Joint analysis of prevalence and incidence data using conditional likelihood

OLLI SAARELA*
Department of Chronic Disease Prevention,
National Institute for Health and Welfare, Mannerheimintie 166, 00300 Helsinki, Finland
olli.saarela@thl.fi

SANGITA KULATHINAL
Indic Society for Education and Development (INSEED), Nashik 422 011, India and
Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland

JUHA KARVANEN
Department of Chronic Disease Prevention, National Institute for Health and Welfare,
Mannerheimintie 166, 00300 Helsinki, Finland

SUMMARY
Disease prevalence is the combined result of duration, disease incidence, case fatality, and other mortality. If information is available on all these factors, and on fixed covariates such as genotypes, prevalence information can be utilized in the estimation of the effects of the covariates on disease incidence. Study cohorts that are recruited as cross-sectional samples and subsequently followed up for disease events of interest produce both prevalence and incidence information. In this paper, we make use of both types of information using a likelihood, which is conditioned on survival until the cross section. In a simulation study making use of real cohort data, we compare the proposed conditional likelihood method to a standard analysis where prevalent cases are omitted and the likelihood expression is conditioned on healthy status at the cross section.

Keywords: Ascertainment correction; Conditional likelihood; Incidence; Left censoring; Left truncation; Prevalence.

1. INTRODUCTION
In association analysis of covariates and disease incidence, measurements for covariates that do not change over time can be utilized regardless of the time of the measurement. It follows that such fixed covariates, for example, genotypes, can be used to explain disease events both before and after the time of the covariate measurement. Our aim in this paper was to make use of information on disease prevalence to estimate covariate effects on disease incidence.

Prospective follow-up of a birth cohort would be an ideal study design for analysis of disease incidence. An alternative and often more practical study design is to recruit a study cohort as a random sample from some population alive at a given calendar time and thereafter to conduct a prospective follow-up.

*To whom correspondence should be addressed.
We use the term cross section for a cohort recruitment carried out at a specific calendar time (see Keiding, 2006). The cross section and the subsequent follow-up produce 2 different kinds of data, information on disease prevalence at the cross section and on events observed during the follow-up. In addition, covariate information is collected for every individual at the start of the follow-up.

The relation between incidence and prevalence is studied mathematically by Keiding (1991) and Alho (1992). Prevalence data do not give direct information on incidence because they do not include those who have died either due to the disease events of interest or for unrelated reasons. Keiding (1991) considers 2 special situations where age-specific incidence can be estimated from prevalence data only. The first assumes nondifferential mortality between healthy and diseased individuals, while the second uses mortality information from external sources. Examples of such situations are presented by Keiding and others (1989), Ogata and others (2000), and Keiding (2006). When studying covariate effects, the possible dependency of the mortality rates on the covariates has to be taken into account. Because of this, we consider here a situation where the study cohort has a follow-up for all-cause mortality. In this case, the relevant mortality rates for both healthy and diseased people and their dependency on the covariates of interest can be estimated from the available data.

The cross-sectional cohort is subject to 2 kinds of incomplete observation, truncation, and censoring. Left truncation occurs when individuals in the underlying birth cohort die before the start of the follow-up period and therefore can not be observed at all, rendering the cohort unrepresentative of the underlying birth cohort. Left censoring in the context of a cross-sectional cohort occurs if the event times before the start of the follow-up are unknown for the prevalent cases. In addition, right censoring occurs due to loss to follow-up and end of the follow-up period. Analysis methods commonly used for interval censored data are also applicable for data with left- and right-censored failure times. Nonparametric estimation in the presence of left truncation is discussed, for example, by Wang (1991) and Gross and Lai (1996). Likelihood-based survival analysis of left-truncated and right-censored data can be carried out using conditional likelihood (see, e.g. Guo, 1993, p. 229, and the references therein). Here, the likelihood expression for time-to-event data is conditioned on survival without event until the start of the follow-up, which in simple cases results in a convenient likelihood expression.

In the current context, left censoring caused by failure to observe the event times for the prevalent cases can also be handled as left truncation by excluding the prevalent cases and conditioning the likelihood on healthy status at the start of the follow-up (see, e.g. Commenges and others, 1998, p. 1976). The advantage of this approach is the simplicity of the analysis but at the cost of loss of information on the prevalent cases. Inclusion of covariates in a joint analysis of prevalence and incidence data has rarely been studied. A nonparametric approach based on multiple imputation is described by Karvanen and others (2008). In the present paper, we aim for a likelihood-based solution.

