Lifetime Data Anal (2012) 18:261–283
DOI 10.1007/s10985-012-9214-8

Comparison of estimators in nested case–control studies with multiple outcomes

Nathalie C. Støer · Sven Ove Samuelsen

Received: 29 April 2011 / Accepted: 30 January 2012 / Published online: 2 March 2012

Abstract Reuse of controls in a nested case–control (NCC) study has not been considered feasible since the controls are matched to their respective cases. However, in the last decade or so, methods have been developed that break the matching and allow for analyses where the controls are no longer tied to their cases. These methods can be divided into two groups; weighted partial likelihood (WPL) methods and full maximum likelihood methods. The weights in the WPL can be estimated in different ways and four estimation procedures are discussed. In addition, we address modifications needed to accommodate left truncation. A full likelihood approach is also presented and we suggest an aggregation technique to decrease the computation time. Furthermore, we generalize calibration for case-cohort designs to NCC studies. We consider a competing risks situation and compare WPL, full likelihood and calibration through simulations and analyses on a real data example.

Keywords Nested case–control · Competing risks · Weighted partial likelihood · Maximum likelihood for nested case–control · Calibration

1 Introduction

When analyzing infrequent events in a cohort study, the cohort may need to be very large to obtain reliable estimates. Retrospective sampling designs like nested case–control (NCC) (Thomas 1977) or case-cohort (CC) (Prentice 1986) can then be very useful alternatives. In both designs, some or all covariates are obtained only for a subset of the original cohort. In the CC design, this subset usually includes all cases and a

N. C. Støer (✉) · S. O. Samuelsen
Department of Mathematics, University of Oslo, P.O. Box 1053, 0316 Oslo, Norway
e-mail: nathalcs@math.uio.no

S. O. Samuelsen
e-mail: osamuels@math.uio.no

subcohort sampled at the outset of the study. In a NCC design, the subset again usually includes all cases, but at each event time m controls are sampled from those at risk at that time. In a prospective cohort study, on the other hand, covariates are collected for every subject in the cohort. NCC and CC are therefor useful alternatives when some covariates, for instance analyzed blood samples, are too expensive to obtain for a large group of people. Moreover, because covariates only need to be obtained for cases and controls more effort can be used to ensure high quality of the collected data.

Traditionally, the risk sets in a NCC design only contain the case together with the time-matched controls, on the other hand in a CC design, controls contribute to a number of risk sets while the cases either are included at their event times (Prentice 1986) or whenever they are at risk (Kalbfleisch and Lawless 1988). A main advantage of the CC design has been the possibility to reuse controls for other endpoints. This has not been feasible with the NCC design since the controls are matched to their respective cases. However, in the last decade or so, methods have been developed that allow for analyses where the controls are no longer tied to their cases (Samuelsen 1997; Chen 2001; Scheike and Juul 2004; Saarela et al. 2008; Salim et al. 2009). This opens up for the possibility to reuse controls also within the NCC design.

Two main strategies have been proposed for reusing controls in a NCC design. One method is using weighted partial likelihoods (WPLs) (Samuelsen 1997; Chen 2001). Another is full likelihood approaches (Scheike and Juul 2004; Saarela et al. 2008), where the whole cohort is used and information not obtained for individuals outside the case–control set is treated as missing. A full likelihood is computationally demanding, but we will suggest an aggregation technique that reduces the complexity.

Reusing controls within the NCC design can be useful in many event-history situations with general types of endpoints. We will only consider a competing risks setting where the cause specific hazards are given by Cox's proportional hazards model. In particular, we will consider competing risks settings with one relatively common endpoint and one much less common endpoint. We are then able to use controls from the common endpoint as additional controls for the less common endpoint, and this can potentially give large efficiency improvements.

Recently, a method of finding more efficient weights in CC designs has been proposed (Breslow et al. 2009a,b), referred to as calibration of weights. The idea is to adjust the weights in such a way that known totals of some auxiliary variables are estimated exactly, but at the same time the weights stay as close to the population based weights as possible. We will suggest a generalization of this approach to the NCC design and investigate if it can improve the efficiency of the WPL approach further.

The aim of this article is mainly to compare WPL with different weighing schemes and Saarela's full likelihood in the sense of bias and efficiency through simulations. In Sect. 2 we look at WPL with four different weighing schemes and Saarela's full likelihood. How to deal with left truncation is also reviewed, and needed to be extended for some of the estimators. In Sect. 3 we generalize the calibration technique to NCC. The estimation methods are compared through simulations in Sect. 4, while in Sect. 5 we look at the methods through a real data example. Finally we conclude with a discussion in Sect. 6.

2 NCC with competing risks/multiple outcomes

2.1 The cohort

Our basic framework is a competing risks situation where each subject can experience at most one out of K different events. This is a special case of the multiple endpoints or general life-history setting where each subject can experience several or even all K events.

Let the cohort $\mathscr{C} = \{1, \ldots, n\}$ consist of n individuals who either experience one out of K different events, $E_i = k$, or are censored, $E_i = 0$, at time t_i. This means that E_i is an event indicator taking values in $\{0, 1, \ldots, K\}$, where 0 corresponds to being censored, and for each individual we observe the pair (t_i, E_i). Let $\mathscr{E}_k = \{i \in \mathscr{C}, E_i = k\}$ be the set of all cases experiencing event k and denote the set of all cases by $\mathscr{E} = \bigcup_{k=1}^{K} \mathscr{E}_k$.

Let $\beta = (\beta_1', \ldots, \beta_k', \ldots, \beta_K')$ and $\gamma = (\gamma_1', \ldots, \gamma_k', \ldots, \gamma_K')$, where β_k' and γ_k' are vectors of regression coefficients connected to endpoint k. Let (x_i, Z_i) be our covariates, assumed to be time constant. Here the x_i is considered to be non-random, and assumed known for the entire cohort, while the Z_i is a vector of covariates only known for cases and controls. For the full likelihood it will be useful to consider the Z_i as random variables, which is emphasized by the capital letter.

We assume outcome specific proportional hazard models where we model the hazard function for events of type k for individual i as

$$\alpha_{ki}(t|x_i, Z_i, \beta_k', \gamma_k') = \alpha_{0k}(t) \exp(\beta_k' x_i + \gamma_k' Z_i). \qquad (1)$$

Here $\alpha_{0k}(t)$ is the baseline hazard related to endpoint k.

We restrict our attention to time constant covariates, but at least for WPL it is possible to handle covariates that can be ascertained at every event time an individual is at risk. The standard NCC-method on the other hand, handle time-dependent covariates that are determined for the cases and their time-matched controls only at the event times of the cases.

A common situation with survival data is left truncation. Let l_i denote the entry time of individual i and let t_i be the exit time. A subject is then observed from l_i to t_i, where t_i is an event time or a censoring time depending on the event indicator. The WPL and the full likelihood are first described without left truncation in Sects. 2.3 and 2.4, but in Sect. 2.5 we describe how left truncation may be taken into consideration.

2.2 The NCC study

We consider independent NCC studies for each of the K different endpoints in the same cohort. At each event time, m controls are sampled from the individuals still at risk. Let O_i be a binary variable indicating whether or not individual i is a member of a case–control set, and let $\mathscr{O} = \{i \in \mathscr{C} : O_i = 1\}$ be the collection of all cases and sampled controls.

Estimation in NCC designs with only one endpoint has traditionally been based on a partial likelihood similar to the usual Cox-likelihood (Thomas 1977)

$$L_k(\beta_k', \gamma_k') = \prod_{E_i=k} \frac{\exp(\beta_k' x_i + \gamma_k' Z_i)}{\sum_{j \in \tilde{\mathscr{R}}_i} \exp(\beta_k' x_j + \gamma_k' Z_j)}.$$

Here $\tilde{\mathscr{R}}_i$ denote the m sampled controls together with the failing individual i at time t_i and $\prod_{E_i=k}$ is taken to mean the product over all i where $E_i = k$. Computationally, the likelihood can be handled as a stratified version of a Cox-likelihood, where each $\tilde{\mathscr{R}}_i$ constitute a stratum. Inference can be based on standard large sample theory so that the estimator is approximately normally distributed and the variance is obtained from the inverse of the information matrix (Borgan et al. 1995).

