STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:5991–6008
Published online 12 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3416

Nested case–control data utilized for multiple outcomes: A likelihood approach and alternatives

Olli Saarela[1,*,†], Sangita Kulathinal[2,3], Elja Arjas[1,3] and Esa Läärä[4]

[1]*National Public Health Institute, Helsinki, Finland*
[2]*Indic Society for Education and Development (INSEED), Nashik, India*
[3]*Department of Mathematics and Statistics, University of Helsinki, Finland*
[4]*Department of Mathematical Sciences, University of Oulu, Finland*

SUMMARY

Suppose a nested case–control design has been applied for collecting covariate data when studying a specific disease. With possible new outcomes of interest it would be sensible to utilize the previously selected control group instead of (or in addition to) a new control selection, given that the same covariate data were relevant and available, and that their measurements had adequate stability and quality. We formulate this problem in the framework of the competing risks survival model. In this approach covariate information collected for all outcomes can be utilized in the analysis. We not only propose likelihood-based parameter estimation but we also review alternative methods based on weighted partial/pseudolikelihoods. The methods discussed here are closely related to the analysis of a case–cohort design, where the control group is not tied to cases of a specific disease. The different methods are compared in a simulation study. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: case–cohort design; case–control design; competing risks; full likelihood; nested case–control design; pseudolikelihood; weighted partial likelihood

1. INTRODUCTION

The nested case–control (NCC) and case–cohort (CC) designs are the two major cost-efficient sampling schemes employed to collect exposure data in studies where large cohorts are needed to observe enough events of interest, making it impractical to collect all exposure data for the

*Correspondence to: Olli Saarela, Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Mannerheimintie 166, 00300 Helsinki, Finland.
†E-mail: olli.saarela@ktl.fi

Contract/grant sponsor: European Commission
Contract/grant sponsor: Academy of Finland; contract/grant numbers: 114786, 122883
Contract/grant sponsor: EU Network of Excellence *Cancer Control Using Population-based Registries and Biobanks* (CCPRB)
Contract/grant sponsor: Academy of Finland; contract/grant number: 120146

Received 24 October 2007
Accepted 21 July 2008

complete cohort. In both designs, measurements of some or all risk factors of interest are obtained only for a subset of the original study cohort. Typically, the risk factor information is collected on all cases, i.e. the subjects with an observed event of interest during the follow-up, and on a sample of the study cohort that acts as a control group.

Although the term 'nested case–control design' is sometimes used more generally in the epidemiologic literature to refer to all kinds of case–control sampling from a closed or otherwise enumerated source population (see e.g. [1, 2]), here we use the term more restrictively for a design where the control group is obtained by a procedure synonymously called as risk set sampling [3] or time-matched sampling [1]. Here for each new incident case diagnosed, a number of control subjects (usually 1–5 [4]) are selected randomly from the cohort members (excluding the case itself) who are at risk at the time of diagnosis of the case. (Incidence) density sampling [2] or concurrent sampling [5], as first outlined in [6], are more general terms, where the sampling from the population at risk is not necessarily tied to the times when the cases are diagnosed. Another variation is that the risk sets can be formed in grouped time rather than continuous time [7]. A common practice is to apply additional matching or stratification on a few selected background variables, such as region of residence, gender, age (or time of birth) and time of entry to the cohort. In a CC study the control group, called the subcohort, is a random sample of the original cohort, selected either at the outset of the study or retrospectively [8–10]. As in the NCC design, the efficiency of the sampling can sometimes be improved by making use of the covariate and follow-up data collected for the complete cohort, for example by determining the subcohort selection probabilities with a parametric model [11] or within strata [12]. For a broader discussion on the relative merits of the NCC and CC designs, we refer to e.g. [10, 13–16].

Analysis of time-to-event data collected under the cohort sampling designs has traditionally been based on the proportional hazards model and variations of the Cox partial likelihood [17]. In the NCC situation the risk sets in the partial likelihood contributions contain only the case and the time-matched controls [18], whereas in the pseudolikelihood expression proposed by Prentice [10] for the CC situation, the subcohort members contribute to multiple risk sets whereas the cases outside the sampled subcohort contribute to the risk sets only at their event times. More recent methodological developments have made it possible to analyse data from these designs using similar methods, in which the controls from NCC selection are no longer tied to their respective cases [11, 12, 19–24].

The possibility of using the same control group for several disease outcomes is commonly thought to be absent in NCC studies, in which controls are selected for cases of one particular disease [2]. This property is considered to be an important advantage of the CC design, as the same subcohort, being sampled without reference to cases of any specific outcome, is readily available for cases of several types [25]. With a possible new outcome of interest, it would be more economical in the NCC design, too, to utilize an existing control group from a previous NCC study instead of (or in addition to) an entirely new control selection, given that the same covariates were available and relevant, and the quality of their measurements remains sufficiently stable over time. Also, it would be desirable to be able to use the cases of the primary disease as additional controls in the analysis of the new outcome of interest. The current problem of reusing controls of a previous NCC selection is motivated by, e.g. Nordic biobank cancer cohorts [26], where NCC designs have been used in studying the associations of certain serologically assessed factors with specific cancer types.

Samuelsen [19] noted the possibility of using controls from a previous NCC selection in the analysis of other diseases, but the idea was not studied further. The above-mentioned developments

DOI: 10.1002/sim

in statistical methods for analysing data coming from NCC and CC studies, where in particular the controls from NCC selection are freed from being tied to their time-matched cases, motivate us for a closer examination of the possibility of reuse of controls. In this paper we propose a new likelihood-based approach for this problem, closely related to that of Kulathinal and Arjas [23]. For comparison, we also extend the existing weighted partial likelihood methods [11, 19, 24] for multiple endpoints and apply them to the same problem. The methods considered here allow one to choose the particular time scale, which is used as a basis of the survival models. Here we specifically consider the issue of left truncation arising from using age in this role. In our illustration we make use of real cohort data as a basis for a simulation study for the comparison of the different methods.

The plan of the paper is as follows. In Section 2 we consider the likelihood-based approach for the survival analysis of cohort sampling designs and the assumptions under which such an approach is applicable. In Section 3 we briefly review the alternative methods based on weighted partial/pseudolikelihoods. Section 4 discusses the choice of time scale. In Section 5 we compare the proposed methods in the context of a real cohort. The paper concludes with a discussion in Section 6.

2. LIKELIHOOD-BASED ANALYSIS WITH PARTIALLY OBSERVED COVARIATE DATA

Following Kulathinal and Arjas [23] we first set up notations for the general cohort sampling design with multiple outcomes using the competing risks setting. Consider a cohort $\mathscr{C}=\{1,\ldots,n\}$ of n subjects, followed up for the incidence of K different diseases of interest. The variables collected on all individuals $i\in\mathscr{C}$ are denoted as X_i. An individual i enters the study as healthy and is followed up until the first event of interest or right censoring caused by death, loss to follow-up, or the end of the study. Let T_i be the observed time for this event or right censoring and let E_i be an indicator of the type of observed event at T_i. The variable E_i assumes one of the values in $\{0,1,\ldots,K\}$, with 0 referring to right censoring at T_i.

