 Vol 33: 25–36, 2006

Bayesian Inference from Case–cohort Data with Multiple End-points

SANGITA KULATHINAL

Department of Epidemiology and Health Promotion, National Public Health Institute (KTL)

ELJA ARJAS

Department of Mathematics and Statistics, University of Helsinki

ABSTRACT. In a case–cohort design a random sample from the study cohort, referred as a subcohort, and all the cases outside the subcohort are selected for collecting extra covariate data. The union of the selected subcohort and all cases are referred as the case–cohort set. Such a design is generally employed when the collection of information on an extra covariate for the study cohort is expensive. An advantage of the case–cohort design over more traditional case–control and the nested case–control designs is that it provides a set of controls which can be used for multiple end-points, in which case there is information on some covariates and event follow-up for the whole study cohort. Here, we propose a Bayesian approach to analyse such a case–cohort design as a cohort design with incomplete data on the extra covariate. We construct likelihood expressions when multiple end-points are of interest simultaneously and propose a Bayesian data augmentation method to estimate the model parameters. A simulation study is carried out to illustrate the method and the results are compared with the complete cohort analysis.

Key words: Bayesian data augmentation, case–cohort design, cause-specific hazards, follow-up data, genotype data, incomplete data, likelihood, multiple end-points

1. Introduction

In epidemiology, cardiovascular diseases (CVD) have always been a major concern, in spite of the considerable knowledge which has accumulated regarding the CVD risk factors, because of their high incidence and mortality rates. In the present era of genomics, epidemiologists are turning towards genetics in order to understand the role of predisposing genes better, which may then be strongly associated with either the known CVD risk factors or directly to CVD.

Often, in epidemiological studies, risk factor data and disease or event follow-up data are collected for a well-defined population cohort. Also it is now becoming popular to augment genotype data to carry out the systematic association study of the genes, risk factors and the disease. It is expensive to genotype all the cohort members if the cohort is large, which is often the case. Various designs have been suggested where it is not necessary to genotype the whole cohort. One such design is the case–cohort design proposed by Prentice (1986), where genotypes are obtained for a subcohort, a random sample of the original cohort, and for all cases, i.e. those who experience an event of interest. Statistical methods are mostly developed for Cox's proportional hazards model by using either pseudo-partial likelihood (Prentice, 1986; Self & Prentice, 1988; Lin & Ying, 1993; Barlow, 1994) or a class of estimating equations based on the partial likelihood score function (Chen & Lo, 1999; Kong *et al.*, 2004). In Sorensen & Andersen (2000) study, competing risks data sampled using the case–cohort design assuming a Cox regression model for cause-specific hazards. The main focus in their study is on investigating the correlation between the relative risk estimates for a given exposure on different types of end-points when estimated, following Self & Prentice (1988), using a pseudo-likelihood defined for competing causes. In a recent independent work,

Scheike & Martinussen (2004) considered maximum likelihood estimation of relative risks, assuming Cox's regression model, under case–cohort sampling where covariates are observed only for the case–cohort members while event follow-up data are observed for the entire cohort (see also, Scheike & Juul, 2004 for nested case–control sampling). They considered only a single end-point and proposed the expectation maximization (EM)-algorithm to be used to maximize the likelihood based on the entire cohort. They observed that the maximum likelihood estimator performs better compared with the standard case–cohort analysis in all situations, although the gain in efficiency is minor in case of a rare disease. Such performance of their maximum likelihood estimator is due to the information about the missing covariates provided only by the disease. Here, we consider a situation where the risk factors, or equivalently, covariates as well as disease follow-up data, are collected for the entire cohort and genotype data are collected under the setting of a case–cohort design. If the association between the risk factors like lipid measurements and the considered gene(s) is strong, then the risk factors provide supplementary information about the missing genotype, in addition to the association between the gene and the disease. In such a situation, the maximum likelihood estimates of the relative risks obtained using the full likelihood of Scheike & Martinussen (2004) can be expected to be more accurate than those obtained using the standard case–cohort analysis because of the supplemented information even when the considered disease is rare. In our experience the effect of the covariates which are observed for the entire cohort is also estimated more accurately using the full likelihood approach than using the pseudo-partial likelihood for case–cohort data.

In this paper, we consider a case–cohort design where the covariates and the event follow-up data are observed for the entire cohort and genotype data are collected for the case–cohort set, i.e. for a subcohort drawn from the original cohort and for all cases. We define an individual as a case if any of the several possible events of interest is observed during the follow-up. We propose a full likelihood expression for the case–cohort data for multiple end-points under a general survival model and describe a Bayesian data augmentation method for the analysis. Our approach can be seen as an extension of the full likelihood method proposed independently by Scheike & Martinussen (2004) to multiple end-points and to a general survival model. Most of the observations made by Scheike & Martinussen (2004) are applicable here even though, in addition to the follow-up information, we also consider risk factor phenotypes which are observed for the entire cohort. We refer to Scheike & Martinussen (2004) for a detailed comparison of the maximum likelihood estimator obtained using full likelihood with other existing estimators.

This paper is structured as follows. In section 2, we define the case–cohort setting when multiple end-points are considered. In section 3, likelihood expressions are constructed using the complete cohort when the interest is in studying the association between the multiple end-points and baseline variables including genotypes. As the definition of a case at the design level includes multiple end-points, we specify likelihoods when all possible end-points are modelled. We also propose a flexible Bayesian data augmentation method to carry out the parametric or non-parametric survival analysis in section 3. In section 4, a simulation study is performed to illustrate the method proposed in section 3. The paper ends with discussion. The present work was inspired by an ongoing research project MORGAM (http://www.ktl.fi/morgam), in which a genetic study is based on a similar case–cohort design.

