6-19-2006

Hierarchical Lévy Frailty Models and a Frailty Analysis of Data on Infant Mortality in Norwegian Siblings

Tron Anders Moger
University of Washington/University of Oslo, tronm@u.washington.edu

Odd O. Aalen
University of Oslo, o.o.aalen@medisin.uio.no

1 INTRODUCTION

Traditionally, shared frailty models have been used to analyze survival data on families (see e.g. Hougaard, 2000). Popular shared frailty models use the gamma, inverse Gaussian and positive stable distributions. Hougaard (1986) considered a three-parameter family of distributions, the power variance function (PVF) distributions, and showed that it included the former three distributions as special cases. Aalen (1992) extended the PVF family, to also include the compound Poisson distributions, generated by independent gamma variables. The compound Poisson distribution has a positive probability of zero frailty, yielding an immune proportion of the population.

Shared frailty models can be insufficient in some cases, especially when combined with a parametric baseline hazard. 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 *et al.* (2004), we introduced a compound Poisson-PVF model, with both family and individual frailty. The model was further discussed in Moger and Aalen (2005). In this paper, the model is extended by applying the same technique to the time parameter, ρ, in the more general family of distributions determined by non-negative Lévy processes. This family covers all the distributions mentioned above, among others. The model is fairly easily extended into handling more general dependence structures, e.g. several levels of dependence, and genetic models for parents and children.

As an illustration, a compound Poisson-gamma version of the model, as used in Moger *et al.* (2004) and Moger and Aalen (2005), will be applied to general infant mortality data in siblings from the Medical Birth Register of Norway. Several studies show an increased risk of recurrence of infant deaths in siblings. The cause of death which has received particular attention the last few decades is sudden infant death syndrome (SIDS). In a previous analysis of data from the Medical Birth Registry of Norway, Øyen *et al.* (1996) found a relative risk of recurrence of post-perinatal SIDS in the second birth when SIDS occurred in the first birth of 5.9. The corresponding relative risk for recurrence of post-perinatal non-SIDS death was 6. Specifically, all causes of non-SIDS infant deaths showed high relative risks of recurrence, except infections. In a study in Oregon, Guntheroth *et al.* (1990) found relative risks of 5.4 and 6 for recurrence of SIDS and non-SIDS death in subsequent siblings, respectively. In a population based case-control study in the UK, Leach *et al.* (1999) found an odds ratio of 3.82 for a prior infant death in the SIDS group. In addition, infants with explained deaths were significantly more likely than control subjects to have had a previous sibling death in infancy, with an odds ratio of 5.96.

The familial aggregation of infant mortality could be due to possibly unknown genetic or environmental causes. SIDS may be associated with strong environmental risk factors, such as poor living conditions, sleeping position or maternal smoking (Beckwith (1990), Fleming *et al.* (1990) and Beal (1992)). A genetic risk factor may be sleep apnea syndrome (Pillar and Lavie (1995)), which shows strong familial aggregation, indicating that it is an inherited syndrome. Øyen *et al.* (1996) suggests that SIDS can be caused by an interaction between genetic susceptibility and an environmental factor. Other important causes of infant death are infections, congenital malformations and various birth-related causes. Both genetic and environmental factors interact in the etiology of congenital malformations (Lie *et al.* (1994)). Birth-related infant deaths may be explained by pregnancy and labour complications specific to certain mothers (Kåregård and Gennser (1986)). Infections and random accidents are not expected to contribute to the familial aggregation.

Since frailty models deal with heterogeneity due unknown factors, it is tempting to analyze survival data on infants from a frailty point of view. In the papers mentioned above, all obtain simple estimates for the relative risks by basically comparing the mortality incidence in the sibships who have experienced infant deaths to the incidence in the general population. We further expand on the previous analysis by Øyen *et al.* (1996) by including covariates in the frailty model. There are several different options on where to include the covariates, each yielding different interpretations of

the regression coefficients. In the analysis of the data, results from some of the different approaches are compared. One may also estimate the relative risk due to unobserved factors, which we call the frailty relative risk, calculated as the risk of dying given that a sibling has died in infancy, compared to the risk of dying given that a sibling has survived this period. The advantage of using the relative risk as a measure of the dependence in the data, is that it is a measure most people working with epidemiological data is familiar with. Depending on what covariates one includes in the model, the relative risk due to unobserved factors can either decrease or increase. One may also calculate the conditional survival function given that a sibling has died in infancy, and the survival function given that a sibling has survived, to give a graphical presentation of the results.

In Section 2, the infant mortality data are presented. Section 3 gives a brief introduction to frailty models and the Lévy distribution, and the techniques for constructing hierarchical Lévy frailty models, are presented in Section 4. In Section 5 we present different options on how to model the covariates, and introduce the frailty relative risk as a way of measuring the dependence due to unobserved factors. Results from the analysis of the infant mortality data follows in Section 6, and a discussion is given in Section 7.

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 31th December 1998, 1986576 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 size of the sibships is about two, slightly higher than the average number of children per woman in Norway, which is around 1.8. The proportion of women without any children by the age of 40, has increased from ca. 10% for the 1935-cohort to ca. 13% for the 1960-cohort (from Statistics Norway's web pages). Since we do not have access to specific causes of death for the purpose of this study, we are only able to analyze the data with general mortality as the outcome. However, we do not find it unreasonable to assume that many of the deaths occurring during the first year are due to some genetic or common environmental factors which have a great impact on the infants survival. The study by Øyen *et al.* (1996) indicates that the proportion of deaths due to e.g. infections, which is not expected to show familial aggregation, is only around 11%.

Following Øyen *et al.* (1996), we analyze post-perinatal deaths (7-364 days). This means that an infant have 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, length, mother's age, parity and gender. The proportion of missing data is fairly small, ranging from 0.2% for birth weight and 2.4% for length, to 5.9% for gestational age and 6.0% for gender. There are no missing values for the other covariates. For missing data in the continuous covariates, the mean value are imputed in the analyses. However, 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, was 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. 99 sibships have two deaths and six sibships have three deaths. Due to the demanding computational time involved when analyzing the full database, we will apply the methods in Moger *et al.* (in revision) to analyze a case-cohort sample.

3 FRAILTY MODELS AND LÉVY FRAILTY DISTRIBUTIONS

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

$$h(t|Z) = Z\lambda(t) \tag{2.1}$$

The simplest frailty model for survival data where individuals in groups may be correlated, is the shared frailty model, introduced in Clayton (1978). The frailty variable Z varies over groups, and all individuals in a group share the same frailty, creating positive dependence between survival times. The individuals are independent given Z. Let $L_Z(\bullet)$ denote the Laplace transform of Z and let $\Lambda(t) = \int \lambda(u)du$ be the cumulative baseline hazard function. If a group consists of k individuals, the unconditional joint survival function is

$$S(t_1, ..., t_k) = \mathrm{E}(e^{-Z(\Lambda(t_1)+...+\Lambda(t_k))}) = L_Z(\textstyle\sum_{i=1}^{k} \Lambda(t_i))$$

The density is found by differentiation with respect to $t_1, ..., t_k$. Hence, it is advantageous to apply frailty distributions with a simple Laplace transform, which is then easy to differentiate for any number of events. Mainly because of this, the most common distribution for Z is the gamma distribution. A more flexible choice is the PVF distribution, which includes the gamma, stable, inverse Gaussian and compound Poisson distributions as special cases. Common parametric choices for the baseline hazard $\lambda(t)$ are the Weibull, exponential and Gompertz distributions. Hougaard (2000) gives an extensive overview of shared frailty models.