An individual is said to be a case of type k if an outcome event k is observed first during the follow-up, and a group of such cases is denoted as $\mathscr{E}_k=\{i\in\mathscr{C}:E_i=k\}$. Using the follow-up data, the set of all types of cases $\mathscr{E}=\bigcup_{k=1}^{K}\mathscr{E}_k$ is identified. The data observed on all cohort members are abbreviated as $\mathscr{F}=\{(T_i,E_i,X_i):i\in\mathscr{C}\}$. The risk factors of interest are denoted by Z_i. Because of the study design, data on these factors are collected only on a subset of the whole cohort. Let O_i be an indicator for this: $O_i=1$ if Z_i is observed on individual i, and $O_i=0$ if not. Let $\mathscr{O}=\{i\in\mathscr{C}:O_i=1\}$ represent the set of individuals on whom such covariate data are obtained. Typically, $\mathscr{E}\subset\mathscr{O}\subset\mathscr{C}$.

In a typical NCC study, the covariate data are initially collected on the cases of one particular type, say $\mathscr{E}_k$, and on a group of time-matched controls $\mathscr{S}_k$. The sampling of controls is done from the risk sets at times T_i when $E_i=k$ occurs. For each such time T_i, a control set $\mathscr{S}_i$ is sampled without replacement from $\mathscr{R}_i(X_i)\setminus\{i\}$, where $\mathscr{R}_i(X_i)$ is the risk set at time T_i, after possibly applying some predefined stratification criteria on the X_i covariates. All sets $\mathscr{S}_i$ are sampled independently, and the pooled set of controls is $\mathscr{S}_k=\bigcup_{i\in\mathscr{E}_k}\mathscr{S}_i$. The control group can also include some cases, because risk set sampling is done without regard to the future case status. It is possible that a separate time-matched selection of controls has been carried out for several endpoints $k\in\mathscr{D}\subseteq\{1,\ldots,K\}$. The combined set of all controls is then denoted by $\mathscr{S}=\bigcup_{k\in\mathscr{D}}\mathscr{S}_k$.

DOI: 10.1002/sim

and the set of individuals on whom covariate data are collected is $\mathscr{C} = \mathscr{E} \cup \mathscr{S}$. Of the main interest here is the situation where previously selected control sets are reused in the analysis of endpoints $k \in \{1, \ldots, K\} \setminus \mathscr{D}$ without a new control selection.

Scheike and Martinussen [21] and Kulathinal and Arjas [23] considered likelihood expressions for the complete cohort $\mathscr{C}$, when the set $\mathscr{C}$ is obtained from a CC selection and the covariate of interest is missing for the set $\mathscr{C} \setminus \mathscr{C}$. Scheike and Juul [22] used a similar approach in the standard NCC setting and noted that the likelihood expression is of a similar form in both designs. Here we show that a similar likelihood expression results from a general cohort sampling design, including the NCC situation described above. As the covariate Z_i is unobserved for $i \in \mathscr{C} \setminus \mathscr{C}$, the likelihood-based approach requires modelling of the distribution of Z_i. For notational convenience we assume here discrete Z_i, but the distribution might as well be continuous. The covariates X_i are observed on all $i \in \mathscr{C}$ and can be included in the conditioning of the likelihood. Since the probability for $O_i = 1$ depends on the observed data $\mathscr{F}$, it may not be obvious that the likelihood expression can be written as a product of independent contributions over individuals $i \in \mathscr{C}$. Let $O = (O_1, \ldots, O_n)$, $Z = (Z_1, \ldots, Z_n)$, $E = (E_1, \ldots, E_n)$, $T = (T_1, \ldots, T_n)$ and $X = (X_1, \ldots, X_n)$ be the vector notations for the variables defined before, and let θ contain all the model parameters. In Appendix A we show formally that the likelihood expression can be written in the product form under the following two assumptions:

A1. Random vectors (T_i, E_i, Z_i, X_i), $i \in \mathscr{C}$, are independent.
A2. The conditional distribution of O depends only on the data observed for all $i \in \mathscr{C}$: $p(O|T, E, Z, X; \theta) = p(O|\mathscr{F})$.

Here the former is the standard assumption of independence across observations in $\mathscr{C}$, which is considered as a random sample. The latter assumption appears to include most of the commonly used cohort sampling designs, including the ones where X contains correlates of Z. An example of such a situation would be the counter-matching design of Langholz and Borgan [27], if the counter-matching is done using a surrogate measurement of the exposure of interest. The second assumption also corresponds to the assumption that the missingness in Z is ignorable in the sense that, given $\mathscr{F}$, it does not depend on the missing values themselves (missing at random) or on the model parameters. Note, however, that the likelihood-based approach requires careful modelling of the dependencies between Z and $\mathscr{F}$ to correctly capture the distribution of Z. We write the vector of parameters into the form $\theta = (\beta, \mu)$, where $\beta = (\beta_1, \ldots, \beta_K)$ and β_k represent the parameters characterizing the hazard function $\lambda_k(t|Z_i, X_i; \beta_k)$ specific to event type k, $k = 1, \ldots, K$, and μ the parameters characterizing the distribution of covariates Z. Then the likelihood (A3) of Appendix A can be written in the form

$$L(\beta, \mu) \propto \prod_{i \in \mathscr{C}} p(T_i, E_i|Z_i, X_i; \beta) p(Z_i|X_i; \mu) \times \prod_{i \in \mathscr{C} \setminus \mathscr{C}} \sum_{z_i} p(T_i, E_i|z_i, X_i; \beta) p(Z_i = z_i|X_i; \mu) \tag{1}$$

Here the likelihood expression for (T_i, E_i) can be defined in terms of the outcome specific hazard functions:

$$p(T_i, E_i|Z_i, X_i; \beta) \propto \prod_{k=1}^{K} [\lambda_k(T_i|Z_i, X_i; \beta_k)]^{I(E_i = k)} \exp\left\{ - \int_0^{T_i} \sum_{k=1}^{K} \lambda_k(t|Z_i, X_i; \beta_k) \, dt \right\} \tag{2}$$

DOI: 10.1002/sim

Chen and Little [28], Martinussen [29], Scheike and Martinussen [21], and Scheike and Juul [22] recommended use of the EM-algorithm to maximize the Cox partial likelihood in different situations involving missing covariate data. Kulathinal and Arjas [23] proposed a fully parametric solution using Bayesian data augmentation. The Bayesian approach may be preferred at least in situations where the partially observed covariate is continuous; in this case the EM-solution requires either discretization of the covariate or numerical integration in the expectation step. For maximum likelihood estimation, instead of EM-algorithm, we propose direct maximization of the observed data likelihood of the form (1) with respect to parameters β and μ, defining (2) using a parametric survival model. This has the advantage that asymptotic standard errors can be obtained directly by inverting the observed information matrix. Also, as is seen in Section 4, the same approach is directly applicable to a likelihood expression resulting from left truncated data.