In a competing risks setting the total partial likelihood is a product over single endpoint likelihoods. When estimating k regression coefficients connected to a given endpoint, all products except the k-th are constants.The information matrix will subsequently be a block diagonal matrix and one Cox-regression per endpoint can be carried out.

2.3 Weighted partial likelihood

In the partial likelihood above, the controls are tied to their respective cases, and it has been considered impossible to reuse them for other types of events. In addition, the case and its time-matched controls are only used in the estimation at the event time of the case. A seemingly more efficient way of using the available information is with a WPL (Samuelsen 1997; Saarela et al. 2008) in which all sampled risk sets are pooled together. In this way, controls can be reused for other event times.

A WPL for event k is of the form

$$L_k(\beta_k', \gamma_k') = \prod_{E_i=k} \frac{\exp(\beta_k' x_i + \gamma_k' Z_i)}{\sum_{j \in \mathscr{O}_i} \exp(\beta_k' x_j + \gamma_k' Z_j) w_j}, \tag{2}$$

where $\mathscr{O}_i$ is the collection of all cases and controls at risk at time t_i. We will refer to $\mathscr{O}_i$ as a subcohort, using a term from CC studies in a slightly different way. The w_j is a weight assigned to the j-th individual in the subcohort. If we ignore the weights, the cases will be over-represented since all cases, but only a fraction of the non-cases, are included in $\mathscr{O}$. A sensible way of dealing with this is to weight the individuals by the inverse of their probability of actually being included in the subcohort,

$$w_j = \begin{cases} 1 & E_j \neq 0 \\ 1/p_j & E_j = 0 \end{cases}.$$

Different ways of estimating these sampling probabilities have been proposed. Samuelsen (1997) suggested a "Kaplan–Meier like" estimator;

$$p_j = 1 - \prod_{i \in \mathscr{E},\, t_i \le t_j} \left\{ 1 - \frac{m}{n(t_i) - 1} \right\},$$

where $n(t_i)$ is the number at risk at t_i. A similar estimator was considered by Suissa et al. (1998). Other weighting options include logistic regression models, where we model the inclusion probabilities as

$$p_j = \mathrm{E}[O_j | t_j] = \frac{\exp(\xi + f(t_j))}{1 + \exp(\xi + f(t_j))}. \tag{3}$$

The most general case is when $f(t)$ is some smooth function of t, referred to as GAM (generalized additive model), this has been tried out by Samuelsen et al. (2007). We have used smoothing splines to estimate p_j, but other smoothers should also work. A special case of this logistic regression framework is a model with $f(t) = \eta t$ which correspond to standard logistic regression, referred to as GLM, tried out by Saarela et al. (2008). It is important to note that only non-cases are included when using (3) to estimate the inclusion probabilities. A fourth option, called local averaging was proposed by Chen (2001) for generalized CC designs. The method involves choosing a partition of the time axis and calculate separate weights for controls censored in different time intervals. Let t_B be the upper time limit for the study, let $0 = t^0 < t^1 < \cdots < t^B$ be a partition of the follow-up time, and define $\mathscr{I}_b = (t^{b-1}, t^b]$. Then

$$w_b = \frac{\sum_{j=1}^{n} I(t_j \in \mathscr{I}_b,\ E_j = 0,\ j \in \mathscr{C})}{\sum_{j=1}^{n} I(t_j \in \mathscr{I}_b,\ E_j = 0,\ j \in \mathscr{O})} \tag{4}$$

where $I(\cdot)$ is an indicator function. The numerator in (4) counts the number of individuals censored in $\mathscr{I}_b$, while the denominator counts how many of them that were sampled as controls. Individual j is then given weight w_b if censored in $\mathscr{I}_b$. This can be seen as a special case of GLM with $\xi + f(t) = \mathrm{logit}(1/w_b)$ for $t \in \mathscr{I}_b$. Samuelsen et al. (2007) also pointed out that this technique can be interpreted as post-stratification on censoring times.

The variance estimation is not straightforward with WPL since the controls enter the likelihood at all event times whenever they are at risk. Samuelsen (1997) proposed a variance estimator for his weights. Chen (2001) has a different variance estimator for the local averaging weights, but the variance estimator for post-stratification in Samuelsen et al. (2007) may also be used. However for GAM-weights, to our knowledge, there has not yet been derived a variance estimator. A possibly conservative solution is to use robust variance (Lin and Wei 1989; Barlow 1994). This is also applicable with the other weights and is what we have used in our simulations.

2.4 Full likelihood

Another way of dealing with NCC data and multiple outcomes is a full maximum likelihood approach (MLE) where the entire cohort is used in the estimation, treating individuals not included in the subcohort as missing data (Saarela et al. 2008). We now specify the baselines $\alpha_{0k}(t_i) = \alpha_{0k}(t_i;\, \psi_k)$ parametrically. Then $\theta_k = (\beta_k,\, \gamma_k,\, \psi_k)$ are the parameters characterizing the hazard functions $\alpha_k(t_i|Z_i,\, x_i;\, \theta_k)$ and all the parameters are summarized by $\theta = (\theta_1, \ldots, \theta_K)$. Since Z_i is not fully observed, we model it as a random variable with a parametric distribution characterized by parameters μ. Saarela et al. (2008) showed that the full likelihood can be written as

$$L(\theta, \mu) \propto \prod_{i \in \mathscr{O}} p(T_i,\, E_i|Z_i,\, x_i;\, \theta)p(Z_i|x_i;\, \mu) \\ \times \prod_{i \in \mathscr{C}\setminus\mathscr{O}} \int p(T_i,\, E_i|z,\, x_i;\, \theta)p(z|x_i;\, \mu)dz \qquad (5)$$

with $p(T_i,\, E_i|Z_i,\, x_i;\, \theta)$ being the distribution of $(T_i,\, E_i)$ conditional on $(Z_i,\, x_i)$, $p(z_i|x_i;\, \mu)$ is the distribution of Z_i conditional on x_i, and the integral is over all possible values z of Z_i. The likelihood is made up of two parts. This is due to the fact that we have different information about the fully observed subset of the cohort; $\mathscr{O}$, and non-sampled individuals; $\mathscr{C}\setminus\mathscr{O}$. Saarela et al. (2008) state two assumptions; (i) the random vectors $(T_i,\, E_i,\, x_i,\, Z_i)$ for $i \in \mathscr{C}$ are independent, and (ii) the conditional distribution of the indicator of being sampled, $p(O|T,\, E,\, x,\, Z)$, only depend on data observed for all $i \in \mathscr{C}$, hence $p(O|T,\, E,\, x,\, Z) = p(O|T,\, E,\, x)$. They then prove that under assumptions (i) and (ii) the likelihood expression is the same regardless of the sampling procedure, hence the sampling distribution can be disregarded.

The likelihood (5) is based on more modeling assumptions than the WPL (2). First, we have to specify a parametric baseline. Then, the conditional distribution of $(T_i,\, E_i)$ given $(Z_i,\, x_i)$ takes the form

$$p(T_i,\, E_i|Z_i,\, x_i;\, \theta) \\ \propto \prod_{k=1}^{K} [\alpha_k(T_i|Z_i,\, x_i;\, \theta)]^{I(E_i=k)} \exp\left\{-\int_0^{T_i} \sum_{k=1}^{K} \alpha_k(t|Z_i,\, x_i;\, \theta)dt\right\}. \qquad (6)$$

Secondly, we have to assume a parametric conditional distribution for Z given x. Another problem is that the integral in (5) may be hard or even impossible to evaluate analytically. Therefore one often has to resort to approximation methods like Monte Carlo integration/importance sampling or Markov Chain Monte Carlo methods.

