The Efficiency of Simple and Counter-matched Nested Case-control Sampling

ØRNULF BORGAN

University of Oslo

ESPEN F. OLSEN

Kommunal Landspensjonskasse, Oslo

ABSTRACT. This paper presents a study of the performance of simple and counter-matched nested case-control sampling relative to a full cohort study. First we review methods for estimating the regression parameters and the integrated baseline hazard for Cox's proportional hazards model from cohort and case-control data. Then the asymptotic distributional properties of these estimators are recapitulated, and relative efficiency results are presented both for regression and baseline hazard estimation.

Key words: asymptotic relative efficiency, baseline hazard estimation, counter-matching, Cox regression, nested case-control studies, risk set sampling, survival analysis.

1. Introduction

In Cox regression (Cox, 1972) one compares the covariate value of a failing individual to the covariate values of all the individuals at risk at the time of the failure. In large epidemiological cohort studies of a rare disease, Cox regression requires collection of covariate information on all individuals in the cohort even though only a small fraction of these actually get diseased. Nested case-control studies, in which each failing individual, or case, is compared to a small sample of controls selected from those at risk at the time of the failure, may therefore give a substantial reduction in the resources that need to be allocated to a study. Furthermore, as most of the information is contained in the cases, such a study may still be sufficient to give reliable answers to the questions of interest.

In the classical form of such case-control studies nested within a cohort, the controls are sampled randomly from the risk sets (Thomas, 1977; Oakes, 1981). Often some information is available for all cohort members, e.g. for an occupational cohort a surrogate measure of the exposure to some hazardous substance, like type of work or duration of employment, may be available for everyone. Inspired by work on “two-stage” designs for unmatched case-control studies (e.g. Breslow & Cain, 1988), Langholz & Borgan (1995) developed a stratified version of the simple nested case-control design which makes it possible to incorporate such information into the sampling process in order to obtain a more informative sample of controls. This design, called counter-matching, may also be useful when exposure information is available for everyone, while information on a possible confounder has to be collected for the cases and their controls. For a long time it was a common belief that the integrated baseline hazard could not be estimated from nested case-control data. Borgan & Langholz (1993) showed, however, how this may be done provided one has information on the numbers at risk at the times of the failures. A thorough study of the asymptotic properties of the estimators of the regression parameters and the integrated baseline hazard from sampled risk set data, also including other sampling schemes for the controls than simple and stratified random sampling, was provided by Borgan et al. (1995).
The purpose of the present paper is to investigate the efficiency of the simple and counter-
matched nested case-control designs. We do this by studying the asymptotic efficiencies for
estimating the regression parameters and the integrated baseline hazard based on case-control
data relative to estimation based on full cohort data. For the regression parameters our
asymptotic calculations supplement those of Goldstein & Langholz (1992), Breslow et al.
(1983) and Langholz & Borgan (1995), while the relative efficiency of the case-control designs
for baseline hazard estimation has not been studied earlier. The plan of the paper is as follows.
In section 2 we describe the Cox model and review how regression parameters and integrated
baseline hazard may be estimated from cohort as well as case-control data. A summary of the
asymptotic properties of the estimators are provided in section 3, while in section 4 these results
are specialized to the model we consider in our efficiency investigations in sections 5 and 6.
Some concluding remarks are given in section 7.

2. Estimation in the Cox model

In this section we review how the regression parameters and the integrated baseline hazard
in the Cox model may be estimated from cohort data (e.g. Andersen et al., 1993, sect.
VII.2.1), as well as from simple and counter-matched case-control studies nested within a
cohort (Borgan & Langholz, 1993; Borgan et al., 1995; Langholz & Borgan, 1995).

We start out by describing the Cox model and the data available in a cohort study. Let the
cohort under study consist of $n$ individuals, indexed by $i = 1, 2, \ldots, n$, and assume that the
hazard rate at time $t$ for the $i$th individual with covariates $Z_i(t) = (Z_{i1}(t), \ldots, Z_{ip}(t))^T$ takes the
form

$$
\alpha(t; Z_i) = \alpha_0(t) \exp(\beta^T Z_i(t)).
$$

Here $\alpha_0(t)$ is the baseline hazard, i.e. the hazard for an individual with $Z_i \equiv 0$, and
$\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of regression parameters measuring how the covariates
influence the hazard rate. We introduce $Y_i(t)$ for the indicator taking the value 1 if the $i$th
individual is at risk \textit{“just before”} time $t$ and the value zero otherwise, and let
$\mathcal{R}(t) = \{ i : Y_i(t) = 1 \}$ be the risk set at $t$. Further, let $t_1 < t_2 < \cdots$ denote the times when
failures are observed, and, assuming no tied failures, let $i_j$ be the index of the individual
failing at $t_j$. For the cohort study it is assumed that covariate information is collected for
all individuals at risk at the failure times, i.e. $Z_i(t_j)$ is registered for all $t_j$ and all
$i \in \mathcal{R}(t_j)$.

Let us then describe how one may perform a case-control study nested within the cohort. We
concentrate on the counter-matched design; the simple design being a special case. To this end,
assume that at each time point $t$, the risk set $\mathcal{R}(t)$ may be divided into $L$ (possibly time-
dependent) strata, and let $C_i(t) \in \{1, \ldots, L \}$ be the stratum indicator for individual $i$. The
stratification must be based on information available \textit{“just before”} time $t$. We introduce
$\mathcal{R}(t) = \{ i : Y_i(t) = 1, C_i(t) = l \}$; the subset of $\mathcal{R}(t)$ which belongs to stratum $l$, and let
$n_l(t) = |\mathcal{R}_l(t)|$ be the number at risk in this stratum at time $t$. We also introduce $n(t) = \sum n_l(t)$
for the total number of individuals at risk at $t$. Then, in a counter-matched study, at each failure
time $t_j$, $m_l$ controls are sampled randomly without replacement from stratum $l$, except for the
case’s stratum from which only $m_l - 1$ controls are sampled. The case is always included in the
sample so that there is a total of $m_l$ from each stratum $l$. We let $m = \sum m_l$, and denote by $\mathcal{R}(t_j)$
the sampled risk set consisting of the case $i_j$ and the $m - 1$ controls sampled at $t_j$. Covariate
information is collected for the cases and the controls at the relevant failure times, but are not
needed for the other individuals in the cohort. Thus we register $Z_i(t_j)$ for all $t_j$ and all
$i \in \mathcal{R}(t_j)$. The simple nested case-control study is the special case with only one stratum, i.e. $L = 1$. 