3. METHODS BASED ON WEIGHTED PARTIAL/PSEUDOLIKELIHOODS

In addition to the likelihood-based analysis of the complete cohort $\mathscr{C}$, another potentially useful method for reusing controls is based on weighted Cox partial likelihood expressions for the set of cases and controls $\mathscr{O}$, where the original time-matched control sets are pooled together. This possibility was first noted in the discussion by Samuelsen [19]. In a general form, the weighted Cox partial likelihood expression for multiple endpoints can be written as

$$\prod_{k=1}^{K}\prod_{i\in\mathscr{E}_k}\frac{w_i(T_i)\lambda_k(T_i|Z_i,X_i;\beta_k)}{\sum_{j\in\mathscr{O}}w_j(T_i)Y_j(T_i)\lambda_k(T_i|Z_j,X_j;\beta_k)} \tag{3}$$

where $Y_j(T_i)$ is the at-risk indicator at event time T_i and $w_j(T_i)$ is a possibly time-dependent weight for individual j. Here the denominator includes the cases of all types and the controls selected for cases of types $k\in\mathscr{D}$. The cases are overrepresented in the set $\mathscr{O}$, and thus using unit weights would result in biased estimates. Under the CC design the subcohort is in itself a representative sample from the original study cohort. For this design Prentice [10] proposed a weighting scheme where the cases not included in the subcohort are removed from the risk sets and the subcohort members are given unit weights. A slightly different approach, applicable in both the CC and NCC settings, is to give the cases unit weights and weigh the remaining controls by the inverses of their probabilities of being included in a sample as a control. These weights correspond to conditional probabilities $\pi_i=p(O_i=1|\mathscr{F})$, that is, the probabilities of being included in the set $\mathscr{O}$, given all the data observed on the cohort. This probability is one for all cases of interest, and for the non-cases it is the inclusion probability in the group of controls. Using such weights, the weighted partial likelihood (3) can be written in the form

$$\prod_{k=1}^{K}\prod_{i\in\mathscr{E}_k}\frac{\lambda_k(T_i|Z_i,X_i;\beta_k)}{\sum_{j\in\mathscr{E}}Y_j(T_i)\lambda_k(T_i|Z_j,X_j;\beta_k)+\sum_{j\in\mathscr{S}\setminus\mathscr{E}}\dfrac{1}{\pi_j}Y_j(T_i)\lambda_k(T_i|Z_j,X_j;\beta_k)} \tag{4}$$

If only the parameters β_k associated with a specific disease k are of interest, the partial likelihood contributions from the other outcomes can be omitted and the resulting weighted partial likelihood expression is a product over the set of cases $\mathscr{E}_k$. However, cases of the other types still contribute

DOI: 10.1002/sim

to the risk sets with weight one. In a situation where the covariate of interest is collected only for one outcome, say k, the weighted partial likelihood can be written using only the set $\mathscr{E}_k \cup \mathscr{S}$ as

$$\prod_{i\in\mathscr{E}_k} \frac{\lambda_k(T_i|Z_i, X_i; \beta_k)}{\sum_{j\in\mathscr{E}_k} Y_j(T_i)\lambda_k(T_i|Z_j, X_j; \beta_k) + \sum_{j\in\mathscr{S}\setminus\mathscr{E}_k} \frac{1}{\pi_j} Y_j(T_i)\lambda_k(T_i|Z_j, X_j; \beta_k)} \tag{5}$$

where the cases of the other types do not contribute to the risk sets. This is the standard semiparametric pseudolikelihood expression for a single endpoint, the set $\mathscr{S}$, and the inclusion probabilities depending on the cohort sampling design used [19, 30].

Two different approaches have been proposed for estimating the inclusion probabilities. The first estimates the theoretical inclusion probability for the sampling procedure used in the control selection. For example, Samuelsen [19] suggested an expression for the inclusion probability in risk set sampling. The generalization of this for multiple endpoints with separate control selections can be written as

$$\hat{\pi}_i = 1 - \prod_{k\in\mathscr{D}} \prod_{j\in\mathscr{E}_k\setminus\{i\}} \left[1 - \frac{m_k Y_i(T_j)}{\sum_{l\in\mathscr{C}} Y_l(T_j) - 1}\right] \tag{6}$$

Here m_k is the number of controls selected per case of type k and $\sum_{l\in\mathscr{C}} Y_l(T_j)$ is the size of the risk set in the cohort at event time T_j. It should be noted that the use of formula (6) does not require that the sets $\mathscr{E}_k$, $k \in \mathscr{D}$, are mutually exclusive, like they are in our competing risks style notation. With stratified sampling, estimator (6) could be applied within each sampling stratum. When very close matching has been applied and the sampling strata are small, it may not be sensible to try to estimate the theoretical sampling probabilities. An extreme case would be a deterministic nearest neighbour type of matching where, given the cohort data, sampling probabilities are one for all selected controls and zero for the others.

An approach of a different kind would be to discard the procedure originally used in the control selection, but use in the estimation of the inclusion probabilities only the observed data in $\mathscr{F}$ and the realized sample of controls. Kim and De Gruttola [11] suggested using predictive probabilities from a parametric model, like logistic regression. Samuelsen *et al.* [24] proposed poststratification where the set of cases define their own stratum and the remaining controls are stratified according to follow-up time and possible matching covariates. Using the current notations, the inclusion probability in the poststratum $\mathscr{P}$ for individual $i \in \mathscr{P}$ is given by the sampling fraction

$$\hat{\pi}_i = \frac{\sum_{j\in\mathscr{S}\setminus\mathscr{E}} 1_{\{j\in\mathscr{P}\}}}{\sum_{j\in\mathscr{C}\setminus\mathscr{E}} 1_{\{j\in\mathscr{P}\}}} \tag{7}$$

The above stratification corresponds to the multiple endpoint form (4) of the weighted partial likelihood while using a set $\mathscr{E}_k$ in the place of $\mathscr{E}$ would correspond to the single endpoint form (5). For the estimates (7) to remain precise, it seems clear that the stratification should not be too fine. Therefore, parametric models may be better suited to situations where several matching factors need to be considered.

In addition to semiparametric weighted partial likelihood expressions discussed above, the same weighting approach can also be applied for constructing a parametric pseudolikelihood expression

DOI: 10.1002/sim

[19, 30]. Here the true log-likelihood for set $\mathscr{C}$ is approximated with a weighted log-likelihood for set $\mathscr{C}$:

$$\sum_{i\in\mathscr{C}} \frac{1}{\pi_i} \log p(T_i, E_i|Z_i, X_i; \beta)$$

$$= \sum_{i\in\mathscr{C}} \left(\sum_{k=1}^{K} 1_{\{E_i=k\}} \log \lambda_k(T_i|Z_i, X_i; \beta_k) - \int_0^{T_i} \sum_{k=1}^{K} \lambda_k(t|Z_i, X_i; \beta_k)\, \mathrm{d}t \right)$$

$$- \sum_{i\in\mathscr{S}\setminus\mathscr{C}} \frac{1}{\pi_i} \int_0^{T_i} \sum_{k=1}^{K} \lambda_k(t|Z_i, X_i; \beta_k)\, \mathrm{d}t \tag{8}$$

It should be noted that none of the expressions discussed in this section possess the properties of a likelihood and therefore general asymptotic results are not available for the resulting estimators. Possibilities for variance estimation are briefly discussed in Appendix B.

