1-13-2006

Case-cohort Methods for Survival Data on Families from Routine Registers

Tron Anders Moger
University of Washington/University of Oslo, tronmo@medisin.uio.no

Yudi Pawitan
Karolinska Institutet, Stockholm, Sweden, pawitan@meb.ki.se

Ørnulf Borgan
University of Oslo, borgan@math.uio.no

1 INTRODUCTION

Case-cohort (Prentice [1]) and nested case-control (Thomas [2]) study designs are highly useful to estimate the covariate effects on survival times within a cohort, without having to collect data on covariates for each member of the cohort. In the Nordic region, there are several population-based registers (e.g. the national cancer registers, the Swedish Multi-Generation Register, The Medical Birth Register of Norway and Statistics Norway) which routinely store information on both diseases and possible covariates. By combining information from different registers, one can construct databases which include survival times and times of onset of specific diseases for millions of individuals. By use of the personal identification numbers, the data can be linked into families, providing a great opportunity to study whether a disease shows a significant familial aggregation, and if this aggregation can be explained by some covariates. However, using traditional cohort methods for multivariate survival analysis when handling such vast amounts of data would be computationally extremely time consuming. It would also not be time-efficient, since the diseases of interest are very rare, and one might be interested in including other covariates those readily available in the registers. Since more and more information is stored in today's society, routine registers should also be expected to appear in countries outside the Nordic region. Hence, sampling methods for handling this type of data are needed.

Commonly, models for handling multivariate survival data fall into one of the following two categories: Marginal models and frailty models. When using marginal models, measuring the dependence of individuals within families is not of interest, and the dependence is treated as a nuisance parameter. Instead, one estimates the population average effect of the covariates under the working assumption that all individuals, both within and between families, are independent. One then corrects the standard errors of the parameter estimates by using a variance estimator which takes the dependence into account. In frailty models, both measuring the covariate effects and the the level of dependence is of interest. A frailty variable (which can be constructed as a sum or product of several variables) describes the unobserved random variation and creates dependence between lifetimes. It is usually assumed to act multiplicatively on the hazard function for an individual,

$$\text{Individual hazard rate} = Z \times \lambda(t) \tag{1}$$

where Z is the frailty variable, and $\lambda(t)$ is the baseline hazard. Conditional on the frailty variable, individuals within a family are assumed to be independent. The models may be formulated conditional on the frailty variable, or by means of the marginal distributions, where the frailty is usually integrated out. The latter approach is often called copula models. For copula models, one assumes proportional hazards in the marginal distributions instead of proportional hazards in the conditional distribution. One important difference between the two parameterizations, is that one gets population average effects for the covariates in a copula model (corresponding e.g. to a univariate Cox analysis), while in a frailty model, the regression coefficients are calculated conditional on the value of the frailty variable. For a review of the methods for multivariate survival data, see [3].

For marginal models for multivariate survival data, Lu and Wang [4] propose a case-control method for semi-parametric Cox models. In this case, the challenge is to sample controls in a way that secures independence between the cases and the controls in each sampled cases-control set. If one wishes to measure the strength of the dependence empirically, without any modelling assumptions, Hsu *et al.* ([5] and [6]) propose a case-control method for family data based on the cross-ratio ([7] and [8]). The method is easily implemented in standard software, such as S-Plus. Pfeiffer *et al.* [9] present a marginal survival model for the analysis of first degree relatives of case and control probands sampled from a population register. These methods enable estimation of the population average effect of the proband's risk status on the hazard of disease in the relatives.

If one wishes to model the dependence by a specific probability distribution, Li *et al.* [10] propose a case-control method for parametric copula models. In the paper, they use a gamma distribution for the copula (corresponding to a shared gamma frailty model). Shih and Chatterjee [11] propose a semi-parametric counterpart to the method.

An important question is whether to do the sampling on an individual or family level. In the papers mentioned so far, sampling is done on an individual level, but with the exception of [9], they deal with situations where family registers are unavailable. The sampling of families is done conditional upon first sampling case and control probands. In the case of population based registers, however, we have information on all families, regardless of whether they contain any cases. Here, it would clearly be more logical to sample families instead of individuals. Two papers by Andersen ([12] and [13]) propose a nested case-control method for copula models applied to family register data, illustrated by simulations.

We present two case-cohort alternatives to frailty models, and an extensive simulation study to examine the results. Since one knows the size of the cohort, one can base the case-cohort method on pseudo-likelihoods. The method should be flexible enough to allow for an arbitrary number of members in each family and different kinds of frailty models, including multivariate frailty models for analyzing more general pedigrees with complicated dependence structures. Although frailty models frequently yield quite complicated likelihood functions, the resulting case-cohort likelihood should not be much more complex than it is for cohort data. One should also get a good efficiency compared to a cohort analysis by just sampling a very small (say, 10% or below) proportion of the control families. The idea behind this paper is to apply existing methodology on case-cohort methods for univariate survival data, where all individuals are assumed to be independent, to family data, where families are assumed to be independent, but individuals within a family are dependent. We specifically use methodology from the papers of Kalbfleisch and Lawless [14] and Borgan *et al.* [15]. The main difference is that we sample families instead of individuals.

In Section 2, we present an unstratified case-cohort method with independent Bernoulli sampling (with replacement). A case-cohort method without replacement and stratified sampling is presented in Section 3. Using a different combination (Stratified Bernoulli sampling/ unstratified sampling without replacement) is straightforward. In Section 4 the simulation study is presented, and a small application to data from the Medical Birth Registry of Norway is given in Section 5. A discussion is given in the final Section 6, while derivations of some of the results in Sections 2 and 3 are given in an Appendix. The paper deals with fully parametric models, and a conditional parametrization of the frailty. One should be able to use exactly the same methods for parametric copula models, since they are basically frailty models with a different parameterization.