Section 2 of the paper describes the study setup, which produces both prevalence and incidence data. In Section 3, we extend the standard conditional likelihood methods for left-truncated and right-censored time-to-event data to incorporate also left-censored event times. By doing this, we aim to utilize the information in the prevalent cases, which are commonly omitted from the analysis. We have been motivated by association studies with genotypic covariates where the requirement of constancy of covariates over time is naturally fulfilled, but our methods can be applied also for other situations where the covariate of interest is sufficiently stable over time. In Section 4, in a simulation study making use of real cohort data, we compare the utility of the new conditional likelihood approach to the standard approach in which prevalent cases are excluded from the analysis. The paper concludes with a discussion in Section 5.

2. Observed data

We consider a cohort \( C = \{1, \ldots, n\} \), which is recruited as a random sample from some defined population alive at a given calendar time. Individual \( i \in C \) enters the follow-up at age \( b_i \) and is followed up for
incidence of an event of interest, which can be either fatal or nonfatal, and for all-cause mortality until age $c_i$. Let $T_i$ denote the age at the time of the first event of interest, death, or right censoring. If a nonfatal event of interest occurs at $T_i$, the follow-up continues for subsequent mortality until age $T_i + U_i$, where $U_i$ represents the time elapsed from the first nonfatal event of interest to death due to any cause or right censoring. Exact event times can only be observed during the window of the follow-up period, $[b_i, c_i)$. In addition, at the start of the follow-up, it is observed if an individual has a prevalent condition corresponding to an event of interest, which has occurred before the start of the follow-up, that is, $T_i < b_i$. The exact times of the first event for the prevalent cases are unknown. In addition to the time-to-event data, a vector of covariates $z_i$ is observed for each individual $i \in \mathcal{C}$. It is assumed that these covariates remain constant over time, so that the possible occurrence of an event before the covariate measurements does not alter the covariate values.

Let $E_i = 1$ denote an event of interest and $E_i = 2$ death due to some other reason at age $T_i$. For simplicity, we consider only right censoring ($E_i = 0$) of type I due to the end of the follow-up period at age $T_i = c_i$ (see, e.g. Kalbfleisch and Prentice, 2002, p. 53). As is common in cardiovascular epidemiology (see, e.g. Thorvaldsen and others, 1995), we consider separately the immediate (e.g. within 28 days) mortality after the event of interest, termed as case fatality. The reason for this is that the factors affecting the fatality of the disease event may be very different to those affecting the underlying disease of interest or the later long-term mortality. In order to define the observable cohort, we need to define a random variable, say $D_i$, which represents case fatality for events of type $E_i = 1$. In addition, we define an indicator taking value $F_i = 1$ if a death is observed at age $T_i + U_i$ and $F_i = 0$ if the subject is right censored at age $T_i + U_i = c_i$. The different possible event histories are illustrated in Figure 1. Now individuals $\{i : T_i < b_i, E_i = 2\}$ and $\{i : T_i < b_i, E_i = 1, D_i = 1\}$ (fatal first event before the start of the follow-up, lines 8 and 9 in Figure 1) are left truncated and are not observed when the cohort is recruited. In addition, individuals $\{i : T_i < b_i, E_i = 1, D_i = 0, U_i < b_i - T_i, F_i = 1\}$ (a nonfatal event and a subsequent death before the start of the follow-up, line 10 in Figure 1) are also left truncated. Prevalent cases $\{i : T_i < b_i, E_i = 1, D_i = 0, U_i \geq b_i - T_i\}$ are observed, but the event times are left censored (lines 6 and 7

![Fig. 1. Illustration of the different event histories. One row represents an individual. Solid lines represent follow-up for the first event and dotted lines follow-up for total mortality after first nonfatal event.](https://academic.oup.com/biostatistics/article-abstract/10/3/575/293574/download)

**Fig. 1.** Illustration of the different event histories. One row represents an individual. Solid lines represent follow-up for the first event and dotted lines follow-up for total mortality after first nonfatal event.
in Figure 1). Exact event times for the first events are observed for \( \{ i : b_i \leq T_i < c_i \} \). Using notation \( L_i \equiv [T_i < b_i, E_i = 1, D_i = 0, U_i \geq b_i - T_i] \), the cohort ascertainment rule can be written as

\[ [T_i \geq b_i] \lor L_i. \quad (2.1) \]

3. Likelihood expressions

3.1 Single time scale, type I censoring

In the following, age is considered as the only time scale of the analysis of the first event and the event time \( T_i \) refers to the age at the time of the first event. This corresponds to the assumption of homogeneity of the relevant hazard rates over calendar time. The more general case of multiple time scales is discussed in the next section. The probability model for the first event can be defined using cause-specific hazard functions. Now we can write the second term in (3.1) in terms of 2 separate survival models and the case-fatality model as

\[
p(T_i, E_i | z_i; \theta) = \prod_{k=1}^{2} [\lambda_k(T_i | z_i; \theta_k)]^{1_{E_i = k}} S(T_i | z_i; \theta), \quad (3.1)
\]

where \( S(T_i | z_i; \theta) = \exp \{ - \int_0^{T_i} \sum_{k=1}^{2} \lambda_k(t | z_i; \theta_k) \, dt \} \) and \( 1_A \) is an indicator function of \( A \). Here, we consider the proportional hazards assumption \( \lambda_k(t | z_i; \theta_k) = \lambda_{0k}(t | \alpha_k) \exp(\beta_k z_i) \), where \( \theta_k = (\alpha_k, \beta_k) \).