Throughout this paper, Z is assumed to follow frailty distributions defined by non-negative Lévy processes, which in this paper is a process with non-negative, independent, time-homogeneous increments. The Laplace transform of the frailty variable following such a process $Z = \{Z(\rho) : \rho \geq 0\}$, is

$$L_Z(s) = \mathrm{E} \exp[-sZ(\rho)] = \exp[-\rho\Psi(s)] \tag{2.2}$$

by the Lévy-Khintchine formula. Here, ρ corresponds to the time parameter t in a Lévy process $Z(t)$, s is the argument of the Laplace transform, and $\Psi(s)$ is the Laplace exponent or cumulant generating function. The family of Lévy distributions covers most of the common frailty distributions, including the PVF distribution. This formulation may allow for frailties that develop over time, for which ρ is not a constant, but this is not considered here. For more information on frailty models derived from Lévy processes, see Aalen and Hjort (2002) and Gjessing *et al.* (2003).

4 HIERARCHICAL LÉVY FRAILTY MODELS

An important limitation of the shared frailty models is the fact that all members of a family have the same frailty. This can be inappropriate, since one would also expect some individual variation due to non-shared genes and environmental factors, and different degrees of dependence for different types of relatedness, e.g. siblings, families, neighborhoods. In Moger *et al.* (2004), we introduced a frailty model based on the compound Poisson distribution with random scale. By applying a PVF distribution to a scale parameter in the compound Poisson frailty model, one gets a model with variation on both family and individual level. Some further discussion of the properties of the model is found in Moger and Aalen (2005). Since the compound Poisson distribution is included in the family of Lévy frailty distributions, a small extension of the compound Poisson-PVF frailty model can be accomplished by using the more general Lévy frailty distributions for the individual heterogeneity. This yields a hierarchical Lévy model, and it will be discussed here. We will not give any details on the likelihood construction for the different models in this paper. However, by

using the same techniques as in Moger and Aalen (2005), this should be straightforward also for more complex models.

In Moger and Aalen (2005), the compound Poisson-PVF model was constructed by applying a PVF distribution on ρ in (2.2). The compound Poisson distribution models the heterogeneity on the the individual level, where all individuals have independent frailties. The PVF distribution on ρ models the family heterogeneity, so that all individuals in a family share a common value of ρ, thus creating dependence between relatives. Individuals from different families are independent. More generally, let Z_1 be the frailty variable for the individual level. This variable will often have independent values for all individuals. Let $Z_2, ..., Z_k$ be the frailty variables for higher levels, which will typically be independent for some members of a family, but shared for others. The variable Z_k may have the same value for all individuals in a family. Let each Z_i follow a Lévy distribution with Laplace transform $L_{Z_i}(s) = \exp(-\rho_i \Psi_i(s))$ and probability distribution f_{Z_i}. Denote the total frailty by Y. Consider only the variation at the bottom level, and let all the other levels be given. Add a new level of frailty by randomizing ρ_1 by Z_2. The Laplace transform of Y will be

$$L_Y(s) = \mathrm{E}\left(L_{Z_1}(s)|Z_2\right) = \int \exp\left(-\rho_1 \Psi_1(s)\right) f_{Z_2}(\rho_1) d\rho_1 = \exp\left[-\rho_2 \Psi_2(\Psi_1(s))\right] \tag{3.1}$$

This is a more general version of the model in Moger and Aalen (2005). When combined with a parametric baseline hazard $\lambda(t)$ in (2.1), one may get a large improvement in fit compared to the simpler shared frailty models, since these models use separate distributions for individual and family variation. For non-parametric $\lambda(t)$'s, the model will be equivalent to a shared frailty model, since the individual heterogeneity will be subsumed in $\lambda(t)$. However, there could perhaps be situations where one would like to model the individual frailty by a specific probability distribution, even when using a non-parametric baseline hazard. The model can be useful for family data on diseases that are hypothesized to be caused by strong, unknown genetic or environmental effects, for which it is impossible to collect covariate information, but for which there exist biological theories on how the disease mechanism works. This can be hinting at using specific parametric distributions for modelling the baseline hazard and the individual and family heterogeneity in a frailty model, as discussed for testicular cancer in Moger *et al.* (2004).

The two-level Lévy model applies to data on groups where the genetic or environmental association is expected to be equal for all individuals, for instance litters, siblings or brothers. To extend the model to more general pedigrees where subgroups of individuals are more closely correlated than others, add another level to the model by randomizing ρ_2 by Z_3. This yields the Laplace transform

$$\begin{aligned} L_Y(s) &= \mathrm{EE}\left(L_{Z_1}(s)|Z_2, Z_3\right) = \int\int \exp\left[-\rho_1 \Psi_1(s)\right] f_{Z_2}(\rho_1) d\rho_1 f_{Z_3}(\rho_2) d\rho_2 \\ &= \exp\left[-\rho_3 \Psi_3(\Psi_2(\Psi_1(s)))\right] \end{aligned} \tag{3.2}$$

and so on for further levels. Hence, the structure is generated by applying function iteration to the Laplace exponent. The model described by $L_3(s)$ could be used on data with two levels of dependence, consisting e.g. of families in a neighborhood. The distribution Z_3 could then describe common environmental factors shared by all individuals in the neighborhood, while Z_2 corresponds to factors which are shared by a family, but independent for different families. The distribution Z_1 models individual environmental factors which are independent for all.

An interesting special case applies when the positive stable distributions are used, that is, when $\Psi_i(s) = s^{\alpha_i}$(the scale parameter of the distribution, usually called δ, will play the role of ρ_i). The Laplace transform of Y is then

$$L_Y(s) = \exp\left[-\rho_3 s^{\alpha_1 \alpha_2 \alpha_3}\right]$$

which again is the Laplace transform of a stable distribution. This result is presented e.g. in Hougaard (2000), pp. 354-362, in the section on the multiplicative stable model. Moreover, he suggests a trivariate model for the lifetimes (T_1,T_2,T_3) of a sibling group, where individuals 2 and 3

are monozygotic twins, and individual 1 is a singleton. Hence, sibling 2 is more strongly correlated to sibling 3 than to sibling 1. For the hierarchical Lévy models, this can be constructed as follows. All siblings share the same value of Z_2 (and hence of ρ_1). Siblings 2 and 3 will have the same value of Z_1, whereas sibling 1 will have an independent value of Z_1. Let different superscripts denote independent values of the Z_i's. The joint Laplace transform for the sibling group will then be

$$\begin{aligned} L(s_1,s_2,s_3) &= \mathrm{EE}\left(\exp\left[-Z_1^1 s_1 - Z_1^2 s_1 - Z_1^2 s_1\right]|Z_1,Z_2\right) \\ &= \mathrm{E}\left(\exp\left[-\rho_1\Psi_1(s_1) - \rho_1\Psi_1(s_2+s_3)\right]|Z_2\right) \\ &= \exp\left[-\rho_2\Psi_2\left(\Psi_1(s_1)+\Psi_1(s_2+s_3)\right)\right] \end{aligned}$$

since all siblings are independent given Z_1, Z_2. This gives the joint survival function

$$S(t_1,t_2,t_3) = \exp\left[-\rho_2\Psi_2\left(\Psi_1(\Lambda(t_1)) + \Psi_1(\Lambda(t_2)+\Lambda(t_3))\right)\right]$$

and generalizes the formula of Hougaard (2000), p. 356.