2 UNSTRATIFIED CASE-COHORT WITH BERNOULLI SAMPLING

Following the notation in [14], let the data consist of N families. Let $\mathbf{X}_1, ..., \mathbf{X}_N$ be independent vectors of random variables, where $\mathbf{X}_i$ contains all available information on events, censorings, covariates etc. for all members in family i, and define $S = \{1, 2, ..., N\}$ as the set of all families. Let $\mathbf{X}_i$ have density $f_i(\mathbf{x})$, the joint density for the individuals in family i, where $f_i(\mathbf{x})$ depends on a vector of unknown parameters $\boldsymbol{\theta}$. This includes parameters for the frailty distribution, the covariates, and in the parametric case, for the baseline hazard. Suppose that S can be divided into 2 mutually exclusive subspaces, $S = S_0 \cup S_1$, where S_0 is the subspace of families with at least one case, and S_1 is the subspace of control families (families with no cases). If $\mathbf{X}_i \in S_j$, then $\mathbf{X}_i$ is observed with probability p_j. Corresponding to the case-cohort scenario of Prentice [1], all case families will be included in the sample with probability $p_0 = 1$. Control families will be included with probability $p_1 < 1$. Of course, one may also wish to sample case families, so that $p_0 < 1$. In that case, the following calculations still apply. The observed data consists $\{\mathbf{X}_i,\ i \in D_j\}$, $j = 0, 1$, where D_j is the set of families sampled from S_j. If $\mathbf{X}_1, ..., \mathbf{X}_N$ are completely observed, the log likelihood is

$$l = \sum_{i=1}^{N} \log f_i(\mathbf{X}_i; \boldsymbol{\theta}) \tag{2}$$

An estimator of the full log likelihood (2) is the log pseudo-likelihood

$$l_p = \sum_{j=0}^{1} \frac{1}{p_j} \sum_{i \in D_j} \log f_i(\mathbf{X}_i; \boldsymbol{\theta}) \tag{3}$$

Define the cohort history $\mathcal{G}$, generated by $\mathbf{X}_1, ..., \mathbf{X}_N$. Let $R_i = I(i \in D_0 \cup D_1)$, and assume that they are independent Bernoulli variates corresponding to sampling with replacement. Define $\pi_i = P(R_i = 1|\mathcal{G})$. The connection to the $p_j^{'}$s in [14] is $\pi_i = p_j I(i \in S_j)$. Notice that

$$l_p = \sum_{j=0}^{1} \sum_{i \in S_j} \frac{R_i}{p_j} \log f_i(\mathbf{X}_i; \boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} \log f_i(\mathbf{X}_i; \boldsymbol{\theta})$$

Then (3) can be rewritten as

$$l_p = l + \sum_{i=1}^{N} \left(\frac{R_i}{\pi_i} - 1 \right) \log f_i(\mathbf{X}_i; \boldsymbol{\theta}) \tag{4}$$

Notice also that $\mathrm{E}(R_i/\pi_i - 1|\mathcal{G}) = 0$, since $\mathrm{E}(R_i|\mathcal{G}) = \pi_i$. Hence, $\mathrm{E}(l_p|\mathcal{G}) = l$, where the expectation is taken over the sampling. By the details in the Appendix, the maximum pseudo-likelihood estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, follows an approximate multinormal distribution with expected value equal to the true $\boldsymbol{\theta}_0$ and covariance matrix $\mathbf{A}(\boldsymbol{\theta}_0)^{-1} + \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-1}$. Here, $\mathbf{A}(\boldsymbol{\theta}_0)$ is estimated by

$$\widehat{\mathbf{A}}(\widehat{\boldsymbol{\theta}}) = \sum_{j=0}^{1} \frac{1}{p_j} \sum_{i \in D_j} I_i(\widehat{\boldsymbol{\theta}}),$$

where $\mathbf{I}_i(\boldsymbol{\theta}) = -\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}' \log f_i(\mathbf{X}_i; \boldsymbol{\theta})$, the observed information matrix for family i, and $\mathbf{B}(\boldsymbol{\theta}_0)$ is estimated by

$$\widehat{\mathbf{B}}(\widehat{\boldsymbol{\theta}}) = \sum_{j=0}^{1} \frac{1 - p_j}{p_j^2} \sum_{i \in D_j} \mathbf{s}_i(\widehat{\boldsymbol{\theta}}) \mathbf{s}_i(\widehat{\boldsymbol{\theta}})^{'}$$

where $\mathbf{s}_i(\boldsymbol{\theta}) = \partial/\partial\boldsymbol{\theta} \log f_i(\mathbf{X}; \boldsymbol{\theta})$, the score function for family i. Hence, the estimator for the covariance matrix of the parameter estimates is the same sandwich-estimator as for univariate data. This completes the case-cohort method with Bernoulli sampling.

The results presented in this section and in the Appendix are also valid when the π_i's are allowed to vary between the control families in S_2. Specifically, to improve efficiency, one may have stratified Bernoulli sampling. The π_i's would then stay constant within strata, but vary between them.

3 STRATIFIED CASE-COHORT WITH SAMPLING WITHOUT REPLACEMENT

As is well-known in case-cohort methods for univariate survival data, one can improve the efficiency of the parameter estimates by stratifying according to additional information, usually covariates, of the members in the cohort (see e.g. [15]). In the case of family survival data, information on family size is another important characteristic available for all. The idea is to divide S into $k+1$ strata with regard to covariate values and family size. In univariate survival data, one would do the stratification on the covariates according to the individual values, but as we now operate on a family level, the stratification on the covariates is done on the mean covariate values of a family. In a frailty analysis of dependent multivariate data, the dependence in survival times is due to unknown, possibly heritable factors, and this dependence is modeled by the frailty variable. It is important to include covariates that explain parts of the dependence, resulting in smaller variance of the frailty variable, and thus less dependence due to unobserved factors. Since such covariates often will have correlated values (or be common to all

members of a family), the mean covariate values should be useful surrogate measures for the exposure level of a family, on which to base the stratification.

As mentioned in the previous section, it is not a problem to stratify the data when doing Bernoulli sampling. However, one may further improve the results by sampling without replacement, which is considered here. As before, the total number of families is N. Let $S = S_0 \cup S_1 \cup ... \cup S_k$, where S_0 is the stratum of all case families, and $S_1, ..., S_k$ are strata of the control families based on the cohort history $\mathcal{G}$. Which stratum a control family belongs to is decided by the covariate values and the family size. One could also divide the case families into different strata, if it is impossible to include all. We then select, by random sampling, m_j families without replacement from the n_j families in stratum j in the cohort, $j = 1, ..., k$. The probability p_j of being sampled from stratum j will now be

$$p_j = \begin{cases} 1 \text{ if } j = 0 \\ \frac{m_j}{n_j} \text{ if } j > 0 \end{cases}$$