Our main interest is in the parameters \( \beta_1 \) describing the relationship between the covariates \( z_i \) and the events of type \( E_i = 1 \).

All conditional likelihood expressions for the data described in Section 2 are written conditional on the covariate data and the ascertainment rule (2.1). The probability for the ascertainment rule given the covariate data can be written as

\[
p(T_i \geq b_i \lor L_i | z_i; \theta, \eta, \mu)
\]

\[
= p(T_i \geq b_i | z_i; \theta) + p(T_i < b_i, E_i = 1, D_i = 0, U_i \geq b_i - T_i | z_i; \theta, \eta, \mu)
\]

\[
= C(b_i, z_i; \theta, \eta, \mu). \quad (3.2)
\]

Here \( \mu \) is a vector of parameters defining the probability model for case fatality and \( \eta \) is a vector of parameters defining the hazard rate for all-cause mortality after the first nonfatal event. These parameters are not of primary interest but are needed for evaluating expression (3.2). This is the probability that individual \( i \) does not die before age \( b_i \), that is, the probability of not becoming excluded from the cohort due to left truncation. The different state transitions and the corresponding transition probabilities needing to be modeled are illustrated in Figure 2.

By defining a hazard rate \( \lambda^*(u | T_i, z_i; \eta) \) for all-cause mortality after the first nonfatal event at age \( T_i \), we can write the second term in (3.2) in terms of 2 separate survival models and the case-fatality model as

\[
p(T_i < b_i, E_i = 1, D_i = 0, U_i \geq b_i - T_i | z_i; \theta, \eta, \mu)
\]

\[
= \int_0^{b_i} \int_{b_i-t}^{\infty} p(t, E_i = 1, D_i = 0, u | z_i; \theta, \eta, \mu) \, du \, dt
\]

\[
= \int_0^{b_i} p(U_i \geq b_i - t | t, E_i = 1, D_i = 0, z_i; \eta) p(D_i = 0 | t, E_i = 1, z_i; \mu) p(t, E_i = 1 | z_i; \theta) \, dt
\]

\[
= \int_0^{b_i} S^*(b_i - t | t, z_i; \eta) p(D_i = 0 | t, E_i = 1, z_i; \mu) \lambda_1(t | z_i; \theta_1) S(t | z_i; \theta) \, dt,
\]

where \( S^*(b_i - t | t, z_i; \eta) = \exp \{ - \int_0^{b_i-t} \lambda^*(u | t, z_i; \eta) \, du \}. \)
Joint analysis of prevalence and incidence data

For $i \in C$, 4 kinds of likelihood contributions can be identified corresponding to the sets

\[ \mathcal{X} = \{ i : T_i \geq b_i, (E_i = 0 \lor E_i = 2) \}, \]
\[ \mathcal{Y} = \{ i : T_i \geq b_i, E_i = 1, D_i = 1 \}, \]
\[ \mathcal{Z} = \{ i : T_i \geq b_i, E_i = 1, D_i = 0 \}, \]

and \[ \mathcal{W} = \{ i : L_i \}. \] The likelihood contribution for individuals $i \in \mathcal{X}$ who either die during the follow-up or are right censored at the end of the follow-up period (lines 1 and 2 in Figure 1) is

\[
p(T_i, E_i \mid z_i, [T_i \geq b_i \lor L_i]; \theta, \eta, \mu) = \frac{p(T_i, E_i, [T_i \geq b_i \lor L_i] \mid z_i; \theta, \eta, \mu)}{p(T_i \geq b_i \lor L_i \mid z_i; \theta, \eta, \mu)} = p(T_i, E_i \mid z_i; \theta)/C(b_i, z_i; \theta, \eta, \mu). \tag{3.3}
\]

The likelihood contribution for individuals $i \in \mathcal{Y}$ with a fatal event of interest (line 3 in Figure 1) includes the probability model for case fatality,

\[
p(T_i, E_i = 1, D_i = 1 \mid z_i, [T_i \geq b_i \lor L_i]; \theta, \eta, \mu) \]
\[ = p(D_i = 1 \mid T_i, E_i = 1, z_i; \mu)p(T_i, E_i = 1 \mid z_i; \theta)/C(b_i, z_i; \theta, \eta, \mu). \tag{3.4}
\]