The model above has a nested dependence structure, where all individuals are correlated. One may also construct a simple genetic model, for data consisting of parents and children. Here, the parents are assumed to be independent, but the children are related both to the parents and to each other. Let the parents have independent values of both Z_1 and Z_2, whereas the children will have independent values of Z_1, but their value of ρ_1 assigned by Z_2 is determined by those of the parents. In a simple additive model, assume that it is the mean of the parents' values, $(\rho_1^1+\rho_1^2)/2$. Let s_1 and s_2 be the argument of the Laplace transform for the parents, and $s_3,...,s_k$ the arguments for the children. The joint Laplace transform of this model is given by

$$\begin{aligned} L(s_1,...,s_k) &= \mathrm{E}\left(\exp\left[-\rho_1^1\Psi_1(s_1) - \rho_1^2\Psi_1(s_2) - \textstyle\sum_{j=3}^k \frac{\rho_1^1+\rho_1^2}{2}\Psi_1(s_j)\right]|Z_2\right) \\ &= \mathrm{E}\left(\exp\left\{-\rho_1^1\left[\Psi_1(s_1) + \frac{1}{2}\textstyle\sum_{j=3}^k\Psi_1(s_j)\right] - \rho_1^2\left[\Psi_1(s_2) + \frac{1}{2}\textstyle\sum_{j=3}^k\Psi_1(s_j)\right]\right\}|Z_2\right) \\ &= \exp\left\{-\rho_2\left[\Psi_2\left(\Psi_1(s_1) + \frac{1}{2}\textstyle\sum_{j=3}^k\Psi_1(s_j)\right) + \Psi_2\left(\Psi_1(s_2) + \frac{1}{2}\textstyle\sum_{j=3}^k\Psi_1(s_j)\right)\right]\right\} \end{aligned}$$

A drawback with this model, is that the frailty distribution of the children will have a different variance than the parents' distribution.

If the baseline hazard $\lambda(t)$ in (2.1) includes a scale parameter, one often sets the expectation of the frailty distribution equal to one, to assure identifiability. Simple results are valid for the expectation and variance of the hierarchical Lévy frailty model, provided that they exist for the model in question. Assume that the expectation of the Z_i's equals one when $\rho_i = 1$, that is, we have $\Psi_i'(0) = 1$ for all i. For the variable in (3.1), we then have

$$\mathrm{E}Y = \Psi_2'(\Psi_1(0))\Psi_1'(0) = 1$$

and

$$\begin{aligned} \mathrm{Var}Y &= \Psi_2''(\Psi_1(0))(\Psi_1'(0))^2 + \Psi_2'(\Psi_1(0))\Psi_1''(0) \\ &= \Psi_2''(0) + \Psi_1''(0) = \mathrm{Var}Z_1 + \mathrm{Var}Z_2 \end{aligned}$$

By induction it follows that the random variable in (1) has expectation and variance

$$\mathrm{E}Y = 1, \quad \mathrm{Var}Y = \mathrm{Var}Z_1 + \mathrm{Var}Z_2 + \mathrm{Var}Z_3$$

and similarly for higher levels. Hence, the variance of a hierarchical Lévy frailty variable can be decomposed into a sum coming from different sources, without affecting the expectation. This is very useful in a frailty context, where the expectation often should be kept constant and just the variance be decomposed. As an example, $\mathrm{Var}Z_1$ can be interpreted as the frailty variance related to individual factors, $\mathrm{Var}Z_2$ is the frailty variance related to common genetic and environmental factors within a family, and $\mathrm{Var}Z_3$ is the frailty variance relating to common environmental factors in the neighborhood.

5 COVARIATES AND FRAILTY RELATIVE RISK

The frailty variable describes heterogeneity due to unknown covariates. One would usually like to include covariates in the model, e.g. in a Cox regression term $\exp(\boldsymbol{\beta}^T\mathbf{X})$, to be able to explain some of the unobserved heterogeneity. One usually distinguishes between common covariates, which have the same value for all individuals in a family, and individual covariates, which have different values for different individuals in a family. In a frailty analysis of family data, many covariates with an effect on survival will have values that are correlated within family members, creating some of the dependence in survival time within families. When including such covariates in the model, the dependence due to unobserved factors should go down. By using a surrogate measure, such as the mean value of a covariate within a family, one may also treat highly correlated, but still individual, covariates as common covariates.

Table 1: Overview of some of the different covariate modelling options in a two-level Lévy model.

Model	Where to include β	Interpretation of $\exp(\beta)$
1. Fully conditional	In baseline hazard $\lambda(t)$	Individual-specific relative risk
2. Conditional on family only	In ρ_1	Family-specific relative risk
3. Conditional on subgroup	In ρ_2	Larger subgroup-specific relative risk
4. Marginal	In marginal distribution	Population average relative risk
5. Accelerated failure times	In baseline hazard $\lambda(t)$	Accelerated failure times

In the hierarchical Lévy model, there are several options on where to include the covariates. We will only consider the two-level sibling/litter model (3.1) in this section, where Z_1 is independent for all, and Z_2 assigns values of ρ_1 that are shared by all individuals in a family, but independent between families. It will be similar for higher-level models, but with even more options. Table 1 shows the different options for this two-level model. As a specific example, consider the compound Poisson-gamma model in the applications in Moger *et al.* (2004) and Moger and Aalen (2005). The individual frailty Z_1 is compound Poisson distributed, with $\Psi_1(s) = \{1 - [\nu/(\nu + s)]^\eta\}$ and the family frailty Z_2 is gamma distributed, $\Psi_2(s) = [\ln(\delta + s) - \ln\delta]$. The parameters ν and η are the scale and shape parameter of the CP-distribution of Z_1, while δ is the scale parameter of the gamma distribution of Z_2. This gives the following joint survival function for k siblings:

$$S(t_1, ...t_k) = \left(\frac{\delta}{\delta + \sum_{i=1}^{k} \{1 - [\nu/(\nu + \Lambda(t_i))]^\eta\}} \right)^{\rho_2} \tag{4.1}$$

In this expression, one may include covariates in the baseline hazard, $\lambda'(t) = \exp(\boldsymbol{\beta}^T\mathbf{X})\lambda(t)$. In this case, the regression coefficients $\boldsymbol{\beta}$ are conditional on both Z_1 and Z_2, giving proportional hazards conditional on the full frailty Y. That is, the estimated regression effects are interpreted conditional on having the same value of both the family and individual frailty. This yields individual specific regression effects, which may be of interest when counselling a patient with both a family history of a disease, and exposure to unmeasured individual risk factors. The second alternative is to include covariates on the family parameter $\rho_1' = \exp(\boldsymbol{\beta}^T\mathbf{X})\rho_1$. This option yields proportional hazards conditional on Z_2, meaning that the regression coefficients are interpreted as the effects of the covariates when comparing different individuals within a family. The covariates will then appear outside or just inside the sum in (4.1), depending on whether the covariates are common or not. This approach can be preferred when the family-specific hazard is of interest, e.g. in genetic counselling. This is the same interpretation of the regression coefficients as the conditional parameterization in a shared frailty model. One should think that this option is most relevant for common covariates, since ρ_1 has the same value for all members of the family in this model. A

third alternative is to include the covariates in ρ_2. This approach could be relevant for covariates that yield different levels of family frailty, and are shared by large subgroups of people, for instance ethnicity. The covariate relative risk $\exp(\beta)$ can then be interpreted in terms of the relative change in $\mathrm{E}(Z_2)$ for one subgroup compared to another. A fourth alternative is to have proportional hazards marginally. This yields population average effects of the covariates, and is often of interest from the public health perspective. One then parameterize (4.1) by means of the marginal survival function, $S(t_i) = \exp\left[-\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\Omega(t_i)\right]$, where $\Omega(t_i)$ is the integrated hazard in the marginal distribution, after first having found the inverse relation between the marginal survival function and the conditional integrated hazard (from (3.1)):

$$\begin{aligned} S(t_i) &= \exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t_i)))\right] \\ \Lambda(t_i) &= \Psi_1^{-1}(\Psi_2^{-1}(-\ln S(t_i)/\rho_2)) \end{aligned}$$