Let D_0 denote the families in the case stratum, and let $D_1, ..., D_k$ denote the sampled families from the k control strata. Define $R_i = I(i \in D_0 \cup D_1 \cup ... \cup D_k)$, and let π_i be as defined for Bernoulli sampling. Since the sampling is done with replacement, the R_i's are now dependent variables. With these modifications, the log pseudo-likelihood in (4) will still apply, as this is unaffected by the sampling. However, when calculating the covariance matrix of $\widehat{\boldsymbol{\theta}}$, the fact that the R_i's are dependent has to be taken into consideration. From the derivations given in the Appendix, the maximum pseudo-likelihood estimator $\widehat{\boldsymbol{\theta}}$ follows an approximate multinormal distribution with expected value $\boldsymbol{\theta}_0$ and covariance matrix $\mathbf{A}(\boldsymbol{\theta}_0)^{-1} + \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}_{st}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-1}$. Here, $\mathbf{A}(\boldsymbol{\theta}_0)$ is estimated by

$$\widehat{\mathbf{A}}(\widehat{\boldsymbol{\theta}}) = \sum_{j=0}^{k} \frac{n_j}{m_j} \sum_{i \in D_j} \mathbf{I}_i(\widehat{\boldsymbol{\theta}}),$$

and $\mathbf{B}_{st}(\boldsymbol{\theta}_0)$ is estimated by

$$\widehat{\mathbf{B}}_{st}(\widehat{\boldsymbol{\theta}}) = \sum_{j=0}^{k} \frac{n_j(n_j - m_j)}{m_j^2} \sum_{i \in D_j} (\mathbf{s}_i(\widehat{\boldsymbol{\theta}}) - \overline{\mathbf{s}}_j(\widehat{\boldsymbol{\theta}}))(\mathbf{s}_i(\widehat{\boldsymbol{\theta}}) - \overline{\mathbf{s}}_j(\widehat{\boldsymbol{\theta}}))',$$

where $\overline{\mathbf{s}}_j(\boldsymbol{\theta}) = m_j^{-1} \sum_{i \in D_j} \mathbf{s}_i(\widehat{\boldsymbol{\theta}})$, the estimated average value of the score function in stratum j. This is the same result as in [15] for univariate survival data.

4 A SIMULATION STUDY

To compare the methods, we use a shared gamma frailty model with a Weibull baseline hazard. The standard multiplicative model in (1) is used. The frailty variable Z is gamma distributed with scale and shape parameter δ so that $\mathrm{E}(Z) = 1$, to make sure the model can be identified. The baseline hazard corresponds to a Weibull distribution with scale parameter α and shape parameter κ, of the form $\exp(\alpha)\kappa t^{\kappa-1}$. In addition, a Cox-term for one covariate is included in the model, so that $\lambda(t) = \exp(\alpha + \beta W)\kappa t^{\kappa-1}$, where W denotes a dichotomous covariate in the simulatons. To see how the methods behave for different degrees of correlation within families, we simulate cohorts for three different values of δ: δ =0.1, 0.6 and 2. For a measure of correlation like Kendall's τ, this gives the values τ =0.83, 0.45 and 0.20, corresponding to high, moderate and low dependence within families. In the estimation, we used $\log(\delta)$, which is more stable. The parameter $\boldsymbol{\theta}$ is then $\boldsymbol{\theta} = \{\log(\delta), \alpha, \kappa, \beta\}$.

For each value of δ, 500 cohorts are simulated, with 10000 families in each. The number of members is assigned randomly to the families, so that 30% have 1 member, 45% have two members, 15% have three members, 7% have four members, 2% have five members and 1% have six members (corresponding e.g. to sibships). Each individual is assigned a value of a binary covariate, where $W = 0$ with 70% probability,

and $W = 1$ with 30% probability. We then simulate the survival time of each individual according to the model mentioned above, and draw censoring times from a normal distribution with a mean of 75 and a standard deviation of 10. This yields a range for the simulated survival times from just above zero to 128. The values of the other parameters are $\alpha = -4.50$, $\kappa = 0.46$ and $\beta = -0.693$ ($\exp(\beta) = 0.5$) for all simulations. The censoring rate is 95% for $\delta = 0.1$, 93.7% for $\delta = 0.6$ and 93.5% for $\delta = 2$.

There are f_i members in family i. Let c_{li} indicate whether the survival time t_{li} for individual l in family i is censored ($c_{li} = 0$) or not ($c_{li} = 1$). If we define $c_{\cdot i} = \sum c_{li}$ as the number of events in family i, the log pseudo-likelihood is

$$l_p = \sum_{j=0}^{k} \frac{1}{p_j} \sum_{i \in D_j} \log \left\{ \prod_{l=1}^{f_i} \left[\exp(\alpha + \beta W_{li}) \kappa t_{li}^{\kappa - 1} \right]^{c_{li}} (-1)^{c_{\cdot i}} L_Z^{(c_{\cdot i})} \left(\sum_{l=1}^{f_i} \exp(\alpha + \beta W_{li}) t_{li}^{\kappa} \right) \right\}$$

Here, $L_Z^{(c_{\cdot i})}(\bullet)$ denotes the $c_{\cdot i}$-th derivative of the Laplace transform of Z, $L_Z(s) = [\delta/(\delta + s)]^{\delta}$ (see e.g. Hougaard, 2000, p. 221-222 for details). The number of strata, k, the number of families in D_j, and the value of the p_j's, all depend on the sampling design. From l_p one may calculate the score function and sandwich-estimator for the different sampling methods in the simulations. For each simulated cohort, parameters are estimated both from cohort data and from different case-cohort data. For the case-cohort data, all case families are included, and a certain proportion of the control families. Standard errors for the cohort estimates are calculated from the observed information matrix, and for the case-cohort estimates, from the respective sandwich-estimator.

Table 1: Simulation results when $\log(\delta) = -2.303$, Kendall's τ=0.83. ESE=Empirical standard error, MSE=mean estimated standard error, Eff.=Efficiency compared to cohort estimates, 95% CP=Coverage probability for 95% confidence intervals. Bern=Bernoulli sampling, Strat1=Stratification on family size only, Strat2=Stratification on family size & covariate.