Optimization of the full likelihood can be very time consuming since the integrals in (5) need to be evaluated for every individual in $\mathscr{C}\setminus\mathscr{O}$. However, the integrals only differ with respect to the x_i and the T_i, and often there will be several individuals with (approximately) equal x_i and T_i. Assume that there are $Q = (1, \ldots, j, \ldots, q)$ such patterns and that there are S_j individuals with pattern $(x_j,\, T_j)$. Then the last term in (5) can be written as

$$\prod_{j=1}^{q}\left[\int p(T_j, E_j|z, x_j; \theta)p(z|x_j; \mu)dz\right]^{S_j}$$

and there will only be q integrals to evaluate.

Often it will not matter much whether follow-up time is measured in days or weeks or if a somewhat cruder scale is being used for covariates. It can therefore be sensible to group follow-up time and covariates to increase the number of equal likelihood contributions. Then substantial reduction in computation time can be achieved. After follow-up times and covariates have been grouped, a sensitivity analysis can be done by, for instance, fitting a WPL to both grouped and non-grouped data. Additionally, one could fit the MLE using only fully observed covariates both with original and grouped data. If considerable difference is observed, the grouping is too coarse.

Scheike and Juul (2004) have, by a somewhat different argument, arrived at the same likelihood as Saarela et al. (2008). However, instead of modeling the distribution of Z parametrically, they model it non-parametrically through strata defined by the fully observed covariates. They thereby manage to keep the baseline unspecified as in a Cox-likelihood. The likelihood is difficult to optimize directly, but they suggest using the Expectation-Maximization algorithm instead. We have not included this likelihood in our simulations due to the extensive programming it would require.

2.5 WPL and full likelihood with left truncated survival times

Since left truncation is a common situation with survival data, we need to be able to deal with this when estimating sampling probabilities. It turns out that all estimating methods described above can be modified to deal with left truncation. Samuelsen (1997) noted that the only modification needed for his weights was to restrict the product to the event times when the subject was at risk,

$$p_j = 1 - \prod_{i\in\mathscr{E}, l_j<t_i\leq t_j}\left\{1 - \frac{m}{n(t_i)-1})\right\}.$$

The logistic regression models can be modified as

$$\mathrm{E}(O_j|t_j, l_j) = \frac{\exp(\xi + f(t_j, l_j))}{1+\exp(\xi + f(t_j, l_j))}$$

for some smooth function $f(t_j, l_j)$. As a simplification we have used $f(t_j, l_j) = f_t(t_j) + f_l(l_j)$. If $f(t_j, l_j)$ is a linear function of t and l, we are back to a standard logistic regression model. Let $0 = l^0 < l^1 < \cdots < l^A$ be a partition of the left truncation times, then the local averaging weights (Chen 2001) can be extended to

$$w_{ab} = \frac{\sum_{j=1}^{n} I(l_j \in \mathscr{J}_a,\ t_j \in \mathscr{I}_b,\ E_j = 0,\ j \in \mathscr{C})}{\sum_{j=1}^{n} I(l_j \in \mathscr{J}_a,\ t_j \in \mathscr{I}_b,\ E_j = 0,\ j \in \mathscr{O})},$$

where $\mathscr{I}_b$ is defined as before and $\mathscr{J}_a = (l^{a-1}, l^a]$. A subject with an entry time in $\mathscr{J}_a$ and a censoring time in $\mathscr{I}_b$ will be given weight w_{ab}. The intervals need to be chosen with some caution in order not to get too few individuals in an interval.

To deal with left truncation in the full likelihood we need to condition on $T_i > l_i$, and Saarela et al. (2008) point out that the full likelihood is given by

$$L(\theta, \mu) \propto \prod_{i \in \mathscr{O}} \frac{p(T_i, E_i | Z_i, x_i; \theta) p(Z_i | x_i; \mu)}{\int p(T_i \geq l_i | z, x_i; \theta) p(z | x_i; \mu) dz} \times \prod_{i \in \mathscr{C} \setminus \mathscr{O}} \frac{\int p(T_i, E_i | z, x_i; \theta) p(z | x_i; \mu) dz}{\int p(T_i \geq l_i | z, x_i; \theta) p(z | x_i; \mu) dz}.$$

As before the integrals are over all possible values z of Z_i, while $p(T_i \geq l_i | z_i, x_i; \theta) = S(l_i | z_i, x_i; \theta)$ is the survival function up to the entry time, and $p(T_i, E_i | Z_i, x_i; \theta)$ is given by (6).

3 Generalization of calibration to NCC

3.1 Two-phase stratified sampling

Assume that n individuals are sampled from an infinite population. These individuals constitute the cohort and are referred to as the Phase 1 sample. Furthermore, let the cohort be classified into K strata with n_k individuals in stratum k. The strata are defined through information known for everyone, and $n = n_1 + n_2 + \cdots + n_K$. At Phase 2 we use stratified sampling, i.e. we sample $m_k \leq n_k$ individuals at random without replacement from the kth stratum. It is natural to let the cases be an additional stratum where the whole stratum is sampled with probability 1. Additional covariates are then obtained from the Phase 2 individuals.

CC studies with stratified sampling are two-phase studies. NCC can also be seen as a two-phase sampling design, even though the Phase 2 sampling is a bit more complicated since the controls are matched on time. Thereby the sampling probabilities/weights also depend on time. However, the local averaging (Chen 2001) deal with this time dependence by assigning all subjects censored in the same time interval the same weight. The NCC sampling is then approximated by stratified random sampling where the subjects are stratified according to censoring times (Samuelsen et al. 2007).

3.2 Calibration in general

Let us introduce some additional notation: With $p_i = Pr(i \in \mathscr{O})$ being the sampling probabilities, let now $d_i = 1/p_i$ be the corresponding weights. Further let $w_i = d_i g_i$ be the so-called calibrated weights (Breslow et al. 2009a,b), and let $A_i = (A_{i1}, \ldots, A_{ip})$ be auxiliary variables known for the entire cohort and correlated with the regression covariates.

The calibration technique (Deville and Särndal 1992) originate from survey sampling as a method to improve the Horvitz–Thompson estimator $\hat{y}_{\text{H–T}} = \sum_{i \in \mathcal{O}} d_i y_i$ of a population total $y_{\text{tot}} = \sum_{i=1}^{n} y_i$. The improvement is obtained by using auxiliary variables A, to obtain weights that satisfy the calibration equation

$$\hat{A}_{\text{tot}} = \sum_{\mathcal{O}} w_i A_i = \sum_{i=1}^{n} A_i = A_{\text{tot}}. \tag{7}$$

It states that the total of some auxiliary variables should be estimated exactly. Intuitively, if A and y have high correlation, $\hat{y}_{\text{tot}} = \sum_{i \in \mathcal{O}} w_i y_i$ will probably be closer to y_{tot} than $\hat{y}_{\text{H–T}}$.

Using the same line of thought we want to estimate β from a suitable regression model with individual score contributions U_i and information matrix I. From a first order Taylor approximation of the score function around β_0, the true value of β, we can write

$$\hat{\beta} \approx \beta_0 + \sum_{\mathcal{O}} d_i I^{-1}(\beta_0) U_i(\beta_0)$$

both within a CC- and a NCC design with a WPL. We therefor want to calibrate with respect to auxiliary variables correlated with $I^{-1}(\beta_0)U_i(\beta_0)$ (see below for a good choice of auxiliary variables).