4. CHOICE OF TIME SCALE

In the previously cited literature the choice of the main time scale in a time-to-event analysis has received relatively little attention. In the methods discussed above the controls are no longer tied to their respective cases. This allows one to model and analyse the data from a perspective of a time scale, which is different from the original one that was used in the risk set sampling of the controls. Korn *et al.* [31] and Thiébaut and Bénichou [32] have strongly argued that in epidemiologic applications age should be used as the main time scale instead of time-on-study. This is a reasonable suggestion, as most variation in the hazard rate for many diseases happens over age and invariably any model would require proper adjustment for age. However, some additional complications are encountered due to left truncation at the ages of the eligible subjects when they are recruited to the cohort and their follow-up starts. In likelihood-based analysis left truncation is equivalent to conditioning the likelihood expression with a selection rule. When age is used as the main time scale, this condition is $T_i \geqslant b_i$, that is, the observed age T_i, when the first event or censoring for individual i occurs, is greater than his/her age b_i at the start of the follow-up. The conditional likelihood contribution for individual $i \in \mathscr{C}$ becomes

$$p(T_i, E_i, Z_i|X_i, T_i \geqslant b_i; \beta, \mu) = \frac{p(T_i, E_i, Z_i|X_i; \beta, \mu)}{p(T_i \geqslant b_i|X_i; \beta, \mu)}$$

$$= \frac{p(T_i, E_i|Z_i, X_i; \beta)\, p(Z_i|X_i; \mu)}{\sum_{z_i} p(T_i \geqslant b_i|z_i, X_i; \beta)\, p(Z_i = z_i|X_i; \mu)} \tag{9}$$

and the complete likelihood expression is then

$$L(\beta, \mu) \propto \prod_{i\in\mathscr{C}} \frac{p(T_i, E_i|Z_i, X_i; \beta)\, p(Z_i|X_i; \mu)}{\sum_{z_i} p(T_i \geqslant b_i|z_i, X_i; \beta)\, p(Z_i = z_i|X_i; \mu)}$$

$$\times \prod_{i\in\mathscr{C}\setminus\mathscr{C}} \frac{\sum_{z_i} p(T_i, E_i|z_i, X_i; \beta)\, p(Z_i = z_i|X_i; \mu)}{\sum_{z_i} p(T_i \geqslant b_i|z_i, X_i; \beta)\, p(Z_i = z_i|X_i; \mu)} \tag{10}$$

DOI: 10.1002/sim

The numerator in (9) is defined in (2) and the denominator is the probability of survival without events of interest up to age b_i, a probability that obviously depends on the model parameters.

If the covariates Z_i were observed on every individual $i \in \mathscr{C}$, parameters μ would not need to be estimated and the likelihood expression for time-to-event data could be written conditionally on all covariate data. In this case the conditional likelihood expression simplifies, and left truncation becomes equivalent to excluding all follow-up time before b_i (see, for example, Guo [33, p. 229] and the references therein). For this reason, the methods described in Section 3, which utilize only the set $\mathscr{C}$, require no special adjustment for left truncation; the risk sets are defined from the age at the start of the follow-up till the age when an event or right censoring occurs. For likelihood-based analysis with missing covariate data for set $\mathscr{C} \setminus \mathscr{C}$, the likelihood expression for left truncated data remains in a non-standard form. However, an expression such as (10) can still be used as a likelihood as long as the denominator term can be numerically evaluated. The details of the parameter estimation are described in Appendix B.

5. ILLUSTRATION

In this section we use a real cohort as a basis of a simulation study with which we illustrate and compare different methods. Given the observed follow-up data (T_i, E_i) and other covariate data X_i for this cohort, we simulate the values of an additional covariate Z_i and carry out NCC sampling from the cohort. This approach should provide a more realistic setting for comparison than completely simulated data.

5.1. Example cohort

The cohort we consider in our illustration consists of 5073 men from southern and western Finland, who originally belonged to the placebo group of the ATBC cancer prevention study [34]. The cohort is included in the MORGAM Project, an international pooling of cardiovascular cohorts [35]. Collection of whole blood samples was carried out in 1992–1993 when the men were 54–77 years old, and the cohort was followed up for cardiovascular endpoints until the end of the year 1999. Measurements of classical risk factors for cardiovascular diseases, such as blood pressure, cholesterol, and body composition, are available on all cohort members. Here we consider two types of endpoints, acute myocardial infarction (MI) and ischaemic stroke (IS). After exclusion of 680 individuals with documented or self-reported history of cardiovascular disease at cohort baseline and 25 individuals with incomplete covariate measurements, event type classification was done based on the type of the first outcome event that occurred during the follow-up, using only follow-up from age 55 to age 80. This resulted in 361 MI cases ($E_i = 1$), 192 IS cases ($E_i = 2$), and 3815 right censored observations ($E_i = 0$), either due to the end of the follow-up period or death due to cause other than MI or IS.

Even though the MORGAM Project has opted for the CC design for its genetic substudy, in part because the analysis of several endpoints had already been planned at the design stage, here we consider a scenario where a NCC design has been applied for MI to collect genotypic or biomarker covariates, and the same set of controls is then used for an analysis of IS ($K = 2$ and $\mathscr{D} = \{1\}$). Because the two diseases share many risk factors and can be assumed to be highly dependent, we examine the possible advantages of analysing them simultaneously in a situation where the same covariate would be collected for both outcomes. Here we consider covariate effects in a situation in

DOI: 10.1002/sim

which the most important known risk factors for cardiovascular diseases have been already taken into account by including them in the X_i covariates.

5.2. *Simulation study*

As described above, the values of the additional covariate Z_i were simulated from a conditional distribution, where the conditioning was based on the observed data from the example cohort. For this, we used the probability model (9) and defined the survival model (2) as the proportional hazards Weibull regression model where the cause-specific hazard function has the form

$$\lambda_k(t|Z_i, X_i; \beta_k) = \alpha_k \kappa_k (\alpha_k t)^{\kappa_k - 1} \exp\{\gamma_k Z_i + \eta_k' X_i\} \tag{11}$$

where β_k is a collection of all model parameters. The population distribution of the simulated covariate is defined as $p(Z_i|X_i; \mu) = \mu^{Z_i}(1-\mu)^{1-Z_i}$. In covariates X_i we included smoking status at the time of baseline examination, mean of systolic and diastolic blood pressure, non-HDL cholesterol, HDL cholesterol, and body mass index. Given the observed data on (T_i, E_i, X_i) and parameters (β, μ), the binary covariate for individual $i \in \mathscr{C}$ can be sampled from the conditional distribution

$$p(Z_i|T_i, E_i, X_i; \beta, \mu) = \frac{p(T_i, E_i|Z_i, X_i; \beta)p(Z_i|X_i; \mu)}{\sum_{z_i \in \{0,1\}} p(T_i, E_i|z_i, X_i; \beta)p(Z_i = z_i|X_i; \mu)} \tag{12}$$

Fixing the regression coefficients γ_k in (11) and the frequency μ, parameters η_k, κ_k and α_k and covariates Z_i were simulated using Markov chain Monte Carlo sampling. When simulating the data, the regression coefficient for the partially observed covariate Z_i was set equal for the MI and IS endpoints. Thousand data sets of covariates Z_i were produced with different values of γ_k and μ and sampling of controls and parameter estimation were carried out for each of these. All analyses of data used the same observed covariates X_i.

