Regression models for infant mortality data in Norwegian siblings, using a compound Poisson frailty distribution with random scale

TRON ANDERS MOGER*, ODD O. AALEN

Institute of Basic Medical Sciences, Department of Biostatistics, University of Oslo, PO Box 1122 Blindern, N-0317 Oslo, Norway
tronmo@medisin.uio.no

SUMMARY

The power variance function distributions, which include the gamma and compound Poisson (CP) distributions among others, are commonly used in frailty models for family data. In a previous paper, we presented a frailty model constructed by randomizing the scale parameter in a CP distribution. When combined with a parametric baseline hazard, this yields a model with heterogeneity on both the individual and the family level and a subgroup with zero frailty, corresponding to people not experiencing the event. In this paper, we discuss covariates in the model. Depending on where the covariates are inserted in the model, one may have proportional hazards at the individual level, the family level, and a larger group level (for covariates shared by many families, e.g. ethnic groups) or get accelerated failure times. Each of these alternatives gives a specific interpretation of the covariate effects. An application to data infant mortality in siblings from the Medical Birth Registry of Norway is included. We compare the results for some of the different covariate modeling options.

Keywords: Family data; Frailty; Infant mortality; Relative risk; Survival analysis.

1. INTRODUCTION

Traditionally, shared frailty models have been used to analyze survival data on sibships (see, e.g. Hougaard, 2000). Popular shared frailty models use the gamma, inverse Gaussian, and positive stable distributions. Hougaard (1986) considered a 3-parameter family of distributions, the power variance function (PVF) distributions, and showed that it included the former 3 distributions as special cases. Aalen (1992) extended the PVF family to also include the compound Poisson (CP) distributions. The CP distribution has a positive probability of zero frailty, yielding a subgroup of the population that cannot experience the event.

When combined with a parametric baseline hazard, shared frailty models can be insufficiently flexible in some cases. By design, the shared frailty models use a single random variable to model both individual variation due to unobserved individual covariates and variation due to unobserved common covariates. It would be beneficial to have a model which included one random factor for each type of variation to...
improve the fit. In Moger and others (2004), we introduced a CP–PVF model, with both family and individual frailty. The model was further discussed in Moger and Aalen (2005) but without mentioning covariates, which is one of the main topics here. In frailty analysis, covariates are usually included to explain a part of the dependence in survival times modeled by the frailty variable. The CP–PVF model is more flexible in this regard than a shared frailty model, and we will discuss the different alternatives on how to model the covariates in detail.

As an illustration, a CP–gamma version of the model is used to analyze infant mortality in siblings, where we compare results from the different covariate modeling options. We also estimate the relative risk due to unobserved factors, which we call the frailty relative risk (FRR). It is calculated as the risk to one sibling dying given that another sibling has died during the first year of life compared to the risk to one sibling dying given that another sibling has survived this period. The relative risk is a measure that most people working within epidemiology are familiar with, and this is an advantage of the FRR. A graphical presentation of the results is provided by using the conditional survival function. One then plots the survival function given that a sibling has died in infancy and the survival function given that a sibling has survived to see how the dependence within sibships affects the survival for different combinations of the covariates.

In Section 2, the infant mortality data are presented. Section 3 gives a brief description of the CP–PVF model. In Section 4, we present different options on how to model the covariates and introduce the FRR as a way of measuring the dependence due to unobserved factors. Results from the analysis of the infant mortality data follow in Section 5, and a discussion is given in Section 6. The paper considers parametric models, but a comment on semiparametric models is given in Section 6.

2. THE DATA ON INFANT MORTALITY IN SIBLINGS

The Medical Birth Registry of Norway has recorded all births in Norway (population around 4.5 million) since 1967, from the 16th week of gestation onward. By December 31, 1998, 1.8 million births were recorded. Information on all deaths occurring during the first year of life registered by Statistics Norway is linked to the birth records. By use of the national identification number on the mothers, the births may be linked into sibships. We do not consider the father. The average number of children per woman in Norway is around 1.8, including women without children. The proportion of women without any children by the age of 40 has increased from 9.6% for the 1935 cohort to 11.9% for the 1960 cohort (from Statistics Norway’s web pages). The average sibship size in the Birth Registry data is around 2. We do not have access to specific causes of death for the purpose of this study.

Several studies show an increased risk of recurrence of infant deaths in siblings (e.g. Øyen and others, 1996; Guntheroth and others, 1990; Leach and others, 1999), with relative risks and odds ratios around 4–6. Important causes of infant death are sudden infant death syndrome, infections, congenital malformations, and various birth-related causes. Øyen and others (1996) indicate that all these are expected to contribute to the familial aggregation in infant death, except infections, which are mainly due to random factors. However, only around 11% of the total number of deaths was due to infections. Following Øyen and others (1996), we analyze post-perinatal deaths (7–364 days). This means that an infant has to survive the first week to be included in the data. Infant mortality is very rare in Norway, and the prevalence of post-perinatal mortality has dropped from 0.5% in 1967 to 0.2% in 1998.

The database includes some covariates, most of which are known to have an influence on infant mortality. These are birth weight, gestational age, infant’s birth year, mother’s birth year, birth length, maternal age, parity reported by the mother, and gender. Many of the recurrent deaths are probably due to unknown genetic and environmental factors, so these covariates are not expected to reduce the dependence in survival times much. The proportion of missing data is fairly small, it is 0.1% for birth weight, 0.2% for gender, 2.3% for birth length, and 6.1% for gestational age. There are no missing values for the other
covariates. For missing data in the continuous covariates, the mean values are imputed in the analyses. We have excluded all infants with missing gender and 793 infants with unknown gender from the data. In addition, 20 infants where the mother’s identification number was missing were excluded. Multiple births are also excluded, since they are expected to be more closely correlated than siblings in general. The final cohort includes 1814188 infants, with 6551 deaths occurring in 6440 sibships. Ninety-nine sibships have 2 deaths and 6 sibships have 3 deaths.

3. THE CP–PVF MODEL

As usual, we use the multiplicative frailty model, where the hazard for each individual is given as the product of a frailty variable $Z_1$ and a basic rate $\lambda(t)$ common to all individuals. Conditionally on $Z_1$, the individual hazard $h(t)$ is given by

$$h(t|Z_1) = Z_1 \lambda(t).$$

(3.1)

The simplest frailty model for multivariate survival data is the shared frailty model introduced in Clayton (1978). The frailty variable $Z_1$ varies over sibships, and all individuals in a sibship share the same frailty, creating positive dependence between survival times within sibships. The individuals are independent given $Z_1$. The unconditional survival function is given by $S(t) = L_{Z_1}(\Lambda(t))$, where $L_{Z_1}(\bullet)$ is the Laplace transform of $Z_1$ and $\Lambda(t) = \int \lambda(u)du$ is the cumulative baseline hazard function. The gamma distribution is the most popular choice for $Z_1$, mainly because of a simple Laplace transform. Another choice is the CP distribution, which consists of a discrete proportion at 0, corresponding to people not experiencing the event or having zero frailty, and a continuous curve of positive frailties. For some diseases (e.g. purely genetic diseases), having zero frailty can be interpreted as being immune. In the standard parameterization used in Aalen (1992), the CP distribution has scale parameters $\rho_1$ and $\nu$ and a shape parameter $\eta$. A more flexible choice is the PVF distribution, which also has 2 scale parameters, $\rho_2$ and $\theta$, and a shape parameter $\alpha$. If $\theta = 0$, one gets the stable distribution, and if $\alpha = 0$, one gets the gamma distribution. If $\alpha < 0$, one gets the CP distribution in an alternative parameterization. Common parametric choices for the baseline hazard $\lambda(t)$ are the Weibull, exponential, and Gompertz distributions. Hougaard (2000) gives an extensive overview of frailty models.