2. Case–cohort data with multiple end-points

Consider a cohort $\mathcal{C}$ which is followed up for k end-points, e.g. coronary heart disease, stroke and death from other causes. Baseline variables such as date of the baseline examination,

age at the baseline examination, blood pressure, smoking status, total cholesterol, etc. are collected for the entire cohort. Considering an individual indexed by $i \in \mathcal{C}$, let x_i denote the baseline variables of interest and let G_i denote the corresponding genotypes at the considered loci. Let T_i be the time to the first event of interest and let E_i be an indicator of the type of observed event during the follow-up. The variable E_i assumes one of the values in $\{0, 1, \ldots, k\}$, with 0 referring to a right-censored observation.

We refer to an individual as a case if any event of interest is observed during the follow-up, i.e. $E_i \neq 0$.

A subcohort $\mathcal{S}$, a sample from the original cohort $\mathcal{C}$, is selected randomly by using equal or known unequal sampling probabilities that may depend on the baseline variables. Various sampling techniques have been suggested for efficient sampling of the subcohort (see Prentice, 1986; Kim & Gruttola, 1999; Wacholder, 1991). Using the follow-up data, the set of cases $\mathcal{E}$ is identified and the case–cohort set $\mathcal{O}$ is defined as the union of $\mathcal{S}$ and $\mathcal{E}$. Genotypes are then observed for the individuals belonging to $\mathcal{O}$.

Let $S_i = 1$ if individual i was selected into the subcohort and $S_i = 0$ otherwise. Let O_i be the indicator of whether genotyping is performed: $O_i = 1$ if the genotype is observed and $O_i = 0$ if it is not observed. Note that $\mathcal{S} = \{i \in \mathcal{C} : S_i = 1\}$, $\mathcal{E} = \{i \in \mathcal{C} : E_i \neq 0\}$ and $\mathcal{O} = \{i \in \mathcal{C} : O_i = 1\} = \mathcal{S} \cup \mathcal{E}$. It is important to note that $\mathcal{S}$ will generally include some members of $\mathcal{E}$ also. In the classical case–cohort design, the data (T_i, E_i, x_i) are available only for the subjects in the case–cohort set, i.e. if $O_i = 1$.

3. Likelihood expressions and Bayesian data augmentation

3.1. Likelihood expressions

Here, we construct likelihood expressions to analyse case–cohort data as cohort data with missing genotypes. Let $\lambda_j^\theta(t \mid x, g)$, $j = 1, \ldots, k$, be the cause-specific conditional hazards corresponding to the k event types given the baseline covariates and the genotype information. As our approach to the statistical inference is based on likelihood, we have indicated the dependence of these hazards on the interest parameter θ by a superscript.

In the likelihood we consider contributions from all individuals in the cohort $\mathcal{C}$, not just from those in the case–cohort set $\mathcal{O}$. Hence, we need to model the genotype distribution and the effect of gene on the covariates X, so as to deal with the unobserved genotypes of the individuals in $\mathcal{C} \setminus \mathcal{O}$. These model parameters are denoted by π and η, respectively, and the joint distribution is written as $\mathrm{pr}(g_i \mid \pi)\mathrm{pr}(x_i \mid g_i; \eta)$. We assume that censoring is non-informative that is, the hazard corresponding to the censoring, denoted by $\lambda_0(t \mid x)$, does not depend on G or on the parameters of interest θ and η. The parameter of primary interest is θ. According to the study design, the probabilities $\mathrm{pr}(O_i \mid x_i, T_i, E_i; \theta, \eta)$ do not depend on θ and η. We consider now the construction of the likelihood expression in the situation in which all events are of interest.

When $O_i = 1$, there are three possibilities for (S_i, E_i), viz. $(S_i = 1, E_i \neq 0)$, $(S_i = 0, E_i \neq 0)$ and $(S_i = 1, E_i = 0)$, and when $O_i = 0$ only $(S_i = 0, E_i = 0)$ is possible. In the first two situations we have $O_i = 1$ and $E_i = j \neq 0$, and then the likelihood contribution from the individual is given by

$$
\begin{aligned}
&\mathrm{pr}(T_i \in \mathrm{d}t, E_i = j, O_i = 1, G_i = g_i, X_i = x_i \mid \theta, \eta, \pi) \\
&\quad = \mathrm{pr}(T_i \in \mathrm{d}t, E_i = j \mid X_i = x_i, G_i = g_i; \theta)\mathrm{pr}(X_i = x_i \mid G_i = g_i; \eta)\mathrm{pr}(G_i = g_i \mid \pi) \\
&\quad = \lambda_j^\theta(t \mid x_i, g_i)\,\mathrm{d}t \exp\left\{ - \int_0^t \sum_{r=0}^{k} \lambda_r^\theta(u \mid x_i, g_i)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g_i; \eta)\mathrm{pr}(G_i = g_i \mid \pi) \\
&\quad \propto \lambda_j^\theta(t \mid x_i, g_i)\,\mathrm{d}t \exp\left\{ - \int_0^t \sum_{r=1}^{k} \lambda_r^\theta(u \mid x_i, g_i)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g_i; \eta)\mathrm{pr}(G_i = g_i \mid \pi). \quad (1)
\end{aligned}
$$

The first equality follows because $E_i = j$ implies $O_i = 1$ and the proportionality is due to assumed non-informative censoring.