However, this approach causes all the parameters in the individual frailty Z_1 to cancel out:

$$\begin{aligned} S(t_1,...,t_k) &= \exp\left[-\rho_2\Psi_2(\textstyle\sum_{i=1}^k \Psi_1(\Psi_1^{-1}(\Psi_2^{-1}(-\ln S(t_i)/\rho_2))))\right] \\ &= \exp\left[-\rho_2\Psi_2(\textstyle\sum_{i=1}^k \Psi_2^{-1}(-\ln S(t_i)/\rho_2)\right] \end{aligned}$$

A scale parameter in $\Psi_2(\bullet)$ will also cancel out. This means that the model becomes identical to the marginal parameterization of a shared frailty model in this case. Finally, one may also formulate the model as an accelerated failure time model. The integrated baseline hazard is then $\Lambda(t/\exp(\boldsymbol{\beta}^T\mathbf{X}_i))$, which may be inserted in (4.1). When the baseline hazard is Weibull, of the form $\lambda(t) = \alpha\kappa t^{\kappa-1}$ the model will be the same as the fully conditional model (first model in Table 1) with $\boldsymbol{\beta} = -\boldsymbol{\beta}_{acc}\kappa$, but with an accelerated failure time interpretation of the regression coefficients. As the options all give different interpretations of the regression effects, one has to consider each application individually, to find out which interpretation one would prefer.

The relative risk is often used as a measure of the strength of genetic association in a family. For the conditional parameterizations (1-3 in Table 1), the relative risk may be calculated as the risk of dying within a time t if a sibling has died compared to the risk of dying if a sibling has survived:

$$RR = \frac{P(\text{Sib 1 dies within time } t|\text{Sib 2 dead within time } t)}{P(\text{Sib 1 dies within time } t|\text{Sib 2 has survived up to time } t)}$$

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 expression for the relative risk:

$$RR = \frac{\{1 - 2\times\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)))\right] + \exp\left[-\rho_2\Psi_2(2\times\Psi_1(\Lambda(t)))\right]\}\times\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)))\right]}{\{1-\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)))\right]\}\times\{\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)))\right] - \exp\left[-\rho_2\Psi_2(2\times\Psi_1(\Lambda(t)))\right]\}} \tag{4.2}$$

This generalizes the formula (16) in Moger and Aalen (2005). This expression will be the relative risk due to unobserved factors, in this paper called the frailty relative risk. In models Cond1 and Cond2 with covariates, the frailty relative risk will to a neglible degree depend on the values of the covariates, as they do not cancel out. The frailty relative risk is then calculated as the risk when all covariate values are equal for the individual and her sibling (typically 0). For model Cond3, however, the covariates enter ρ_2, and the frailty relative risk will largely depend of the covariate values. We then use the mean value of the covariate, for comparison with models Cond1 and Cond2. Note that 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.

Another measure of dependence commonly used for frailty models is Kendall's τ. Earlier attempts on calculating Kendall's τ for the compound Poisson-PVF model have failed, because of the non-continuity of the compound Poisson distributions. However, the dependence in the two-level model is constructed in much the same way as for shared frailty models. The variable

Z_1 only creates individual heterogeneity, and has little to do with the dependence, whereas the dependence is modelled by Z_2. Hence, in situations where the distribution of Z_1 is non-continuous, but distribution of Z_2 is, one should think that Kendall's τ for the two-level Lévy model can be approximated by the corresponding formulas from standard shared frailty models.

By similar calculations as for (4.2), one may derive the conditional survival given that a sibling has died within time $t = t_0$, and the conditional survival given that a sibling has survived at $t = t_0$. For the two-level Lévy model, they are given as

$$\begin{aligned} & S_{\text{Sib 1}}(t|\text{Sib 2 dead within } t_0) \\ = & \frac{\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)))\right] - \exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)+\Psi_1(\Lambda(t_0))))\right]}{1-\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)))\right]} \\ & S_{\text{Sib 1}}(t|\text{Sib 2 alive at } t_0) \\ = & \frac{\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t))+\Psi_1(\Lambda(t_0)))\right]}{\exp\left[-\rho_2\Psi_2(\Psi_1(\Lambda(t)))\right]} \end{aligned} \tag{4.3}$$

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

6 APPLICATION TO THE INFANT MORTALITY DATA

This application is meant to be an illustration of different aspects when working with hierarchical Lévy frailty models, more than finding the best model for the infant mortality data. We will show results from several of the different modelling options presented in the previous section. Because of the extremely large number of data, it would be computationally impossible to analyze the full database. By using the methods shown in Moger *et al.* (in revision), 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 family size before sampling, and exactly 5% are randomly sampled without replacement from each stratum. There are four strata for the control families, for sibships of size 1, 2, 3 and >4. This yields 45750 sibships with 89745 individuals in the control 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 are 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 Moger *et al.* (in revision), 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 pseudo-likelihood. We use the compound Poisson (CP)-gamma model in (4.1) to analyze the data. Let there be k_l members in family l. Let c_{il} indicate whether the survival time t_{il} for individual i in family l is censored ($c_{il} = 0$) or not ($c_{il} = 1$). Define $c_{\cdot l} = \sum_i c_{il}$ as the number of events in family l. This yields the following pseudo-likelihood in the conditional parameterization:

$$L(\boldsymbol{\theta}) = \prod_{j=0}^{4} \frac{1}{p_j} \prod_{l\in D_j} \left[\prod_{i=1}^{k_l} \left(\frac{\eta\nu^\eta \lambda(t_{il})}{(\nu+\Lambda(t_{il}))^{\eta+1}} \right)^{c_{il}} \right] (-1)^{c_{\cdot l}} L_{\rho_1}^{(c_{\cdot l})} \left(\sum_{i=1}^{k_i} \{1-[\nu/(\nu+\Lambda(t_{il}))]^\eta\} \right)$$

where $\boldsymbol{\theta}$ is the vector of parameters to be estimated, and $L_{\rho_1}^{(c_{\cdot l})}(\cdot)$ is the $c_{\cdot l}$-th derivative of the Laplace transform of ρ_1, $L_{\rho_1}(s) = [\delta/(\delta+s)]^{\rho_2}$. The pseudo-likelihood is identical to the cohort likelihood on p. 54 in Moger and Aalen (2005), except for the sampling weights p_j for the case families ($j = 0$) and the four strata of control families, and that the sum is over the case-cohort sample D_j instead of over the full cohort. The baseline hazard is assumed to follow a Weibull distribution, $\lambda(t) = \kappa(t-6)^{\kappa-1}$, for $t > 6$. The scale parameter of the Weibull distribution is subsumed in

the frailty distribution. For the fully conditional model (Cond1), the covariates enter the pseudo-likelihood in the three places where the baseline hazard $\lambda(t)$ appears. For model Cond2, the covariates enter in the numerator $\eta\nu^{\eta}\lambda(t_{il})\exp(\boldsymbol{\beta}^T\mathbf{X}_{il})$ and in $\sum_{i=1}^{k_l}\exp(\boldsymbol{\beta}^T\mathbf{X}_{il})\left\{1-[\nu/(\nu+\Lambda(t_{il}))]^{\eta}\right\}$, while for model Cond3, the covariates enter in ρ_2 only. For the marginal model Marg, the likelihood will be equal to a shared gamma frailty model in the marginal parameterization, see e.g. equation (7.35), p.234 in Hougaard (2000). As both the parameters of the individual distribution and the scale parameter δ in the gamma distribution are cancelled out, the model Marg includes an additional scale parameter in the marginal Weibull baseline hazard $\Omega(t)=\alpha(t-6)^{\kappa}$. To estimate the standard errors of the parameters, we use a sandwich-type estimator (see Moger *et al.*, in revision, for details) $\mathbf{A}(\boldsymbol{\theta})^{-1}+\mathbf{A}(\boldsymbol{\theta})^{-1}\mathbf{B}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^{-1}$. Here, $\mathbf{A}(\boldsymbol{\theta})$ is estimated by