\textcopyright Board of the Foundation of the Scandinavian Journal of Statistics 1999.
The estimators based on cohort data and nested case-control data are of the same form. For both situations the vector of regression parameters is estimated by $\hat{\beta}$, the value of $\beta$ which maximizes a partial likelihood of the form

$$L(\beta) = \prod_{t_j} \frac{\exp\{\beta^T Z_i(t_j)\} w_i(t_j)}{\sum_{i \in \mathcal{R}(t_j)} \exp\{\beta^T Z_i(t_j)\} w_i(t_j)}.$$  \hfill (1)

Furthermore, the integrated baseline hazard $A_0(t) = \int_0^t a_0(u) \, du$ is estimated by

$$\hat{A}_0(t) = \sum_{t_j < t} \frac{1}{\sum_{i \in \mathcal{R}(t_j)} \exp\{\hat{\beta}^T Z_i(t_j)\} w_i(t_j)}.$$  \hfill (2)

The difference between the two situations is reflected in the choice of sets $\mathcal{R}(t_j)$ and weights $w_i(t_j)$ in (1) and (2). For cohort data, the $\mathcal{R}(t_j)$ are the risk sets $\mathcal{R}(t_j)$, and the weights $w_i(t_j)$ are all equal to one. For case-control data, the $\mathcal{R}(t_j)$ are the sampled risk sets $\tilde{\mathcal{R}}(t_j)$, and the weights are inversely proportional to the proportions of the strata which are in the sampled risk sets, i.e. $w_i(t_j) = n_i(t_j)/m_i$ if individual $i$ belongs to stratum $l$ at time $t_j$. Both for cohort and case-control data, the covariance matrix of $\hat{\beta}$ may be estimated by $\mathcal{J}(\hat{\beta})^{-1}$, where

$$\mathcal{J}(\beta) = -\frac{\partial^2}{\partial \beta^T} \log L(\beta)$$

is the observed information matrix. Further, the covariance between $\hat{A}_0(s)$ and $\hat{A}_0(t)$ may be estimated by

$$\hat{\sigma}^2(s, t) = \hat{\sigma}^2(s \wedge t) + \hat{B}(s)^T \mathcal{J}(\hat{\beta})^{-1} \hat{B}(t),$$  \hfill (3)

where

$$\hat{\sigma}^2(t) = \sum_{t_j < t} \left[ \left( \sum_{i \in \mathcal{R}(t_j)} \exp\{\hat{\beta}^T Z_i(t_j)\} w_i(t_j) \right)^2 \right]^{-1}$$

and

$$\hat{B}(s) = \sum_{t_j < s} \frac{Z_i(t_j) \exp\{\hat{\beta}^T Z_i(t_j)\} w_i(t_j)}{\sum_{i \in \mathcal{R}(t_j)} \exp\{\hat{\beta}^T Z_i(t_j)\} w_i(t_j)}.$$  \hfill (4)

Here the leading term on the right-hand side of (3) is due to the variability in estimating the integrated hazard while the second term accounts for the variability due to the estimation of $\hat{\beta}$.

### 3. Asymptotic properties

We then briefly review the asymptotic properties of the estimators described in the preceding section. The results for the full cohort are extracted from Andersen et al. (1993, sect. VII.2.2), while those for nested case-control data are taken from Borgan et al. (1995). In both cases the results are valid under the regularity conditions stated in these references.

We restrict ourselves to the i.i.d. case where $(Y_i(\cdot), Z_i(\cdot), C_i(\cdot)); i = 1, 2, \ldots, n$; are indepen-
dent copies of \((Y(\cdot), Z(\cdot), C(\cdot))\), and introduce \(p(t) = P(Y(t) = 1)\) and \(\tau = \inf\{t: p(s) = 0 \text{ for } s \geq t\}\). For both the cohort study and the case-control designs we have that
\[
\sqrt{n}(\hat{\beta} - \beta) \overset{d}{\rightarrow} \mathcal{N}(0, \Sigma^{-1}),
\]
and
\[
n^{-1}(\hat{\beta} - \beta) \overset{p}{\rightarrow} \Sigma
\]
as \(n \to \infty\). Moreover, the asymptotic information matrix \(\Sigma\) takes the form
\[
\Sigma = \int_0^t G(s)p(s)\alpha_0(s) \, ds.
\]
The difference between the two situations is reflected in the form of the \(p \times p\) matrix \(G(t)\). Explicit expressions are given in (10) and (14) below. Furthermore, for both types of data,
\[
\sqrt{n}(\tilde{A}_0(\cdot) - A_0(\cdot)) \overset{d}{\rightarrow} U(\cdot)
\]
on \(D[0, \tau]\), as \(n \to \infty\). Here \(U(\cdot)\) is a Gaussian process with covariance function of the form
\[
\sigma^2(s, t) = \omega^2(s \wedge t) + B(s)^T \Sigma^{-1} B(t),
\]
with
\[
\omega^2(t) = \int_0^t \{\gamma(s)/p(s)\} \alpha_0(s) \, ds
\]
and
\[
B(t) = \int_0^t e(s)\alpha_0(s) \, ds.
\]
Furthermore (6) may be uniformly consistently estimated by \(n\) times (3). Here cohort and case-control data give rise to different expressions for \(\gamma(s)\) and the \(p\)-dimensional vector \(e(t)\); cf. (11), (12), (15) and (16).