In addition to the above, individuals $i \in \mathcal{Z}$ with a nonfatal event of interest (lines 4 and 5 in Figure 1) are followed up for all-cause mortality after the first event and the conditional likelihood contribution includes the corresponding survival model,

\[
p(T_i, E_i = 1, D_i = 0, U_i, F_i \mid z_i, [T_i \geq b_i \lor L_i]; \theta, \eta, \mu) \]
\[ = p(U_i, F_i \mid T_i, E_i = 1, D_i = 0, z_i; \eta) \]
\[ \times p(D_i = 0 \mid T_i, E_i = 1, z_i; \mu)p(T_i, E_i = 1 \mid z_i; \theta)/C(b_i, z_i; \theta, \eta, \mu). \tag{3.5}
\]
Finally, the conditional likelihood contribution for the prevalent cases \( i \in \mathcal{W} \) (lines 6 and 7 in Figure 1) is

\[
p(T_i < b_i, E_i = 1, D_i = 0, U_i, F_i \mid z_i, [T_i \geq b_i \lor L_i]; \theta, \eta, \mu) \\
= p(T_i < b_i, E_i = 1, D_i = 0, U_i, F_i \mid z_i; \theta, \eta) / C(b_i, z_i; \theta, \eta, \mu) \\
= \int_0^{b_i} p(U_i, F_i \mid t, E_i = 1, D_i = 0, z_i; \eta) p(D_i = 0 \mid t, E_i = 1, z_i; \mu) \\
\times p(t, E_i = 1 \mid z_i; \theta) \, dt / C(b_i, z_i; \theta, \eta, \mu).
\]  

(3.6)

By combining expressions (3.3), (3.4), (3.5), and (3.6), the complete conditional likelihood expression becomes

\[
L_1(\theta, \eta, \mu) \\
= \prod_{i \in \mathcal{C}} \frac{1}{C(b_i, z_i; \theta, \eta, \mu)} \prod_{i \in \mathcal{X}} p(T_i, E_i \mid z_i; \theta) \prod_{i \in \mathcal{Y}} p(D_i = 1 \mid T_i, E_i = 1, z_i; \mu) p(T_i, E_i = 1 \mid z_i; \theta) \\
\times \prod_{i \in \mathcal{Z}} p(U_i, F_i \mid T_i, E_i = 1, D_i = 0, z_i; \eta) p(D_i = 0 \mid T_i, E_i = 1, z_i; \mu) p(T_i, E_i = 1 \mid z_i; \theta) \\
\times \prod_{i \in \mathcal{W}} \int_0^{b_i} p(U_i, F_i \mid t, E_i = 1, D_i = 0, z_i; \eta) p(D_i = 0 \mid t, E_i = 1, z_i; \mu) p(t, E_i = 1 \mid z_i; \theta) \, dt.
\]

(3.7)

The use of (3.7) as a likelihood function requires estimation of the cause-specific hazard functions for the event of interest and other mortality and modeling of case fatality and total mortality after first nonfatal event, all of which may depend on the covariates \( z_i \). In practice, this requires that the cohort in question has a follow-up for all-cause mortality and an age range sufficiently wide for the data to give enough information on the cause-specific hazard rates over age.

For comparison, the standard analysis approach would be to discard the prevalent cases. This is equivalent to handling left censoring as left truncation. The ascertainment rule is then \( T_i \geq b_i \) and the individuals \( i \in \mathcal{W} \) are excluded from the analysis. The terms involving \( \eta \) and \( \mu \) can be dropped from the likelihood and the conditional likelihood expression becomes

\[
L_2(\theta) = \prod_{i \in \mathcal{C} \setminus \mathcal{W}} p(T_i, E_i \mid z_i, T_i \geq b_i; \theta)
\]

\[
= \prod_{i \in \mathcal{C} \setminus \mathcal{W}} \frac{p(T_i, E_i \mid z_i; \theta)}{p(T_i \geq b_i \mid z_i; \theta)}
\]

\[
= \prod_{i \in \mathcal{C} \setminus \mathcal{W}} \frac{\prod_{k=1}^2 [\lambda_k(T_i \mid z_i; \theta_k)]^{1[E_i=1]} \exp \left\{ - \int_{b_i}^{T_i} \sum_{k=1}^2 \lambda_k(t \mid z_i; \theta_k) \, dt \right\}}{\exp \left\{ - \int_{b_i}^{T_i} \sum_{k=1}^2 \lambda_k(t \mid z_i; \theta_k) \, dt \right\}}
\]

\[
= \prod_{i \in \mathcal{C} \setminus \mathcal{W}} \prod_{k=1}^2 [\lambda_k(T_i \mid z_i; \theta_k)]^{1[E_i=1]} \exp \left\{ - \int_{b_i}^{T_i} \sum_{k=1}^2 \lambda_k(t \mid z_i; \theta_k) \, dt \right\},
\]

(3.8)

which is of similar form to expression (8) in Guo (1993). It should also be noted that the terms involving the hazard rate for other-cause mortality factor out of the likelihood and the related parameters \( \theta_2 \) need
Joint analysis of prevalence and incidence data

not be estimated if they are not of primary interest. The above is a much more convenient expression than (3.7) as it avoids estimation of several nuisance parameters. However, using (3.8) will result in less efficient parameter estimation because the information contained in the left-censored observations is lost. In Section 4, we will compare the performance of (3.7) and (3.8) in the context of 2 real cohorts to determine whether the added efficiency justifies the more complex analysis.