In Moger and Aalen (2005), we presented a 2-level frailty model, where $Z_1$ was CP distributed with scale parameter $\rho_1$. Unlike the shared frailty models, $Z_1$ does not have a common value within sibships. Instead, it models extra individual heterogeneity and is independent within sibships. The second level of the model is constructed by letting $\rho_1$ follow a PVF distribution, $Z_2$. The frailty $Z_2$ models sibship heterogeneity. All individuals in a sibship share a common value of $Z_2$, thus creating dependence between siblings. Individuals from different sibships are independent. The Laplace transforms for $Z_1$ and $Z_2$ are $\exp(-\rho_1 \Psi_1(s))$ and $\exp(-\rho_2 \Psi_2(s))$, respectively. Since all individuals are independent given $Z_2$, the Laplace transform for the total model is

$$L_{Z_1, Z_2}(s) = E(L_{Z_1}(s)|Z_2) = \int \exp(-\rho_1 \Psi_1(s))f_{Z_2}(\rho_1)d\rho_1 = \exp[-\rho_2 \Psi_2(\Psi_1(s))].$$

(3.2)

The Laplace exponents are given by $\Psi_1(s) = \{1 - [\nu/(\nu + s)]^\eta\}$ and $\Psi_2(s) = 1/\alpha[(\theta + s)^\alpha - \theta^\alpha]$. Since the model is constructed by randomizing a scale parameter on each level, it does not fit into the classes of multiplicative ($Z = Z_1 Z_2$) or additive ($Z = Z_1 + Z_2$) multivariate frailty models discussed, for example, in Hougaard (2000, Chapter 10). The CP–PVF model is constructed more in the way of overdispersion models (see, e.g. Hougaard and others, 1997). Since $Z_1$ is CP distributed, one individual may have a high value of $Z_1$ and thus a high risk of experiencing the event (death/disease), while her sibling may have zero frailty. Alternatively, one could let $Z_1$ be PVF distributed and $Z_2$ be CP distributed. This would imply that a proportion of the sibships have zero frailty, instead of a proportion of the individuals, meaning, for
example, that all individuals in a sibship inherit zero frailty from their parents. The model can be extended to further levels (e.g. individuals, families, neighborhoods) by randomizing $\rho_2$ by another distribution $Z_3$, and so on.

By using the property that all individuals are independent conditional on $Z_2$ and integrating $Z_2$ out, we get the following joint survival function for a sibship of $k$ individuals:

$$S(t_1, \ldots, t_k) = \exp \left\{ -\rho_2 \Psi_2 \left[ \sum_{i=1}^{k} \Psi_1(\Lambda(t_i)) \right] \right\}. \quad (3.3)$$

For the discussion of covariates in Section 4, it is important to note that the subscript $i$ refers to individuals, not sibships. However, some covariates, like mother’s birth year, can take the same value for all members in a sibship, whereas others will be specific to each member. By inserting the functions $\Psi_1(s)$ and $\Psi_2(s)$ of the CP–PVF model, the survival functions are given by

$$S(t_1, \ldots, t_k) = \exp \left\{ -\frac{\rho_2}{\alpha} \left[ \left( \theta + \sum_{i=1}^{k} \left[ 1 - \frac{\nu}{\nu + \Lambda(t_i)} \right] \right) - \theta^\alpha \right] \right\}, \quad \text{if } \alpha \neq 0,$$

$$S(t_1, \ldots, t_k) = \left( \frac{\theta}{\theta + \sum_{i=1}^{k} \left[ 1 - \frac{\nu}{\nu + \Lambda(t_i)} \right]} \right)^{\rho_2}, \quad \text{if } \alpha = 0.$$

The formula for $\alpha = 0$ is found by using the Laplace transform of a gamma distribution $L(s) = \theta / (\theta + s)^{\rho_2}$. The density is found by differentiation with respect to the events. The probability that an individual has zero frailty given $Z_2$, $P(Z_1 = 0|Z_2 = \rho_1)$, is given by $\exp(-\rho_1)$. From Moger and Aalen (2005), the unconditional probability that an individual has zero frailty can then be found by integrating over $Z_2$ and is given by

$$P(Z_1 = 0) = L_{Z_2}(1) = \exp \left\{ -\frac{\rho_2}{\alpha} ((\theta + 1) - \theta^\alpha) \right\}, \quad \text{if } \alpha \neq 0,$$

$$P(Z_1 = 0) = L_{Z_2}(1) = \left( \frac{\theta}{\theta + 1} \right)^{\rho_2}, \quad \text{if } \alpha = 0.$$

### 4. Covariates and FRR

One usually distinguishes between common covariates, which have the same value for all individuals in a sibship, and individual covariates, which have different values for the individuals in a sibship. Common covariates will account for some of the dependence in the survival times within sibships. When including such covariates in the model, the dependence due to frailty should go down.

Table 1 gives an overview of the different models, where the covariates will appear in the joint survival function, and the interpretation of the regression coefficients. The first alternative is to have proportional hazards at the individual level (M1). This yields individual-specific regression effects. The second alternative gives proportional hazards at the sibship level (M2) and gives sibship-specific regression effects. The third alternative yields proportional hazards at the group level (M3, for covariates shared by many sibships, e.g. ethnic groups), giving group-specific regression effects. Finally, one may have an accelerated failure times interpretation of the covariates (M4).

In M1, covariates are included in the baseline hazard $\lambda'(t_i) = \exp(\beta^T X) \lambda(t_i)$ in (3.3). In this case, the regression coefficients $\beta$ are conditional on both $Z_1$ and $Z_2$, giving proportional hazards conditional on both levels. That is, the estimated regression effects are interpreted conditional on having the same value of both the sibship and the individual frailty. For example, one may study a disease, which is caused by
Table 1. Overview of the different covariate modeling options in the CP–PVF model