Parameter:	Cohort	10%Bern	5%Bern	5%Strat1	2.5%Strat1	2.5%Strat2	1%Strat2
$\log(\delta)$=-2.303	-2.308	-2.306	-2.306	-2.307	-2.307	-2.307	-2.306
ESE	0.076	0.085	0.095	0.078	0.078	0.077	0.079
MSE	0.080	0.090	0.100	0.081	0.082	0.081	0.082
Eff.	100%	80.7%	64.7%	96.3%	94.7%	97.6%	93.9%
95% CP	96.6%	96.8%	96.4%	96.0%	95.8%	96.8%	95.8%
α=-4.500	-4.503	-4.501	-4.503	-4.503	-4.503	-4.503	-4.502
ESE	0.068	0.074	0.081	0.069	0.073	0.068	0.069
MSE	0.072	0.078	0.085	0.073	0.075	0.072	0.073
Eff.	100%	84.0%	70.5%	97.7%	88.1%	99.7%	98.5%
95% CP	95.6%	96.2%	95.8%	96.6%	95.8%	96.0%	95.8%
κ=0.460	0.461	0.461	0.461	0.461	0.461	0.461	0.461
ESE	0.013	0.013	0.013	0.013	0.013	0.013	0.013
MSE	0.014	0.014	0.014	0.014	0.014	0.014	0.014
Eff.	100%	98.9%	98.6%	99.8%	99.3%	100.0%	100.0%
95% CP	96.8%	97.2%	96.8%	97.0%	97.0%	96.8%	97.0%
β =-0.693	-0.700	-0.700	-0.700	-0.698	-0.697	-0.699	-0.702
ESE	0.089	0.093	0.098	0.103	0.111	0.094	0.098
MSE	0.088	0.093	0.099	0.099	0.109	0.093	0.100
Eff.	100%	91.8%	83.3%	74.8%	64.1%	90.1%	82.5%
95% CP	94.2%	95.4%	95.4%	94.2%	94.6%	95.4%	95.2%

The results are shown in Tables 1-3. There is one table for each value of $\log(\delta)$. The rows of the tables show, for each parameter, the mean value of the parameter from the simulations, the empirical standard error (ESE), the mean estimated standard error (MSE), the efficiency compared to the cohort

estimates, and the empirical proportion of the 95% confidence intervals ($\widehat{\boldsymbol{\theta}} \pm 1.96 \times \text{SE}(\widehat{\boldsymbol{\theta}})$) which cover the true value of the parameter (95% CP). The columns of the tables show results from seven different sampling methods: Cohort, Bernoulli sampling with 10% or 5% of the control families included (10% Bern and 5% Bern), stratified sampling with stratification on family size only and 5% or 2.5% of the control families included from each strata (5% Strat1 and 2.5% Strat1), and stratified sampling with stratification on both family size and covariates and 2.5% respectively 1% of the control families included from each strata (2.5% Strat2 and 1% Strat2).

Table 2: Simulation results when $\log(\delta) = -0.511$, Kendall's τ=0.45. ESE=Empirical standard error, MSE=mean estimated standard error, Eff.=Efficiency compared to cohort estimates, 95% CP=Coverage probability for 95% confidence intervals. Bern=Bernoulli sampling, Strat1=Stratification on family size only, Strat2=Stratification on family size & covariate.

Parameter:	Cohort	10%Bern	5%Bern	5%Strat1	2.5%Strat1	2.5%Strat2	1%Strat2
$\log(\delta)$= -0.511	-0.499	-0.500	-0.488	-0.500	-0.492	-0.499	-0.502
ESE	0.147	0.164	0.189	0.149	0.158	0.151	0.157
MSE	0.142	0.159	0.178	0.146	0.150	0.146	0.150
Eff.	100%	80.2%	60.4%	97.4%	87.0%	95.5%	87.9%
95% CP	95.0%	94.8%	95.0%	95.0%	94.6%	94.8%	93.6%
α =-4.500	-4.505	-4.507	-4.504	-4.505	-4.503	-4.505	-4.506
ESE	0.062	0.069	0.079	0.063	0.069	0.064	0.064
MSE	0.063	0.070	0.077	0.065	0.067	0.064	0.065
Eff.	100%	81.2%	62.5%	97.1%	82.7%	94.4%	93.6%
95% CP	94.2%	95.2%	95.0%	94.8%	94.4%	94.2%	95.0%
κ =0.460	0.461	0.461	0.461	0.461	0.461	0.461	0.461
ESE	0.013	0.013	0.013	0.013	0.013	0.013	0.013
MSE	0.013	0.013	0.013	0.013	0.013	0.013	0.013
Eff.	100%	99.3%	99.8%	100.0%	100.0%	100.8%	97.4%
95% CP	95.0%	94.8%	95.2%	95.4%	95.4%	95.2%	94.6%
β =-0.693	-0.694	-0.692	-0.693	-0.692	-0.697	-0.695	-0.693
ESE	0.076	0.086	0.092	0.094	0.113	0.083	0.099
MSE	0.074	0.084	0.093	0.093	0.109	0.084	0.095
Eff.	100%	77.9%	68.2%	64.4%	44.6%	82.7%	58.5%
95% CP	94.6%	94.6%	95.6%	93.6%	94.6%	95.4%	94.6%

For the case-cohort method with Bernoulli sampling, the efficiency is quite good compared to the cohort estimates for all parameters, when 10% of the control families are included in the sub-cohort. However, it drops significantly when the sampling rate declines to 5%. Since the methods will be used on cohorts consisting of up to millions of families, even a 10% sampling rate may be too high to allow for a time-efficient analysis. The efficiency is also affected by the level of dependence in the data. One gets more precise estimates when the dependence is high. However, the agreement between the empirical standard errors and the standard errors from the sandwich-estimator is good in all cases, and the bias is low. This indicates that the methods for univariate survival data do indeed also work for multivariate data.

For the stratified case-cohort analyses, where the stratification is done according to family size only (Strat1), there are four strata: Families with one, two, three and four+ members. When the sampling rate is 5%, there is a large improvement in the efficiency for the frailty parameter δ and the scale parameter α, compared to the corresponding Bernoulli sampling. This is expected, since these parameters are mainly decided by the level of dependence and the prevalence of the disease. The efficiency is still very good when the sampling rate drops to 2.5%, except for β. Again, there is a good agreement between the empirical standard errors and the standard errors from the sandwich-estimator.

Table 3: Simulation results when $\log(\delta) = 0.693$, Kendall's τ=0.20. ESE=Empirical standard error, MSE=mean estimated standard error, Eff.=Efficiency compared to cohort estimates, 95% CP=Coverage probability for 95% confidence intervals. Bern=Bernoulli sampling, Strat1=Stratification on family size only, Strat2=Stratification on family size & covariate.