Secondly, if $O_i = 1$ but the event of interest is right censored, $E_i = 0$, then the likelihood contribution is given by

$$\begin{aligned}
&\mathrm{pr}(T_i \in \mathrm{d}t, E_i = 0, O_i = 1, G_i = g_i, X_i = x_i \mid \theta, \eta, \pi) \\
&\quad = \mathrm{pr}(T_i \in \mathrm{d}t, E_i = 0, S_i = 1, G_i = g_i, X_i = x_i \mid \theta, \eta, \pi) \\
&\quad = \mathrm{pr}(S_i = 1)\mathrm{pr}(T_i \in \mathrm{d}t, E_i = 0, G_i = g_i, X_i = x_i \mid S_i = 1; \theta, \eta, \pi) \\
&\quad \propto \mathrm{pr}(T_i \in \mathrm{d}t, E_i = 0, G_i = g_i, X_i = x_i \mid \theta, \eta, \pi) \\
&\quad = \lambda_0(t \mid x_i)\,\mathrm{d}t \exp\left\{ - \int_0^t \sum_{r=0}^{k} \lambda_r^{\theta}(u \mid x_i, g_i)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g_i; \eta)\mathrm{pr}(G_i = g_i \mid \pi) \\
&\quad \propto \exp\left\{ - \int_0^t \sum_{r=1}^{k} \lambda_r^{\theta}(u \mid x_i, g_i)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g_i; \eta)\mathrm{pr}(G_i = g_i \mid \pi). \qquad (2)
\end{aligned}$$

The first equality follows because in a situation in which the event $(E_i = 0, O_i = 1)$ can happen only if the genotype is observed because of randomization into the subcohort, i.e. $S_i = 1$. The first proportionality follows as the randomization probability does not depend on the parameters of interest and the second proportionality is due to the assumed non-informative censoring.

Finally consider the case in which the event of interest is right censored, $E_i = 0$, and $O_i = 0$ because $S_i = 0$. In this situation the genotype information is missing and the reasoning as in (2) gives the likelihood contribution as

$$\begin{aligned}
&\mathrm{pr}(T_i \in \mathrm{d}t, E_i = 0, O_i = 0, X_i = x_i \mid \theta, \eta, \pi) \\
&\quad \propto \sum_g \mathrm{pr}(T_i \in \mathrm{d}t, E_i = 0 \mid X_i = x_i, G_i = g; \theta)\mathrm{pr}(X_i = x_i \mid G_i = g; \eta)\mathrm{pr}(G_i = g \mid \pi) \\
&\quad = \sum_g \lambda_0(t \mid x_i)\,\mathrm{d}t \exp\left\{ - \int_0^t \sum_{r=0}^{k} \lambda_r^{\theta}(u \mid x_i, g)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g; \eta)\mathrm{pr}(G_i = g \mid \pi) \\
&\quad \propto \sum_g \exp\left\{ - \int_0^t \sum_{r=1}^{k} \lambda_r^{\theta}(u \mid x_i, g)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g; \eta)\mathrm{pr}(G_i = g \mid \pi). \qquad (3)
\end{aligned}$$

Assuming now that (T_i, E_i, G_i, X_i) are independent across different individuals implies that the likelihood arising from the complete cohort can be obtained as a product of the appropriate individual contributions, each being one of the above forms (1)–(3). The full likelihood for parameters θ, η and π is proportional to

$$L(t, e, g, x \mid \theta, \eta, \pi) = L_1 L_2 L_3, \qquad (4)$$

where

$$\begin{aligned}
L_1 = &\prod_{i \in \mathcal{E}} \prod_{j=1}^{k} \{\lambda_j^{\theta}(t_i \mid x_i, g_i)\}^{I(E_i = j)} \\
&\exp\left\{ - \int_0^{t_i} \sum_{r=1}^{k} \lambda_r^{\theta}(u \mid x_i, g_i)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g_i; \eta)\mathrm{pr}(G_i = g_i \mid \pi), \qquad (5)
\end{aligned}$$

$$L_2 = \prod_{i \in \mathcal{O} \setminus \mathcal{E}} \exp\left\{ - \int_0^{t_i} \sum_{r=1}^{k} \lambda_r^{\theta}(u \mid x_i, g_i)\,\mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g_i; \eta)\mathrm{pr}(G_i = g_i \mid \pi), \qquad (6)$$

$$L_3 = \prod_{i \in \mathcal{C} \setminus O} \sum_g \exp\left\{ - \int_0^{t_i} \sum_{r=1}^{k} \lambda_r^{\theta}(u \mid x_i, g)\, \mathrm{d}u \right\} \mathrm{pr}(X_i = x_i \mid G_i = g; \eta) \mathrm{pr}(G_i = g \mid \pi). \quad (7)$$

In practice, genotyping of the individuals belonging to the case–cohort set was performed only after the cases and the censored individuals are identified from the follow-up data. However, as can be seen from (1)–(3), the inference based on the complete cohort likelihood would be the same if genotype G was observed first, although missing for a part of the cohort. The likelihood contribution from individuals outside the subcohort and with right-censored time-to-event data involves a summation over unknown genotypic and using genotypic frequencies $\mathrm{pr}(X_i = x_i \mid G_i = g; \eta)\mathrm{pr}(G_i = g \mid \pi)$ of the study population as weights. Note that the likelihood expressions (1) and (2) factorize into terms involving θ, η and π, but this is not true for the expression (3). Hence, even if (η, π) are not of primary interest, in a full-cohort analysis their estimation has to be carried out simultaneously.