To give explicit expressions for \(G(t)\), \(\gamma(t)\) and \(e(t)\) for cohort data, introduce the covariate vector \(Z_Y(t)\) distributed as \(Z(t)\) conditional on \(Y(t) = 1\), and let \(\tilde{Z}_Y(t)\) be the random vector with distribution specified by requiring
\[
E[h(\tilde{Z}_Y(t))] = E\left(h(Z_Y(t)) \exp\{B^T Z_Y(t)\}/E[\exp\{B^T Z_Y(t)\}]\right),
\]
for all bounded measurable functions \(h\). Then
\[
G_{co}(t) = \text{cov}\{\tilde{Z}_Y(t)\}E[\exp\{B^T Z_Y(t)\}],
\]
\[
\gamma_{co}(t) = 1/E[\exp\{B^T Z_Y(t)\}],
\]
and
\[
e_{co}(t) = E\{\tilde{Z}_Y(t)\}.
\]
For counter-matched case-control data, we introduce \(p_i(t) = P\{C(t) = i| Y(t) = 1\}\), and let \(Z_{Y,k,i}(t), i = 1, \ldots, m_k, k = 1, \ldots, L\) be mutually independent covariate vectors with
Let $Z_{Y,l,j}(t)$ distributed as $Z(t)$ conditional on $Y(t) = 1$ and $C(t) = l$. Furthermore, given $Z_{Y,1,1}(t), \ldots, Z_{Y,L,m_L}(t)$, let

$$E(Z_y(t)|Z_{Y,1,1}(t), \ldots, Z_{Y,L,m_L}(t))$$

and

$$\text{cov}(Z_y(t)|Z_{Y,1,1}(t), \ldots, Z_{Y,L,m_L}(t))$$

be the vector of expectations and the covariance matrix, respectively, of the covariate vector $Z_y(t)$ one obtains by selecting one of the $Z_{Y,l,j}(t)$ with probability

$$\frac{\exp\{\beta^T Z_{Y,l,j}(t)\} p(t)/m_l}{\sum_{k=1}^L \sum_{j=1}^{m_k} \exp\{\beta^T Z_{Y,k,j}(t)\} p_k(t)/m_k}. \quad (13)$$

Then

$$G_{cc}(t) = E\left(\text{cov}(Z_y(t)|Z_{Y,1,1}(t), \ldots, Z_{Y,L,m_L}(t)) \sum_{l=1}^L \sum_{j=1}^{m_l} \frac{p_l(t)}{m_l} \exp\{\beta^T Z_{Y,l,j}(t)\}\right), \quad (14)$$

$$\gamma_{cc}(t) = E\left(\frac{1}{\sum_{l=1}^L \sum_{j=1}^{m_l} \frac{p_l(t)}{m_l} \exp\{\beta^T Z_{Y,l,j}(t)\}}\right) \quad (15)$$

and

$$e_{cc}(t) = EE\{Z_y(t)|Z_{Y,1,1}(t), \ldots, Z_{Y,L,m_L}(t)\}. \quad (16)$$

The simple nested case-control design is the special case with $L = 1$ and $p_1(t) \equiv 1$.

In the next section we describe how the above result specialize for a simple model with two binary covariates.

### 4. Calculation of asymptotic efficiencies

In sections 5 and 6 we present a large sample comparison of the efficiency of regression parameter and integrated baseline hazard estimation for the simple and counter-matched nested case-control designs relative to the full cohort study. Following Langholz & Borgan (1995), we consider two situations where counter-matched case-control sampling may be beneficial. In section 5 we consider the situation where sampling strata are based on a surrogate associated with a covariate measuring exposure to some risk factor, and true exposure measurements are collected for the cases and their controls. In section 6 strata are based on true exposure measurements, and information is collected for the sample about a possible confounding factor. Langholz & Borgan (1995) assumed that the distribution of the covariates for those at risk remains constant over time. We do not make such an approximating assumption, and in section 7 we discuss how our results for regression parameter estimation compares with those of Langholz & Borgan (1995).

As described in the succeeding sections, the two situations we consider are both covered by the following generic model. Assume an i.i.d. model for the full cohort with two binary covariates $Z_1$ and $Z_2$ with joint distribution $\pi_{ij} = P(Z_1 = i, Z_2 = j); i, j = 0, 1$. Given the covariates $Z = (Z_1, Z_2)^T$ failures occur according to the proportional hazards model

$$\alpha(t; Z) = \alpha_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2). \quad (17)$$
Failure times may be right censored by a censoring time independent of the covariates having distribution function $H$.

We assume that $Z_1$ is known for the entire cohort. For the cohort study we also observe $Z_2$ for all cohort members. For the case-control designs, however, information on $Z$ having distribution function $F$ is not available.

Failure times may be right censored by a censoring time independent of the covariates $X$.

For the full cohort the expressions for these quantities follow readily by (10)–(12) using that $H(t) = 1 - S(t)$, where $S(t)$ is the survival function. The other quantities $G(t)$, $J(t)$ and $E(t)$ entering into equations (5), (7) and (8) depend on the study design.

Let us indicate how the asymptotic (co)variances of $\hat{\beta}$ and $\hat{\gamma}(t)$ given by (5) and (6) may be evaluated for this model. To this end we first note that

$$p(t) = P(Y(t) = 1) = (1 - H(t)) \sum_{i=0}^{1} \sum_{j=0}^{1} S_0(t)^{\exp(i\beta_1 + j\beta_2)} \pi_{ij},$$

with $S_0(t) = \exp(-A_0(t))$ the baseline survival function. The other quantities $G(t)$, $J(t)$ and $E(t)$ entering into equations (5), (7) and (8) depend on the study design.