Parameter:	Cohort	10%Bern	5%Bern	5%Strat1	2.5%Strat1	2.5%Strat2	1%Strat2
$\log(\delta)$=0.693	0.723	0.740	0.745	0.726	0.742	0.731	0.739
ESE	0.339	0.393	0.417	0.357	0.367	0.360	0.362
MSE	0.331	0.379	0.419	0.343	0.359	0.345	0.357
Eff.	100%	74.6%	66.5%	90.5%	85.5%	89.0%	87.9%
95% CP	94.6%	93.8%	96.4%	94.8%	95.0%	93.6%	95.6%
α =-4.500	-4.503	-4.503	-4.504	-4.503	-4.504	-4.503	-4.502
ESE	0.062	0.068	0.074	0.062	0.065	0.061	0.066
MSE	0.061	0.069	0.077	0.064	0.066	0.062	0.064
Eff.	100%	78.1%	66.6%	93.7%	84.7%	98.2%	82.4%
95% CP	96.6%	96.8%	97.0%	96.4%	95.4%	97.0%	94.6%
κ =0.460	0.461	0.461	0.461	0.461	0.461	0.461	0.461
ESE	0.012	0.012	0.012	0.012	0.012	0.012	0.012
MSE	0.012	0.012	0.012	0.012	0.012	0.012	0.012
Eff.	100%	99.5%	99.3%	100.0%	99.1%	98.2%	96.9%
95% CP	95.2%	95.4%	95.2%	95.2%	95.2%	95.0%	95.6%
β =-0.693	-0.695	-0.693	-0.692	-0.694	-0.689	-0.694	-0.699
ESE	0.076	0.088	0.098	0.098	0.116	0.088	0.106
MSE	0.071	0.083	0.093	0.093	0.112	0.083	0.096
Eff.	100%	73.9%	59.7%	60.4%	43.2%	75.3%	51.8%
95% CP	93.4%	92.6%	93.8%	93.6%	94.0%	93.4%	92.4%

For the stratified case-cohort analyses, where the data are stratified both according to family size and the covariate value (Strat2), there are eight strata in total. Since the mean value of the covariate in each cohort is 0.3, stratification is done according to whether the mean value of the covariate in each family is above or below 0.3. Stratification on family size is unchanged. This gives a large gain in the efficiency for the regression coefficient β, compared to the analyses where stratification was done on family size only. This gain in efficiency is expected, especially since the covariate has a significant effect on the survival. There is also a consistent improvement in the efficiency for δ and α, which perhaps is more surprising. The efficiency is good when just 1% of the control families are sampled.

5 APPLICATION TO DATA ON INFANT MORTALITY IN MULTIPLE BIRTHS

This section shows some simulations on data from the Medical Birth Registry of Norway. The data set has previously been used in [16]. It contains information on deaths during the perinatal (7-364 days) period for all multiple births (twins, triplets etc.) in Norway since 1967. There are 48357 infants in 24077 sibships, and 443 deaths. 23 sibships have two deaths, and one sibship has three deaths.

When working with frailty models in a cohort setting, important questions arise regarding the choice of frailty distribution, and how to model the baseline hazard. In [16], a shared gamma frailty model and a compound Poisson-gamma model, both with Weibull baseline hazards, were fitted to the data. In the compound Poisson-gamma model, individual heterogeneity is modelled by a compound Poisson

distribution, while family heterogeneity is modelled by a gamma distribution. The shared gamma model gave a bad fit to the data, whereas the compound Poisson-gamma model fitted the data quite well.

As a simpler alternative to the compound Poisson-gamma model, one may construct a gamma-gamma model in a similar way. Both family and individual heterogeneity is then described by gamma distributions. Let the individual frailty be modelled by a gamma distribution with scale parameter ν and shape parameter η, with Laplace transform $L(s) = [\nu/(\nu+s)]^{\eta}$. The parameter η describes family heterogeneity and is also gamma distributed, with scale parameter θ and shape parameter δ. Individuals within families are independent given the value of η, and dependence is created by letting η have the same value for individuals within a family. By standard frailty theory, the survival function given η for each individual will be $S(t|\eta) = [\nu/(\nu+\Lambda(t))]^{\eta}$, and the hazard function given η will be $h(t|\eta) = \eta\lambda(t)/(\nu+\Lambda(t))$. Here, $\Lambda(t)$ is the integrated baseline hazard. By using the same notation as in Section 4, and the same likelihood construction as in [16], the log pseudo-likelihood for the gamma-gamma model is

$$l_p = \sum_{j=0}^{k} \frac{1}{p_j} \sum_{i \in D_j} \log \left\{ \prod_{l=1}^{f_i} \left(\frac{\lambda(t_{li})}{\nu + \Lambda(t_{li})} \right)^{c_{li}} (-1)^{c_{\cdot i}} L_{\eta}^{(c_{\cdot i})} \left(\sum_{l=1}^{f_i} \{ \log [\nu + \Lambda(t_{li})] - \log(\nu) \} \right) \right\}$$

where $L_{\eta}^{(c_{\cdot i})}(\bullet)$ denotes the $c_{\cdot i}$-th derivative of the Laplace transform of η, $L_{\eta}(s) = [\delta/(\delta+s)]^{\delta}$. The gamma-gamma model is used because the compound Poisson-gamma model yielded singular information matrices when fitted to the data. It turns out that the gamma-gamma model fits the data almost as well as the compound Poisson-gamma model, indicating that the problem of singularity could be due to over-parameterization (the compound Poisson-gamma has one additional parameter).

In this application, we fit a shared gamma model and a gamma-gamma model to the data, and examine how the case-cohort estimates of the parameters compare to the cohort estimates for these two models. The case-cohort estimates should be close to the cohort estimates regardless of whether the model fits the data or not. Cohort estimates, and case-cohort estimates based on 200 samples from the data, are presented. In the analysis, the scale parameter α of the Weibull baseline hazard is subsumed in the frailty distribution. This is to simplify the derivation of the information matrix and sandwich estimator for the gamma-gamma model. Hence, the baseline hazard $\lambda(t)$ is of the form $\lambda(t) = \kappa t^{\kappa-1}$. The gamma distribution describing family heterogeneity in both models has two parameters; a scale parameter θ and a shape parameter δ. For the gamma-gamma model, one gets an additional parameter from the distribution describing individual heterogeneity, the scale parameter ν. There are no covariates in this application.