In some situations, it may be possible to split the vector of covariates into two components: (i) say X_1, is not expected to be associated with the considered gene(s); and (ii) say X_2, is expected to be associated with the gene(s) and possibly with X_1 as well. In this case, let ψ denote the parameter of $\mathrm{pr}(x_1; \psi)$ and η denote the model parameters of $\mathrm{pr}(x_2 \mid g, x_1; \eta)$. Note that the model parameters are now $(\pi, \psi, \eta, \theta)$. Such special situations can be easily incorporated in the likelihood expression (4). For example, in CVD epidemiology studies, X_1 can be smoking status or age at the time of baseline examination and X_2 can be lipid measurements directly reflecting the gene effect. This model is described using a directed acyclic graph in Fig. 1. The outer plate refers to the population ($\mathcal{P}$) while the inner one refers to the cohort ($\mathcal{C}$) selected from this population. The double arrows are deterministic links between the nodes, for example, the set ($\mathcal{E}$) of cases is fully determined by the individual $E_i's$. The $G_i's$ are missing for those belonging to $\mathcal{C} \setminus \mathcal{O}$. The parameters $(\pi, \psi, \eta, \theta)$ do not depend on i and hence they are outside the inner plate. These parameters, estimated using

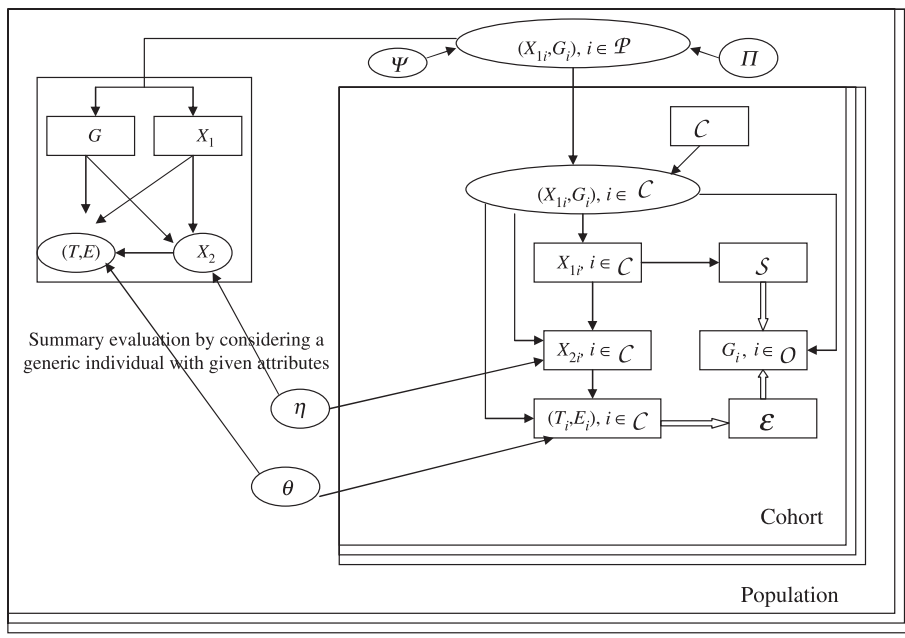

Fig. 1. Directed acyclic graph representing the model ($\mathcal{P}$, population; $\mathcal{C}$, cohort; $\mathcal{S}$, subcohort; $\mathcal{E}$, cases).

the cohort data, are then used in a summary evaluation of the considered dependencies. For this purpose, we suggest using event-specific predictive subdistribution functions (see section 3.2) of a generic individual from the population, with specified values for the covariate(s) and the genotype. This link is explained by a small box on the left side of the outer plate.

The case–cohort set is typically enriched by the causative genotypes and hence, a straightforward estimation of the population genotype frequencies using only the case–cohort set tends to be biased. This bias is corrected by using an appropriate survival model along with the conditional distribution of X given G. The proposed likelihood approach enables a simultaneous estimation of the genotype frequencies and of the effect of the covariates and genes on the events. Procedures like the EM algorithm (see Scheike & Martinussen, 2004) and Bayesian data augmentation designed for handling incomplete data can be employed here. The numerical computations can be elaborate particularly if the cohort is large, and they will depend on the total number of possible genotypes. On the contrary, considerable savings in the computations can be made if many individuals share the same covariate values x, in which case they give rise to similar likelihood contributions. In the simulation example reported in section 4, we also use straightforward frequency estimates, obtained from the subcohort, for the population genotype probabilities appearing in (3). This would seem to be a reasonable alternative for estimating θ in a full-cohort analysis if the subcohort is not too small in relation to the degree of polymorphism in the considered loci.

3.2. Bayesian data augmentation

Depending on the survival model the parameter θ is often specified in an explicit way. For example, in the case of a multiplicative hazards model, θ is the parameter of k cause-specific baseline hazards and of the relative risks associated with covariates x and g. Similarly, η can be defined depending on the model used for $\mathrm{pr}(x \mid g;\eta)$.

Adopting the Bayesian approach to inference, the posterior distribution of the parameters are proportional to the product likelihood (4) and the prior for θ, η and π. However, the posterior distribution of (θ, η) will generally have a complex form because of incomplete data. Below, we describe a Bayesian data augmentation method which is based on Gibbs sampling.