<table>
<thead>
<tr>
<th>Proportional hazards at</th>
<th>Survival function, (S(t_1, \ldots, t_k))</th>
<th>Interpretation, (\exp(\beta))</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1: Individual level</td>
<td>(\exp\left{-\rho_2 \Psi_2 \left[ \sum_{i=1}^{k} \Psi_1 (\exp(\beta^T X_i) \Lambda(t_i)) \right]\right})</td>
<td>Individual-specific relative risk</td>
</tr>
<tr>
<td>M2: Sibship level</td>
<td>(\exp\left{-\rho_2 \Psi_2 \left[ \sum_{i=1}^{k} \exp(\beta^T X_i) \Psi_1 (\Lambda(t_i)) \right]\right})</td>
<td>Sibship-specific relative risk</td>
</tr>
<tr>
<td>M3: Groups of sibships level</td>
<td>(\exp\left{-\exp(\beta^T X_i) \rho_2 \Psi_2 \left[ \sum_{i=1}^{k} \Psi_1 (\Lambda(t_i)) \right]\right})</td>
<td>Group-specific relative risk</td>
</tr>
<tr>
<td>M4: Accelerated failure times</td>
<td>(\exp\left{-\rho_2 \Psi_2 \left[ \sum_{i=1}^{k} \Psi_1 (\Lambda(t_i) / \exp(\beta^T X_i)) \right]\right})</td>
<td>Accelerated failure times</td>
</tr>
</tbody>
</table>

both genetics and some important individual risk factors. These risk factors might be well known, but it is expensive to collect information on all of them. Hence, your data only include information on some of these risk factors. One would then like to know the effect of the observed risk factors, adjusted for both the levels of individual and shared frailty modeled by \(Z_1\) and \(Z_2\). Note that the genetic component is unknown, creates dependence, and is modeled by \(Z_2\). The missing risk factors are modeled in \(Z_1\).

The second alternative (M2) is to include covariates on the sibship frailty parameter \(\rho_1 = \exp(\beta^T X) \rho_1\) to get proportional hazards conditional on the sibship frailty \(Z_2\). This gives the estimated covariate effects when comparing 2 individuals within a sibship and can be preferred in genetic counseling. It is then interesting to know the effect of a risk factor given that you have a family history of a specific disease, instead of using, for example, population average estimates. A shared frailty model gives exactly the same interpretation of the covariates. As the covariates in M2 are put at sibship level, it is mainly relevant for covariates that are highly correlated within sibships (remember that the frailty \(Z_2\), and hence \(\rho_1\), is shared within sibships on this level). If a covariate is common to all individuals in sibships, it can be put outside the sum in the survival function of M2 in Table 1. In M2, the proportion of individuals with zero frailty will depend on the covariates. The formulas are

\[
P(Z_1 = 0) = L_{Z_2}(\exp(\beta^T X)) = \exp\left\{-\frac{\rho_2}{\alpha} [\theta + \exp(\beta^T X) - \theta^\alpha]\right\}, \quad \text{if } \alpha \neq 0,
\]

\[
P(Z_1 = 0) = L_{Z_2}(\exp(\beta^T X)) = \left(\frac{\theta}{\theta + \exp(\beta^T X)}\right)^{\rho_2}, \quad \text{if } \alpha = 0.
\]

This model is used in Moger and others (2004) in an analysis of testicular cancer patients and their brothers. Testicular cancer incidence has increased as a function of birth cohort. The incidence in 5-year birth cohort intervals is almost a common covariate within brothers and was included in \(\rho_1\).

A third alternative (M3) is to include the covariates in \(\rho_2 = \exp(\beta^T X) \rho_2\). This approach is relevant for risk factors that are common to many sibships, for instance mother’s birth year or ethnicity. It is then interesting to know the covariate effect given that you belong to a certain ethnic group or have a mother born in 1954. Of course, to model unobserved risk factors of this type, one has to apply another level of frailty. This frailty would then be shared by all individuals belonging to the same group and be independent between groups. Covariates at the group level would be interpreted as conditional on the group frailty. In our data, we have no dependence structure above the sibship level and, hence, no group frailty to condition on. However, we have a covariate at the group level, as mother’s birth year is shared by many sibships, and it can be modeled at this level as an illustration. In this case, the relative risk \(\exp(\beta)\) can be interpreted as the change in mean sibship frailty, \(E(Z_2)\), for one group compared to another. If \(Z_2\) is gamma distributed, \(E(Z_2) = \rho_2 \exp(\beta^T X) / \theta\). Hence, an \(\exp(\beta)\) of 2.2 means that the sibship frailty of one group is 2.2 times...
higher than the baseline group or 120% higher. Note that in M1, the covariates can be seen as a scale transformation of \( \lambda(t) \), in M2 as a scale transformation of \( Z_1 \), and in M3 as a scale transformation of \( Z_2 \). Similar to model M2, the proportion of individuals with zero frailty will depend on the covariates:

\[
P(Z_1 = 0) = \exp \left\{ -\frac{\rho_2 \exp(\beta^T X)}{\alpha} (\theta + 1)^\alpha - \theta^\alpha \right\}, \quad \text{if } \alpha \neq 0,
\]

\[
P(Z_1 = 0) = \left( \frac{\theta}{\theta + 1} \right)^{\rho_2 \exp(\beta^T X)} , \quad \text{if } \alpha = 0.
\]

A fourth alternative is to formulate the model as an accelerated failure time model (M4). The integrated baseline hazard is then \( \Lambda(t_i / \exp(\beta^T X_i)) \), which may be inserted in (3.3). Hence, the covariates represent a scale transformation of the time. Model M4 only applies if the baseline hazard in (3.1) is Weibull of the form \( \lambda(t_i) = \alpha \kappa t_i^{\kappa - 1} \). The model is then the same as model M1 with \( \beta = -\beta_{\text{AFT}} \kappa \). The accelerated failure times interpretation is suitable for wear-related processes. Instead of interpreting the effects as relative risks of getting a disease, one might be more interested in how much longer (or shorter) individuals are expected to live for different covariate values. For model M4, a covariate yielding a relative risk of 1.5 would mean a 50% increase in lifetime (i.e. it is positive to have relative risks above 1). As we will see in the application, there can be a fairly large discrepancy between relative risks and accelerated failure times. For example, if the hazard is decreasing, a modest reduction in relative risks may be equivalent to a large increase in lifetime.

The models above are all based on the conditional parameterization of the hazard function (3.1), that is, one gets proportional hazards conditional on the frailty. One may have proportional hazards marginally instead, as in a standard Cox model for univariate data. This will yield population average effects of the covariates and is often of interest from the public health perspective. This approach is closely related to copula models. The relationship between the conditional and marginal parameterization in shared frailty models is described in Hougaard (2000, pp 221–224). Specifically, (3.3) is parameterized by means of the univariate marginal survival functions \( S(t_i) = \exp(-\exp(\beta^T X_i) \Omega(t_i)) \), where \( \Omega(t_i) \) is the integrated hazard in the marginal distribution. First, the inverse relation between the marginal survival function and the conditional integrated hazard in (3.2) has to be found:

\[
S(t_i) = \exp(-\rho_2 \Psi_2(\Psi_1(\Lambda(t_i))) ,
\]

\[
\Lambda(t_i) = \Psi_1^{-1}(\Psi_2^{-1}(\Psi_1^{-1}(\Psi_2^{-1}(\Psi_1^{-1}(\Lambda(t_i)))))).
\]

We then insert this expression into (3.3) to find the multivariate survival function, but this causes all the parameters in the individual frailty \( Z_1 \) to cancel out:

\[
S(t_1, \ldots, t_k) = \exp \left\{ -\rho_2 \Psi_2 \left( \sum_{i=1}^k \Psi_1^{-1}(\Psi_2^{-1}(\Psi_1^{-1}(\Psi_2^{-1}(\Psi_1^{-1}(\Lambda(t_i)))))) \right) \right\}
\]

\[
= \exp \left\{ -\rho_2 \Psi_2 \left( \sum_{i=1}^k \Psi_2^{-1}(\Psi_1^{-1}(\Psi_2^{-1}(\Psi_1^{-1}(\Lambda(t_i)))))) \right) \right\}.
\]

Hence, the CP–PVF model becomes identical to the marginal parameterization of a shared PVF frailty model and will not give a better fit to the data in this case.