Table 4: Results from the analysis of the multiple birth data, using shared gamma and Gamma-gamma frailty models. For the cohort analysis: Par=parameter estimates, SE=estimated standard error. For the 200 10% Bernoulli case-cohort analyses: MPar=mean parameter estimates, MSE=mean estimated standard error, 95%ECI=95% empirical confidence intervals of the parameters.

	Cohort		10% Bernoulli case-cohort			
Parameter	Par	SE	Mpar	95%ECI	MSE	95%ECI
Shared gamma model						
δ	0.0679	0.0147	0.0681	(0.0656, 0.0710)	0.0148	(0.0143, 0.0155)
θ	94.38	23.72	94.49	(93.02, 95.83)	23.77	(23.32, 24.17)
κ	0.4469	0.0210	0.4469	(0.4467, 0.4471)	0.0210	(0.0210, 0.0210)
Gamma-gamma model						
δ	0.0766	0.0169	0.0767	(0.0739, 0.0800)	0.0170	(0.0164, 0.0179)
θ	27.39	9.42	27.42	(26.91, 27.87)	9.43	(9.25, 9.60)
ν	7.59	1.55	7.59	(7.58, 7.61)	1.546	(1.543, 1.550)
κ	0.9367	0.1209	0.9367	(0.9361, 0.9372)	0.1209	(0.1206, 0.1211)

The results are shown in Table 4. The table shows cohort estimates of the parameters (Par) with

standard errors from the observed information matrix (SE), as well as mean parameter estimates from 200 Bernoulli case-cohort samples (MPar) where 10% of the control sibships are included, mean standard errors from the sandwich estimator (MSE), and 95% empirical confidence intervals (95%ECI, based on the 2.5% and 97.5% percentiles of the empirical distribution) for the parameters and the estimated standard errors. Both the shared gamma and the gamma-gamma model are fitted to each sample. The results show that the parameter estimates from the case-cohort analysis are very close to the cohort estimates, whether the model used to analyze the data provides a good or a bad fit. This indicates that the problem of choosing the "right" distribution for the frailty variable or baseline hazard in a case-cohort analysis, should not be any harder than for a cohort analysis. Notice also that the estimates for δ are quite similar for both models. This is expected, since it is the shape parameter of the distribution describing family heterogeneity, and thus the dependence in the data, in both the shared gamma and the gamma-gamma model. For the cohort application, using a shared gamma model yields a log-likelihood value of -4865.00, whereas the log-likelihood is -4840.98 for the gamma-gamma model. For the 200 case-cohort samples, the mean log pseudo-likelihood values are -4864.64 for the shared gamma model, and -4840.61 for the gamma-gamma model. The minimum difference between the log-likelihood values is 23.98. Thus, even though the likelihood ration test does not apply in the usual manner for pseudo-likelihoods, this indicates a rejection of the shared gamma model in all 200 samples. For more details regarding the interpretation of the cohort results from this analysis, see [16].

6 DISCUSSION

When dealing with very large multivariate data sets, sampling techniques are important to keep the computation time manageable, or, to make computations possible at all. The case-cohort methods presented in this paper show that one can achieve very good efficiencies by using only a small proportion of the data set. Both methods demand fairly simple modifications to the likelihood functions and variance estimators of parametric frailty and copula models for family data. As the results show, vast improvements can be made in the efficiency of the parameters by stratifying according to family size and covariates. Stratification is expected to improve efficiency both when doing Bernoulli sampling and sampling without replacement. Stratified sampling works really well in this setting, since register cohorts are huge. This means that one can divide the cohort into almost as many strata as one would like, without fearing that a stratum has too few families in it. On the other hand, unstratified methods are much easier to implement than a stratified method with a large number of strata. For several applications, one might be satisfied with unstratified sampling.

When using the methods considered in this paper, there are four possible combinations of sampling designs: Bernoulli sampling with/without stratification, and sampling without replacement with/without stratification. We chose to show results for unstratified Bernoulli sampling and stratified sampling without replacement, since the first method is expected to give the lowest efficiency, while the second is expected to give the best efficiency out of the four combinations. To see how unstratified sampling without replacement compares to Bernoulli sampling, we did some additional simulations on the cohorts generated in Section 4. When sampling 5% of the control families, the efficiencies of the estimated parameters increased from 65% to 85% for $\log(\delta)$ and from 71% to 89% for α, when the true value of δ was 0.1. When the true value of δ was 0.6, the efficiency increased from 60% to 72% for $\log(\delta)$, and from 63% to 95% for α. Hence, it is beneficial to sample without replacement, at least for the frailty and baseline hazard parameters. There was no improvement for β.

The censoring rate in the simulations in this paper was set to 93.5%-95%. For many diseases the censoring rate is above 99%. Other simulations indicate that the efficiency could depend on the censoring rate. When it is 98% and δ is 0.1 (very high dependence), the efficiency of $\log(\delta)$ and α for unstratified Bernoulli sampling improves to 76% and 81% when sampling 5% of the control families, and 95% and 98% when sampling 10% of the control families. However, since the simulated cohorts consist of just 10000 families, too few familial cases are generated in data with low dependence and an censoring rate of

98%, resulting in a large over-estimation of δ from both the cohort and case-cohort analyses (and worse efficiency for the case-cohort estimates). This is also the reason why the bias of the estimated δ's is larger when the true $\delta = 2$ in the simulations. A problem with the nested case-control method presented in Andersen ([12] and [13]), is that no indications on the efficiency of the estimators were given when a very small proportion of the data were used. In her simulations, the censoring rate was 90%, and one or three control families were sampled per case family, meaning that the proportion of the data actually used in the simulations would be too large for most practical purposes.

The value of the Weibull shape parameter in the simulation study is set to 0.46, meaning that the hazard is decreasing as a function of time. This might appear as a bit odd. Although we have not tried any increasing hazard, the precision of κ seems to be almost unaffected by the sampling. Also, our experience from earlier applications of frailty models with Weibull baseline hazard says that estimating this parameter never causes any problems.

In Borgan *et al.* [15], which concerns stratified case-cohort designs for univariate survival data, the stratification was intended for one covariate of particular interest. For register data, it is not a problem to stratify on several covariates. In the simulation study, the stratification was done on whether or not the mean value of the covariate for a family was above or below the mean value of the empirical covariate distribution. Other possibilities are to construct strata from the median or the quartiles of the empirical distribution of the covariate. Even though the covariate was not correlated within families in the simulation study, stratification on the covariate proved to have a good effect. The effect could be even greater when the covariate value is correlated within families, as the surrogate measure (the mean covariate value in a family) would be more precise. If one for instance considers a common covariate like the birth year of the mother in a sibling study, stratification in the multivariate case would be the same as in the univariate case. The simulations clearly show that one should stratify on the covariates to get precise estimates for the β's, especially when sampling less than 5% of the control families.