There would be several alternative ways to conduct a risk set sampling from the cohort. Here we consider the simple design where a single control was sampled for each of the 361 MI cases from the cohort at risk at the calendar time of each event, using date of birth as a matching factor. Two different matching alternatives are considered: random sampling within five-year date of birth groups and nearest neighbour matching for the date of birth. In the latter case, if the same nearest date occurred several times, the control was selected randomly from these.

For the survival analysis of the above data we used both the likelihood-based approach described in Section 2 and the weighted partial likelihoods described in Section 3. In all cases the model was a proportional hazards survival model either for a single outcome (IS) or for both IS and MI. The latter approach makes use of the covariate information collected for both endpoints whereas the former approach, where the MI events are treated as right censorings, would have to be used in a situation where the covariate of interest is not measured on the MI cases. For the likelihood-based approach we considered two different parametric survival models, the Weibull model and the piecewise constant hazard model using five-year age groups. These models are discussed in more detail in e.g. [36, 37]. For a comparison, we also applied the standard Cox regression analysis in the situation where the covariate of interest would be collected from all members of the cohort.

Of the weighting methods reviewed in Section 3, in addition to Samuelsen's estimate (6), we applied a logistic model and poststratification to estimate the inclusion probabilities in the control group. The logistic model was fitted using the date of birth, and the time under study as

DOI: 10.1002/sim

continuous covariates. Poststratification was carried out using case status, date of birth, and time under study. The intervals used in the stratification of date of birth were [1915, 1920), [1920, 1925), [1925, 1930), [1930, 1935) and [1935, 1940], and of the time under study (in years) [0, 2), [2, 4), [4, 6), and [6, 8]. With all weighting methods we applied both the multiple outcome form of the partial likelihood (4), where also the MI cases contribute to the risk sets with weight one, and the single outcome partial likelihood (5), where the MI cases do not contribute to the risk sets. The same methods were applied with both stratified sampling and nearest neighbour selection, with the exception of formula (6), which is valid only in the former case. Details of parameter estimation are given in Appendix B.

5.3. Results

The main parameter of interest here is the regression coefficient γ_2 for the effect of the partially observed covariate Z_i on the hazard of IS. The results for this parameter from 1000 replications using stratified sampling are shown in Table I, and the results using nearest neighbour matching in Table II. When estimated, results on the population frequency parameter μ are also displayed. Simulations were carried out with different values of μ but the results showed similar behaviour, and therefore only the results for $\mu=0.2$ are reported. All considered weighting alternatives gave very similar results in terms of the variance of the estimates. Being able to estimate the theoretical sampling probabilities for the original sampling procedure did not seem to be important, as the weights estimated with Samuelsen's estimate (6) did not show better performance than the weights based on logistic modelling and poststratification. The latter two also produced reasonable results with the nearest neighbour matching. For a comparison, we also tried poststratification using only the case status and case status with date of birth (results not shown). The less accurate estimation of the inclusion probabilities did not seem to affect the variances of the point estimates but it seemed to induce some bias away from zero. Overall, the two outcome partial likelihood of the form (4) gave slightly lower variances for the point estimates compared with the single outcome partial likelihood (5), suggesting that there is some benefit from utilizing both types of cases in the weighted partial likelihood analysis.

Of the likelihood-based methods, both the Weibull model and the piecewise constant model performed well. The piecewise constant model resulted in slightly smaller variances of the point estimates, which suggests that it fits better to these data. The single outcome models, where the MI cases were treated as right censorings and the simulated covariate was treated as missing for the MI cases, underestimated the population frequency parameter μ when Z_i had an effect on MI. However, this did not seem to cause any bias in the regression coefficient estimates with the parameter values used here. In the special case of $\gamma_1=0$ (MI events are independent of Z_i), it is clear that MI events could be handled as non-informative censoring, without any harm in the estimation of γ_2 and μ. However, when γ_1 is different from zero, MI events give information on the missing Z_i values and if this dependency is not modelled, estimation of μ is biased. With more extreme values of γ_1 this could also cause bias in the estimates of γ_2. To avoid this, simultaneous modelling of multiple endpoints is needed when these have similar risk factors.

In all instances, the likelihood-based analysis gave smaller variances of the regression coefficient estimates than the comparable analysis based on weighted partial likelihood. To check whether this is a result of using parametric survival models, we also repeated the analysis with weighted parametric pseudolikelihoods of the form (8), using both Weibull and piecewise constant models

DOI: 10.1002/sim

Table I. Summary statistics from 1000 replications with stratified sampling: sample mean and standard deviation of point estimates, sample mean of standard error estimates, mean estimate of μ (true value = 0.2).

$\gamma_1 = \gamma_2$	Method	Model/weights	Mean $\hat{\gamma}_2$	St. dev. $\hat{\gamma}_2$	Mean $\widehat{\mathrm{SE}}$	Mean $\hat{\mu}$
Single endpoint models						
0.0	Cohort	Cox	−0.009	0.184	0.182	—
	MLE	Weibull	−0.004	0.227	0.225	0.200
	MLE	Piecewise constant	−0.004	0.225	0.223	0.200
	WPL	Samuelsen [19]	−0.006	0.235	0.234	—
	WPL	Logistic	−0.006	0.235	0.233	—
	WPL	Poststratification	−0.006	0.237	0.232	—
0.3	Cohort	Cox	0.296	0.174	0.172	—
	MLE	Weibull	0.304	0.223	0.219	0.190
	MLE	Piecewise constant	0.299	0.220	0.217	0.189
	WPL	Samuelsen [19]	0.308	0.230	0.228	—
	WPL	Logistic	0.306	0.230	0.227	—
	WPL	Poststratification	0.310	0.232	0.226	—
0.6	Cohort	Cox	0.599	0.170	0.164	—
	MLE	Weibull	0.616	0.231	0.217	0.179
	MLE	Piecewise constant	0.604	0.226	0.215	0.177
	WPL	Samuelsen [19]	0.624	0.237	0.226	—
	WPL	Logistic	0.621	0.237	0.225	—
	WPL	Poststratification	0.629	0.239	0.223	—
Multiple endpoint models						
0.0	Cohort	Cox	−0.009	0.184	0.182	—
	MLE	Weibull	−0.001	0.225	0.220	0.200
	MLE	Piecewise constant	−0.001	0.221	0.218	0.200
	WPL	Samuelsen [19]	−0.004	0.233	0.231	—
	WPL	Logistic	−0.004	0.232	0.229	—
	WPL	Poststratification	−0.004	0.235	0.229	—
0.3	Cohort	Cox	0.296	0.174	0.172	—
	MLE	Weibull	0.308	0.219	0.212	0.199
	MLE	Piecewise constant	0.297	0.214	0.209	0.196
	WPL	Samuelsen [19]	0.308	0.227	0.224	—
	WPL	Logistic	0.308	0.226	0.223	—
	WPL	Poststratification	0.308	0.229	0.222	—
0.6	Cohort	Cox	0.599	0.170	0.164	—
	MLE	Weibull	0.620	0.220	0.205	0.198
	MLE	Piecewise constant	0.596	0.214	0.203	0.192
	WPL	Samuelsen [19]	0.622	0.231	0.220	—
	WPL	Logistic	0.621	0.230	0.219	—
	WPL	Poststratification	0.622	0.233	0.218	—

MLE = maximum likelihood estimation; WPL = weighted partial likelihood estimation.