Breslow et al. (2009a,b) and Lumley (2010) have suggested calibration as a way of reducing the variability due to the sampling in stratified CC designs, and we conjecture that this property will carry over to the NCC design. The key point is that when some information is known for the entire cohort, weights that do not take this into account are generally not efficient. Calibration of weights is a way to incorporate the additional information in order to increase efficiency.

The restriction from the calibration equation (7) does not uniquely specify the weights. It is thus required that they stay as close as possible to the original weights. Consequently, w_i are weights that make the estimates of the population totals of the auxiliary variables exact, while they remain "as close as possible" to the original weights (discussed in Sect. 2.3). The term "as close as" requires a measure of distance, $G(w, d)$. Breslow et al. (2009a) used two alternatives; $G_1(w, d) = (w - d)^2/2d$ and $G_2(w, d) = w\log(w/d) - w + d$. See Deville and Särndal (1992) and Deville et al. (1993) for more distance measures and a discussion of their properties.

3.3 Implementation of calibration for NCC studies with several outcomes

In order to do the calibration in practice, the `survey` package (Lumley 2010) in `R` can be used. One way of incorporating time-dependent weights into the calibration is to define strata according to follow-up time. Then we implicitly assume that the

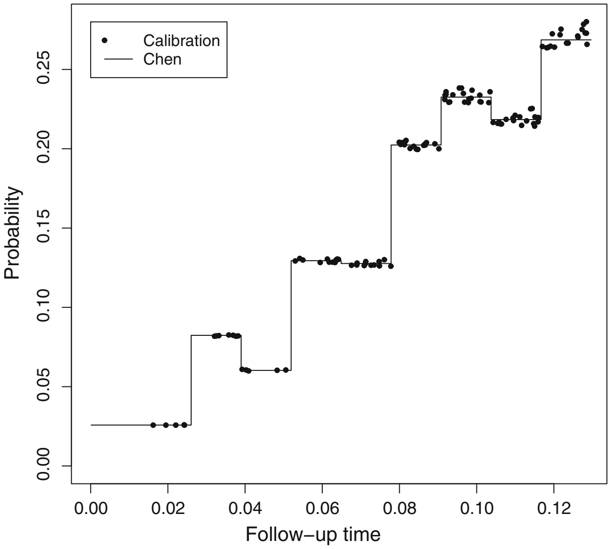

Fig. 1 Inclusion probabilities for controls estimated with calibration and local averaging from the model in Simulation III without left truncation and with nine intervals based on censoring time. Chen correspond to weights estimated with local averaging

original weights are constant within each interval, but vary between intervals. This amounts to require the weights to be as close as possible to local averaging weights (Chen 2001). One example is given in Fig. 1 where the sampling probabilities from one simulation are plotted. It is also possible to explicitly specify the original weights, or more precisely, the sampling probabilities when defining the two-phase design. From simulations we have experienced that the two methods are similar with respect to estimated regression coefficients and standard errors. We have chosen the method based on defining strata because it fits more naturally into the originally framework with stratified CC designs.

As already mentioned, we want to obtain auxiliary variables that are fully observed and correlated with $I^{-1}(\beta_0)U_i(\beta_0)$. A natural choice would then be the dfbetas

$$A_i = I^{-1}(\tilde{\beta})U_i(\tilde{\beta}),$$

where $\tilde{\beta}$ is the cohort estimate. However, the cohort dfbetas are unknown and have to be approximated.

The entire calibration procedure with one endpoint is as follows:

(1) Predict covariates that are not known for the entire cohort. Breslow et al. (2009a,b) refer to the plug-in procedure of Kulich and Lin (2004) where the prediction is done with a weighted regression with the fully observed covariates as explanatory variables. Use weights estimated with one of the methods described in Sect. 2.3.

(2) Do a Cox-regression on the full cohort where fully observed covariates are used as they are recorded. The predicted values from the regression in step 1 are

imputed for all cohort members for the partially observed covariates. Extract the dfbetas from this Cox-regression.

(3) Either: make strata on the basis of follow-up time and left truncation time and specify a stratified two-phase design. Or: specify a non-stratified two-phase design where the sampling probabilities for the second phase are included.

(4) Carry out calibration with the dfbetas as auxiliary variables to obtain new weights w_i.

(5) Do a weighted Cox-regression on NCC data with calibrated weights to obtain the final estimates.

The generalization to multiple endpoints is straight forward, only step 2, 4 and 5 need minor modifications: Instead of 2, do one Cox-regression for each endpoint and obtain one set of dfbetas per endpoint. Instead of 4, do one calibration per endpoint with the corresponding sets of dfbetas as auxiliary variables, and obtain one unique vector of calibrated weights for each endpoint. Instead of 5, do one weighted Cox-regression on NCC data per endpoint using the corresponding calibrated weights.

It is through the imputation that the information known for the entire cohort is used to improve the weights. It is therefor important to choose a good prediction model. We follow Breslow et al. (2009a,b) and recommend single imputation with a regression model suited for the situation at hand.

The variance is estimated by the `survey` package. Since the NCC design can be approximated with a stratified CC design, we believe that the variance should be appropriate. Nevertheless, a theoretical justification is desirable, but it will likely be even more complex than the justification for the CC design. Anyhow, the similarities between the designs are large, so we think it is important to investigate calibration within the NCC framework.

4 Simulation studies

4.1 Main simulation

In order to compare the MLE and the WPL approach with different weighing schemes, we have done three simulation experiments with a cohort of size $n = 2{,}000$, two endpoints, one relatively common that about 10% experienced and another that only about 3% experienced. We sampled $m = 1$ control per case, matched only on time, and each simulation was carried out 1,000 times. The focus here will be on the rare endpoint since the advantage of WPL is greatest in this case.

The simulation model was as follows; the survival times were exponentially distributed with outcome specific hazards as defined in (1). We chose regression parameters $\beta = \gamma = 1$ and by tuning the baseline hazards we controlled the number of cases. The censoring times were drawn from a uniform distribution, which corresponds to random censoring. We had two covariates, the fully observed x was uniformly distributed from zero to one in all three simulations. The partially observed Z was binary distributed with $E[Z|x] = 0.5$ in Simulation I, $E[Z|x] = x$ in Simulation II, and $E[Z|x] = F(x)$ in Simulation III, where $F(x)$ is the cdf of a beta distribution with parameters $(50, 50)$. The difference between the simulations is how strong the correlation between x and Z are. In Simulation I, x and Z are independent, while in Simulation II and III, Z and x are dependent with correlation coefficients of 0.577 and 0.850, respectively. We chose 10 time intervals for the Chen-weights. For calibration we chose to use the stratification approach with nine strata based on censoring time and one additional strata for cases. The partially observed covariate Z was predicted with a weighted logistic regression, using GAM-weights. The remaining parts of the calibration were carried out as explained in Sect. 3.3.

Table 1 displays the results of the simulations for the rare endpoint. A corresponding table for the common endpoint is Table 4 in the Appendix. The first thing to notice is the large efficiency improvements for any of the methods compared to the traditional estimator. This is because we get several additional controls by using both cases experiencing the common endpoint and their controls, as extra controls. Note that the traditional estimator has a larger bias than any of the other estimators. This is likely due to small samples which could affect the traditional estimator more due to fewer effective controls.

There is little difference between the four weighing schemes: The estimates are almost identical and the standard errors, both the empirical and the robust, are very similar except for the Chen-weights that have somewhat higher standard errors compared to the other three. Furthermore, the robust standard errors are all in good agreement with the empirical standard errors. Comparing the efficiency gains for WPL between the common and the rare endpoint (Tables 1, 4) we see that there is more to gain with WPL for the rare endpoint. This is natural since compared to the number of cases there are fewer extra controls for cases of the common endpoint. However, there are efficiency improvements for WPL for the common endpoint compared to the traditional estimator as well.