The relative risk is often used as a measure of the strength of genetic association in a sibship. Let \( T_1 \) be the survival time of the first sib, while \( T_2 \) is the survival time of the second sib. Define \( F_1(s) = P(T_2 \leq s | T_1 \leq t) \), the probability that sib 2 dies within time \( s \) given that sib 1 has died within time \( t \),
and \( F_0(s) = P(T_2 \leq s | T_1 > t) \), the probability that sib 2 dies within time \( s \) given that sib 1 has survived at time \( t \). The relative risk can be calculated as \( RR = F_1(s)/F_0(s) \). Hence, it is a function of time, and it measures the relative risk due to unobserved factors. We call it the FRR.

The simplest case is when \( s = t \), for example, the relative risk of dying within the first year if your sibling has died/survived the first year. As in Moger and Aalen (2005), by using the Laplace transforms of \( Z_1 \) and \( Z_2 \) and the fact that siblings are independent given \( Z_2 \), one gets the following FRR for the CP–PVF model:

\[
FRR = \frac{1 - 2 \times \exp[-\rho_2 \Psi_2(K(t))] + \exp[-\rho_2 \Psi_2(2 \times K(t))]}{1 - \exp[-\rho_2 \Psi_2(K(t))]} \times \frac{\exp[-\rho_2 \Psi_2(K(t))] - \exp[-\rho_2 \Psi_2(2 \times K(t))]}{\exp[-\rho_2 \Psi_2(K(t))]},
\]

where \( K(t) = \Psi_1(\Lambda(t)) \). In models M1, M2, and M4 with covariates, the FRR will to a negligible degree depend on the values of the covariates, as they do not cancel out. The FRR is calculated as the risk when all covariate values are equal for the individual and her sibling (typically 0). For model M3, however, the covariates enter \( K(t) \), and the FRR will largely depend on the covariate values. We then use the mean value of the covariate for comparison with models M1 and M2. Different covariate values for an individual and her sibling can be included to see how certain combinations of risk factors will affect the relative risk.

By similar calculations as for (4.2), one may derive the conditional survival for an individual at time \( t \) (e.g. surviving the first year), given that a sibling has died within \( t \), \( 1 - F_1(t) \), and the conditional survival for an individual at time \( t \), given that a sibling has survived at \( t \), \( 1 - F_0(t) \). They are given as

\[
1 - F_1(t) = \frac{\exp[-\rho_2 \Psi_2(\Psi_1(\Lambda(t)))]}{1 - \exp[-\rho_2 \Psi_2(\Psi_1(\Lambda(t)))]} - \frac{\exp[-\rho_2 \Psi_2(\Psi_1(\Lambda(t)) + \Psi_1(\Lambda(t_0)))]}{1 - \exp[-\rho_2 \Psi_2(\Psi_1(\Lambda(t)))]},
\]

\[
1 - F_0(t) = \frac{\exp[-\rho_2 \Psi_2(\Psi_1(\Lambda(t))) + \Psi_1(\Lambda(t_0))]}{\exp[-\rho_2 \Psi_2(\Psi_1(\Lambda(t)))]}.
\]

These curves can then be plotted with or without covariates included in the model and compared to the population survival function.

5. APPLICATION TO THE INFANT MORTALITY DATA

We show results from the models M1, M2, and M4, as presented in Section 4. For the common covariate mother’s birth year, we also show model M3. The CP–PVF model yields a subgroup of infants with zero frailty, which for many diseases can be interpreted as immunity. For infant death, however, this is obviously unreasonable from a biological point of view, and the subgroup is not so interesting. Everyone may die in infancy as the result of an accident. In our data, the subgroup with zero frailty can be interpreted as an approximation to the proportion of infants not experiencing the event. The risk of dying of an accident during infancy in Norway is small compared to the risk of dying due to genetic disorders, for example, malformations. Hence, “healthy” infants are estimated to be in the subgroup with zero frailty. To save space, we skip the discussion on how this proportion changes for different models and covariates.

Because of the extremely large amount of data, an analysis of the full database would be computationally very time consuming (around 5 h for maximizing a single likelihood with one covariate in R, more than a week in S-Plus). By using the methods given in the technical report at www.bepress.com/uwbiostat/paper 277, we analyze a case–cohort sample of the data. All sibships with one or more cases are included in the sample. The control sibships are stratified according to sibship size before sampling and exactly 5% are randomly sampled without replacement from each stratum. There are 4 strata for the control sibships, for sibships of size 1, 2, 3, and >4. This yields 45750 control sibships with 89745 and 109762 infants in the total sample. Stratifying according to sibship size is important to get good precision in the estimated
frailty and baseline hazard parameters. The precision of these parameters is mainly decided by the number of familial cases and the prevalence of the disease, and one gets a more precise estimate of the latter by the stratification. According to the results in the technical report, this should give an efficiency of almost 100% for the frailty and baseline hazard parameters compared to a cohort analysis using the same model. The precision of the regression effects will naturally be much lower, perhaps around 70–75%, but this is sufficient as an illustration of the model. To account for the fact that a case–cohort sample is analyzed, sampling weights will enter the likelihood, yielding a standard pseudolikelihood. We use the CP–gamma model in (3.3) to analyze the data. Let there be $k_l$ members in sibship $l$. Let $c_{il}$ indicate whether the survival time $t_{il}$ for individual $i$ in sibship $l$ is censored ($c_{il} = 0$) or not ($c_{il} = 1$). Define $c_s = \sum ic_{il}$ as the number of events in sibship $l$. By combining the likelihood in Section 8 in Moger and Aalen (2005) with the pseudolikelihood in the technical report, we get

$$L(\theta) = \prod_{j=0}^{4} \prod_{p_j} \prod_{l \in D_j} \left[ \prod_{i=1}^{k_l} \left( \frac{\eta^T \lambda(t_{il})}{(v + \Lambda(t_{il}))^{\eta+1}} \right)^{c_{il}} \right] \left( -1 \right)^{c_s} \frac{1}{p_j} \frac{1}{l_{\rho_1}} \left( \sum_{i=1}^{k_s} \left[ 1 - \frac{v}{(v + \Lambda(t_{il}))^\eta} \right] \right), \quad (5.1)$$