For the full cohort the expressions for these quantities follow readily by (10)–(12) using that $Z(t) = (Z_{Y,1}(t), Z_{Y,2}(t))^T$ has the distribution

$$\pi_{ij}(t) = P(Z_{Y,1}(t) = i, Z_{Y,2}(t) = j) = p(t)^{-1}(1 - H(t)) S_0(t)^{\exp(i\beta_1 + j\beta_2)} \pi_{ij};$$

with $i, j = 0, 1$; while the distribution of $Z(t) = (Z_{Y,1}(t), Z_{Y,2}(t))^T$ by (9) takes the form

$$P(Z_{Y,1}(t) = i, Z_{Y,2}(t) = j) = \frac{\pi_{ij}(t) \exp(i\beta_1 + j\beta_2)}{\sum_{k=0}^{1} \sum_{l=0}^{1} \pi_{kl}(t) \exp(k\beta_1 + l\beta_2)}.$$
\[
\gamma_{cc}(t) = \mathbb{E}\{1/D(t)\}, \tag{20}
\]
and
\[
e_{cc}(t) = \mathbb{E}\{D(t)\}, \tag{21}
\]
where \(C(t)\) is the \(2 \times 2\) matrix with elements \(C_{ik}(t), \bar{D}(t) = (\bar{D}_1(t), \bar{D}_1(t))^\top\), and the expectation is with respect to the distribution of \(J(t)\). The expected values in (19)–(21) may be evaluated noting that, for simple random sampling of the controls, \(J(t)\) has a multinomial \((\pi_{00}(t), \pi_{01}(t), \pi_{10}(t), \pi_{11}(t); m)\) distribution, while for counter-matched sampling \(J_{01}(t) = m_0 - J_{00}(t)\) and \(J_{11}(t) = m_1 - J_{10}(t)\) are independent and binomially distributed with parameters \((\pi_{01}(t)/\pi_0(t); m_0)\) and \((\pi_{11}(t)/\pi_1(t); m_1)\), respectively.

Finally, if we specify the baseline hazard \(a_0(t)\), the regression parameters \(\beta_1\) and \(\beta_2\), the censoring distribution \(H\), the covariate distribution \{\(\pi_y\)\}, and the sample sizes \(m_0\) and \(m_1\) (with \(m_0 + m_1 = m\)), we may use (5)–(8) and the above relations to evaluate the asymptotic (co)variances of \(\hat{\beta}\) and \(A_0(t)\) by numerical integration. We used Mathematica for the calculation reported in the succeeding sections.

5. Counter-matching on a surrogate measure of exposure

In this section we consider the situation where sampling strata are based on the surrogate \(Z_1\) which is available for all subjects in the cohort. This surrogate is associated with the true exposure \(Z_2\), which is only known for the cases and their sampled controls. It is assumed that the surrogate is uninformative after accounting for \(Z_2\), i.e. \(\beta_1 = 0\) in model (17). We will compare the properties of the estimators for \(\beta_2\) and \(A_0(t)\) for the various study designs by asymptotic calculations. In particular we will present asymptotic relative efficiencies of \(\hat{\beta}_2\) and \(A_0(t)\) computed as the ratio of the asymptotic variances for cohort data to the asymptotic variances for the case-control designs. Note that since we here assume \(\beta_1 = 0\), we in these calculations replace \(\Sigma\) given by (5) by its element (2, 2) and only use the last element of the vector (8). For the case-control designs we consider the situations with one and five controls, i.e. \(m = 2\) and \(m = 6\). In both cases we let \(m_0 = m_1\) so that we have an equal number of individuals in the two strata for the counter-matched design. Finally, we note that for the present situation, the covariate distribution \{\(\pi_y\)\} is most conveniently specified via \(1 - \alpha = \pi_{11}/\pi_1\) and \(1 - \beta = \pi_{00}/\pi_0\), the sensitivity and specificity of the surrogate measure for predicting the true exposure, together with \(\pi_1 = \pi_{01} + \pi_{11}\), the fraction exposed. We first present results for \(\hat{\beta}_2\) and thereafter for \(A_0(t)\).

5.1. Estimation of the regression coefficient

To begin with, we computed asymptotic efficiencies of \(\hat{\beta}_2\) for all the \(2^6 = 64\) combinations of the following factors (each on two levels) which may influence the performance of the estimator:

(a) sensitivity \(1 - \alpha\): 80% or 95%;
(b) specificity \(1 - \beta\): 80% or 95%;
(c) fraction exposed \(\pi_{11}\): 5% or 20%;
(d) baseline hazard \(a_0(t)\): exponential with parameter one or Weibull with parameters 3 and 1.28, i.e. \(a_0(t) = 1\) or \(a_0(t) = 3-1.28(1.28t)^{-3}\);
(e) relative risk \(\exp(\beta_2)\): 2.0 or 5.0 i.e. \(\hat{\beta}_2 = \log 2.0\) or \(\hat{\beta}_2 = \log 5.0\); and
(f) censoring distribution \(H\): uniform(0, 1) or uniform(0, 2).
For the simple nested case-control design the sensitivity and specificity have no impact on the efficiencies, so we only have $2^4 = 16$ factor combinations for this design. The baseline hazards are chosen such that they both correspond to a median lifetime of 0.69 for unexposed individuals (i.e. when $Z_2 = 0$). Thus the censoring distributions have supports which are approximately 1.5 and 3 times the median lifetime for baseline subjects.