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

where $\mathbf{I}_l(\boldsymbol{\theta})=-\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\log L_l(\boldsymbol{\theta})$, the observed information matrix for family l, and $\mathbf{B}(\boldsymbol{\theta})$ is estimated by

$$\widehat{\mathbf{B}}(\widehat{\boldsymbol{\theta}})=\sum_{j=0}^{4}\frac{1-p_j}{p_j^2}\sum_{l\in D_j}\mathbf{s}_l(\widehat{\boldsymbol{\theta}})\mathbf{s}_l(\widehat{\boldsymbol{\theta}})'$$

where $\mathbf{s}_l(\boldsymbol{\theta})=\partial/\partial\boldsymbol{\theta}\log L_l(\boldsymbol{\theta})$, the score function for family i.

To find out how to model the continuous covariates, we first categorized them (e.g. into 500 grams intervals for birth weight, 50 days intervals for gestational age etc., this was done for the full database), and studied how the deaths were distributed in the categories in a cross-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 cross-tabulation. 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, length and the birth year of mother and infant are treated as continuous covariates. Mother's age at birth is categorized into 0-22 years, 23-36 years and 37 years and above. Parity is categorized into 1, 2-3 and 4 and above (this covariate will almost be stratified because of the sampling stratified on family size). The regression coefficients in the tables show the effect relative to the first category for these covariates. The only common covariate is mother's birth year, but infant's birth year and the categorized mother's age are almost common covariates, with correlation of about 0.9 and 0.7, respectively. Parity and gender are clearly individual covariates, while the others have correlated values within sibships. The correlation is around for 0.4 for birth weight, 0.3 for length, and 0.2 for gestational age.

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 three parameters. Although the likelihood ratio test does not apply for pseudo-likelihoods, the better fit of the CP-gamma model is also indicated by the log pseudo-likelihood values.

Tables 2 and 3 show the results of the univariate analyses of the covariates, for the most interesting parameters. We have excluded the different frailty scale parameters to save space. In addition, we show regression effects as relative risks $\exp(\beta)$ with 95% confidence interval $\left(95\%\text{CI}=\exp\left[\widehat{\beta}\pm 1.96\times SE(\widehat{\beta})\right]\right)$. To clarify the relations between the different options on where to place the covariates, we show results for the models 1 (Cond1), 2 (Cond2), and 4 (Marg) from Table 1. For the common covariate mother's birth year, we also show results for model 3 (Cond 3). The frailty relative risk (FRR) measures the dependence that remains in the data after controlling for the covariates, and is calculated by inserting the estimated parameters into the CP-gamma version of (4.2) (see Moger and Aalen, 2005, for further details), with $t=364$. As mentioned in Section 4.2, the marginal model does not give a better fit than a shared gamma model with the

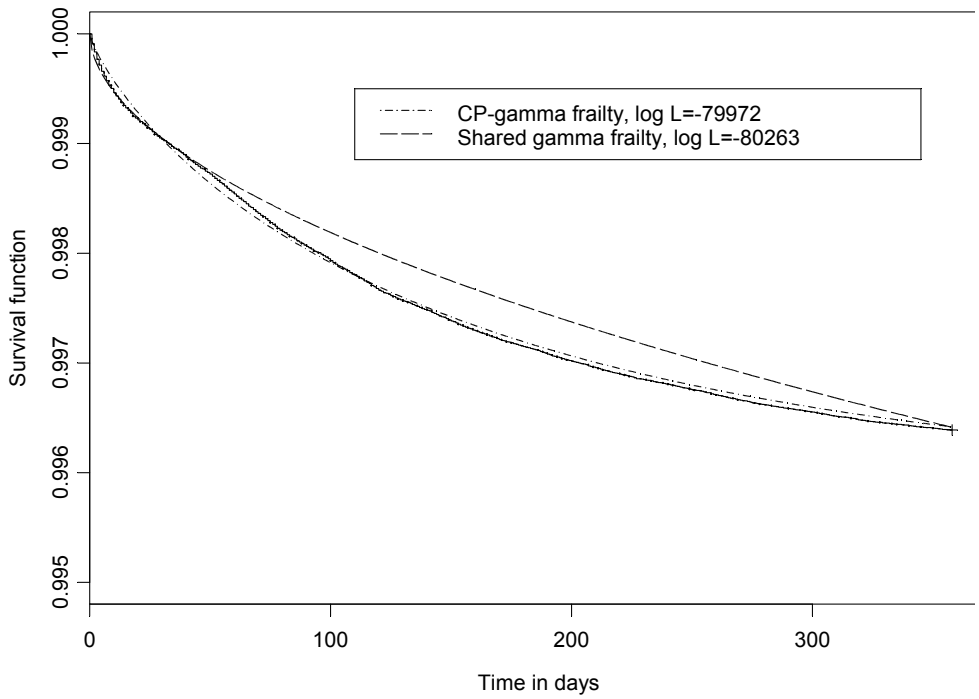

Figure 1: Kaplan-Meier plot of the infant mortality data, with the estimated compound Poisson-gamma model. The shared gamma model is included for reference. Log L=log pseudo-likelihood value.

marginal parameterization (as in Figure 1, without covariates). Hence, it is expected to perform worse in terms of log pseudo-likelihood values than the conditional models.

The CP-gamma model without covariates yields a frailty relative risk of 3.44. Including birth weight in the model reduces the frailty relative risk to 3.13 in model Cond2, but increases it to 3.77 in model Cond1. For model Marg, if we measure the dependence by Kendall's $\tau = 1/(1+2\rho_2)$, the dependence also decreases from 0.56 to 0.50 (ρ_2 was 0.39 for model Marg without covariates). The model Cond2 tend to yield an increased frailty relative risk for all continuous covariates with low correlation within sibships. For the individual covariate parity, however, Cond2 yields a lowered frailty relative risk. This may indicate that it is problematic to put individual covariates in a variable describing the family, or shared, frailty. Putting the individual covariates in a shared gamma frailty model using the conditional parameterization (not shown), yields similar results as for Cond2. Constructing common covariates by using the mean value of the covariates within sibships, however, will in most cases give a reduced dependence, corresponding to the results for Cond1 and Marg. The parameter ρ_2 is similar for models Cond1 and Marg for all covariates. In the two-level model, ρ_2 contributes the most to the estimated dependence (both the frailty relative risk and Kendall's τ), thus indicating that the dependence is similar for the models Marg and Cond2.