(results not shown). Here the empirical standard deviations of the point estimates showed marginally better efficiency than the comparable weighted partial likelihoods, the difference being around 0.001. This indicates that better efficiency of the likelihood-based estimation results mostly from

DOI: 10.1002/sim

Table II. Summary statistics from 1000 replications with nearest neighbour matching: sample mean and standard deviation of point estimates, sample mean of standard error estimates, mean estimate of μ (true value = 0.2).

$\gamma_1=\gamma_2$	Method	Model/weights	Mean $\hat{\gamma}_2$	St. dev. $\hat{\gamma}_2$	Mean $\widehat{\mathrm{SE}}$	Mean $\hat{\mu}$
Single endpoint models						
0.0	Cohort	Cox	−0.009	0.184	0.182	—
	MLE	Weibull	0.003	0.224	0.223	0.199
	MLE	Piecewise constant	0.003	0.222	0.222	0.199
	WPL	Logistic	0.002	0.235	0.231	—
	WPL	Poststratification	0.001	0.238	0.230	—
0.3	Cohort	Cox	0.296	0.174	0.172	—
	MLE	Weibull	0.318	0.220	0.217	0.188
	MLE	Piecewise constant	0.313	0.216	0.216	0.187
	WPL	Logistic	0.322	0.229	0.225	—
	WPL	Poststratification	0.324	0.231	0.224	—
0.6	Cohort	Cox	0.599	0.170	0.164	—
	MLE	Weibull	0.635	0.222	0.215	0.176
	MLE	Piecewise constant	0.623	0.217	0.213	0.175
	WPL	Logistic	0.641	0.231	0.222	—
	WPL	Poststratification	0.647	0.234	0.221	—
Multiple endpoint models						
0.0	Cohort	Cox	−0.009	0.184	0.182	—
	MLE	Weibull	0.004	0.220	0.218	0.199
	MLE	Piecewise constant	0.003	0.216	0.216	0.199
	WPL	Logistic	0.002	0.232	0.227	—
	WPL	Poststratification	0.001	0.234	0.227	—
0.3	Cohort	Cox	0.296	0.174	0.172	—
	MLE	Weibull	0.313	0.213	0.209	0.199
	MLE	Piecewise constant	0.302	0.209	0.207	0.195
	WPL	Logistic	0.316	0.225	0.220	—
	WPL	Poststratification	0.315	0.227	0.220	—
0.6	Cohort	Cox	0.599	0.170	0.164	—
	MLE	Weibull	0.621	0.211	0.202	0.198
	MLE	Piecewise constant	0.598	0.206	0.200	0.191
	WPL	Logistic	0.628	0.226	0.216	—
	WPL	Poststratification	0.628	0.228	0.216	—

MLE = maximum likelihood estimation; WPL = weighted partial likelihood estimation.

factors other than the use of parametric survival models. The likelihood-based approach makes use of the covariate values X_i collected from all cohort members, which means that the regression coefficients for these are estimated more precisely. This may in part explain the smaller variance of the coefficients for covariate Z_i. Further efficiency gains with the likelihood-based approach can be achieved if the X_i covariates include correlates of Z_i and this dependency is modelled as a part of the likelihood [23].

DOI: 10.1002/sim

6. DISCUSSION

In this paper we formulated the problem of utilizing existing data from a previous NCC study for the analysis of a new disease of interest in the framework of the competing risks survival model. In our approach covariate information collected on the controls and on the cases of all diseases under study can be used in the analysis. We proved a general theorem stating that, under a set of commonly satisfied assumptions, the complete cohort likelihood has the same form irrespective of the cohort sampling design used for collecting the covariate data. In a likelihood-based analysis of the complete cohort, the cohort sampling design is handled as a missing data problem. Even though we applied maximum likelihood parameter estimation using the observed data likelihood for the purpose of comparing the different methods, the likelihood-based approach would also allow a fully Bayesian analysis.

Our illustration considered reusing a previously selected control set in the analysis of a new endpoint of interest without a new control selection. Because of this we did not apply the traditional NCC estimator where the risk sets include only the case-specific time-matched controls. However, the methods described in this paper are also applicable to a situation where a separate time-matched selection of controls has been carried out for the new endpoint of interest. In this case utilizing all of the control sets would give efficiency gains compared with the traditional NCC estimator.

We found that the likelihood-based approach gave slightly better efficiency compared with the weighted partial likelihood estimators. It may also be free of some undesirable small sample properties associated with the weighted estimators (cf. the bias away from zero observed in the simulation study of Kulathinal *et al.* [38]). On the other hand, it requires modelling of the distribution of the partially observed covariate, which can be avoided by using the weighted partial likelihood that involves no missing data. Because of this, the weighted partial likelihood methods may have an advantage when time-dependent covariate data are collected under the NCC design, as missing data in time-dependent covariates cannot be easily handled in likelihood-based survival analysis. It depends on the application whether the possible advantages of the likelihood-based analysis outweigh the more demanding implementation. The likelihood-based approach can be recommended in situations where the covariates observed on all cohort members include correlates of the covariate of interest, in which case the extra data that can be utilized in the likelihood-based analysis will further improve the efficiency of the parameter estimation.

In the analysis using weighted partial likelihood, the inclusion probabilities in the pooled control set need to be estimated. Different approaches proposed for estimating these probabilities can be divided into estimators of theoretical sampling probabilities in the procedure used in the original control selection [19] and approximate methods that only make use of the realized control sample and other observed data [11, 24]. We did not find large differences in the performance of the two approaches in a situation where both were applicable. The approximate methods are also applicable in a situation where the controls have been selected using very close or deterministic matching.

When the controls are no longer tied to their original time-matched cases, the time scale to be used can be chosen freely. Here, we have specifically considered the additional complication of left truncation that emerges when age is used as the main time scale. In the weighted partial likelihood analysis, which involves no missing covariate data, no special adjustments are needed for left truncation. In the likelihood-based analysis, the likelihood expression needs to be conditioned on the selection rule that the age at event is greater than the age at the start of the follow-up. This results in an additional correction term in the likelihood contributions. The conditional likelihood

DOI: 10.1002/sim

expression can be used for parameter estimation like any likelihood as long as the correction term can be numerically evaluated.

Finally, we wish to emphasize that reusing an existing control group as reference subjects for cases of another disease is reasonable only if the necessary risk factors, confounders, and modifiers have been measured on both the old controls and the new cases, with a similar quality of measurements, and with sufficient stability of the material on which measurements are done. For example, the comparability of measurements between cases and controls is a concern in studies addressing the effects of biomarkers measured from frozen biological material. In a NCC study when fresh time-matched controls are sampled for each case, and the necessary assays are performed simultaneously for the whole case–control set, the effects of analytic batch, storage time, and freeze–thaw cycles on these measurements can be controlled. When these effects are present, Rundle *et al.* [39] argue that the CC design is less suitable. This argument would also apply to the reuse of controls from a previous NCC study with cases of a new disease.