Further theoretical work would be to develop a semi-parametric counterpart to the methods presented here. Andersen ([12] and [13]) has proposed a nested case-control method both for parametric and semi-parametric copula models. The weights p_j could enter the pseudo-likelihood in a more complicated manner when the model is semi-parametric. In addition, getting convergence for complicated frailty models with non-parametric baseline hazards can be a bit tricky, at least when working with the conditional parameterization of the frailty.

ACKNOWLEDGEMENTS

Many thanks to the Medical Birth Registry of Norway for providing the multiple birth data, and to the Department of Biostatistics, University of Washington, where most of this work was done.

APPENDIX

In this Appendix, the derivations for the asymptotic properties of $\boldsymbol{\theta}$ are reviewed. First, for case-cohort with Bernoulli sampling of control families, and then for case-cohort where the sampling of control families is without replacement.

From (2), the cohort score function is

$$\mathbf{s}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}} \log f_i(\mathbf{X}_i; \boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{s}_i(\boldsymbol{\theta})$$

For the Bernoulli sampling in Section 2, the pseudo-score function from (4) becomes

$$\mathbf{s}_p(\boldsymbol{\theta}) = \mathbf{s}(\boldsymbol{\theta}) + \sum_{i=1}^{N} \left(\frac{R_i}{\pi_i} - 1 \right) \mathbf{s}_i(\boldsymbol{\theta}) \tag{5}$$

Notice that $\mathbf{s}_p(\boldsymbol{\theta}) = 0$ is an unbiased estimating equation, since the expected value $\mathbf{s}(\boldsymbol{\theta}_0)$ is zero by general likelihood theory, and, by the same argument as for (4), the expected value of the second term in (5) is also zero. To find the covariance matrix of $\mathbf{s}_p(\boldsymbol{\theta}_0)$, $\mathrm{Cov}(\mathbf{s}_p(\boldsymbol{\theta}_0))$, notice that

$$\mathrm{Cov}(\mathbf{s}(\boldsymbol{\theta}_0)) = \sum_{i=1}^{N} \mathrm{E}\left(-\frac{\partial^2 \log f_i(\mathbf{X}_i;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = \sum_{i=1}^{N} \mathbf{I}_i(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0) \tag{6}$$

as usual. Furthermore,

$$\begin{aligned}
&\mathrm{Cov}\left[\sum_{i=1}^{N}\left(\frac{R_i}{\pi_i}-1\right)\mathbf{s}_i(\boldsymbol{\theta}_0)\right] = \sum_{i=1}^{N}\mathrm{E}\left[\left(\frac{R_i}{\pi_i}-1\right)^2 \mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right] \\
= &\sum_{i=1}^{N}\mathrm{E}\left\{\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\mathrm{E}\left[\left(\frac{R_i}{\pi_i}-1\right)^2\middle|\mathcal{G}\right]\right\} = \sum_{i=1}^{N}\mathrm{E}\left(\frac{1-\pi_i}{\pi_i}\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right) \\
= &\sum_{j=1}^{0}\frac{1-p_j}{p_j}\mathrm{E}\left(\sum_{i\in S_j}\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right) = \mathbf{B}(\boldsymbol{\theta}_0),
\end{aligned}$$

exactly the same expression as in [14] for univariate data. Also, the two terms in (5) are uncorrelated, since

$$\mathrm{E}\left[\mathbf{s}(\boldsymbol{\theta})\times\sum_{i=1}^{N}\left(\frac{R_i}{\pi_i}-1\right)\mathbf{s}_i(\boldsymbol{\theta})'\right] = \mathrm{E}\left\{\sum_{i=1}^{N}\mathbf{s}(\boldsymbol{\theta})\times\mathbf{s}_i(\boldsymbol{\theta})'\mathrm{E}\left[\left(\frac{R_i}{\pi_i}-1\right)\middle|\mathcal{G}\right]\right\} = 0$$

Hence, we have showed that $\mathrm{Cov}(\mathbf{s}_p(\boldsymbol{\theta}_0))=\mathbf{A}(\boldsymbol{\theta}_0)+\mathbf{B}(\boldsymbol{\theta}_0)$. Let $\widehat{\boldsymbol{\theta}}$ solve the equation $\mathbf{s}_p(\boldsymbol{\theta}) = 0$. By the usual Taylor expansion around the true value $\boldsymbol{\theta}_0$, one gets

$$\begin{aligned}
0 &= \mathbf{s}_p(\widehat{\boldsymbol{\theta}}) \approx \mathbf{s}_p(\boldsymbol{\theta}_0) + \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{s}_p(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \\
\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0 &\approx \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{s}_p(\boldsymbol{\theta}_0) \\
\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0 &\approx \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{s}_p(\boldsymbol{\theta}_0)
\end{aligned}$$

where $\mathbf{I}(\boldsymbol{\theta}_0)$ is the observed information matrix. By the central limit theorem, $\mathbf{s}_p(\boldsymbol{\theta}_0)$ is approximately normal, with expected value 0 and covariance matrix $\mathbf{A}(\boldsymbol{\theta}_0)+\mathbf{B}(\boldsymbol{\theta}_0)$ by the previous results. The covariance matrix of $\widehat{\boldsymbol{\theta}}$ then becomes

$$\mathbf{A}(\boldsymbol{\theta}_0)^{-1}(\mathbf{A}(\boldsymbol{\theta}_0)+\mathbf{B}(\boldsymbol{\theta}_0))\mathbf{A}(\boldsymbol{\theta}_0)^{-1} = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}+\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-1}$$