where $\theta$ is the vector of parameters to be estimated and $L_{\rho_1}^{(c_{il})} (\cdot)$ is the $c_s$ th derivative of the Laplace transform of $\rho_1$, $L_{\rho_1} (s) = \left[ \theta / (\theta + s) \right]^{p_2}$. The pseudolikelihood is identical to the cohort likelihood in Moger and Aalen (2005, p 54), except for the sampling weights $p_j$ for the case sibships ($j = 0$) and the 4 strata of control sibships, and that the sum is over the case–cohort sample $D_j$ instead of over the full cohort. The baseline hazard in (3.1) is assumed to follow a Weibull distribution, $\lambda(t) = \kappa(t - 6)^{x-1}$, for $t > 6$. The scale parameter of the Weibull distribution is subsumed in the frailty distribution. For the model conditional on sibship frailty only (M2), the covariates appear in the numerator in (5.1), as $\eta^T \lambda(t_{il}) \exp(\beta^T X_{il})$, and in the Laplace transform of $\rho_1$, as $\sum_{i=1}^{k_s} \exp(\beta^T X_{il}) \left[ 1 - \frac{v}{(v + \Lambda(t_{il}))^\eta} \right]$. For the group-specific model M3, the covariates appear in $\rho_2$ only, as $\exp(\beta^T X_{il}) \rho_2$. Standard asymptotic methods are not applicable due to the use of a pseudolikelihood. To estimate the standard errors of the parameters, we use a sandwich-type estimator (see www.bepress.com/uwbiostat/paper277 for details) $A(\theta)^{-1} + A(\theta)^{-1} B(\theta) A(\theta)^{-1}$. Here, $A(\theta)$ is estimated by

$$\hat{A}(\theta) = \sum_{j=0}^{4} \frac{1}{p_j} \sum_{l \in D_j} \hat{I}_l(\theta),$$

where $\hat{I}_l(\theta) = -\hat{c}_l^2 / \hat{c}_l^2 \hat{c}_l^2 \log L_j(\theta)$, the observed information matrix for sibship $l$, and $B(\theta)$ is estimated by

$$\hat{B}(\theta) = \sum_{j=0}^{4} \frac{1}{p_j} \sum_{l \in D_j} \hat{s}_l(\hat{\theta}) \hat{s}_l(\hat{\theta})',$n

where $\hat{s}_l(\theta) = \hat{c}_l / \hat{c}_l \hat{c}_l \log L_j(\theta)$, the score function for sibship $i$.

To find out how to model the continuous covariates, we first categorized them (e.g. into 500-g intervals for birth weight, 50-day intervals for gestational age, etc., this was done for the full database), and studied how the deaths were distributed in the categories in a tabulation. For instance, low birth weight is known to be a risk factor for infant death. One might believe that a very high birth weight also could increase the risk of death, but this did not seem to be the case from the tabulation (because we only consider post-perinatal deaths, and the very heavy babies die within the first week). There could, of course, be interactions between some of the covariates, but this is not considered in this illustration. Hence, the
covariates birth weight, gestational age, birth length, and the birth year of mother and infant are treated as continuous covariates. Maternal age is categorized into $\leq 22$ (reference category), 23–36, and $\geq 37$ years. Parity is categorized into 1 (reference category), 2–3, and $\geq 4$. We do not account for litter relationships in this evaluation.

First, consider an analysis without covariates. Figure 1 shows a Kaplan–Meier plot of the data, with the estimated CP–gamma frailty model. For reference, a shared gamma model with Weibull baseline is also included in the plot. As visually seen from the plot, the CP–gamma model gives a vast improvement in fit compared to the shared gamma model with 3 parameters (2 Weibull parameters + 1 gamma parameter). Although the likelihood ratio test does not apply for pseudolikelihoods, the better fit of the CP–gamma model is also indicated by the log pseudolikelihood values.

### 5.1 Univariate covariate models

Tables 2 and 3 show the results of the univariate analyses of the covariates, for the most interesting parameters. We have excluded the frailty scale parameters to save space. The parameter $\rho_2$ is the shape parameter of the gamma distribution describing sibship heterogeneity. Smaller values for $\rho_2$ means more skewly distributed frailty among sibships. In addition, we show regression effects as relative risks/accelerated failure times as $\exp(\beta)$ with 95% confidence interval ($95\% \text{ CI} = \exp(\hat{\beta} \pm 1.96 \times \text{SE}(\hat{\beta}))$. The FRR measures the dependence in the data with and without covariates and is calculated by inserting the estimated parameters into the CP–gamma version of (4.2) with $t = 364$. Model M4 is the same as model M1, but it uses a different parameterization for the regression effect.

The Weibull shape parameter $\kappa$ has stable values within each model for all covariates, yielding a decreasing baseline hazard. This is due to the fact that we only consider deaths during the first year, where the risk is greatest during the first weeks and then gradually declines. For model M1, a 500-g increase in birth weight yields a relative risk of 0.52, meaning a 48% reduced risk of death during the post-perinatal...
Table 2. Log pseudolikelihood values, parameter estimates/standard errors (SEs), and univariate covariate effects for different model options: M1 = conditional on full frailty, M2 = conditional on $Z_2$, M3 = in $\rho_2$, and M4 = accelerated failure times. See text for details