It turned out that choice of baseline hazard and censoring distribution had very little impact on the results. For simple nested case-control sampling with $m = 2$ the asymptotic efficiencies for the other three combinations of baseline hazard and censoring distribution were always within 97.5% and 100.1% of those obtained with unit exponential baseline and uniform(0, 1) censoring distribution. For simple sampling with $m = 6$ and the counter-matched designs the differences were even smaller. Therefore, when discussing the effect of the other four factors, i.e. sensitivity, specificity, fraction exposed and relative risk, we may restrict our attention to the situation where the baseline is unit exponential and the censoring distribution is uniform(0, 1).

For this situation Table 1 presents the asymptotic efficiencies of $\hat{\beta}_2$ for simple and counter-matched case-control sampling relative to the full cohort for the various combinations of sensitivity, specificity, fraction exposed and relative risk. The highest relative efficiencies are obtained with a high fraction exposed ($\pi_1 = 0.20$) and a low relative risk ($\exp(\hat{\beta}_2) = 2.0$), while the situation with a small fraction exposed ($\pi_1 = 0.05$) and a high relative risk ($\exp(\hat{\beta}_2) = 5.0$) gives the lowest efficiencies. Furthermore a substantial gain is obtained by counter-matching on the surrogate measure for exposure. As noted by Langholz & Borgan (1995), the efficiency gain increases with sensitivity and specificity, but the increase is more marked with increasing specificity than with sensitivity.

A further illustration of the relative efficiency of $\hat{\beta}_2$ is given in Fig. 1. This presents the asymptotic relative efficiency of $\hat{\beta}_2$ as a function of the relative risk $\exp(\hat{\beta}_2)$ when the fraction exposed is 1%, 5% and 20%, sensitivity and specificity are both 80%, baseline hazard is unit exponential and censoring is uniform(0, 1). It is well known that for testing the association between a single exposure and failure, the efficiency of the simple nested case-control design is $(m - 1)/m$ when $m - 1$ controls are selected per case, and that this result holds independently

Table 1. Asymptotic relative efficiencies of $\hat{\beta}_2$ for the simple and counter-matched nested case-control designs when stratification for the counter-matched design is based on a surrogate measure of exposure$^a$

<table>
<thead>
<tr>
<th></th>
<th>$\exp(\hat{\beta}_2) = 2.0$</th>
<th>$\exp(\hat{\beta}_2) = 5.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_1 = 0.05$</td>
<td>$\pi_1 = 0.20$</td>
</tr>
<tr>
<td>(a) $m = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple sampling</td>
<td>0.347</td>
<td>0.388</td>
</tr>
<tr>
<td>Counter-matched</td>
<td>0.567</td>
<td>0.601</td>
</tr>
<tr>
<td>sampling</td>
<td>0.649</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>0.737</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td>0.866</td>
<td>0.879</td>
</tr>
<tr>
<td>(b) $m = 6$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple sampling</td>
<td>0.728</td>
<td>0.767</td>
</tr>
<tr>
<td>Counter-matched</td>
<td>0.826</td>
<td>0.852</td>
</tr>
<tr>
<td>sampling</td>
<td>0.868</td>
<td>0.888</td>
</tr>
<tr>
<td></td>
<td>0.884</td>
<td>0.902</td>
</tr>
<tr>
<td></td>
<td>0.948</td>
<td>0.956</td>
</tr>
</tbody>
</table>

$^a$For unit exponential baseline, uniform(0, 1) censoring and $m_0 = m_1$ for the counter-matched designs.
of censoring and covariate distributions (Breslow et al., 1983; Goldstein & Langholz, 1992). This explains why the curves for simple random sampling start out in 1/2 = 0.50 and 5/6 = 0.83, respectively, for m = 2 and m = 6. It is seen, however, that as the relative risk departs from unity, the efficiencies decrease, and that the decrease is larger for a rare risk factor than for a common one. Figure 1 further illustrates the gain one may obtain by counter-matching on a surrogate for exposure.

5.2. Estimation of integrated baseline hazard

Let us then consider the estimator $\hat{A}_0(t)$ for the integrated baseline hazard. We first studied the asymptotic relative efficiencies of the case-control designs relative to the full cohort for the $2^6 = 64$ factor-level combinations described at the beginning of section 5.1 at two points in time, $t_1$ and $t_2$, corresponding to survival probabilities of 90% and 50% for

Fig. 1. Asymptotic relative efficiencies of $\hat{\beta}_2$ as a function of relative risk for the nested case-control designs when stratification is based on a surrogate measure of exposure: fraction exposed 1% (---); fraction exposed 5% (-----) fraction exposed 20% (- - -). Sensitivity and specificity are 80%, baseline is unit exponential, censoring is uniform(0, 1), and $m_0 = m_1$ for the counter-matched designs.
baseline subjects. More precisely we let \( t_1 = 0.11 \) when the baseline is unit exponential, and \( t_1 = 0.37 \) when the baseline is Weibull with parameters 3 and 1.28. In both cases \( t_2 = 0.69 \). Also for baseline hazard estimation, the form of the baseline and the censoring distribution had little influence on the results. For both time-points and all case-control designs, the asymptotic efficiencies for the other three combinations of baseline hazard and censoring distribution were always within 98.7% and 101.3% of those obtained with unit exponential baseline and uniform(0, 1) censoring distribution. Again the largest differences were observed for simple random sampling of the controls with \( m = 2 \).

Table 2 presents the relative efficiencies of \( \hat{A}_0(t) \) for simple and counter-matched case-control sampling for the same combinations of the other factors as in Table 1. We see that all the efficiencies are quite high, in fact most of them are well above 90%. As in Table 1, the efficiencies take higher values for a relative risk of 2.0 than for one of 5.0. For baseline hazard estimation, however, the efficiencies decrease when a larger fraction of the cohort is exposed, while the opposite was the case for regression parameter estimation. Finally, counter-matching on the surrogate improves the efficiencies for baseline hazard estimation in a similar way as was the case in Table 1. But as even the efficiencies of the simple case-control design are quite high, this improvement is of less importance.