Assuming that the genotypes are observed fully, the likelihood expression is simply the product of the appropriate individual contributions, each now being the forms of (1) and (2). In this case, the posterior distribution of the model parameters θ, η and π is proportional to the product of the prior and the complete data likelihood L_{c} which is given as follows,

$$
\begin{aligned}
L_{\mathrm{c}}(t,e,g,x \mid \theta,\eta,\pi) &= \prod_{i\in\mathcal{E}}\prod_{j=1}^{k} \mathrm{pr}(X_i = x_i \mid G_i = g_i;\eta)\mathrm{pr}(G_i = g_i \mid \pi) \\
&\{\lambda_j^{\theta}(t_i \mid x_i, g_i)\}^{I(E_i=j)} \exp\left\{-\int_0^{t_i}\sum_{r=1}^{k}\lambda_r^{\theta}(u \mid x_i, g_i)\,\mathrm{d}u\right\} \\
&\prod_{i\in\mathcal{C}\setminus\mathcal{E}} \mathrm{pr}(X_i = x_i \mid G_i = g_i;\eta)\mathrm{pr}(G_i = g_i \mid \pi)\exp\left\{-\int_0^{t_i}\sum_{r=1}^{k}\lambda_r^{\theta}(u \mid x_i, g_i)\,\mathrm{d}u\right\} \\
&= \prod_{i\in\mathcal{C}} \mathrm{pr}(X_i = x_i \mid G_i = g_i;\eta)\mathrm{pr}(G_i = g_i \mid \pi) \\
&\prod_{i\in\mathcal{C}}\prod_{j=1}^{k}\{\lambda_j^{\theta}(t_i \mid x_i, g_i)\}^{I(E_i=j)} \exp\left\{-\int_0^{t_i}\sum_{r=1}^{k}\lambda_r^{\theta}(u \mid x_i, g_i)\,\mathrm{d}u\right\} \\
&= L_{1\mathrm{c}}(\theta)L_{2\mathrm{c}}(\eta)L_{3\mathrm{c}}(\pi),
\end{aligned}
$$

where $L_{1c}(\theta)$ is the part of L_c involving θ only, $L_{2c}(\eta)$ is the part of L_c involving η only, and $L_{3c}(\pi)$ is the part of L_c involving π only.

The Gibbs sampler, applied to the posterior distribution of (θ, η, π) and missing genotypes involves specification of the following full conditional probabilities:

$$\begin{aligned}\mathrm{pr}(G_i = g \mid \cdot) \propto\ & \mathrm{pr}(G_i = g \mid \pi)\mathrm{pr}(X_i = x_i \mid G_i = g; \eta) \\ & \prod_{j=1}^{k} \{\lambda_j^{\theta}(t_i \mid x_i, g)\}^{I(E_i = j)} \exp\left\{ - \int_0^{t_i} \sum_{r=1}^{k} \lambda_r^{\theta}(u \mid x_i, g)\, \mathrm{d}u \right\},\end{aligned} \tag{8}$$

$$\mathrm{pr}(\theta \mid \cdot) \propto \mathrm{pr}(\theta) L_{1c}(\theta), \tag{9}$$

$$\mathrm{pr}(\eta \mid \cdot) \propto \mathrm{pr}(\eta) L_{2c}(\eta), \tag{10}$$

$$\mathrm{pr}(\pi \mid \cdot) \propto \mathrm{pr}(\pi) L_{3c}(\pi). \tag{11}$$

To implement the Gibbs sampler, the missing genotypes should be sampled from (8) for each $i \in \mathcal{C} \setminus O$, and the parameters θ, η and π should be sampled from (9) to (11), where all variables are being conditioned on their most recently sampled values. Note that the conditional distributions (9)–(11) are in fact marginals of the joint posterior distribution of (θ, η, π) given the complete data on genotypes. If sampling from any of these full conditional probabilities is difficult then a Metropolis–Hastings step can be used. The normalizing factor for the probability (8) is a finite sum over possible genotypes and hence the corresponding normalized probability can be used to sample the missing genotypes. For the application of the EM-algorithm, Scheike & Martinussen (2004) discussed the conditional distribution similar to (8). Of course, their EM-algorithm can be generalized to multiple end-points, too. The Bayesian data augmentation algorithm using the Gibbs sampler can be summarized as follows:

Step 1: specify initial values of θ, η and π
Step 2: simulate unknown genotypes using (8)
Step 3: update π using (11)
Step 4: update η using (10)
Step 5: update θ using (9)
Step 6: repeat steps 2–5 till the convergence is achieved.

Note that steps 3–5 use genotype data generated in step 2. The numerical approximation of the posterior distribution of the cause-specific hazards can easily be obtained using the samples generated from the posterior distribution of the parameters, as an outcome of the above algorithm.