3.2 Generalizations

In Section 3.1, we derived the likelihood expressions for the case where age is used as the only time scale in time-to-event analysis. This is reasonable, as in many applications, most variation in the hazard rates occurs over age and the models would in any case require adjustment for age. However, in principle, it is straightforward to generalize the likelihood expressions for multiple time scales. For example, defining the date of birth $d_i$ for individual $i$, the hazard rates in the survival model (3.1) can be allowed to depend on calendar time; hence,

$$p(T_i, E_i | d_i, z_i; \theta) = \prod_{k=1}^{2} \left[ \lambda_k(T_i, d_i + T_i | z_i; \theta_k) \right]^{1[E_i = k]} S(T_i | d_i, z_i; \theta),$$

where $S(T_i | d_i, z_i; \theta) = \exp \left\{ -\int_{0}^{T_i} \sum_{k=1}^{2} \lambda_k(t, d_i + t | z_i; \theta_k) dt \right\}$. In the proportional hazards model, if age is still considered as the most important time scale, the hazard function may be defined as

$$\lambda_k(t, d_i + t | z_i; \theta_k) = \lambda_{0k}(t | a_k) \exp \{ \beta_k \theta_k (d_i + t) \},$$

where $\gamma_k$ is the calendar time trend for endpoint $k \in \{1, 2\}$. It should be noted that the follow-up period for the study cohort should be long enough to deliver reasonable estimates of calendar time trends in hazards rates. Alternatively, external information could be utilized to obtain “plug-in” estimates for the calendar time effects. As noted by Keiding (1991), the assumption of time homogeneity is often implicitly made in applications. This is understandable since commonly the main interest is in estimating covariate effects rather than hazard functions over time, and the modeling of several time scales makes the task cumbersome. Also, compared with age, calendar time probably acts less often as a confounder. In this case, omitting the calendar time scale from the analysis is similar to omitting a covariate.

In Section 3.1, we assumed type I right censoring due to the end of the follow-up period of the study cohort. The model could easily be generalized to allow random censoring due to loss to follow-up (e.g. emigration) or other reasons. In this case, the random censoring would have to be modeled similarly as all-cause mortality, with an additional endpoint in the competing risks survival model (3.1). However, in many cases, a sensible approach would be to define the background population of interest as the birth cohort corresponding to the population in the sampling frame at the time of recruitment to the study cohort. Possible loss to follow-up after the cohort recruitment is then handled similarly as type I censoring. For a population-based study cohort this means that we are primarily interested in making inference on the population that lives in the study area at the time of the cross section. For similar reasons, we do not consider the effects of immigration or time-varying birth rates in our approach.

4. Illustration

In this section, we use follow-up data from 2 real cohorts to study the performance of the proposed conditional likelihood expression (3.7) in comparison to the standard conditional likelihood analysis with the likelihood expression (3.8). The main interest is in determining how much efficiency is gained by including the prevalent cases in estimation of covariate effects on the endpoint of interest. For this purpose,
given the observed follow-up data for the cohorts, we simulate an additional binary covariate with varying effects and population frequencies and carry out maximum likelihood parameter estimation using both the new and the standard conditional likelihood expression. Making use of real cohort data in the simulation study should provide more realistic comparison than fully simulated data. The data-based simulation procedure used here is similar to that in Saarela and others (2008) and is aimed at comparing the efficiency of the 2 alternative likelihood methods. As the actual inference is based on standard maximum likelihood techniques, we elaborate less on the other properties of the resulting estimators.

4.1 Example cohorts

The data we consider in our illustration are men from 2 nonoverlapping Finnish population-based cohorts, FINRISK 92 (FR92) and FINRISK 97 (FR97), which are part of the National FINRISK Study (Salomaa and others, 1996; Vartiainen and others, 2000) and are included in the MORGAM project, an international pooling of cardiovascular cohorts (Evans and others, 2005; Kulathinal and others, 2005). These cohorts are recruited as random samples from defined geographical populations. They have been used, for example, in candidate gene (Silander and others, 2008) and replication studies (Karvanen and others, 2009) in which the method proposed in the present paper would be directly applicable. In addition to cardiovascular endpoints, the cohorts are followed up for all-cause mortality. The endpoint of interest in our illustration is the first (fatal or nonfatal) coronary heart disease (CHD) event. Event times for those with prevalent CHD status at the start of the follow-up are considered left censored, with the age at the first event being less than the age at the start of the follow-up. Event numbers and some characteristics of the cohorts are presented in Table 1. The FR97 cohort has a shorter follow-up period and wider age range, which explains the higher ratio of prevalent CHD cases to incident cases observed during the follow-up. In addition to the data shown in the table, FR92 and FR97 follow-up data included 76 and 108 deaths, respectively, during the follow-up, which had occurred for individuals with a previous nonfatal CHD event (including prevalent CHD cases).