For a relative risk of 2.0, the efficiencies of \( \hat{A}_0(t_2) \) were almost the same as those at time \( t_1 \) shown in Table 2, the largest difference being 0.6% for simple random sampling with \( m = 2 \) and \( \pi_1 = 0.20 \). However, when the relative risk increased to 5.0, the efficiencies of \( \hat{A}_0(t_2) \) did not decrease as in Table 2, in fact they were all slightly larger than the corresponding ones for a relative risk of 2.0. This phenomenon is further illustrated in Fig. 2, which shows the relative efficiency of \( \hat{A}_0(t) \) as a function of time when one control is selected per case. When the relative risk is 2.0 the efficiencies have a bathtub shape, while for a relative risk of 5.0 they are increasing with time.

### 6. Counter-matching on exposure

In this section we assume that \( Z_1 \) is an exposure variable which is known for the full cohort, while \( Z_2 \) is a confounding factor to be collected for the sampled subjects. For the

<table>
<thead>
<tr>
<th>( m = 2 )</th>
<th>( 1 - \alpha )</th>
<th>( 1 - \beta )</th>
<th>( \exp(\beta_2) = 2.0 )</th>
<th>( \exp(\beta_2) = 5.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple sampling</td>
<td>( 0.80 )</td>
<td>( 0.95 )</td>
<td>0.975</td>
<td>0.908</td>
</tr>
<tr>
<td>Counter-matched sampling</td>
<td>( 0.95 )</td>
<td>( 0.95 )</td>
<td>0.990</td>
<td>0.964</td>
</tr>
<tr>
<td>( \pi_1 = 0.05 )</td>
<td>( \pi_1 = 0.20 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple sampling</td>
<td>( 0.80 )</td>
<td>( 0.80 )</td>
<td>0.934</td>
<td>0.908</td>
</tr>
<tr>
<td>Counter-matched sampling</td>
<td>( 0.95 )</td>
<td>( 0.95 )</td>
<td>0.990</td>
<td>0.964</td>
</tr>
<tr>
<td>( \pi_1 = 0.05 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \pi_1 = 0.20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( a \) See footnote to Table 1. Time \( t_1 = 0.11 \) corresponds to 90% survival probability for baseline subjects.
counter-matched designs stratification is based on the true exposure $Z_1$. We assume that both exposure and confounder influence the risk of failure, so the full model (17) applies. Usually, for such a situation, $\beta_1$ is the parameter of main importance. This measures the effect of the exposure after controlling for the effect of the confounding factor. One may, however, also be interested in the effect of the confounder measured by $\beta_2$ and the integrated baseline hazard $A_0(t)$. Below we will compare asymptotic relative efficiencies of estimators for these parameters for the case-control designs relative to the full cohort.

For the case-control studies, we consider as earlier the situations with one and five controls per case and with an equal number of individuals in the two strata for the counter-matched design. It is here convenient to specify the covariate distribution $\{\pi_{ij}\}$ by the odds ratio $\theta = (\pi_{11}\pi_{00})/(\pi_{10}\pi_{01})$ together with $\pi_{11} = \pi_{10} + \pi_{11}$ and $\pi_{01} = \pi_{00} + \pi_{11}$, the fraction exposed and the prevalence of the confounding factor.

We first present asymptotic efficiencies for estimating the regression parameters and thereafter results for the baseline hazard.

Fig. 2. Asymptotic relative efficiencies of $\hat{A}_0(t)$ as a function of time $t$ for the nested case-control designs with one control per case when stratification is based on a surrogate measure of exposure: fraction exposed 1% (---); fraction exposed 5% (—) fraction exposed 20% (--.--). Relative risk is $\exp(\beta_2) = 2.0$ in the upper panel and $\exp(\beta_2) = 5.0$ in the lower panel. Sensitivity and specificity are 80%, baseline is unit exponential, and censoring is uniform(0, 1)
6.1. Estimation of the regression coefficients

We first computed asymptotic relative efficiencies of $\hat{\beta}_1$ and $\hat{\beta}_2$ for the combinations of the fraction exposed (here $\pi_1$), baseline hazard, relative risk of exposure (here $\exp(\beta_1)$), and censoring distribution described in section 5.1. Throughout we assumed that the odds ratio is $\theta = 2.0$, the prevalence of the confounder is $\pi_1 = 0.30$, and its relative risk is $\exp(\beta_2) = 2.5$. The results for unit exponential baseline and uniform(0, 1) censoring are shown in Table 3. For simple random sampling of the controls, the efficiencies for $\hat{\beta}_1$ are similar to, but slightly smaller than those for the one-parameter case shown in Table 1. Also the values obtained for $\hat{\beta}_2$ are of the same order of magnitude as the efficiencies 0.377 and 0.764 one obtains with one and five controls per case, respectively, in the one parameter situation when the relative risk is $\exp(\beta_2) = 2.5$ and $\pi_1 = 0.30$. Counter-matching ensures an equal number of exposed ($Z_1 = 1$) and non-exposed ($Z_1 = 0$) subjects in each sampled risk set. This substantially increases the efficiencies of $\hat{\beta}_1$, which are above 80% even when only one control is collected per case. Note that, in contrast to what is the case for simple sampling, the efficiencies for the counter-matched designs are increasing with the relative risk of exposure. The price paid for the efficiency gain for $\hat{\beta}_1$, usually the estimator of prime importance, is a reduced efficiency for estimating $\hat{\beta}_2$, the effect of the confounding factor, in particular when there are few exposed individuals in the cohort ($\pi_1 = 0.05$).