When estimating β, Saarela's full likelihood gives impressive efficiency gains. This is very natural since x is known for the entire cohort and the full likelihood utilizes this. When estimating γ we see that when x and Z are independent there is almost nothing to gain by using the full likelihood compared to WPL. However, when we induce correlation between x and Z the efficiency increases and it becomes close to fully efficient. It should be noted that in Simulations II and III we have modeled, somewhat unrealistic, the distribution of Z without any parameters μ. In practice it is natural to model the dependence between Z and x with some type of regression model, as is done in the data example in the next section. Our simulation results show higher efficiency gains for the full likelihood than reported by Saarela et al. (2008) for the partially observed covariate. This is likely due to the correlation we have induced between x and Z, as we saw that there was little to gain when x and Z were independent.

We see similar results with calibration as for MLE, it is more efficient than WPL when estimating β, which is natural since it utilize that x is known for the entire cohort, although maybe in a more indirect way. One might also expect that the calibration should be more efficient than WPL when Z is correlated with x. However, this is only the case in Simulation III where the correlation is 0.850, which means that a high correlation is needed before the calibration improves the efficiency. Breslow et al. (2009b) have experienced that an R^2 between x and Z of at least 0.5 is needed to substantially decrease the variance.

Table 1 Results from Simulation I–III for the rare endpoint

Method	β Mean est.	Mean est. se	Emp. se	Eff.	γ Mean est.	Mean est. se	Emp. se	Eff.
Simulation I								
Cox cohort	0.986	0.438	0.439	–	1.016	0.288	0.299	–
Trad. NCC	1.032	0.698	0.725	0.367	1.077	0.438	0.483	0.383
Samuelsen	0.995	0.490	0.494	0.790	1.018	0.312	0.319	0.879
GAM	0.996	0.490	0.494	0.790	1.017	0.313	0.320	0.873
GLM	0.996	0.490	0.495	0.787	1.018	0.312	0.319	0.879
Chen	0.999	0.492	0.497	0.780	1.018	0.314	0.322	0.862
MLE	1.000	0.440	0.439	1.000	1.017	0.309	0.316	0.889
Calibration	0.993	0.453	0.452	0.796	1.018	0.316	0.321	0.868
Simulation II								
Cox cohort	1.031	0.565	0.571	–	1.030	0.355	0.352	–
Trad. NCC	1.027	0.906	0.957	0.356	1.090	0.545	0.564	0.390
Samuelsen	1.033	0.630	0.649	0.774	1.039	0.382	0.387	0.827
GAM	1.031	0.631	0.650	0.772	1.039	0.383	0.388	0.823
GLM	1.034	0.631	0.650	0.772	1.040	0.383	0.388	0.823
Chen	1.034	0.634	0.651	0.769	1.040	0.384	0.390	0.814
MLE	1.028	0.570	0.575	0.990	1.019	0.362	0.358	0.967
Calibration	1.033	0.593	0.603	0.897	1.041	0.386	0.391	0.810
Simulation III								
Cox cohort	1.005	0.882	0.838	–	1.040	0.551	0.528	–
Trad. NCC	1.015	1.477	1.513	0.307	1.101	0.878	0.921	0.329
Samuelsen	1.000	0.996	0.985	0.724	1.045	0.605	0.605	0.762
GAM	1.001	0.997	0.989	0.718	1.045	0.606	0.606	0.759
GLM	1.001	0.996	0.985	0.724	1.045	0.605	0.605	0.762
Chen	1.001	1.002	0.995	0.710	1.045	0.608	0.611	0.747
MLE	1.001	0.887	0.844	0.986	1.043	0.555	0.533	0.981
Calibration	1.014	0.924	0.886	0.894	1.043	0.579	0.567	0.867

β is the log-relative risk connected to x, observed for the entire cohort, γ is log-relative risk connected to Z only observed for the cases and controls

MLE maximum likelihood estimation, *Mean est.* mean of estimates, *Mean est. se* mean of estimated standard error, *Emp. se* empirical standard error, *Eff.* efficiency compared to Cox-regression on the full cohort, calculated with the empirical variance. Efficiency for MLE is calculated by comparing empirical se with empirical se from MLE on full cohort

4.2 Left truncation

Table 2 displays the result of a second simulation where also left truncation is taken into account. These results are for the rare endpoint and as in Sect. 4.1 the results for the common endpoint are reported in Table 5 in the Appendix. The entering times l are drawn from $U[0,\ 0.05]$, the censoring times are then drawn from $U[l,\ 0.13]$ and

Table 2 Results for the rare endpoint from simulation with left truncated survival times

	β				γ			
Method	Mean est.	Mean est. se	Emp. se	Eff.	Mean est.	Mean est. se	Emp. se	Eff.
Simulation I								
Cox cohort	1.012	0.466	0.463	–	1.018	0.289	0.290	–
Trad. NCC	1.078	0.751	0.806	0.332	1.067	0.446	0.459	0.400
Samuelsen	1.041	0.524	0.530	0.763	1.025	0.315	0.317	0.837
GAM	1.039	0.526	0.532	0.757	1.026	0.316	0.318	0.831
GLM	1.044	0.526	0.530	0.763	1.028	0.316	0.317	0.837
Chen	1.042	0.531	0.539	0.738	1.033	0.318	0.321	0.816
MLE	1.048	0.470	0.467	0.979	1.025	0.310	0.314	0.853
Calibration	1.047	0.487	0.492	0.886	1.045	0.321	0.333	0.758
Simulation II								
Cox cohort	1.003	0.585	0.600	–	1.017	0.367	0.385	–
Trad. NCC	1.080	0.949	0.998	0.361	1.057	0.564	0.612	0.396
Samuelsen	1.013	0.652	0.664	0.817	1.019	0.396	0.421	0.836
GAM	1.013	0.655	0.666	0.812	1.019	0.398	0.423	0.828
GLM	1.015	0.655	0.667	0.809	1.021	0.398	0.422	0.832
Chen	1.025	0.662	0.673	0.795	1.024	0.401	0.426	0.817
MLE	1.018	0.584	0.590	1.065	1.028	0.372	0.384	1.021
Calibration	1.003	0.620	0.652	0.847	1.026	0.402	0.426	0.817
Simulation III								
Cox cohort	1.027	0.890	0.894	–	1.011	0.543	0.549	–
Trad. NCC	1.134	1.493	1.547	0.334	1.012	0.868	0.908	0.366
Samuelsen	1.027	1.007	1.008	0.787	1.017	0.600	0.604	0.826
GAM	1.026	1.011	1.013	0.779	1.019	0.602	0.604	0.826
GLM	1.029	1.012	1.014	0.777	1.020	0.602	0.605	0.823
Chen	1.040	1.022	1.021	0.767	1.020	0.607	0.607	0.818
MLE	0.979	0.897	0.905	0.971	1.046	0.549	0.555	0.978
Calibration	1.020	0.936	0.924	0.936	1.035	0.575	0.574	0.915

β is the log-relative risk connected to x, observed for the entire cohort, γ is log-relative risk connected to Z only observed for the cases and controls

MLE maximum likelihood estimation, *Mean est.* mean of estimates, *Mean est. se* mean of estimated standard error, *Emp. se* empirical standard error, *Eff.* efficiency compared to Cox-regression on the full cohort, calculated with the empirical variance. Efficiency for MLE is calculated by comparing empirical se with empirical se from MLE on full cohort

individuals with an event time smaller than the left truncation time are excluded. The baseline hazards are tuned to obtain approximately 10 and 3% cases that survive their truncation time. The rest of the simulation setup is the same as in Sect. 4.1.

When the survival times are left truncated we need to make intervals both with respect to censoring times and left truncation times to calculate Chen-weights. In order to get enough subjects in each interval we had to decrease the number of intervals based on censoring times to 4 and we chose 3 intervals based on left truncation time, resulting in 12 intervals in total.