<table>
<thead>
<tr>
<th>No Covariates</th>
<th>log $L$</th>
<th>$\kappa$ (SE)</th>
<th>$\rho_2$ (SE)</th>
<th>$\beta$ (SE)</th>
<th>exp($\beta$) (95% CI)</th>
<th>FRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weight (per 500 g)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>-78186</td>
<td>0.61 (0.01)</td>
<td>0.47 (0.07)</td>
<td>-0.65 (0.004)</td>
<td>0.52 (0.52–0.53)</td>
<td>3.13</td>
</tr>
<tr>
<td>M2</td>
<td>-78199</td>
<td>0.81 (0.01)</td>
<td>0.36 (0.05)</td>
<td>-0.47 (0.009)</td>
<td>0.63 (0.61–0.64)</td>
<td>3.77</td>
</tr>
<tr>
<td>M4</td>
<td>-78186</td>
<td>0.61 (0.01)</td>
<td>0.47 (0.07)</td>
<td>1.06 (0.007)</td>
<td>2.89 (2.85–2.93)</td>
<td>3.13</td>
</tr>
<tr>
<td>Gestational age (per 30 days)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>-79067</td>
<td>0.73 (0.01)</td>
<td>0.46 (0.06)</td>
<td>-1.18 (0.019)</td>
<td>0.31 (0.30–0.32)</td>
<td>3.18</td>
</tr>
<tr>
<td>M2</td>
<td>-79143</td>
<td>0.81 (0.01)</td>
<td>0.35 (0.05)</td>
<td>-0.59 (0.020)</td>
<td>0.55 (0.54–0.57)</td>
<td>3.97</td>
</tr>
<tr>
<td>M4</td>
<td>-79067</td>
<td>0.73 (0.01)</td>
<td>0.46 (0.06)</td>
<td>1.63 (0.021)</td>
<td>5.09 (4.90–5.30)</td>
<td>3.18</td>
</tr>
<tr>
<td>Infant’s birth year (per 5 years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>-79886</td>
<td>0.82 (0.02)</td>
<td>0.41 (0.06)</td>
<td>-0.20 (0.006)</td>
<td>0.82 (0.81–0.83)</td>
<td>3.43</td>
</tr>
<tr>
<td>M2</td>
<td>-79837</td>
<td>0.81 (0.01)</td>
<td>0.41 (0.06)</td>
<td>-0.11 (0.007)</td>
<td>0.90 (0.88–0.91)</td>
<td>3.45</td>
</tr>
<tr>
<td>M4</td>
<td>-79886</td>
<td>0.82 (0.02)</td>
<td>0.41 (0.06)</td>
<td>0.24 (0.007)</td>
<td>1.27 (1.25–1.29)</td>
<td>3.43</td>
</tr>
<tr>
<td>Birth length (per cm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>-79040</td>
<td>0.66 (0.01)</td>
<td>0.46 (0.07)</td>
<td>-0.25 (0.004)</td>
<td>0.79 (0.78–0.79)</td>
<td>3.13</td>
</tr>
<tr>
<td>M2</td>
<td>-79282</td>
<td>0.81 (0.01)</td>
<td>0.36 (0.05)</td>
<td>-0.10 (0.003)</td>
<td>0.90 (0.90–0.91)</td>
<td>3.84</td>
</tr>
<tr>
<td>M4</td>
<td>-79040</td>
<td>0.66 (0.01)</td>
<td>0.46 (0.07)</td>
<td>0.37 (0.005)</td>
<td>1.45 (1.43–1.46)</td>
<td>3.13</td>
</tr>
<tr>
<td>Mother’s birth year (per 5 years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>-79956</td>
<td>0.80 (0.02)</td>
<td>0.41 (0.06)</td>
<td>-0.07 (0.010)</td>
<td>0.94 (0.92–0.95)</td>
<td>3.43</td>
</tr>
<tr>
<td>M2</td>
<td>-79930</td>
<td>0.81 (0.01)</td>
<td>0.41 (0.06)</td>
<td>-0.06 (0.007)</td>
<td>0.95 (0.93–0.96)</td>
<td>3.43</td>
</tr>
<tr>
<td>M3</td>
<td>-79930</td>
<td>0.81 (0.02)</td>
<td>0.76 (0.06)</td>
<td>-0.06 (0.007)</td>
<td>0.95 (0.93–0.96)</td>
<td>3.43</td>
</tr>
<tr>
<td>M4</td>
<td>-79956</td>
<td>0.80 (0.02)</td>
<td>0.41 (0.06)</td>
<td>0.08 (0.011)</td>
<td>1.09 (1.07–1.10)</td>
<td>3.43</td>
</tr>
</tbody>
</table>

period. Model M2 yields a 37% reduced risk. There is a similar picture for the other covariates, except for parity, with model M1 yielding stronger absolute values for the covariate effects than M2. For model M4, a 500-g increase in birth weight yields a relative effect on the timescale of 2.89 or a 189% estimated increase in lifetime within the first year. This is reasonable from the data, since most deaths occur during the first few weeks. Hence, an increase in birth weight is expected to cause a large relative increase in lifetime. The regression coefficient in model M3 for mother’s birth year gives the following interpretation: a 5-year increase in mother’s birth year reduces the level of sibship frailty by 5% (relative risk of 0.95). Models M2 and M3 coincidentally appear to be identical in the table. As explained in the previous section, they are not the same model.

The CP–gamma model without covariates yields an FRR of 3.44. Including birth weight in the model reduces the FRR to 3.13 in model M1, but it increases to 3.77 in model M2. In contrast to models M1/M4, model M2 yields increased FRRs also for gestational age and birth length. Birth weight, birth length, and gestational age are all continuous covariates with low correlation within sibships. They reduce the frailty variance due to unobserved individual factors (as modeled by $Z_1$). However, since they are slightly correlated within sibships, they also reduce some of the frailty variance due to common factors (as modeled by $Z_2$). If these covariates capture the individual frailty variation better when included in $\rho_1$ (M2) than in $\lambda(t)$ (M1/M4), the dependence could very well increase in model M2 and decrease in models M1/M4. The effect of the unknown, common covariates would then appear stronger in M2. It seems likely that this
Table 3. Log pseudolikelihood values, parameter estimates/standard errors (SEs), and univariate covariate effects for different model options: M1 = conditional on full frailty, M2 = conditional on \( Z_2 \), and M4 = accelerated failure times. For maternal age and parity, \( \beta \)s are shown for 2 categories compared to the reference group. See text for details

<table>
<thead>
<tr>
<th></th>
<th>( \log L )</th>
<th>( \kappa ) (SE)</th>
<th>( \rho_2 ) (SE)</th>
<th>( \beta ) (SE)</th>
<th>( \exp(\beta) ) (95% CI)</th>
<th>FRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal age (0–22 [reference category], 23–36, &gt;36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>−79941</td>
<td>0.86 (0.03)</td>
<td>0.41 (0.06)</td>
<td>−0.72, −0.22</td>
<td>0.48, 0.80 (0.001), (0.002)</td>
<td>0.48–0.49, (0.79–0.80)</td>
</tr>
<tr>
<td>M2</td>
<td>−79957</td>
<td>0.81 (0.01)</td>
<td>0.41 (0.06)</td>
<td>−0.45, −0.20</td>
<td>0.63, 0.82 (0.04), (0.08)</td>
<td>0.59–0.68, (0.71–0.95)</td>
</tr>
<tr>
<td>M4</td>
<td>−79941</td>
<td>0.86 (0.03)</td>
<td>0.41 (0.06)</td>
<td>0.85, 0.26</td>
<td>3.42 (0.005), (0.006)</td>
<td>2.33–2.36, (1.25–1.33)</td>
</tr>
<tr>
<td>Parity (1 [reference category], 2–3, &gt;3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>−79972</td>
<td>0.81 (0.01)</td>
<td>0.41 (0.06)</td>
<td>0.01, 0.05</td>
<td>1.01, 1.05 (0.008), (0.012)</td>
<td>0.99–1.02, (1.03–1.08)</td>
</tr>
<tr>
<td>M2</td>
<td>−79957</td>
<td>0.81 (0.01)</td>
<td>0.44 (0.06)</td>
<td>0.11, 0.25</td>
<td>1.12, 1.28 (0.03), (0.06)</td>
<td>1.07–1.18, (1.13–1.44)</td>
</tr>
<tr>
<td>M4</td>
<td>−79972</td>
<td>0.81 (0.01)</td>
<td>0.41 (0.06)</td>
<td>0.005, −0.05</td>
<td>3.41 (0.009), (0.013)</td>
<td>0.99–1.02, (0.93–0.98)</td>
</tr>
<tr>
<td>Gender (reference category male)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>−79940</td>
<td>0.79 (0.01)</td>
<td>0.41 (0.06)</td>
<td>−0.35 (0.013)</td>
<td>3.44 (0.70, 0.69–0.72)</td>
<td>0.70–0.72</td>
</tr>
<tr>
<td>M2</td>
<td>−79927</td>
<td>0.81 (0.01)</td>
<td>0.41 (0.06)</td>
<td>−0.24 (0.026)</td>
<td>3.43 (0.79, 0.75–0.83)</td>
<td>0.79–0.83</td>
</tr>
<tr>
<td>M4</td>
<td>−79940</td>
<td>0.79 (0.01)</td>
<td>0.41 (0.06)</td>
<td>0.44 (0.013)</td>
<td>3.44 (1.55, 1.51–1.59)</td>
<td>1.55–1.59</td>
</tr>
</tbody>
</table>

happens here due to the fact that M2 provides a better fit as indicated by the log pseudolikelihood values. Of course, if we instead use a common value of the covariate within sibships in model M2 (by using, e.g. the mean value of the covariate), we remove the individual variation in the covariate and get an from M2 that is comparable to models M1/M4. In ordinary regression, additional covariates can both increase and decrease the effects of covariates already included in the model. In frailty analysis, this also applies to the dependence, which models the effects of unknown covariates.