In practice for a causal analysis (see Arjas & Parner, 2004), the event-specific predictive subdistributions and probability of event-free survival are of interest, where we consider a generic individual (see Fig. 1) with specified values of the genotype and covariate(s). Such an event-specific predictive subdistribution function, for given covariates and genotype, is obtained by integrating all parameters with respect to their posterior distribution, i.e.

$$\begin{aligned}F_j^{\mathrm{pred}}(t \mid g, x_1, x_2) &= E(\mathrm{pr}(T \le t, E = j \mid g, x_1, x_2; \theta) \mid \mathrm{data}) \\ &= E\left(\int_0^{t} \lambda_j^{\theta}(u \mid g, x_1, x_2) \exp\left\{ - \int_0^{u} \sum_{l=1}^{k} \lambda_l^{\theta}(v \mid g, x_1, x_2)\, \mathrm{d}v \right\} \mathrm{d}u \mid \mathrm{data} \right),\end{aligned} \tag{12}$$

and the overall predictive distribution is given by

$$F^{\text{pred}}(t \mid g, x_1, x_2) = E(\text{pr}(T \le t \mid g, x_1, x_2; \theta) \mid \text{data}) = \sum_{j=1}^{k} F_j^{\text{pred}}(t \mid g, x_1, x_2), \tag{13}$$

where the expectation is with respect to the posterior distribution of θ. If only the genotype g and the covariate x_1 are specified in advance, the corresponding marginalized predictive distributions can be obtained by further integrating out the covariates x_2 and the parameter η,

$$\begin{aligned} F_j^{\text{pred}}(t \mid g, x_1) &= E(\text{pr}(T \le t, E = j \mid g, x_1; \theta, \eta) \mid \text{data}) \\ &= E(E(\text{pr}(T \le t, E = j \mid g, x_1, X_2; \theta, \eta) \mid g, x_1; \theta, \eta) \mid \text{data}), \end{aligned}$$

where the inner expectation is with respect to the distribution of X_2 given G, X_1, θ and η while the outer expectation is with respect to the posterior distribution of (θ, η). The predictive cause-specific hazard of the first event can then be obtained as

$$\lambda_j^{\text{pred}}(t \mid g, x_1, x_2) = \frac{(\text{d}/\text{d}t) F_j^{\text{pred}}(t \mid g, x_1, x_2)}{1 - F^{\text{pred}}(t \mid g, x_1, x_2)} \tag{14}$$

and $\lambda_j^{\text{pred}}(t \mid g, x_1)$ is defined similarly.

We apply the proposed method to the multiplicative hazards model in section 4 using the Bayesian software BUGS, Spiegelhalter *et al.* (1999).

4. Illustration: multiplicative hazards model

Here, we consider a simulation experiment which is comparable in terms of size, design and goals, to situations generally encountered in large epidemiological studies. We considered a cohort of size 2500 with approximate targeted numbers of events of three types; 140 of type 1 (5%), 50 type 2 (2%) and 90 type 3 (3%), and a subcohort of size 250 (10% of the cohort size). We considered a biallelic (A or a) marker with population genotype frequencies (0.44, 0.44, 0.12) corresponding to genotypes (AA, Aa, aa), respectively, a binary covariate X_1 with distribution $\text{pr}(X_1 = 0) = 0.7$ and $\text{pr}(X_1 = 1) = 0.3$, and another covariate X_2 which is associated with the marker gene via the conditional distribution given in Table 1 and is independent of X_1 for given G.

The cause-specific hazard rate corresponding to cause j for an individual was specified by

$$\lambda_j^{\theta}(t \mid x_1, x_2, g) = \lambda_j \exp(\beta_{1j} x_1 + z_2^T \beta_j + z_3^T \gamma_j), \tag{15}$$

with $z_2^T = (I(x_2 = 1), I(x_2 = 2))$, $z_3^T = (I(g = \text{Aa}), I(g = \text{aa}))$, $\beta_j^T = (\beta_{2j}, \beta_{3j})$, $\gamma_j^T = (\gamma_{1j}, \gamma_{2j})$, $\beta_1 = (\beta_{11}, \beta_{12}, \beta_{13}) = (0.5, 0.02, 0.6)$, $\beta_2 = (\beta_{21}, \beta_{22}, \beta_{23}) = (0.8, 0.7, 0.1)$ and $\beta_3 = (\beta_{31}, \beta_{32}, \beta_{33}) = (2.0, 0.4, 0.09)$. The values of γ were chosen as $\gamma_1 = (\gamma_{11}, \gamma_{12}, \gamma_{13}) = (0.2, -0.4, -0.1)$ and $\gamma_2 = (\gamma_{21}, \gamma_{22}, \gamma_{23}) = (0.5, -0.3, -0.15)$. The cause-specific baseline intensities were (0.001, 0.001, 0.003), and the censoring intensity were calibrated to get approximately the above-mentioned number of events.

The parameters of interest were $\theta = (\lambda_1, \lambda_2, \lambda_3, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2)$. *A priori*, all the parameters were assumed to be independent. β_1, β_2, β_3, γ_1 and γ_2 were assigned normal priors with

Table 1. *Conditional distribution of X_2 given G*

x_2	0	1	2
$\text{pr}(X_2 = x_2 \mid G = \text{AA})$	0.15	0.6	0.25
$\text{pr}(X_2 = x_2 \mid G = \text{Aa})$	0.32	0.45	0.23
$\text{pr}(X_2 = x_2 \mid G = \text{aa})$	0.36	0.55	0.09

mean 0 and variance 4, and λ_1, λ_2 and λ_3 were assigned gamma priors with shape parameter 0.01 and scale parameter 0.2. The genotype probabilities in the population were assigned a Dirichlet prior with all parameters equal to 1. Note that in practice it should be possible to choose appropriate values for such hyperparameters for the Dirichlet prior depending on the epidemiological background information available.