The calibration was done by first carrying out a weighted logistic regression with the fully observed covariate as explanatory variable to predict the partial observed covariate. Then one Cox-regression per endpoint was done on the full cohort where the partially observed covariate was imputed. The dfbetas were extracted from the Cox-regressions and used as auxiliary variables in the calibration. Finally, the calibrated weights were used in the Cox-regressions to obtain coefficients of interest. In order to obtain weights that depended on the follow-up time and left truncation time, we specified the design to be a stratified two-phase design where the stratification was based on the same intervals as we used when calculating Chen-weights.

Again we see that the weighing methods are far more efficient than the traditional estimator which also here is biased. Apart from in Simulation I all the other estimators are practically unbiased. Without truncation we saw that there was nothing to gain with MLE for the partially observed covariate when it was independent of the fully observed covariate, now we see that MLE is slightly more efficient than WPL in this situation as well. Calibration show similar improvement as it did without left truncation.

4.3 Misspecified models

The full likelihood has two additional modeling assumptions compared to WPL; (a) the distribution of $Z|x$ has to be specified and (b) parametric specification of the baselines has to be given. We wanted to test how vulnerable the likelihood is for misspecification of (a) and (b), and in addition how misspecification of the linear expression of the covariates (c) will affect the estimates and variances. To test (a) we used Simulation III, but specified $p(z|x) = \mu^z(1-\mu)^{(1-z)}$ in the likelihood, hence specified Z independent of x. To check the parametric assumption on the baseline we simulated survival times from a Weibull with a decreasing baseline, but specified an exponential baseline in the likelihood. Finally, we added a square term in the relative risk expression in the simulation, but ignored it when fitting the models to check how that would affect the likelihood.

The results for the rare endpoint can be found in Table 6 in the Appendix. The results for the common endpoint are not shown, but are similar to those for the rare endpoint. We see that correct specification of $Z|x$ is important for the MLE. Both $\hat{\gamma}$ and $\hat{\beta}$ are biased and $\hat{\beta}$ seriously so, which means that a misspecified $p(Z|x)$ does not only affect the estimate connected to Z, but can also highly affect the estimate connected to x. In addition we see that the variances are seriously underestimated for both parameters. However, the misspecification of baseline did not affect the full likelihood much and when the parametric expression was wrongly specified the full likelihood did not do much worse than any of the other estimators.

5 Application to data

Samuelsen et al. (1998) investigated how gestational age and other covariates influenced childhood mortality. The data set consisted of all children born in Norway between 1967 and 1989 who survived their first year and had a gestational age $\geq$16 weeks ($n = 1{,}186{,}655$). The children were originally followed to death, age 15 years or end of 1991, but we are going to limit our selves to follow-up time $\leq$10 years. We do this to make the parametric baseline hazard assumption we need in MLE more valid. Their analysis was cause specific with five different causes, but we are only going to use two; death of cancer, endpoint 1, and death of all other causes, endpoint 2. With follow-up time until 10 years a Weibull baseline hazard was a reasonable choice. We excluded subjects with missing covariates or covariates that were obviously wrongly coded. Because of computational time we needed to reduce to data set further and we therefor only used first born boys in the analysis. We then ended up with a cohort of size $n = 254{,}572$. The study was originally a cohort study and therefor we have all information on every individual, but we are going to do synthetic case–control studies and sample $m = 1$ control per case 200 times.

Childhood mortality is low in Norway, out of 254,572 subjects 868 died, 125 children died of cancer while 743 died from other causes. This means that the subcohort belonging to the cancer endpoint will only consist of 250 subjects, but if we also use cases and controls sampled for endpoint 2, the subcohort increases to 1,736.

The covariates we have used is Z, a dummy for birth weight >3 kg, and x, gestational age in days. As the notation implies, we pretend that birth weight is only known for cases and controls, while gestational age is known for the entire cohort. For a real case–control study this may be considered artificial. However, in our example, gestational age is a natural predictor for birth weight. The correlation between gestational age and birth weight has to be taken into account when $p(Z|x, \mu)$ is specified. We chose a probit model for $\mu_i = P(Z_i = 1|x_i)$, since birth weight conditional on gestational age is approximately normally distributed. Therefore

$$g(\mu_i) = \Phi^{-1}(\mu_i) = \xi_0 + \xi_1 x_i$$
$$p(Z_i|x_i;\, \mu_i) = \mu_i^{Z_i}(1-\mu_i)^{1-Z_i}$$

where $g(\cdot)$ is the link function and Φ is the cumulative probability function for the standard normal distribution.

Table 3 displays the results of the analysis. Birth weight has opposite effect on the endpoints; birth weight above 3 kg. increases the risk of death of cancer while it decreases the risk of death of all other causes. However, it is only significant for other deaths endpoint. The opposite effect of birth weight on our two endpoints is along the findings in the original study of Samuelsen et al. (1998). The effects of gestational age are very small and not significant, but it is interesting to see that, nevertheless, all estimates stay at the same side of zero.

The estimates of gestational age for the cancer endpoint are in good agreement, while for other deaths the ML-estimates both on the full cohort and on the NCC data are smaller than the rest. With birth weight on the other hand, the traditional estimator and WPL seems to add some bias for both endpoints, while the MLE is in better agreement with the cohort analysis.

Table 3 Result from synthetic NCC on birth weight and mortality data

		Cancer endpoint		Other deaths endpoint	
	Method	Gest. age	Birth weight	Gest. age	Birth weight
Estimate	Cohort Cox	−0.0033	0.4813	2.9×10^{-4}	−0.4513
	Cohort MLE	−0.0035	0.4960	0.4×10^{-4}	−0.4380
	Trad. NCC	−0.0036	0.4342	1.9×10^{-4}	−0.4682
	Samuelsen	−0.0033	0.4652	2.7×10^{-4}	−0.4698
	GAM	−0.0032	0.4645	2.7×10^{-4}	−0.4703
	Calibration	−0.0033	0.4657	2.9×10^{-4}	−0.4662
	MLE	−0.0035	0.4969	0.0×10^{-4}	−0.4406
	MLE agg.	−0.0045	0.5152	3.7×10^{-4}	−0.4405
Standard error	Cohort Cox	0.0070	0.3141	0.0027	0.0977
	Cohort MLE	0.0067	0.3136	0.0026	0.0976
	Trad. NCC	0.0106	0.4256	0.0038	0.1489
	Samuelsen	0.0073	0.3343	0.0041	0.1440
	GAM	0.0073	0.3343	0.0041	0.1442
	Calibration	0.0068	0.3342	0.0032	0.1440
	MLE	0.0069	0.3304	0.0028	0.1402
	MLE agg.	0.0067	0.3296	0.0028	0.1399
Simulation based standard error	Trad. NCC	0.0105	0.4071	0.0037	0.1592
	Samuelsen	0.0075	0.3331	0.0038	0.1476
	GAM	0.0075	0.3333	0.0038	0.1485
	Calibration	0.0071	0.3285	0.0031	0.1367
	MLE	0.0067	0.3137	0.0027	0.0987
	MLE agg.	0.0066	0.3138	0.0025	0.0983

Controls are sampled 200 times
MLE maximum likelihood, *Cohort MLE* MLE on cohort

We also report what we call simulation based standard error, $\sqrt{S^2 + V^2}$, as a comparison of the estimated standard error. Here S^2 is the empirical variance over the 200 regression estimates and V^2 is the estimated variance from the cohort analysis. The estimated standard errors seems to be in good agreement with this, perhaps except for the MLE for birth weight with the other deaths endpoint where the simulation based standard error is somewhat smaller. Standard errors for WPL and MLE on the cancer endpoint are very close to the cohort standard errors. This is quite natural since the number of controls per case is quite high. However for the other deaths endpoint there is a difference between WPL and MLE, and especially the much lower standard error of birth weight shows us that there is efficiency to gain when the partially observed covariate depend on a fully observed covariate. It is a bit worrying that the standard error from the calibrated weights for gestational age with cancer endpoint is actually smaller than the cohort standard error, but this is probably due to chance since the sampling was carried out only 200 times and at least the simulation based standard error is somewhat higher.