The covariates gender, mother’s birth year, maternal age, and infant’s birth year all have a significant effect on the mortality but do not seem to have a great influence on the dependence. For model M3, where \( \rho_2^* = \exp(\beta^T X)\rho_2 \) in (4.2), one may use 1955 (the mean birth year of the mothers) as the covariate value to get an FRR that is comparable to the other models. This yields an FRR of 3.43. However, one may also calculate the FRR, for example, for the group of all mothers born in 1945, which gives an FRR of 3.18, or for the group of mothers born in 1970, which gives an FRR of 3.88. The increase in dependence as a function of mother’s birth cohort is probably due to the lower prevalence of nonfamilial infant deaths in more recent birth cohorts. A few familial cases will then have a greater impact on the dependence.

### 5.2 Multivariate covariate models

Table 4 shows the results of a multivariate analysis. Although the log pseudolikelihood values indicate that model M2 fits the data best in this case, one would, in practice, first decide on an interpretation of the covariate effects, then do the analysis. Hence, one would not normally compare the fit between these models. Since there is collinearity between the covariates infant’s birth year, maternal age, and mother’s birth year, we use 1955 as the covariate value for M3. This yields an FRR of 3.43. One may also calculate the FRR for the group of all mothers born in 1945, which gives an FRR of 3.18, or for the group of mothers born in 1970, which gives an FRR of 3.88. The increase in dependence as a function of mother’s birth cohort is probably due to the lower prevalence of nonfamilial infant deaths in more recent birth cohorts. A few familial cases will then have a greater impact on the dependence.
Table 4. Parameter estimates with standard errors (SEs) for the 3 different models, with multivariate estimates of the covariate effects. Est = estimate, M1 = conditional on full frailty, M2 = conditional on Z2, and M4 = accelerated failure times

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model M1</th>
<th></th>
<th></th>
<th>M2</th>
<th></th>
<th></th>
<th>M4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
</tr>
<tr>
<td>κ</td>
<td>0.63</td>
<td>0.007</td>
<td>0.81</td>
<td>0.014</td>
<td>0.63</td>
<td>0.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ρ2</td>
<td>0.57</td>
<td>0.063</td>
<td>0.44</td>
<td>0.065</td>
<td>0.57</td>
<td>0.063</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariate</td>
<td>exp(β)</td>
<td>95% CI</td>
<td>exp(β)</td>
<td>95% CI</td>
<td>exp(β)</td>
<td>95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth weight</td>
<td>0.56</td>
<td>(0.55–0.57)</td>
<td>0.53</td>
<td>(0.51–0.55)</td>
<td>2.52</td>
<td>(2.47–2.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gestational age</td>
<td>0.81</td>
<td>(0.78–0.84)</td>
<td>0.95</td>
<td>(0.91–0.99)</td>
<td>1.40</td>
<td>(1.31–1.48)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infant’s birth year</td>
<td>0.85</td>
<td>(0.84–0.86)</td>
<td>0.88</td>
<td>(0.87–0.89)</td>
<td>1.28</td>
<td>(1.27–1.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth length</td>
<td>0.98</td>
<td>(0.97–0.99)</td>
<td>1.07</td>
<td>(1.06–1.08)</td>
<td>1.03</td>
<td>(1.02–1.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23–36</td>
<td>0.61</td>
<td>(0.58–0.63)</td>
<td>0.65</td>
<td>(0.60–0.70)</td>
<td>2.21</td>
<td>(2.07–2.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;36</td>
<td>0.72</td>
<td>(0.65–0.79)</td>
<td>0.74</td>
<td>(0.64–0.86)</td>
<td>1.69</td>
<td>(1.48–1.92)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–3</td>
<td>1.61</td>
<td>(1.56–1.65)</td>
<td>1.57</td>
<td>(1.49–1.67)</td>
<td>0.47</td>
<td>(0.45–0.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;3</td>
<td>2.00</td>
<td>(1.92–2.08)</td>
<td>1.93</td>
<td>(1.69–2.22)</td>
<td>0.33</td>
<td>(0.31–0.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>0.66</td>
<td>(0.63–0.70)</td>
<td>0.74</td>
<td>(0.70–0.78)</td>
<td>1.92</td>
<td>(1.83–2.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log L</td>
<td>−77678</td>
<td></td>
<td>−77643</td>
<td></td>
<td>−77678</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Birth year (if one knows the value of 2 of these covariates, one also know the value of the third), we have only included the first 2 in the multivariate model. By using (4.2) with all covariates inserted, we find that the FRR has dropped to 2.75 for models M1/M4 and to 3.24 for M2. The effects of birth length and gestational age have become much reduced, almost exclusively due to the confounding effect of birth weight. Being, for example, a tall infant is not an advantage unless the weight is also higher. Gestational age and birth length are still significant, however, and this is of course due to the large size of the data set. There is also a qualitative difference in the effect of birth length between models M1 and M2. Generally, M1 yields stronger covariate effects than M2 in univariate analyses (the same is true when comparing sibship effects to population average effects in a standard shared frailty model). However, the magnitude of the effect difference varies from covariate to covariate. This also applies to the effect of confounders in the different models. Hence, qualitative differences in covariate effects will naturally occur sometimes in multivariate analyses. The effect of parity has increased greatly compared to the univariate analysis, due to confounding with all the other covariates, except gender. For gender, infant’s birth year, and birth weight, there are smaller differences from the univariate analysis.

Figure 2 shows conditional survival functions (4.3) for model M1 with t0 = 364 and different covariate values compared to the Kaplan–Meier plot based on all data. In Figure 2 left, the 2 siblings have birth weight = 3500 g, gestational age = 280 days and birth length = 50 cm, they are girls born in 1981, Sib 1 has parity 1 and Sib 2 has parity 2, and they have a mother in the oldest age group. The figure clearly shows the effect a dead sibling has on survival. Both survival curves are below the Kaplan–Meier curve, since having a mother in the oldest age group has a relatively large negative effect on survival. In Figure 2 right, however, Sib 2 has a birth weight of 4000 g, a gestational age of 300 days, and both are born in 1995 (other covariates are the same). The effect on survival from having a dead sibling is much smaller in this case due to more beneficial values of several covariates. Since both siblings are born in 1995, the curves are now shifted upward compared to the Kaplan–Meier (mean birth year for all data is 1981).
Fig. 2. The effect a dead sibling has on survival, as estimated from the fully conditional model (M1). Left: Siblings 1 and 2 have the same values for all covariates except parity and are born in 1981. Right: Sibling 1 has 500 g higher birth weight and 20 days longer gestational age than Sibling 2, and they are born in 1995. See the text for details.