We first performed the Bayesian computations for the complete data with known genotypes for all individuals, to be used as a reference against which the methods for the incomplete data, could be compared with missing genotype information. We then performed the corresponding analysis using data in which genotype information for those outside the case–cohort set was missing. This was performed in two ways: by simultaneously estimating the population genotype frequencies as a part of the inferences based on the full Bayesian model specified above, and also when the genotype distribution given X_2 was estimated directly in terms of the relative frequencies observed in the subcohort data. As the two methods of incomplete data analysis resulted in similar inferences, we have reported here only the results in which the genotype frequencies were simultaneously estimated.

The analysis was carried out using WinBUGS software (Spiegelhalter *et al.*, 1999). Two chains of 80,000 iterations each where generated to check the convergence and after that samples of size 6000 were generated from the posterior distribution by thinning to every 10th cycle of the simulation.

Table 2 gives the subcohort summary as observed in the simulated data. Forty-five events were included in the subcohort. The size of the case–cohort set was 483. Under the case–cohort design, genotype information was missing from 2017 subjects who were neither selected into the subcohort nor had experienced any of the considered events during the follow-up.

Table 3 gives the posterior median values and 95% credible intervals for genotype frequencies based on the marginal posterior distributions. The credible intervals obtained using complete data are contained in those corresponding to the incomplete data and the parameters used for simulation are close to the posterior median values.

The posterior cumulative distribution of the cause-specific hazard rate $\lambda_1^{\theta}(t \mid x_1, x_2, g)$ corresponding to event type 1 and for $(x_1 = 0, x_2 = 0)$, and the genotype $g = \mathrm{Aa}$ lay in between those for $g = \mathrm{AA}$ and $g = \mathrm{aa}$, using complete and incomplete data. However, the distributions corresponding to the genotypes AA and Aa were close to each other for incomplete data.

The 95% credible intervals of $\lambda_1^{\theta}(t \mid 0, 0, \mathrm{AA})$, (0.0008, 0.0025) for the complete data and (0.0009, 0.0029) for the incomplete data, both contained the true hazard rate (0.001). But

Table 2. *Summary of simulated data*

Event type	1	2	3	0	Total
Subcohort	25	9	11	205	250
Non-subcohort	112	39	82	2017	2250
Cohort	137	48	93	2222	2500

Table 3. *Posterior median values and 95% credible intervals of genotype frequencies*

Genotype	True	Complete	Incomplete	Subcohort[1]
AA	0.44	0.44 (0.42–0.46)	0.41 (0.35–0.46)	0.42 (0.36–0.48)
Aa	0.44	0.44 (0.43–0.46)	0.47 (0.41–0.53)	0.47 (0.41–0.53)
aa	0.12	0.12 (0.11–0.13)	0.13 (0.09–0.17)	0.10 (0.06–0.14)

[1]The relative frequencies and corresponding 95% confidence intervals evaluated using the subcohort data.

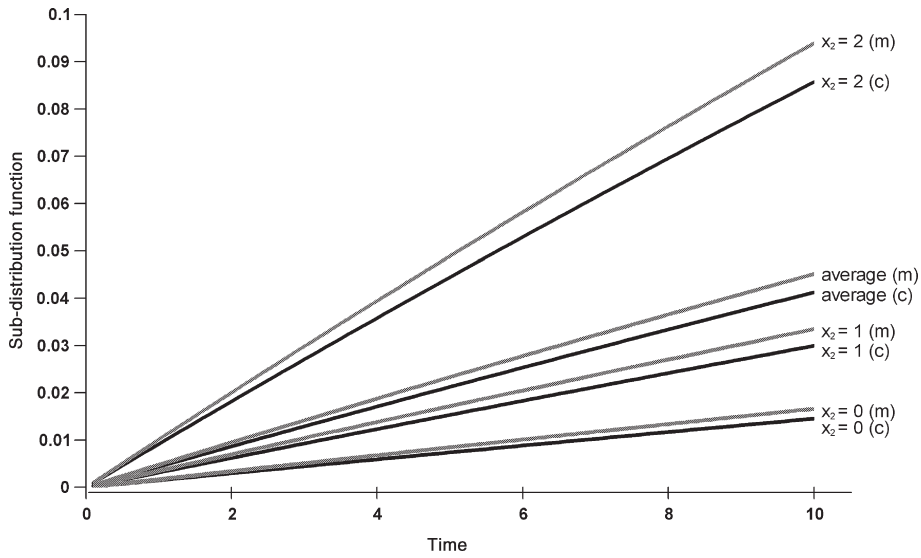

Fig. 2. Comparison of predictive subdistribution functions for event type 1 corresponding to different values of covariate x_2 with x_1 and g fixed at $(x_1 = 0, g = \mathrm{AA})$. Grey lines correspond to the incomplete data while the black lines correspond to the complete data; 'average' corresponds to the marginalized (with respect to x_2) predictive subdistribution function for event 1 with $x_1 = 0$ and $g = \mathrm{AA}$.

the true hazard rate was in the left tail area of both posterior distributions, corresponding to quantiles 0.12 and 0.06 respectively. The posterior median values were 0.0014 and 0.0016 respectively.