From model Cond1, a 500 grams increase in birth weight yields a 48% reduced risk of death during the post-perinatal period. Models Cond2 and Marg, both yield 37% reduced risk. There is a similar picture for the other continuous covariates, with Cond1 always giving the strongest effects of the covariates. For the categorical covariates, however, model Cond1 does not always give

Table 2: Log pseudo-likelihood values, parameter estimates/standard errors, and univariate covariate effects for different model options: Cond1=conditional on full frailty, Cond2=conditional on Z_2, Cond3=in ρ_2, Marg=marginal effects, FRR=Frailty relative risk. See text for details.

	log L	κ (SE)	ρ_2 (SE)	β (SE)	exp(β) (95% CI)	FRR
No Cov.	-79972	0.81 (0.01)	0.41 (0.06)	–	–	3.44
Birth weight (per 500 grams)						
Cond1	-78186	0.61 (0.01)	0.47 (0.07)	-0.65 (0.012)	0.52 (0.51-0.53)	3.13
Cond2	-78199	0.81 (0.01)	0.36 (0.05)	-0.47 (0.009)	0.63 (0.61-0.64)	3.77
Marg	-78509	0.54 (0.01)	0.49 (0.02)	-0.45 (0.008)	0.63 (0.62-0.64)	–
Gestational age (per 30 days)						
Cond1	-79067	0.73 (0.01)	0.46 (0.06)	-1.18 (0.019)	0.31 (0.30-0.32)	3.18
Cond2	-79143	0.81 (0.01)	0.35 (0.05)	-0.59 (0.020)	0.55 (0.54-0.57)	3.97
Marg	-79453	0.54 (0.01)	0.46 (0.02)	-0.56 (0.016)	0.57 (0.55-0.59)	–
Infant's birth year (per 5 years)						
Cond1	-79886	0.82 (0.02)	0.41 (0.06)	-0.20 (0.011)	0.82 (0.80-0.84)	3.43
Cond2	-79837	0.81 (0.01)	0.41 (0.06)	-0.11 (0.007)	0.90 (0.88-0.91)	3.45
Marg	-80131	0.54 (0.01)	0.41 (0.02)	-0.11 (0.007)	0.90 (0.89-0.91)	–
Length (per cm)						
Cond1	-79040	0.66 (0.01)	0.46 (0.07)	-0.25 (0.004)	0.79 (0.78-0.79)	3.13
Cond2	-79282	0.81 (0.01)	0.36 (0.05)	-0.10 (0.003)	0.90 (0.90-0.91)	3.84
Marg	-79589	0.54 (0.01)	0.46 (0.02)	-0.10 (0.003)	0.91 (0.90-0.91)	–
Mother's birth year (per 5 years)						
Cond1	-79956	0.80 (0.02)	0.41 (0.06)	-0.07 (0.010)	0.94 (0.92-0.95)	3.43
Cond2	-79930	0.81 (0.01)	0.41 (0.06)	-0.06 (0.007)	0.95 (0.93-0.96)	3.43
Cond3	-79930	0.81 (0.02)	0.76 (0.06)	-0.06 (0.007)	0.95 (0.93-0.96)	3.43
Marg	-80224	0.54 (0.01)	0.41 (0.02)	-0.06 (0.007)	0.95 (0.93-0.96)	–

the strongest effects. The covariates mother's birth year, mother's age at birth and infant's birth year all have a significant effect on the mortality, but do not have a great influence on the frailty relative risk. The estimate of the regression coefficient in model Cond3 for mother's birth year gives the following interpretation: A five year increase in mother's birth year reduces the expected level of family frailty by 5%, since $\mathrm{E}(Z_2)=\rho_2 \exp(\boldsymbol{\beta}^T\mathbf{X})/\delta$, with $\delta = 96.62$ for this model. With the mean value of the covariate inserted, the FRR is comparable to the other models. However, one may insert a birth year of 1945, to get a FRR of 3.18, or a birth year of 1970, to get a FRR of 3.88. The increase in dependence as a function of mother's birth cohort, is probably due to the lower prevalence of random infant deaths in more recent birth cohorts. A few familial cases will then have a greater impact on the dependence. The Weibull shape parameter κ has stable values within each model for all covariates, yielding a decreasing baseline hazard for all models.

Table 4 shows the results of a multivariate analysis. Overall, the estimated regression effects show similar trends for the three models as in the univariate analyses. Generally, one may combine models Cond1 and Cond2, to include some covariates conditional on the full frailty, and others conditional on the family frailty only. Although the log pseudo-likelihood values indicate that model Cond1 fit the data best, it is probably more correct to view the different parameterizations as equal alternatives, where selection of a specific model should depend on what interpretations are the most interesting for the problem at hand. Since there is collinearity between the covariates infant's birth year, mother's age at birth and mother's birth year (if one knows the value of two of these covariates, one also know the value of the third), we have only included the first two in the multivariate model. With all covariates included, the frailty relative risk has dropped to 2.75 for model Cond1, but is fairly stable at 3.24 for Cond2. Again, the value of ρ_2 is similar for model

Table 3: Log pseudo-likelihood values, parameter estimates/standard errors, and univariate covariate effects for different model options: Cond1=conditional on full frailty, Cond2=conditional on Z_2, Marg=marginal effects, FRR=Frailty relative risk. For mother's age and parity, β's are shown for two categories compared to the reference group. See text for details.

	log L	κ (SE)	ρ_2 (SE)	β (SE)	exp(β) (95% CI)	FRR
No Cov.	-79972	0.81 (0.01)	0.41 (0.06)	–	–	3.44
Mother's age at birth (0-22 (reference category), 23-36, >36)						
Cond1	-79941	0.86 (0.03)	0.41 (0.06)	-0.73, -0.22 (0.001), (0.002)	0.49, 0.80 (0.48-0.50), (0.79-0.81)	3.42
Cond2	-79957	0.81 (0.01)	0.41 (0.06)	-0.45, -0.20 (0.04), (0.08)	0.63, 0.82 (0.59-0.68), (0.71-0.95)	3.40
Marg	-80251	0.54 (0.01)	0.41 (0.02)	-0.45, -0.20 (0.03), (0.07)	0.64, 0.82 (0.60-0.68), (0.72-0.94)	–
Parity (1 (reference category), 2-3, >3)						
Cond1	-79972	0.81 (0.01)	0.41 (0.06)	0.01, 0.05 (0.008), (0.012)	1.01, 1.05 (0.99-1.02), (1.03-1.08)	3.41
Cond2	-79957	0.81 (0.01)	0.44 (0.06)	0.11, 0.25 (0.03), (0.06)	1.12, 1.28 (1.07-1.18), (1.13-1.44)	3.26
Marg	-80250	0.54 (0.01)	0.44 (0.03)	0.11, 0.25 (0.03), (0.06)	1.12, 1.28 (1.07-1.18), (1.13-1.45)	–
Gender (reference category male)						
Cond1	-79940	0.79 (0.01)	0.41 (0.06)	-0.35(0.013)	0.70 (0.69-0.72)	3.44
Cond2	-79927	0.81 (0.01)	0.41 (0.06)	-0.24(0.026)	0.79 (0.75-0.83)	3.43
Marg	-80220	0.54 (0.01)	0.46 (0.08)	-0.24(0.008)	0.79 (0.78-0.80)	–

Cond1 and Marg, indicating that the dependence for these two models is similar. The effects of length and gestational age have become much reduced, mainly because of the confounding effect of birth weight. Being a tall infant is not an advantage unless the weight is also higher. Gestational age is only borderline significant at 5% level for the models Marg and Cond2. The effect of parity has increased greatly compared to the univariate analysis, due to confounding with most of the other covariates. For gender, infant's birth year and birth weight, there are smaller differences from the univariate analysis. Figures 2 and 3 show conditional survival functions (1) for model Cond1 with $t_0 = 364$ and different covariate values, compared to the Kaplan-Meier plot. In Figure 2, the two siblings have values close to the means in the population for most covariates, that is; birth weight=3500 grams, gestational age=280 days, length=50 cm, they are girls born in 1981 with parity 1 and 3, and have a mother in the oldest age group. The figure clearly shows the effect a dead sibling has on survival. In Figure 3, however, the second sibling has a birth weight of 4000 grams, and both are born in 1995. The higher birth weight of the second sibling means that the effect on survival of having a dead sibling is much smaller in this case.

