D. and Brookmeyer R. (1989). Alternative variance and efficiency calculations for the case-cohort design. *Biometrika* **76**, 117-123.