The predictive subdistribution function for event type 1, $(x_1 = 0, x_2 = 0)$, and for genotype Aa lay between those corresponding to AA and aa, but the subdistributions for AA and Aa were close to each other. Figure 2 shows the comparison of the predictive subdistribution functions when considering the three values of x_2 and keeping x_1 and g fixed at $x_1 = 0$, $g = \mathrm{AA}$ and also the marginalized (with respect to x_2) predictive subdistribution at $x_1 = 0$ and $g = \mathrm{AA}$. In the former case, the risk of experiencing an event of type 1 increases when the values of covariate x_2 change from $x_2 = 0$ to $x_2 = 2$, which is consistent with the chosen simulation parameter values $\beta_{21} = 0.8$ and $\beta_{31} = 2.0$. The results corresponding to the other values of (x_1, x_2) and event types were similar and hence are not reported here.

5. Discussion

A major advantage of the case–cohort design is its use in studying multiple end-points. The current literature appears to be concerned only with the single end-point case–cohort design and there is no discussion about how one could best make use of data obtained from multiple end-points. The proposed likelihood expressions (1)–(3) make use of the available data and also enable simultaneous estimation of the genotype frequencies, the effect of genes on covariates and the effect of the covariates and genes on the events. Methods based on the full likelihood can be said to have an almost canonical position among all approaches to statistical inference. On the contrary, there are many situations in which full likelihoods are analytically and/or computationally intractable, and this has made, for example, different versions of partial likelihoods or pseudo-likelihoods popular. In most situations, the consequent estimators can then be shown to have asymptotically similar desirable properties as those based on the

full likelihood, as the sample size increases to infinity. In case of more complex sampling of the subcohort, the proposed likelihood can still be used even though it cannot be called 'full likelihood' as sampling probabilities which might be complex functions of the unknown parameters are dropped from the likelihood expressions.

While EM-algorithm provides a way of dealing with the incomplete covariate data very much like we have performed by applying Bayesian inference, our method provides a more complete way of quantifying all the uncertainties involved, in the form of a joint posterior distribution. Generally speaking, if the models are the same, in the sense of having the same likelihood, then the maximum likelihood estimates and the maximum a *posteriori* estimator are comparable point estimates if the priors are vague. Therefore, assessing the methods from the perspective of point estimation, the results of Scheike & Martinussen (2004) are relevant also for our Bayesian method.

In the present illustration, in the standard case–cohort analysis only 250 plus 112 subjects would be included to study the event type 1. However, it is important to note that there is genotype data available additionally on 121 subjects. The results obtained here using the Bayesian data augmentation were close to the complete data analysis. The predictive distributions take into account the uncertainties involved in estimating the model parameters and hence provide a natural measure for comparing the complete data and incomplete data analyses. Considering the marginalized predictive distribution for given values of x_1 and g would be relevant, for example, in a situation in which one would be interested in drawing causal inferences concerning the effects of genotype g and smoking status x_1 on a cardiovascular event. Simultaneous conditioning on a measurement of lipid concentration x_2, as an intermediate phenotype, would be likely to confound such a causal analysis.

In our illustration, the results (not reported here) of the two analyses using incomplete data were close to each other because the population genotype frequencies estimated using the subcohort were close to those estimated simultaneously with other parameters, and also to the true simulation parameters. The convergence of Monte Carlo Markov Chain was faster when the genotype distribution was specified in advance and hence, in practice, if the population allele frequencies are available from some other sources, then they could be employed in the full-cohort analysis. If they are not available, then they can be either estimated, in an unbiased manner by simple frequency estimation from the subcohort, or in a more elaborate way by using the full model.

The survival model (15) that was considered here to illustrate the full likelihood approach and Bayesian data augmentation is one of the many different models that one would encounter in practice, and was used solely to illustrate the approach. Following the same principles, any other survival model could be employed in a similar fashion subject to computational constraints.

When only one specific end-point is of interest, likelihood expressions can similarly be given by appropriate specification of the hazard models according to the time of occurrences of events. This in itself is an independent and vast research area.

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 Resources' of 5th Framework Programme and by the Academy of Finland via its grant number 53646. The research of the second author was supported by the Academy of Finland via its funding of the 'Centre of Population Genetic Analyses'. The authors thank Dr Kari Kuulasmaa and Dr Veikko Salomaa of KTL for useful discussions and comments.

References

Arjas, E. & Parner, J. (2004). Causal reasoning from longitudinal data. *Scand. J. Statist.* **31**, 171–187.
Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics* **50**, 1064–1072.
Chen, K. & Lo, S. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika* **85**, 755–764.
Kim, S. & Gruttola, V. (1999). Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial. *Lifetime. Data. Anal.* **5**, 149–172.
Kong, L., Cai, J. & Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika* **91**, 305–319.
Lin, D. Y. & Ying, Z. (1993). Cox regression with incomplete covariate measurements. *J. Amer. Statist. Assoc.* **88**, 1341–1349.
Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.
Scheike, T. H. & Juul, A. (2004). Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics* **5**, 193–206.
Scheike, T. H. & Martinussen, T. (2004). Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scand. J. Statist.* **31**, 283–293.
Self, S. G. & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **15**, 54–81.
Sorensen, P. & Andersen, P. K. (2000). Competing risks analysis of the case-cohort design. *Biometrika* **87**, 49–59.
Spiegelhalter, D. J., Thomas, A. & Best, N. G. (1999). *WinBUGS version 1.2 user manual*. Medical Research Council Biostatistics Unit, Cambridge, UK.
Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology* **2**, 155–158.

Received September 2004, in final form July 2005

Elja Arjas, Department of Mathematics and Statistics, PO Box 68, FI - 00014 University of Helsinki, Helsinki, Finland.
E-mail: elja.arjas@helsinki.fi