7 DISCUSSION

This paper presents an extension to existing frailty models for family data. The model is hierarchical and is based on the flexible family of Lévy distributions, and includes a large number of possible sub-models. In its simplest form, it is a two-level model, which includes heterogeneity on both the individual and family level. It is fairly easily extended to more levels, which makes it possible to analyze more complicated pedigrees and dependence structures. An extension to a

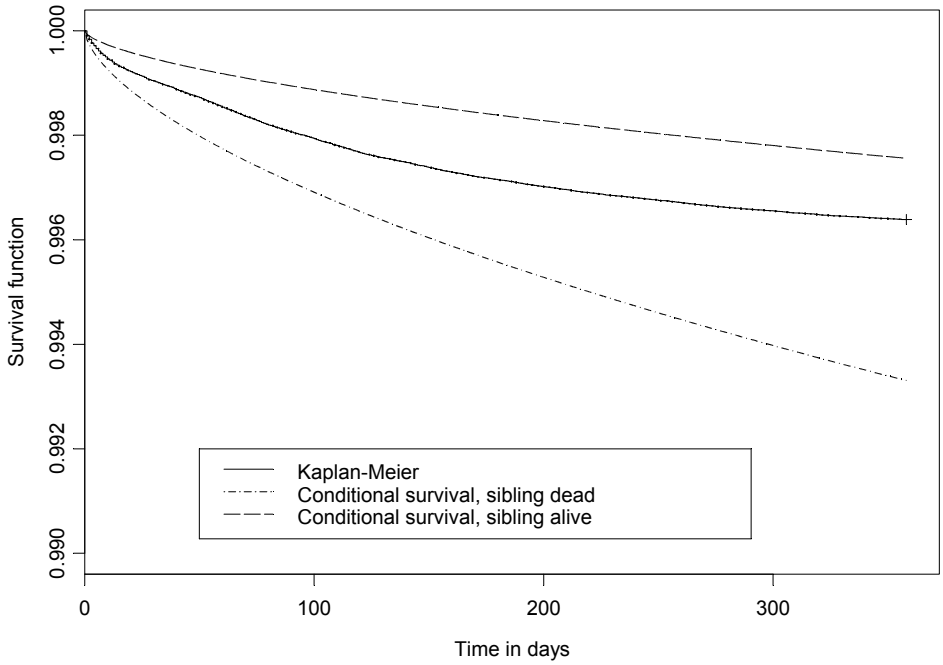

Figure 2: The effect a dead sibling has on survival, as estimated from the fully conditional model (Cond1). Both siblings have covariate values that are close to the mean values in the population and are born in 1981. See the text for details.

nonparametric baseline hazard will make it necessary to use other estimation methods than the ones used for the fully parametric models presented here. This is a challenge, since the likelihood function for the model becomes quite complex, particularly for data on families containing several levels of dependence, and many events in each family. However, complex likelihoods are also the case for other multivariate frailty models, such as additive models (e.g. Petersen, 1998) and the multivariate log-normal model (Ripatti and Palmgren, 2000). In addition, additive models rely on distributions, for which an additive property exists, to make the total frailty tractable. In other words, each Z_i has to have the same type of distribution in an additive model. In the hierarchical model, each Z_i can have a different distribution. This is probably the greatest advantage compared to additive models. Hence, the approach presented here yield more general models. Application of higher-level Lévy models will be the focus of future research. We plan to analyze data on melanoma incidence in two-generation families from the Swedish Multi-Generation Register.

It might seem strange to include results from the marginal parameterization in Tables 3 and 4, since it has a poor fit to data. However, we wanted to include this marginal model to compare the regression coefficients between the different parameterizations, since it naturally appears when you use the marginal parameterization in a two-level Lévy model. A standard semi-parametric Cox regression using the independence working model approach yields approximately the same estimates of both the β's and their standard errors. This indicate that even though the marginal gamma model with Weibull baseline hazard does not fit the data well, the estimated β's are little affected by this. As expected, the fully conditional frailty model gives the strongest effects of most covariates, as these are conditioned on comparing two individuals with the same value of both the

Table 4: Parameter estimates with standard errors for the three different models, with multivariate estimates of the covariate effects. Est.=Estimate, Cond1=conditional on full frailty, Cond2=conditional on Z_2, Marg=marginal effects. See the text for further details.

Model		Cond1		Cond2		Marg	
Parameter		Est.	SE	Est.	SE	Est.	SE
κ		0.63	0.007	0.81	0.014	0.54	0.007
ρ_2		0.57	0.063	0.44	0.065	0.60	0.004
Covariate		$\exp(\beta)$	95% CI	$\exp(\beta)$	95% CI	$\exp(\beta)$	95% CI
Birth weight		0.56	(0.54-0.58)	0.53	(0.51-0.55)	0.54	(0.52-0.56)
Gestational age		0.81	(0.77-0.85)	0.95	(0.91-0.99)	0.96	(0.92-1.01)
Infant's birth year		0.85	(0.84-0.87)	0.88	(0.87-0.89)	0.88	(0.87-0.90)
Length		0.98	(0.97-0.99)	1.07	(1.06-1.08)	1.07	(1.06-1.08)
Mother's age	23-36	0.61	(0.57-0.66)	0.65	(0.60-0.70)	0.65	(0.61-0.70)
	>36	0.72	(0.63-0.84)	0.74	(0.64-0.86)	0.75	(0.65-0.86)
Parity	2-3	1.68	(1.59-1.78)	1.57	(1.49-1.67)	1.56	(1.48-1.66)
	>3	2.16	(1.88-2.49)	1.93	(1.69-2.22)	1.93	(1.68-2.22)
Gender		0.66	(0.63-0.70)	0.74	(0.70-0.78)	0.74	(0.70-0.78)
log L		-77686		-77643		-77951	

family and individual frailty, while the marginal model give the smallest effects of the covariates, as these are population average effects. The model conditioned on the family frailty only, gives effects that are intermediate, but they seem to be closer to the marginal estimates than to the fully conditional estimates.

The estimate for the frailty relative risk, obtained from the analysis without covariates, is somewhat lower than the estimate in Øyen *et al.* (1996). By adding the SIDS deaths and non-SIDS deaths from 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 two or more infants were included, and they only used the first two births in their study. For the cohorts used in our analysis, the sibships consist of anything from 1 to 15 siblings. The sibships of size one also contribute in estimating the frailty parameters, and thus the dependence, since they affect the prevalence of the outcome. Also, the continuation rate among mothers with a first loss is somewhat higher than among those with a first survivor (83% vs 69%, from Øyen *et al.*, 1996), indicating that the one-child survivors come from low-risk families. The ability to include complete sibships of arbitrary size in the analysis, is an advantage of the frailty approach. The frailty relative risks shown in the tables of Section 6 are interpreted for pairs of observations, corresponding to the relative risks obtained from the simple cross-tabulation analysis done in Øyen *et al.* (1996). In a similar manner one may calculate relative risks for triples of observations, for instance the probability of dying given that one out of two siblings has died, compared to the probability of dying given that both have survived. Hence, it is possible to get a more general picture of the relative risks from a frailty analysis, but this is of course dependent on the validity of the model. The fact that frailty models include the aspect of time, means that one may calculate the risk of dying during the first year given that a sibling has died during the first week. This is perhaps not very relevant here, but in other settings, with a longer time-span, it could be. This can also be applied to the conditional survival function (1). Alternatively, one may also plot the relative hazards, defined as the estimated hazard function for an individual at age t, given that a sibling died at t_0 compared to the hazard of a sibling's being alive at t_0, similar to Hougaard (2000) pp. 293-95.