Now, consider the case of stratified sampling without replacement in Section 3. The pseudo-score function in (5) still applies, the two terms in (5) will be uncorrelated, and $\mathbf{A}(\boldsymbol{\theta}_0)$ will be the same as in (6), as these results are unaffected by the sampling. However, $\mathbf{B}(\boldsymbol{\theta}_0)$ is affected by the sampling. To find $\mathbf{B}(\boldsymbol{\theta}_0)$ under stratified sampling without replacement, $\mathbf{B}_{st}(\boldsymbol{\theta}_0)$, we use the finite-population large sample argument from Lehmann [17], as used in [18] and [15]. By using the rule of double variance and the result of Example 3, p. 332-333 in [17] for each stratum, we get

$$\begin{aligned}
\mathbf{B}_{st}(\boldsymbol{\theta}_0) &= \mathrm{Cov}\left[\sum_{i=1}^{N}\left(\frac{R_i}{\pi_i}-1\right)\mathbf{s}_i(\boldsymbol{\theta}_0)\right] = \mathrm{E}\left\{\mathrm{Cov}\left[\sum_{i=1}^{N}\left(\frac{R_i}{\pi_i}-1\right)\mathbf{s}_i(\boldsymbol{\theta}_0)\middle|\mathcal{G}\right]\right\}+0 \\
&= \mathrm{E}\left\{\sum_{j=0}^{k}\mathrm{Cov}\left[\sum_{i\in S_j}\left(\frac{R_i}{p_j}-1\right)\mathbf{s}_i(\boldsymbol{\theta}_0)\middle|\mathcal{G}\right]\right\} = \mathrm{E}\left\{\sum_{j=0}^{k}\mathrm{Cov}\left[\sum_{i\in S_j}\frac{R_i}{p_j}\mathbf{s}_i(\boldsymbol{\theta}_0)\middle|\mathcal{G}\right]\right\} \\
&= \sum_{j=0}^{k}\frac{1}{p_j^2}\mathrm{E}\left\{\frac{m_j(n_j-m_j)}{n_j-1}\boldsymbol{\tau}_j(\boldsymbol{\theta}_0)^2\right\},
\end{aligned}$$

where $\boldsymbol{\tau}_j(\boldsymbol{\theta}_0)^2$ is given by

$$\boldsymbol{\tau}_j(\boldsymbol{\theta}_0)^2 = \frac{1}{n_j} \sum_{i \in S_j} (\mathbf{s}_i(\boldsymbol{\theta}_0) - \overline{\mathbf{s}}_j(\boldsymbol{\theta}_0))(\mathbf{s}_i(\boldsymbol{\theta}_0) - \overline{\mathbf{s}}_j(\boldsymbol{\theta}_0))'$$

This follows from Example 1, p. 328-329 in [17]. Here, $\overline{\mathbf{s}}_j(\boldsymbol{\theta}_0) = n_j^{-1} \sum_{i \in S_j} \mathbf{s}_i(\boldsymbol{\theta}_0)$, the mean value of the score function in stratum j. The quantity $\boldsymbol{\tau}_j(\boldsymbol{\theta}_0)^2$ is estimated by inserting $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$, and by using m_j instead of n_j, and by taking the sum over D_j instead of S_j. Since $p_j = m_j/n_j$, and by using the approximation $n_j/(n_j - 1) \approx 1$, $\mathbf{B}_{st}(\boldsymbol{\theta}_0)$ can be estimated by

$$\widehat{\mathbf{B}}_{st}(\widehat{\boldsymbol{\theta}}) = \sum_{j=0}^{k} \frac{n_j(n_j - m_j)}{m_j^2} \sum_{i \in D_j} (\mathbf{s}_i(\widehat{\boldsymbol{\theta}}) - \overline{\mathbf{s}}_j(\widehat{\boldsymbol{\theta}}))(\mathbf{s}_i(\widehat{\boldsymbol{\theta}}) - \overline{\mathbf{s}}_j(\widehat{\boldsymbol{\theta}}))'$$

Assuming that the normality holds, $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ follows an approximate normal distribution with expected value 0 and covariance matrix $\mathbf{A}(\boldsymbol{\theta}_0)^{-1} + \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}_{st}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-1}$.

REFERENCES

1. Prentice RL. A case-cohort design for epidemiological cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1–11.

2. Thomas DC. Addendum to 'Methods of cohort analysis: Appraisal by application to asbestos mining' by F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *Journal of the Royal Statistical Society, Series* **A** 1977; **140**:469–491.

3. Hougaard P. *Analysis of multivariate survival data*, Springer: New York, 2000.

4. Lu SE, Wang MC. Cohort case-control design and analysis for clustered failure-time data. *Biometrics* 2002; **58**:764–772. DOI:10.1111/j.0006-341x.2002.00764.x.

5. Hsu L, Prentice RL, Zhao LP, Fan JJ. On dependence estimation using correlated failure time data from case-control family studies. *Biometrika* 1999; **86**:743–753. DOI:10.1093/biomet/86.4.743.

6. Hsu L, Prentice RL, Stanford JL. Some further results on incorporating risk factor information in assessing the dependence between paired failure times arising from case-control family studies: an application to prostate cancer. *Statistics in Medicine* 2002; **21**:863–876. DOI:10.002/sim.1055.

7. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*1978; **65**:141–151.

8. Oakes D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 1989; **84**:487–493.

9. Pfeiffer RM, Goldin LR, Chatterjee N, Daugherty S, Hemminki K, Pee D, Li X, Gail MH. Methods for testing familial aggregation of diseases in population-based samples: Application to Hodgkin lymphoma in Swedish registry data. *Annals of Human Genetics* 2004; **68**:498–508. DOI:10.1046/j.1529-8817.2003.00111.x.

10. Li H, Yang P, Schwartz AG. Analysis of age of onset data from case-control family studies. *Biometrics* 1998; **54**:1030–1039.

11. Shih JH, Chatterjee N. Analysis of survival data from case-control family studies. *Biometrics* 2002; **58**:502–509. DOI:10.1111/j.0006-341x.2002.00502x.

12. Andersen EW. Composite likelihood and two-stage estimation in family studies. *Biostatistics* 2004; **5**:15–30. DOI:10.1093/biostatistics/5.1.15.

13. Andersen EW. Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis* 2005; **11**:333–350. DOI:10.1007/s10985-005-2966-7.

14. Kalbfleisch JD, Lawless JF. Likelihood analysis of multistate models for disease incidence and mortality. *Statistics in Medicine* 1988; **7**:147–160.

15. Borgan Ø, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis* 2000; **6**:39–58. DOI:10.1023/A:1009661900674.

16. Moger TA, Aalen OO. A distribution for multivariate frailty based on based on the compound Poisson distribution with random scale. *Lifetime Data Analysis* 2005; **11**:41–59. DOI:10.1007/s10985-004-5639-z.

17. Lehmann E. *Nonparametrics*, Holden-Day: San Fransico, 1975.

18. Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* 1997; **84**:379–394.