The optimization of the full likelihood with these data was time consuming. It took about 11 min to optimize the likelihood once, even though there was no integration involved. This is therefor a situation where the aggregation technique can be useful. We grouped follow-up time into months and gestational age into weeks and ended up with 9,125 groups. The computational time for one optimization then decreased to about 30 s. Furthermore, the simulations also showed that using follow-up time in months instead of days and gestational age in weeks instead of days only changed the estimates slightly, and the standard errors were unaffected.

6 Discussion

We have looked at two main methods for reusing controls from a NCC study; WPL and MLE, both with and without left truncation. In addition we have suggested how to generalize calibration for the NCC design with competing risks, and presented an aggregation technique that may reduce the complexity of the full likelihood.

When two or more endpoints are of interest, being able to use controls sampled for one endpoint as additional controls for another endpoint can improve efficiency quite a lot compared to the traditional estimator. It can especially increase the efficiency if one of the endpoints is quite common and another is rare, or if there has been sampled quite a lot of controls for one endpoint and just one or two for another endpoint. However, if both endpoints are relatively common, or more than three or four controls per case are sampled, the efficiency gain will usually be modest.

WPL (excluding calibration) is easy to use, once you have estimated the sampling probabilities, a regular weighted Cox-regression can be carried out and the choice of weights does not seem to matter much. It is also quite robust for model misspecification since the relative risks obtained with WPL reproduce the cohort results (Scott and Wild 1986, 1991). A slight disadvantage is the variance estimation. An easy, but maybe conservative solution is to use robust variance estimation (Lin and Wei 1989; Barlow 1994), but for some of the weights there exists other possibilities as well (Samuelsen 1997; Chen 2001; Samuelsen et al. 2007).

The MLE approach is more cumbersome since more modeling assumptions and more programming are needed. The extra modeling assumptions make it more vulnerable to model misspecification. We also experienced that it was sometimes hard to optimize the likelihood; different starting values gave slightly different estimates even with a very strict convergence criterion. In addition, if the full cohort is large compared to the subcohort it can sometimes result in convergence and identifiability problems (Saarela and Kulathinal 2007). However, because the full likelihood utilizes all available information it can sometimes estimate regression coefficients connected to fully observed covariates virtually efficiently. The efficiency gain can also be large if there are partially observed covariates correlated with the fully observed covariates. In most optimization programs the information matrix is calculated as a byproduct to find the optimum. The variance estimates from the MLE approach of Saarela et al. (2008) are therefor directly obtained as the inverse of the information matrix. In addition, the aggregation technique can in some situations, when the number of covariates is not too large, reduce the computation time substantially; in our data example the computation time went down from 11 min to 30 s with only slightly changed estimates and unaffected standard errors.

The calibration technique does not rest on more modeling assumptions than WPL with estimated weights. However, one needs to find good prediction models for the partially observed covariates to utilize as much of the information in the fully observed covariates as possible. In our simulations, we also noticed that the calibration seemed to need higher correlation than MLE between fully and partially observed covariates before the efficiency was increased for the partially observed covariate. However, our simulation results are promising and a more formal justification of this approach would be a natural next step.

Salim et al. (2009) considered a competing risks situation where controls are matched on more than follow-up time and are drawn from two partly overlapping cohorts. They suggested weights that are similar to Samuelsen (1997), but modified to accommodate their situation. When all controls are drawn from the same cohort, there is no additional matching and the product is interpreted to be over event times, Salim's weights are equal to Samuelsen's and therefor they have not been considered here.

In a recent paper by Liu et al. (2010) another way of incorporating auxiliary information into the Cox-regression is presented. As with the calibration, the auxiliary information has to be known for the entire cohort and these auxiliary covariates should be correlated with the covariates of interest to improve the efficiency.

In epidemiologic studies matching is often used to try to adjust for confounding effects and to increase efficiency. Controls are then not only sampled so that they are at risk at the event time of the case, but they also have the same values as the case on the matching variables. This will affect the sampling probabilities and the estimation procedures need to be generalized to this situation. If there are not too many matching variables with too many levels, a solution is to calculate the sampling probabilities separately for each value of the matching variable. However, as the number of levels increase the weights may become more and more unstable since less and less data are used to estimate them. Saarela et al. (2008) state that their full likelihood approach can be used with all kinds of sampling schemes as long as the two conditions stated in Sect. 2.4 are fulfilled.

Appendix

The appendix contain Tables 4, 5, 6.

Table 4 Results from Simulation I–III for the common endpoint

	β				γ			
Method	Mean est.	Mean est. se	Emp. se	Eff.	Mean est.	Mean est. se	Emp. se	Eff.
Simulation I								
Cox cohort	1.008	0.242	0.234	–	1.006	0.158	0.158	–
Trad. NCC	1.030	0.375	0.373	0.394	1.015	0.231	0.235	0.452
Samuelsen	1.015	0.322	0.320	0.535	1.008	0.196	0.201	0.618
GAM	1.017	0.323	0.320	0.535	1.007	0.197	0.202	0.612
GLM	1.016	0.322	0.321	0.531	1.008	0.197	0.201	0.618
Chen	1.019	0.325	0.323	0.525	1.008	0.198	0.203	0.606
MLE	1.022	0.246	0.236	0.992	1.007	0.193	0.199	0.630
Calibration	1.014	0.263	0.256	0.836	1.009	0.199	0.204	0.600
Simulation II								
Cox cohort	1.026	0.311	0.309	–	0.992	0.193	0.192	–
Trad. NCC	1.009	0.483	0.488	0.401	1.015	0.285	0.287	0.448
Samuelsen	1.028	0.413	0.405	0.582	1.002	0.239	0.237	0.656
GAM	1.027	0.414	0.405	0.582	1.002	0.240	0.237	0.656
GLM	1.029	0.413	0.405	0.582	1.002	0.239	0.237	0.656
Chen	1.029	0.417	0.408	0.574	1.003	0.241	0.239	0.645
MLE	1.022	0.318	0.312	0.987	0.982	0.204	0.196	0.960
Calibration	1.023	0.352	0.339	0.830	1.003	0.242	0.239	0.645
Simulation III								
Cox cohort	1.000	0.476	0.480	–	1.011	0.295	0.292	–
Trad. NCC	1.004	0.771	0.815	0.347	1.029	0.454	0.462	0.399
Samuelsen	0.992	0.653	0.677	0.503	1.017	0.383	0.387	0.569
GAM	0.993	0.655	0.680	0.498	1.017	0.384	0.388	0.566
GLM	0.993	0.654	0.678	0.501	1.017	0.384	0.387	0.569
Chen	0.993	0.661	0.688	0.487	1.018	0.387	0.393	0.552
MLE	0.996	0.484	0.489	0.964	1.013	0.301	0.299	0.954
Calibration	1.004	0.532	0.546	0.773	1.010	0.337	0.343	0.725

β is the log-relative risk connected to x, observed for the entire cohort, γ is log-relative risk connected to Z only observed for the cases and controls

MLE maximum likelihood estimation, *Mean est.* mean of estimates, *Mean est. se* mean of estimated standard error, *Emp. se* empirical standard error, *Eff.* efficiency compared to Cox-regression on the full cohort, calculated with the empirical variance. Efficiency for MLE is calculated by comparing empirical se with empirical se from MLE on full cohort

Table 5 Results for the common endpoint from simulation with left truncated survival times