APPENDIX A. LIKELIHOOD EXPRESSION FOR COHORT DATA WITH PARTIALLY OBSERVED COVARIATES

Here we show that the likelihood expression for the data described in Section 2 can be written as a product of independent contributions. Let $\widetilde{Z}$ denote the partially observed vector of covariates Z. This notation means that $\widetilde{Z}$ and Z are of same dimension with $\widetilde{Z}_i = Z_i$ for $i \in \mathscr{C}$ and $\widetilde{Z}_i$ unobserved for $i \in \mathscr{C} \setminus \mathscr{C}$. In the following Z is considered discrete, but the formulae could also be generalized to the continuous case. Covariates X are treated as fixed and are always included in the condition of the likelihood. Our aim is to demonstrate that under the assumptions *A1* and *A2*, the likelihood expression for the complete cohort is of the same form regardless of what kind of sampling procedure is used to determine the set $\mathscr{C}$. The following result is needed.

Lemma
$p(Z|T, E, O, X; \theta) = \prod_{i \in \mathscr{C}} p(Z_i|T_i, E_i, X_i; \theta)$.

Proof

$$p(Z|T, E, O, X; \theta) \overset{A2}{=} p(Z|T, E, X; \theta) = \frac{p(Z, T, E, X|\theta)}{p(T, E, X|\theta)}$$

$$\overset{A1}{=} \frac{\prod_{i \in \mathscr{C}} p(Z_i, T_i, E_i, X_i|\theta)}{\prod_{i \in \mathscr{C}} p(T_i, E_i, X_i|\theta)} = \prod_{i \in \mathscr{C}} \frac{p(Z_i, T_i, E_i, X_i|\theta)}{p(T_i, E_i, X_i|\theta)} = \prod_{i \in \mathscr{C}} p(Z_i|T_i, E_i, X_i; \theta) \qquad \square$$

Theorem
The likelihood expression for observed data $(T, E, O, \widetilde{Z})$ is

$$L(\theta) = p(T, E, O, \widetilde{Z}|X; \theta)$$

$$\propto \prod_{i \in \mathscr{C}} p(T_i, E_i, Z_i|X_i; \theta) \prod_{i \in \mathscr{C} \setminus \mathscr{C}} \sum_{z_i} p(T_i, E_i, Z_i = z_i|X_i; \theta)$$

DOI: 10.1002/sim

Proof

$$
\begin{aligned}
L(\theta) &= p(T, E, O, \widetilde{Z}|X; \theta) \\
&= \sum_z p(T, E, O, \widetilde{Z}, Z=z|X; \theta) \\
&= p(O|T, E, X; \theta)p(T, E|X; \theta)\sum_z p(\widetilde{Z}|T, E, O, z, X; \theta)p(Z=z|T, E, O, X; \theta) \\
&\overset{A2}{=} p(O|\mathscr{F})p(T, E|X; \theta)\sum_z p(\widetilde{Z}|T, E, O, z, X; \theta)p(Z=z|T, E, O, X; \theta) \\
&\propto p(T, E|X; \theta)\sum_z p(\widetilde{Z}|T, E, O, z, X; \theta)p(Z=z|T, E, O, X; \theta) \qquad \text{(A1)}
\end{aligned}
$$

The first term inside the sum in (A1) connects the underlying covariate vector Z and observed vector $\widetilde{Z}$ to each other and can be expressed as

$$
\begin{aligned}
p(\widetilde{Z}|T, E, O, z, X; \theta) &= \prod_{i\in\mathscr{C}}[1_{\{O_i=1\}}1_{\{\widetilde{Z}_i=z_i\}} + 1_{\{O_i=0\}}1_{\{\widetilde{Z}_i=\emptyset\}}] \\
&= \prod_{i\in\mathscr{C}} 1_{\{\widetilde{Z}_i=z_i\}} \prod_{i\in\mathscr{C}\setminus\mathscr{C}} 1_{\{\widetilde{Z}_i=\emptyset\}} \qquad \text{(A2)}
\end{aligned}
$$

The latter part includes the indicators for $\widetilde{Z}_i$ taking a missing value for the set $\mathscr{C}\setminus\mathscr{C}$. Using the lemma and the identity (A2), the likelihood (A1) can now be written as

$$
\begin{aligned}
L(\theta) &\propto p(T, E|X; \theta)\sum_z\left[\prod_{i\in\mathscr{C}} 1_{\{\widetilde{Z}_i=z_i\}} \prod_{i\in\mathscr{C}\setminus\mathscr{C}} 1_{\{\widetilde{Z}_i=\emptyset\}} \prod_{i\in\mathscr{C}} p(Z_i=z_i|T_i, E_i, X_i; \theta)\right] \\
&= p(T, E|X; \theta)\prod_{i\in\mathscr{C}}\sum_{z_i} p(Z_i=z_i|T_i, E_i, X_i; \theta)1_{\{\widetilde{Z}_i=z_i\}} \prod_{i\in\mathscr{C}\setminus\mathscr{C}}\sum_{z_i} p(Z_i=z_i|T_i, E_i, X_i; \theta) \\
&= \prod_{i\in\mathscr{C}} p(T_i, E_i|X_i; \theta)\prod_{i\in\mathscr{C}} p(Z_i|T_i, E_i, X_i; \theta)\prod_{i\in\mathscr{C}\setminus\mathscr{C}}\sum_{z_i} p(Z_i=z_i|T_i, E_i, X_i; \theta) \\
&= \prod_{i\in\mathscr{C}} p(T_i, E_i, Z_i|X_i; \theta)\prod_{i\in\mathscr{C}\setminus\mathscr{C}}\sum_{z_i} p(T_i, E_i, Z_i=z_i|X_i; \theta) \qquad \text{(A3)}
\end{aligned}
$$

□

APPENDIX B: DETAILS OF THE PARAMETER ESTIMATION

The use of EM-algorithm would not simplify the denominator terms in the likelihood expression (10) for left truncated data. Because of this, it is more straightforward to carry out the likelihood-based parameter estimation by maximizing numerically the observed data likelihood expressions of the type (10) with respect to the regression coefficients, baseline hazard parameters, and the parameters defining the distribution of the partially observed covariate. The use of (10) as a likelihood requires that it can be evaluated numerically, which can be difficult in practice if Z_i involves several continuous covariates. In that situation the integrals over Z_i could be approximated using Monte Carlo integration by sampling from the distribution $p(Z_i|X_i; \mu)$.

DOI: 10.1002/sim

The weighted expressions of Section 3 are generalizations of the pseudolikelihoods of Kalbfleisch and Lawless [30] and Samuelsen [19] for multiple endpoints. The traditional nested case–control estimator, discussed in e.g. [7], where the risk sets include only the case itself and the case-specific sampled controls, is a true conditional probability. However, for expressions such as (4), where the control sets (and other cases) are pooled, this is true only approximately; given that there is an event at T_i and the corresponding weighted risk set, the weighted partial likelihood contribution approximates the probability of the event occurring to individual i. The asymptotic properties of the weighted partial likelihood estimators will depend on the cohort sampling design used. For example, Kalbfleisch and Lawless [30] discuss the asymptotic properties in Bernoulli sampling designs, Samuelsen [19] in the standard risk set sampling and Kim and De Gruttola [11] in matched risk set sampling designs. A more general solution for variance estimation would be to use robust variances. Lin and Ying [40] and Barlow [41] both proposed the same weighted version of the sandwich-type estimator of Lin and Wei [42]. This is also easily applicable in practice as the sandwich estimator is available in the Cox regression procedures in statistical software such as R and SAS, which also allow the definition of a weight variable. The main motivation for the use of this variance estimator with general weighted partial likelihoods such as (3) comes from Barlow [41], who shows that it is equivalent to a jackknife estimator obtained from the influences in the parameter vector when the weighted observations are removed one at a time. Such a jackknife approach would also be applicable for parametric pseudolikelihoods of the form (8).