The use of the CP-distribution to describe the individual frailty means that a certain proportion

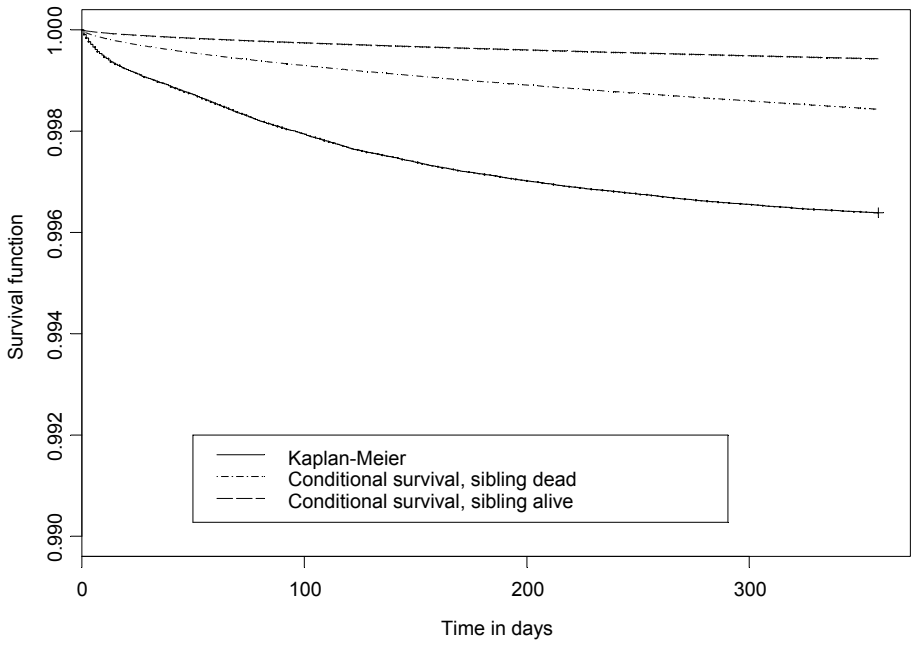

Figure 3: The effect a dead sibling has on survival, as estimated from the fully conditional model (Cond1). Sibling 1 has 500g higher birth weight than Sibling 2, and the birth year is 1995. See the text for details.

of the population (around 99.6% from the estimates) are immune to infant death. This appears unreasonable, and is not meant to be taken literally. However, since infant mortality is very rare, it is no surprise that a CP based model gives a good fit to the data. Fitting a more general PVF-PVF model yields convergence of the individual frailty within the CP-distribution, and, visually, not a better fit than the CP-gamma model used here.

We have not given standard errors for the frailty relative risks in Section 6. If we analyzed cohort data, standard errors and confidence intervals could be found by bootstrap. However, these are case-cohort data, and there does not exist bootstrap methods for multivariate case-cohort survival data yet that we are sure will work well. Although the standard errors for some of the parameters in the model are large, unpublished results in the Ph.D.-thesis by Moger lead us to expect that the confidence intervals for the frailty relative risk will be fairly narrow (perhaps estimated risk±ca.1.5).

We use a Weibull distribution for the basic hazard $\lambda(t)$. There is no biological basis for this choice. For cancer, the classical multistage model of Armitage and Doll (1954) leads to the assumption that $\lambda(t)$ follows a Weibull distribution. One then assumes that a cell has to go through several mutations in order to become malignant. One could imagine a similar reasoning for several of the causes for infant death, but this would be highly speculative. Since we only have general mortality data, the decomposition of the frailty variance, as described in Section 4, is not so relevant here, as it would be difficult to interpret the variance of the frailty components.

ACKNOWLEDGEMENTS

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.This project was sponsored by the Research Council of Norway, grant number 160627/V50.

References

AALEN, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability*, **2**, 951-972.

AALEN, O. O. AND HJORT, N. L. (2002). Frailty models that yield proportional hazards. *Statistics and Probability Letters*, **58**, 335-342.

ARMITAGE, P. AND DOLL, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer,* **8**, 1-12.

BEAL, S. M. (1992). Siblings of infant death syndrome victims. *Clinical Perinatology*, **19**, 839-848.

BECKWITH, J. B. (1990). Sibling recurrence risk of infant death syndrome. *Journal of Pediatrics* (Letter), **118**, 513-514.

CLAYTON, D. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141-151.

FLEMING, P. J., GILBERT, R. E., AZAZ, Y., BERRY, P. J., RUDD, P. T., STEWART, A. AND HALL, E. (1990). The interaction between bedding and sleeping position in sudden infant death syndrome: A population based case-control study. *British Medical Journal*, **301**, 85-89.

GJESSING, H. K., AALEN, O. O. AND HJORT, N. L. (2003). Frailty models based on Lévy processes. *Advances in Applied Probability*, **35**, 532-550.

GUNTHEROTH, W. G., LOHMANN, R. AND SPIERS, P. S. (1990). Risk of sudden infant death syndrome in subsequent siblings. *Journal of Pediatrics*, **116**, 520-524.

HOUGAARD, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387-396.

HOUGAARD, P. (2000). *Analysis of Multivariate Survival Data.* New York: Springer Verlag.

KÅREGÅRD, M. AND GENNSER, G. (1986). Incidence and recurrence rate of abruptio placenta in Sweden. *Obstetrics and Gynecology*, **67**, 523-528.

LEACH, C. E. A., BLAIR, P. S., FLEMING, P. J., SMITH, I. J., PLATT, M. W., BERRY, P. J., GOLDING, J., the CESDI SUDI Research Group (1999). Epidemiology of SIDS and explained sudden infant deaths. *Pediatrics*, **104**, 4/e43.

LIE, R. T., WILCOX, A. J. AND SKJÆRVEN, R. (1994). Recurrence of birth defects. A population based study. *New England Journal of Medicine*, **331**, 1-4.

MOGER, T. A., AALEN, O. O., HEIMDAL, K. AND GJESSING, H. K. (2004). Analysis of testicular cancer data by means of a frailty model with familial dependence, *Statistics in Medicine*, **23**, 617-632.

MOGER, T. A. AND AALEN, O. O. (2005). A distribution for multivariate frailty based on the compound Poisson distribution with random scale. *Lifetime Data Analysis*, **11**, 41-59.

MOGER, T. A., BORGAN Ø. AND PAWITAN, Y. (in revision). Case-cohort methods for survival data on families from routine registers. *Statistics in Medicine.*

PETERSEN, J. H. (1998). An additive frailty model for correlated life times. *Biometrics*, **54**, 646-661.

PILLAR, G. AND LAVIE, P. (1995). Assessment of the role of inheritance in sleep apnea syndrome. *American Journal of Respiratory Critical Care Medicine*, **151**, 688-691.

RIPATTI, S. AND PALMGREN, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 1016-1022.

ØYEN, N., SKJÆRVEN, R. AND IRGENS, L. M. (1996). Population-based recurrence risk of

sudden infant death syndrome compared with other infant and fetal deaths. *American Journal of Epidemiology*, **144**, 300-306.