	β				γ			
Method	Mean est.	Mean est. se	Emp. se	Eff.	Mean est.	Mean est. se	Emp. se	Eff.
Simulation I								
Cox cohort	1.002	0.251	0.247	–	1.013	0.155	0.164	–
Trad. NCC	1.033	0.389	0.387	0.407	1.019	0.230	0.233	0.495
Samuelsen	1.031	0.335	0.331	0.557	1.020	0.195	0.201	0.666
GAM	1.029	0.338	0.332	0.554	1.021	0.196	0.204	0.646
GLM	1.034	0.338	0.332	0.554	1.023	0.196	0.203	0.653
Chen	1.033	0.347	0.343	0.418	1.027	0.200	0.206	0.634
MLE	1.039	0.257	0.251	0.961	1.026	0.191	0.198	0.678
Calibration	1.006	0.277	0.278	0.789	1.020	0.202	0.202	0.659
Simulation II								
Cox cohort	1.003	0.317	0.320	–	1.002	0.197	0.203	–
Trad. NCC	0.997	0.490	0.492	0.423	1.010	0.288	0.289	0.493
Samuelsen	1.013	0.420	0.418	0.586	1.003	0.244	0.246	0.681
GAM	1.012	0.424	0.419	0.583	1.003	0.246	0.246	0.681
GLM	1.015	0.424	0.421	0.578	1.006	0.246	0.247	0.675
Chen	1.024	0.434	0.430	0.554	1.009	0.250	0.252	0.649
MLE	1.017	0.325	0.322	1.012	1.013	0.211	0.208	0.953
Calibration	0.994	0.367	0.364	0.773	1.014	0.252	0.255	0.634
Simulation III								
Cox cohort	1.005	0.479	0.487	–	1.004	0.291	0.307	–
Trad. NCC	1.007	0.775	0.799	0.372	1.026	0.449	0.476	0.416
Samuelsen	1.013	0.656	0.688	0.501	1.005	0.378	0.397	0.598
GAM	1.013	0.662	0.694	0.492	1.006	0.381	0.399	0.592
GLM	1.016	0.663	0.694	0.492	1.007	0.381	0.400	0.589
Chen	1.028	0.680	0.712	0.468	1.008	0.390	0.406	0.572
MLE	0.957	0.492	0.497	0.952	1.038	0.300	0.313	0.950
Calibration	1.014	0.541	0.545	0.798	1.012	0.336	0.353	0.756

β is the log-relative risk connected to x, observed for the entire cohort, γ is log-relative risk connected to Z only observed for the cases and controls

MLE maximum likelihood estimation, *Mean est.* mean of estimates, *Mean est. se* mean of estimated standard error, *Emp. se* empirical standard error, *Eff.* efficiency compared to Cox-regression on the full cohort, calculated with the empirical variance. Efficiency for MLE is calculated by comparing empirical se with empirical se from MLE on full cohort

Table 6 Results from misspecified models for the rare endpoint

	β			γ		
Method	Mean est.	Mean est. se	Emp. se	Mean est.	Mean est. se	Emp. se
Misspecification (a)						
Cox cohort	1.010	0.891	0.893	1.018	0.545	0.565
Cohort MLE	1.010	0.890	0.891	1.019	0.545	0.564
Trad. NCC	1.080	1.492	1.533	1.031	0.866	0.902
Samuelsen	1.024	1.011	1.020	1.016	0.602	0.622
MLE	2.098	0.520	0.460	1.128	0.352	0.343
Misspecification (b)						
Cox cohort	0.993	0.854	0.858	1.022	0.518	0.525
Cohort MLE	0.908	0.853	0.846	0.980	0.518	0.521
Trad. NCC	1.040	1.445	1.557	1.091	0.837	0.914
Samuelsen	0.988	0.957	0.976	1.035	0.566	0.581
MLE	0.898	0.862	0.858	0.985	0.524	0.527
Misspecification (c)						
Cox cohort	0.274	0.922	0.938	0.893	0.542	0.538
Cohort MLE	0.273	0.922	0.938	0.893	0.542	0.539
Trad. NCC	0.142	1.461	1.553	0.992	0.856	0.909
Samuelsen	0.291	1.042	1.048	0.890	0.589	0.598
MLE	0.248	0.933	0.955	0.908	0.549	0.546

Misspecification (a): Z simulated as in Simulation III, but modeled as being independent of x. Misspecification (b): true baseline linearly decreasing (Weibull) while modeled as constant (exponential). Misspecification (c): true risk function $\exp(1 * Z - 1 * x + 1 * x^2)$ modeled as $\exp(1 * Z + 1 * x)$. β is the log-relative risk connected to x, observed for the entire cohort, γ is log-relative risk connected to Z only observed for the cases and controls

MLE maximum likelihood estimation, *Mean est.* mean of estimates, *Mean est. se* mean of estimated standard error, *Emp. se* empirical standard error

References

Barlow WE (1994) Robust variance estimation for the case-cohort design. Biometrics 50(4):1064–1072

Borgan Ø, Goldstein L, Langholz B (1995) Methods for the analysis of samled cohort data in the Cox proportional hazards model. Ann Stat 23(5):1749–1778

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009a) Improved Horvitz–Thompson estimation of model parameters for two-phase stratified samples: applications in epidemiology. Stat Biosci 1(1):32–49

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009b) Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol 169(11):1398–1405

Chen KN (2001) Generalized case-cohort sampling. J R Stat Soc B 63(4):791–809

Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87(418):376–382

Deville JC, Särndal CE, Sautory O (1993) Generalized raking procedures in survey sampling. J Am Stat Assoc 88(423):1013–1020

Kalbfleisch JD, Lawless JF (1988) Likelihood analysis of multi-state models for disease incidence and mortality. Stat Med 7(1–2):149–160

Kulich M, Lin DY (2004) Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc 99(467):832–844

Lin DY, Wei LJ (1989) The robust inference for the Cox proportional hazards model. J Am Stat Assoc 84(408):1074–1078

Liu M, Lu W, Tseng CH (2010) Cox regression in nested case–control studies with auxiliary covariates. Biometrics 66(2):374–381

Lumley T (2010) Complex surveys: a guide to analysis using R. Wiley series in survey mehodology. Wiley, Hoboken

Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73(1):1–11

Saarela O, Kulathinal S (2007) Conditional likelihood inference in a case-cohort design: an application to haplotype analysis. Int J Biostat 3(1):1

Saarela O, Kulathinal S, Arjas E, Läärä E (2008) Nested case–control data utilized for multiple outcomes: a likelihood approach and alternatives. Stat Med 27(28):5991–6008

Salim A, Hultman C, Sparén P, Reilly M (2009) Combining data from 2 nested case–control studies of overlapping cohorts to improve efficiency. Biostatistics 10(1):70–79

Samuelsen SO (1997) A pseudolikelihood approach to analysis of nested case–control studies. Biometrika 84(2):379–394

Samuelsen SO, Magnus P, Bakketeig LS (1998) Birth weight and mortality in childhood in Norway. Am J Epidemiol 148(10):983–991

Samuelsen SO, Ånestad H, Skrondal A (2007) Stratified case-cohort analysis of general cohort sampling designs. Scand J Stat 34(1):103–119

Scheike TH, Juul A (2004) Maximum likelihood estimation for Cox's regression model under nested case–control sampling. Biostatistics 5(2):193–206

Scott AJ, Wild CJ (1986) Fitting logistic models under case–control or choice based sampling. J R Stat Soc B 48(2):170–182

Scott AJ, Wild CJ (1991) Fitting logistic regression models in stratified case–control studies. Biometrics 47(2):497–510

Suissa S, Edwardes MD, Boivin JF (1998) External comparisons from nested-case control designs. Epidemiology 9(1):72–78

Thomas DC (1977) Addendum to "methods of cohort analysis: appraisal by application to asbestos mining" by Liddell FDK, McDonald JC and Thomas DC. J R Stat Soc A 140:469–491