4.2 Simulation procedure

Because the 2 example cohorts described above have follow-up for all-cause mortality, they allow modeling of case fatality, mortality for causes other than CHD, and total mortality after first nonfatal CHD event, all of which are needed for evaluation of the likelihood expression (3.7). In the simulation model, we defined the survival model (3.1) as the proportional hazards Weibull regression where the cause-specific hazard functions are of the form

$$\lambda_k(t \mid Z_i; \theta_k) = a_k \kappa_k (a_k t)^{\kappa_k} \exp\{\beta_k Z_i\}.$$

The model for (28 days) case fatality was defined as a logistic regression model, with age at CHD event as a covariate:

$$p(D_i \mid T_i, E_i = 1, Z_i; \mu) = \frac{\exp[D_i(\mu_1 + \mu_2 T_i + \mu_3 Z_i)]}{1 + \exp[\mu_1 + \mu_2 T_i + \mu_3 Z_i]}.$$

Table 1. Event numbers and characteristics of the example cohorts

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Age range</th>
<th>Follow-up period</th>
<th>n(C)</th>
<th>n(X)</th>
<th>n(Y)</th>
<th>n(Z)</th>
<th>n(W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR92</td>
<td>24–64</td>
<td>1992–2005</td>
<td>2833</td>
<td>2318</td>
<td>167</td>
<td>50</td>
<td>178</td>
</tr>
</tbody>
</table>
Joint analysis of prevalence and incidence data

Total mortality at least 28 days after first nonfatal CHD event was modeled with a Weibull model of the form

\[ \lambda^*(u \mid T_i, Z_i; \eta) = \eta_1 \eta_2 (\eta_1 u)^{\eta_2 - 1} \exp(\eta_3 T_i + \eta_4 Z_i). \]

Note that age is used as the time scale in the model for first events, whereas the time scale in the above model is time elapsed from the first nonfatal CHD event, with age at the CHD event used as a covariate. The population distribution of the simulated covariate was defined as \( p(Z_i \mid \pi) = \pi Z_i (1 - \pi)^{1 - Z_i} \). Thus, the simulated situation corresponds to dominant/recessive effect of a single allele.

In the simulation, the values for variables \( (T_i, E_i, D_i, U_i, F_i) \) were taken from the FINRISK follow-up data. Parameters \( \pi, \beta_1, \beta_2, \mu_3, \) and \( \eta_4 \) were fixed and covariates \( Z_i \) for \( i \in C \), unknown event times \( T_i \) for \( i \in W \), and all other model parameters were simulated from the probability model (3.7) using Markov chain Monte Carlo sampling. Thousand data sets of covariates \( Z_i \) combined with the observed follow-up data were produced with different values of the fixed parameters. Maximum likelihood estimation was then carried out for each of these using the likelihood expressions (3.7) and (3.8).

4.3 Results

The maximization of the likelihood expressions with respect to parameters \( \theta, \eta, \) and \( \mu \) was implemented using the \texttt{optim} function of the R statistical software, applying the Broyden-Fletcher-Goldfarb-Shanno algorithm (R Development Core Team, 2006). Standard error estimates were obtained by inverting the numerically differentiated information matrix at the maximum likelihood point. Compiled functions for evaluation of the likelihoods were written with C programming language to enable fast looping. Integrals in (3.2) and (3.6) were evaluated using numerical integration, applying the \texttt{gsl_integration_qng} function in the GNU Scientific Library (Galassi and others, 2006).

For survival model (3.1), we used both the Weibull regression model defined above and, for comparison, a piecewise constant hazard model with 10-year age groups. The main parameter of interest here is the regression coefficient \( \beta_1 \) describing the association between the simulated covariate and the CHD endpoint. Other regression coefficients \( \beta_2, \mu_3, \) and \( \eta_4 \) were set to 0 in the simulation model but were considered unknown in the maximum likelihood estimation (results not shown). The results for \( \beta_1 \) are given in Table 2. In all cases, the inclusion of the prevalent cases gives a clear gain in efficiency, with the difference to the standard analysis being greater for the FR97 data, which include more prevalent cases. The Weibull and the piecewise constant models gave very similar results. Because the 10-year age grouping is somewhat rough, this suggests some robustness to model misspecification as the data were simulated using the Weibull model. In all cases, point estimators seem to be unbiased, while the asymptotic standard errors seem to be closer to the empirical standard deviations of the point estimates with the higher population frequency \( \pi \) of the simulated covariate.