The choice of baseline hazard and censoring distribution had little impact on the relative efficiency of $\hat{\beta}_1$. For $m = 2$ the asymptotic efficiencies for the other three combinations of baseline hazard and censoring distribution were always within 99.0% and 103.3% of those obtained with unit exponential baseline and uniform(0, 1) censoring. For $m = 6$ the differences were even smaller. For simple sampling and counter-matching with $m = 6$, also the efficiencies of $\hat{\beta}_2$ did not vary much with choice of baseline and censoring distribution; the three other combinations giving values between 94.7% and 100.3% of those reported in Table 3. However, for the counter-matched design with only one control per case, the efficiencies of $\hat{\beta}_2$ were substantially lower when censoring was uniform(0, 2). This reduction in efficiency was most pronounced for $\exp(\beta_1) = 5.0$ and $\pi_1 = 0.05$, were the efficiencies were about 80% of those

---

Table 3. Asymptotic relative efficiencies of the estimators of the regression parameters for the simple and counter-matched nested case-control designs when stratification for the counter-matched design is based on true exposure and information is collected on a confounder. The odds ratio is $2.0$, the prevalence of the confounding factor is $30\%$ and its relative risk is $\exp(\beta_2) = 2.5^a$

<table>
<thead>
<tr>
<th></th>
<th>$\exp(\beta_1) = 2.0$</th>
<th>$\exp(\beta_1) = 5.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_1 = 0.05$</td>
<td>$\pi_1 = 0.20$</td>
</tr>
<tr>
<td>(a) $m = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple sampling</td>
<td>$\hat{\beta}_1$</td>
<td>0.304</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_2$</td>
<td>0.381</td>
</tr>
<tr>
<td>Counter-matched</td>
<td>$\hat{\beta}_1$</td>
<td>0.839</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_2$</td>
<td>0.117</td>
</tr>
<tr>
<td>(b) $m = 6$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple sampling</td>
<td>$\hat{\beta}_1$</td>
<td>0.679</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_2$</td>
<td>0.768</td>
</tr>
<tr>
<td>Counter-matched</td>
<td>$\hat{\beta}_1$</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_2$</td>
<td>0.621</td>
</tr>
</tbody>
</table>

$^a$See footnote to Table 1.
reported in Table 3. The reason for this is that when censoring is reduced more individuals will fail, and with a high relative risk more of these will be exposed individuals. The result is a lower percentage exposed among those still at risk, and hence a lower efficiency of $\hat{\beta}_2$.

Figure 3 shows how the efficiencies of the estimators for the regression coefficients vary as a function of relative risk of exposure $\exp(\beta_1)$ when the fraction exposed is 1%, 5% and 20%, and one control is selected per case. The values shown are for unit exponential baseline and uniform(0, 1) censoring. The figure further illustrates how the efficiencies of $\hat{\beta}_1$ decrease with relative risk for simple sampling of the controls while the opposite is the case for the counter-matched design. Note also that counter-matching with one control per case gives very low efficiencies for $\hat{\beta}_2$ when only 1% of the cohort is exposed.

---

**Fig. 3.** Asymptotic relative efficiencies of the estimators for the regression parameters as a function of the relative risk of exposure for the nested case-control designs with one control per case when stratification for the counter-matched design is based on true exposure and information is collected on a confounder: fraction exposed 1% (…); fraction exposed 5% (——) fraction exposed 20% (---). The efficiencies for $\hat{\beta}_1$ are given in the upper panel and those for $\hat{\beta}_2$ in the lower panel. The odds ratio is 2.0, the prevalence of the confounding factor is 30% and its relative risk is $\exp(\beta_2) = 2.5$. Further baseline is unit exponential and censoring is uniform(0, 1)
6.2. Estimation of integrated baseline hazard

The asymptotic relative efficiencies of the integrated baseline hazard estimator are given in Table 4 for the two points in time $t_1$ and $t_2$ described in section 5.2. The combination of the other factors are as for Table 3. The efficiencies are slightly higher at $t_2$ than at $t_1$. More importantly, all are substantially lower than the efficiencies of section 5.1 for the one-parameter case, so we have to pay a price for estimating one extra parameter. When exposure is rare ($\pi_1 = 0.05$), simple random sampling performs better than counter-matching, while the opposite is the case when exposure is more common ($\pi_1 = 0.20$). Thus for a rare exposure, the improved estimation of $\beta_1$ obtained by counter-matching is more than outweighed by the poorer estimation of $\beta_2$, while this is not the case when exposure is more common. This is further illustrated in Fig. 4 which shows the relative efficiency of the integrated baseline hazard estimator as a function of time when the fraction exposed is 1%, 5%, and 20%. (Note that the scale for the counter-matched design with $m = 2$ differs from the others.) For simple sampling the efficiencies increase as exposure becomes more rare, as was the case for the one-parameter situation of Fig. 2. For counter-matching, however, the efficiencies decrease as the fraction exposed is reduced, and they become very low when one control is selected per case and only 1% of the cohort is exposed.

The efficiencies of $\hat{A}_0(t_1)$ and $\hat{A}_0(t_2)$ did not vary much with choice of baseline and censoring distribution for simple sampling and counter-matching with $m = 6$. The three other combinations all gave efficiencies between 98.3% and 100.1% of those reported in Table 4. With one control per case and uniform(0,2) censoring, however, the efficiencies for the counter-matched design were less than 90% of those of Table 4 when the relative risk is 5.0 and only 5% of the cohort is exposed. This is due to the reduced efficiency of $\hat{\beta}_2$ reported in section 5.1.