6. DISCUSSION

We also did the analyses using the marginal model (4.1). With a Weibull baseline, this model gave a poor fit to the data, as it does not have the extra individual frailty. However, the $\beta$ estimates were confirmed using a semiparametric Cox regression using the independence working model approach. As expected, the marginal model yielded smaller absolute values of the regression coefficients than the other models, as they are population averages. However, the estimates were very close to those from model M2. The estimates from the marginal model might be biased, as it gave a much worse fit to the data. Model M1 is generally expected to yield the largest absolute values for most regression coefficients, as they are conditioned on comparing 2 individuals with the same value of both the sibship and the individual frailty, whereas model M2 and marginal models will give smaller effects. Since the standard errors in most cases are similar or smaller for M1 compared to model M2 and the marginal model, this also means that M1 will give lower $p$-values for the regression coefficients from a Wald test.

The estimate for the FRR, obtained from the analysis without covariates, is somewhat lower than the estimate in Øyen and others (1996). By combining the causes of death in their paper, one gets a total relative risk of 3.74. They calculated relative risks of recurrence in second birth by outcome of first birth. Hence, only sibships of 2 or more infants were included, and they only used the first 2 births in their study. In our analysis, the sibships consist of anything from 1 to 15 siblings. The ability to include complete sibships, survival times, and covariates in the analysis is an advantage of the frailty approach. The sibships with only one child contribute to the frailty parameter estimates, and thus the dependence, since they affect the prevalence of the outcome. The continuation rate among mothers with a first loss is somewhat higher than among those with a first survivor (83% versus 69%, from Øyen and others, 1996). This can be due to the fact that parents want a specified number of (living) children and thus continue until this is obtained, but it could also indicate that the 1-child survivors come from low-risk sibships. Hence, they should be included in the analysis. One may also calculate FRRs for triplets of observations, for instance the probability of dying given that 1 out of 2 siblings has died compared to the probability of dying given that both have survived. This means that it is possible to get a more general picture of the
relative risks in a frailty analysis. As an alternative to the conditional survival functions in (4.3), one may plot the relative hazards (see Hougaard, 2000, pp 293–295).

The FRR greatly depends on the gamma shape parameter $\rho_2$ in all models. The larger the value for this parameter, the lower the FRR becomes, since the frailty is then less skewly distributed over sibships. This property is similar to other dependence measures for shared gamma frailty models. For instance, Kendall’s $\tau$ depends on $\rho_2$ only (Kendall’s $\tau$ is $1/(1+2\rho_2)$ for a shared gamma frailty model). We have not given standard errors for the FRRs in Section 5. If we analyzed cohort data, standard errors and CIs could be found by bootstrap. However, to our knowledge, bootstrap methods for multivariate case–cohort survival data have not yet been developed. Unpublished results in a PhD thesis by Moger lead us to expect that the CIs for the FRR will be fairly narrow (perhaps $\ln(\text{FRR}) \pm 0.5$).

For rare events, the CP–PVF model is a simple way of achieving a good fit to the data due to the subgroup with zero frailty. Also, when using a parametric baseline hazard $\lambda(t)$ in (3.1), one may get a large improvement in fit compared to a shared PVF frailty model. This follows from the fact that the CP–PVF model uses separate distributions for individual and family variation. The shortcomings regarding lack of fit in a shared PVF model can be overcome by fitting a semiparametric model (nonparametric $\lambda(t)$). However, the CP–PVF model is still more flexible with regard to the modeling of covariates. Options M1 and M4 are not possible in a shared PVF model, as it only models family-level frailty. Covariates in a shared PVF model are then interpreted conditional on the family frailty, corresponding to model M2. Similarly, accelerated failure times in a shared frailty model will be conditional on the family frailty only, which is different from M4. In addition, the CP–PVF model can reflect more biological models. An example is the analysis of testicular cancer (Moger and others, 2004, and references therein), where we assumed a Weibull distribution for $\lambda(t)$. The Weibull distribution is a good approximation for the time to tumor in carcinogenesis, and the classical multistage model for cancer development (Armitage and Doll, 1954) leads to the assumption that $\lambda(t)$ follows a Weibull distribution. One further assumes that a small minority of men gets some sort of damage in fetal life, whereas the majority is believed to be nearly immune. This points to a CP distribution for the individual frailty $Z_1$. The damage could be due to factors in the mother, for example, too high estrogen levels, that are skewly distributed among the population of mothers. This indicates the use of a gamma distribution for $Z_2$, and the final model gave a good fit to the data. Note that, in a shared PVF frailty model, immunity (or zero frailty) would be inherited, and hence shared by all individuals in a sibship. The CP–PVF model gives the possibility that immunity is a random, individual event. In the application presented here, we have no biological reason for using a Weibull distribution, other than that the resulting model gave a good fit to the data. Although model M4 relies on the Weibull assumption, the other models do not and may be fitted with other parametric choices for $\lambda(t)$. Note that the FRR (4.2) and conditional survival function (4.3) may also be derived for shared frailty models.

In semiparametric frailty models, the individual heterogeneity will usually be subsumed in $\lambda(t)$. Without covariates, a semiparametric CP–PVF model is then identical to a shared PVF frailty model, as it is not possible to identify the parameters in the distribution of the individual frailty $Z_1$. With covariates, however, it should be possible to identify the individual parameters, although we have not tried this yet. It was not relevant for the application, as we analyzed case–cohort data. Extending the case–cohort methods to semiparametric models is not trivial. As shown in (4.1), the CP–PVF model is identical to a shared PVF model in the case of proportional hazards marginally. If the dependence is not so much of interest and one wants population average effects of the covariates, the model can be fit within a generalized estimation equations frame or by using the independent working model approach. The latter is implemented in R and S-Plus. Several shared frailty models are also implemented in R and S-Plus. Accelerated failure times models are difficult to fit in a semiparametric frame.

Application of higher level models will be the focus of future research. The model is fairly easily extended to more levels by adding another PVF distribution to $\rho_2$. This makes it possible to analyze data
Regression models for infant mortality data

591

with more complicated dependence structures. We plan to analyze data on melanoma incidence in families of parents and children from the Swedish Multi-Generation Register, which includes more than 10 million individuals in the full cohort. The case–cohort methods for family data, briefly described here, will then be very useful.

ACKNOWLEDGMENTS

We are grateful to the Medical Birth Registry of Norway and Rolv Terje Lie for giving us access to their data. Also, thanks to the Department of Biostatistics, University of Washington, where most of this work was done. Conflict of Interest: None declared.

FUNDING

Research Council of Norway (160627/V50).

REFERENCES


[Received March 2, 2007; revised January 14, 2008; accepted for publication January 22, 2008]