The simulated event times for \( i \in W \) allow also consideration of the situation where the left-censored event times would be observed for these individuals. This would be the case, for example, if the event times were later obtained from hospital registers. Thus, we carried out the parameter estimation also using the simulated event times for \( i \in W \) (results not shown). The likelihood contribution for the prevalent cases is then of the same form as (3.6) but without the integral over time. Somewhat surprisingly, knowing the exact event times had little or no effect on the standard errors of the regression coefficient of interest, although the standard errors for the baseline hazard parameters were slightly smaller. This indicates that at least for these 2 cohorts, it is not worth using resources for obtaining the exact event times for the prevalent cases.

Because there is even more motivation to utilize the prevalent cases in the case when only short prospective follow-up exists, we also carried out experiments using only data from 4 initial years of follow-up instead of using the whole length of available follow-up (results not shown). The described estimation
Table 2. Summary statistics from 1000 replications: sample means and SDs of point estimates, sample means of asymptotic SEs, and power of test for $H_0: \beta_1 = 0$ at 1% significance level

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>$\beta_1$</th>
<th>Model</th>
<th>Likelihood expression</th>
<th>FINRISK 92</th>
<th>FINRISK 97</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean $\hat{\beta}_1$</td>
<td>SD $\hat{\beta}_1$</td>
<td>Mean $\hat{SE}$</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>Weibull</td>
<td>$L_2$</td>
<td>-0.010</td>
<td>0.176</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>-0.009</td>
<td>0.144</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piecewise constant</td>
<td>$L_2$</td>
<td>-0.010</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>-0.009</td>
<td>0.142</td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>Weibull</td>
<td>$L_2$</td>
<td>0.302</td>
<td>0.158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.300</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piecewise constant</td>
<td>$L_2$</td>
<td>0.298</td>
<td>0.157</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.297</td>
<td>0.129</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td>Weibull</td>
<td>$L_2$</td>
<td>0.609</td>
<td>0.149</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.606</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piecewise constant</td>
<td>$L_2$</td>
<td>0.601</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.599</td>
<td>0.123</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
<td>Weibull</td>
<td>$L_2$</td>
<td>-0.006</td>
<td>0.137</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>-0.004</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piecewise constant</td>
<td>$L_2$</td>
<td>-0.006</td>
<td>0.136</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>-0.004</td>
<td>0.109</td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>Weibull</td>
<td>$L_2$</td>
<td>0.302</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.303</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piecewise constant</td>
<td>$L_2$</td>
<td>0.299</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.299</td>
<td>0.107</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td>Weibull</td>
<td>$L_2$</td>
<td>0.612</td>
<td>0.135</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.611</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piecewise constant</td>
<td>$L_2$</td>
<td>0.605</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_1$</td>
<td>0.604</td>
<td>0.108</td>
</tr>
</tbody>
</table>

SD, standard deviation; SE, standard error.

The procedure worked well also in this situation, even though the numbers of events were smaller. With the 4 year follow-up, 68% (FR92) and 72% (FR97) of the CHD cases included are prevalent cases, with event times unknown. However, even in this case, we did not see any real improvement in the standard errors of $\beta_1$ estimates when we used the simulated event times in place of the unobserved event times. The probable explanation for this is that both example cohorts still had enough events observed during the follow-up and therefore enough information on both the age-specific incidence of CHD and the left-censored event times. This point is probably also related to the general question of how much information the exact event times add compared with the situation where only disease status is known (see, e.g. Annesi and others, 1989). That being said, knowing the exact event times is helpful in the estimation of some of the other parameters in the model, especially those describing the total mortality after a nonfatal CHD event. It appears that for the $\eta$ parameters to be identifiable, we need at least 1 subject with an observed time for a CHD event and an observed time for subsequent death. When the prospective follow-up is short, there are very few such subjects with both events observed during the window of the follow-up period (with the 4 year follow-up, 3 in FR92 and 1 in FR97). However, it is more likely that some of the prevalent cases die even during the short follow-up period (23 in FR92 and 32 in FR97). Thus, obtaining the exact event times for at least some of these subjects is beneficial in the situation where the prospective follow-up is very short.
5. DISCUSSION

In this paper, we have derived a likelihood expression for joint analysis of prevalence and incidence data, with the aim to improve the efficiency in estimation of covariate effects on disease incidence. Time-to-event data collected for cohorts that are recruited as cross-sectional samples are commonly analyzed by excluding the prevalent cases, which corresponds to conditioning of the analysis by a healthy status at the cross section. In contrast, the proposed likelihood expression is conditioned on survival until the cross section, which requires the estimation of mortality rates for both healthy and diseased individuals. In practice, this requires follow-up for all-cause mortality in addition to the follow-up of the disease events of interest. A wide age range in the study cohort is also required to enable estimation of the age-specific rates. For simplicity, we assumed homogeneity of hazard rates over calendar time. However, with a long enough follow-up period, trends over calendar time could also be estimated. Alternatively, external information could be used to obtain estimates for the calendar time effects.