In our illustration, the maximization of the likelihood expressions of the form (10) was implemented using the `optim` function of the R statistical software, applying the BFGS algorithm [43]. Standard error estimates were obtained by inverting the numerically differentiated information matrix at the maximum likelihood point. Weighted partial likelihood parameter estimation was carried out using the `coxph` function of the R package 'survival'. For the poststratification-based weighting we applied the variance estimation script presented by Samuelsen *et al.* [24]. For the other weighting schemes, the standard errors were obtained using the `robust=TRUE` definition in the `coxph` function.

ACKNOWLEDGEMENTS

The research of the first author was supported by the GenomEUtwin Project grant from the European Commission under the programme 'Quality of Life and Management of the Living Resource' of 5th Framework Programme and part funded through the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project, grant agreement HEALTH-F4-2007-201413. The research of the second author was supported by the Academy of Finland via its grant numbers 114786 and 122883. The work of the fourth author was partly supported by the EU Network of Excellence *Cancer Control Using Population-based Registries and Biobanks* (CCPRB) and Research Grant for Senior Scientists no. 120146 of the Academy of Finland. The authors would like to thank professor Jarmo Virtamo of the National Public Health Institute for permission to use the ATBC data in our illustration and the reviewers for their helpful and thorough comments.

REFERENCES

1. Dos Santos Silva I. *Cancer Epidemiology: Principles and Methods*. IARC: Lyon, 1999.
2. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology* (3rd edn). Lippincott Williams & Wilkins: Baltimore, 2008.
3. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Statistical Science* 1996; **11**:35–53.

DOI: 10.1002/sim

4. Ury HK. Efficiency of case–control studies with multiple controls per case: continuous or dichotomous data. *Biometrics* 1975; **31**:643–649.
5. Rodrigues L, Kirkwood BR. Case–control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. *International Journal of Epidemiology* 1990; **19**:205–213.
6. Miettinen O. Estimability and estimation in case-referent studies. *American Journal of Epidemiology* 1976; **103**:226–235.
7. Langholz B. Case–control study, nested. In *Encyclopedia of Biostatistics* (2nd edn), Armitage P, Colton T (eds). Wiley: Chichester, 2005; 646–655.
8. Kupper LL, McMichael AJ, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association* 1975; **70**:524–528.
9. Miettinen O. Design options in epidemiologic research. An update. *Scandinavian Journal of Work, Environment and Health* 1982; **8**(Suppl. 1):7–14.
10. Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1–11.
11. Kim S, De Gruttola V. Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial. *Lifetime Data Analysis* 1999; **5**:149–172.
12. Borgan Ø, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case–cohort designs. *Lifetime Data Analysis* 2000; **6**:39–58.
13. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case–cohort studies. *The Annals of Statistics* 1988; **16**:64–81.
14. Langholz B, Thomas DC. Nested case–control and case–cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology* 1990; **131**:169–176.
15. Wacholder S. Practical considerations in choosing between the case–cohort and nested case–control designs. *Epidemiology* 1991; **2**:155–158.
16. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case–cohort designs. *Journal of Clinical Epidemiology* 1999; **52**:1165–1172.
17. Cox DR. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
18. Oakes D. Survival times: aspects of partial likelihood. *International Statistical Review* 1981; **49**:235–264.
19. Samuelsen SO. A pseudolikelihood approach to analysis of nested case–control studies. *Biometrika* 1997; **84**: 379–394.
20. Chen K. Generalized case–cohort sampling. *Journal of the Royal Statistical Society, Series B* 2001; **63**:791–809.
21. Scheike TH, Martinussen T. Maximum likelihood estimation for Cox's regression model under case–cohort sampling. *Scandinavian Journal of Statistics* 2004; **31**:283–293.
22. Scheike TH, Juul A. Maximum likelihood estimation for Cox's regression model under nested case–control sampling. *Biostatistics* 2004; **5**:193–206.
23. Kulathinal S, Arjas E. Bayesian inference from case–cohort data with multiple end-points. *Scandinavian Journal of Statistics* 2006; **33**:25–36.
24. Samuelsen SO, Ånestad H, Skrondal A. Stratified case–cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics* 2007; **34**:103–119.
25. Breslow NE. Statistics in epidemiology: the case–control study. *Journal of the American Statistical Association* 1996; **91**:14–28.
26. Pukkala E, Andersen A, Berglund G, Gislefoss R, Gudnason V, Hallmans G, Jellum E, Jousilahti P, Knekt P, Koskela P, Kyyrönen PP, Lenner P, Loustarinen T, Löve A, Ögmundsdóttir H, Stattin P, Tenkanen L, Tryggvadóttir L, Virtamo J, Wadell G, Widell A, Lehtinen M, Dillner J. Nordic biological specimen banks as basis for studies of cancer causes and control—more than 2 million sample donors, 25 million person years and 100 000 prospective cancers. *Acta Oncologica* 2007; **46**:286–307.
27. Langholz B, Borgan Ø. Counter-matching: a stratified nested case–control sampling method. *Biometrika* 1995; **82**:69–79.
28. Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 1999; **94**:896–908.
29. Martinussen T. Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics* 1999; **26**:479–491.
30. Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine* 1988; **7**:149–160.

DOI: 10.1002/sim

31. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American Journal of Epidemiology* 1997; **145**:72–80.
32. Thiébaut ACM, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine* 2004; **23**:3803–3820.
33. Guo G. Event-history analysis for left-truncated data. *Sociological Methodology* 1993; **23**:217–243.
34. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine* 1994; **330**:1029–1035.
35. Evans A, Salomaa V, Kulathinal S, Asplund K, Cambien F, Ferrario M, Perola M, Peltonen L, Shields D, Tunstall-Pedoe H, Kuulasmaa K. MORGAM (an international pooling of cardiovascular cohorts). *International Journal of Epidemiology* 2005; **34**:21–27.
36. Berzuini C, Clayton D. Bayesian analysis of survival on multiple time scales. *Statistics in Medicine* 1994; **13**:823–838.
37. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. *Controlled Clinical Trials* 2003; **24**:682–701.
38. Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case–cohort design in practice—experiences from the MORGAM project. *Epidemiologic Perspectives and Innovations*, vol. 4, 2007. Available from http://www.epi-perspectives.com/content/4/1/15.
39. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:1899–1907.
40. Lin DY, Ying Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* 1993; **88**:1341–1349.
41. Barlow WE. Robust variance estimation for the case–cohort design. *Biometrics* 1994; **50**:1064–1072.
42. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* 1989; **84**:1074–1078.
43. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2006. http://www.R-project.org.

DOI: 10.1002/sim