7. Discussion

In their study of the relative efficiency of regression parameter estimation for the counter-matched design, Langholz & Borgan (1995) assumed that the covariate distribution among

<table>
<thead>
<tr>
<th>Table 4. Asymptotic relative efficiencies of $\hat{A}_0(t_1)$ and $\hat{A}_0(t_2)$ for the simple and counter-matched nested case-control designs when stratification for the counter-matched design is based on true exposure and information is collected on a confounder. The odds ratio is 2.0, the prevalence of the confounding factor is 30% and its relative risk is $\exp(\beta_2) = 2.5$. The times $t_1 = 0.11$ and $t_2 = 0.69$ correspond to 90% and 50% survival probability, respectively, for baseline subjects$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\exp(\beta_1) = 2.0$</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>$\pi_1 = 0.05$</td>
</tr>
<tr>
<td>$\pi_1 = 0.20$</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>$\pi_1 = 0.05$</td>
</tr>
<tr>
<td>$\pi_1 = 0.20$</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>(a) $m = 2$</td>
</tr>
<tr>
<td>Simple sampling</td>
</tr>
<tr>
<td>$\hat{A}_0(t_1)$</td>
</tr>
<tr>
<td>$\hat{A}_0(t_2)$</td>
</tr>
<tr>
<td>Counter-matched sampling</td>
</tr>
<tr>
<td>$\hat{A}_0(t_1)$</td>
</tr>
<tr>
<td>$\hat{A}_0(t_2)$</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>(b) $m = 6$</td>
</tr>
<tr>
<td>Simple sampling</td>
</tr>
<tr>
<td>$\hat{A}_0(t_1)$</td>
</tr>
<tr>
<td>$\hat{A}_0(t_2)$</td>
</tr>
<tr>
<td>Counter-matched sampling</td>
</tr>
<tr>
<td>$\hat{A}_0(t_1)$</td>
</tr>
<tr>
<td>$\hat{A}_0(t_2)$</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
</tbody>
</table>

$^a$See footnote to Table 1.
those at risk remains constant over time. This assumption greatly simplifies the calculations. For then, one does not need to evaluate the integral (5) numerically for cohort and case-control data, but can base the efficiency calculations directly on the matrices $G_{co}$ and $G_{cc}$, which will not depend on $t$; cf. (10) and (14). A similar simplification is, however, not possible for the efficiency calculations for the baseline hazard estimator; cf. (6)–(8).

Typically the relative efficiencies for the regression parameter estimators, obtained by assuming the $\tau_0(t)$ in (18) independent of $t$, are close to those reported in sections 5.1 and 6.1. This should be no surprise, since for most situations considered in this paper, the efficiencies were not much influenced by the choice of baseline hazard and censoring distribution. For the situation of Table 1, the only differences of some importance are found for a common exposure ($\tau_1 = 0.20$) with a high relative risk (exp($\beta_2$) = 5.0). Here the relative efficiencies for simple

---

**Fig. 4.** Asymptotic relative efficiencies of $\hat{A}_0(t)$ as a function of time $t$ for the nested case-control designs when stratification for the counter-matched designs is based on true exposure and information is collected on a confounder: fraction exposed 1% (···); fraction exposed 5% (— —); fraction exposed 20% ( - - -). The efficiencies for $m = 2$ are given in the upper panel and those for $m = 6$ in the lower panel. (Note that the scale for the counter-matched design with $m = 2$ differs from the others.) The odds ratio is 2.0, the prevalence of the confounding factor is 30% and its relative risk is exp($\beta_2$) = 2.5. Further baseline is unit exponential, censoring is uniform(0, 1), and $m_0 = m_1$ for the counter-matched designs.
sampling and counter-matched sampling with specificity 80%, assuming a constant distribution of the covariates over time, are about 5 percentage points larger than those reported in Table 1. For simple sampling, a similar pattern is found for \( \hat{\beta}_1 \) for the situation of Table 3, while the efficiencies for estimating \( \beta_2 \) remain almost unchanged. For counter-matched sampling, however, the differences are more substantial. For \( m = 2 \) the relative efficiencies for \( \beta_2 \), assuming the \( \tau_{ij}(t) \) independent of \( t \), are 5–15 percentage points higher than those reported in Table 3, the largest difference being found for a rare exposure (\( \tau_1 = 0.05 \)) with a high relative risk (\( \exp(\beta_1) = 5.0 \)). This increase is, however, accompanied by a 3–8 percentage points decrease in the efficiencies of \( \hat{\beta}_1 \).

Our study of the asymptotic relative efficiencies of the case-control designs is based on a simple model with two binary covariates (section 4). This makes it feasible to use the analytic expressions for the asymptotic variances in our calculations. Further studies are needed to show whether results similar to ours are valid for continuous covariates and/or in situations with many explanatory variables. Such studies will have to be performed by means of stochastic simulations.

An important result of our study is that the efficiency of the baseline hazard estimator for the case-control designs is quite high, typically 80% or more, relative to the full cohort estimator. This explains why it is possible to combine the baseline hazard estimator with the regression parameter estimators to provide quite reliable estimators of absolute risk for individuals with certain covariate values (Langholz & Borgan, 1997). The baseline hazard estimator performs badly, however, for the counter-matched design when stratification is based on a rare exposure and information is collected on a confounder (section 6.2). For such a situation, if absolute risk estimation is an issue, the counter-matched design is less attractive. A useful alternative may then be to combine the simple and counter-matched designs by sampling some controls by simple random sampling and others by stratified random sampling (Langholz & Goldstein, 1996).

Acknowledgements

Part of the work of Ørnulf Borgan was done during a sabbatical leave at the Department of Statistics, Stanford University, the academic year 1997/98. The department is acknowledged for its hospitality and for providing the best working facilities during this year, while The Norwegian Research Council, Johan and Mimi Wessmann’s foundation and the Norwegian Association of Professional Writers and Translators are acknowledged for their economic support which made this stay possible.

References


*Received December 1997, in final form December 1998*

Ørnulf Borgan, Department of Mathematics, P.O. Box 1053 Blindern, N-0316 Oslo, Norway.