Some authors (e.g. Seaman and Richardson, 2001; Ma and others, 2007) have used conditional likelihoods also under the Bayesian framework. Ma and others (2007) in their appendix obtained a “conditional posterior distribution” simply by replacing the unconditional likelihood function in the Bayes formula by its version conditioned on an ascertainment rule. Accordingly, the conditional likelihood expression for prevalence and incidence data could also be used in Bayesian inference. The Bayesian approach could have some benefits compared with the maximum likelihood approach; for example, we could use an autoregressive smoothing submodel for the age-specific hazard parameters in a piecewise constant model with narrow age groups, thus obtaining something close to a nonparametric modeling of a hazard rate (Berzuini and Clayton, 1994).

Although estimation of incidence rates from prevalence data has been studied earlier (Keiding, 1991, 2006), utilization of covariate measurements of the prevalent cases has received little attention. In part, this may be because of the requirement for time-constant covariates. As the covariate values for the prevalent cases are measured after the first event, and it is of interest to study the effects of the risk factors on the event, not the other way round, in particular, it is required that the event of interest or a possible later intervention does not alter the covariate value. The proposed method is intended to be applied to genetic association studies where this assumption is natural.

In principle, our likelihood-based approach can be extended to accommodate also covariates that may be altered due to a disease event, if such measurements are set as missing for the prevalent cases and the distribution of the covariate is modeled as a part of the likelihood. However, the motivation for including the prevalent cases in the analysis is the availability of measurements for some time-invariant covariates since these are the only measurements that can be utilized for these subjects. Karvanen and others (2008) approached the problem of including possibly time-varying covariates using donor-type multiple imputation, where covariate values for prevalent cases are sampled from random donors who are healthy at the cross section. They used donor-type imputation also for other types of missingness caused by the cross-sectional cohort design, namely the missing event times for prevalent cases and left truncation due to death. In their nonparametric approach, the dependencies in the data are retained by drawing all imputed values from the same donor. For comparison, the fully parametric approach described in the present paper requires estimation of several additional parameters describing such dependencies. However, when parametric assumptions are justified, they are advantageous when the amount of data is small. Applying the multiple imputation approach to estimating the effect of family history of cardiovascular disease, Karvanen and others (2008) also found that the inclusion of prevalent cases gives clear gain in efficiency when compared with standard analysis where these cases are omitted. The nonparametric multiple imputation approach may have some computational advantages over the present likelihood-based method. In practice, the 2 approaches are likely to produce similar results, but the likelihood-based approach is better justified theoretically.
Here, we assumed that covariates are observed for the complete cohort, and the likelihood expressions were always conditioned on the covariate data. If the covariate data are collected for only part of the study cohort, for example, using case–cohort or nested case–control sampling, missing covariate data need to be modeled as part of the likelihood. This requires modeling of the distributions of the covariates. In principle, this is a straightforward extension of the proposed likelihood expression. However, the case of several continuous covariates may be computationally demanding because of the resulting multidimensional integrals.

Using time-to-event data from 2 population-based cohorts, we have demonstrated that inclusion of the prevalent cases gives clear efficiency gains compared with the conditional likelihood analysis where the prevalent cases are omitted. The disadvantages of the proposed conditional likelihood analysis are the need to estimate parameters describing mortality and case fatality and the need for numerical integration in the evaluation of the likelihood expression. It should be emphasized that mortality and case fatality may also be covariate dependent. To allow this dependency, the corresponding covariate effect parameters also need to be estimated. Neither the extra parameters nor the numerical integration in 1 dimension should be computationally burdensome, with the possible exception of genomewide association studies. We conclude that in many applications, the added power justifies the more demanding analysis.

**FUNDING**

This work was supported by the European Community (HEALTH-F4-2007-201413, LSHM-CT-2004-005268) and the Academy of Finland (122883 to S.K.).

**ACKNOWLEDGMENTS**

The authors would like to thank professor Esa Läärä of the University of Oulu for helpful comments and professor Veikko Salomaa of the National Institute for Health and Welfare for permission to use the FINRISK data in our illustration. *Conflict of Interest*: None declared.

**REFERENCES**


[Received November 18, 2008; first revision February 20, 2009; second revision March 24, 2009; accepted for publication April 24, 2009]