Lifetime Data Anal
DOI 10.1007/s10985-010-9188-3

A hierarchical frailty model applied to two-generation melanoma data

Tron Anders Moger · Marion Haugen · Benjamin H. K. Yip · Håkon K. Gjessing · Ørnulf Borgan

Received: 11 September 2009 / Accepted: 15 October 2010

Abstract We present a hierarchical frailty model based on distributions derived from non-negative Lévy processes. The model may be applied to data with several levels of dependence, such as family data or other general clusters, and is an alternative to additive frailty models. We present several parametric examples of the model, and properties such as expected values, variance and covariance. The model is applied to a case-cohort sample of age at onset for melanoma from the Swedish Multi-Generation

T. A. Moger (✉) · M. Haugen
Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
e-mail: tronmo@medisin.uio.no

M. Haugen
e-mail: marion.haugen@medisin.uio.no

T. A. Moger
Institute of Health Management and Health Economics, University of Oslo, Oslo, Norway

B. H. K. Yip
Genome Research Centre, Department of Psychiatry, University of Hong Kong, Pok Fu Lang Road, Hong Kong, Hong Kong
e-mail: yipben@hkucc.hku.hk

H. K. Gjessing
Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway

H. K. Gjessing
Department of Public Health and Primary Health Care, University of Bergen, Bergen, Norway
e-mail: hakon.gjessing@fhi.no

Ø. Borgan
Department of Mathematics, University of Oslo, Oslo, Norway
e-mail: borgan@math.uio.no

Published online: 04 November 2010

Register, organized in nuclear families of parents and one or two children. We compare the genetic component of the total frailty variance to the common environmental term, and estimate the effect of birth cohort and gender.

Keywords Family data · Frailty · Survival analysis · Multivariate · Lévy

1 Introduction

Frailty models for handling multivariate family data try to mimic the correlation structure between family members as seen in real life, e.g. due to genetics. They can be additive, for instance relating to the additive property of independent gamma distributions. The combined frailty can then be constructed as a sum of independent gamma variables, yielding identically distributed gamma frailties for all individuals (e.g. Petersen 1998; Korsgaard and Andersen 1998). Other models use the log-normal distribution, for which any correlation structure may be specified in the covariance matrix (e.g. Ripatti and Palmgren 2000; Yau 2001). However, due to the complexity of the models and the estimation procedures and limited availability of large, good quality family data sets, applications of multivariate frailty models are not very common.

We present a new frailty model based on previous work in Moger and Aalen (2005), where a model was constructed by randomizing a scale parameter in the compound Poisson distribution. The compound Poisson distribution modelled individual heterogeneity, and the random scale parameter modelled effects common to all members in a family. We extend the model by further randomizing a scale parameter on the second level, to get a three-level model, or on the third level, to get a four-level model and so on, giving a hierarchical dependence structure. The model is briefly discussed in Aalen et al. (2008). The heterogeneity on each level is modelled by distributions derived from non-negative Lévy processes. These have simple notation and include all common frailty distributions. The frailty on a specific level can be modelled by a single distribution, or as a sum of several distributions. For the latter approach, we use the same technique as for additive frailty models (Korsgaard and Andersen 1998). The model has several nice properties, for instance the expected value of the combined frailty is constant, while the variance is split into a sum of the variances on each level.

The model is applied to melanoma data from the Swedish Multi-Generation Register. Melanoma has a relatively early age of onset (from around 25 years and up) compared to most cancers. Melanoma incidence has increased fivefold in Sweden during the past 40 years. Both diagnosis period and birth cohort effects can explain the increased incidence (Thörn et al. 1998). An important reason for the increase is the change in sun exposure. However, covariates in the register data are limited to birth cohort and gender. As unobserved heterogeneity (i.e. heterogeneity not explained by observed covariates) is modelled by the frailty variables, it is interesting to split it into different components and study the importance of each. For melanoma, the model might have three components for each individual. The first component describes heterogeneity due to individual environmental factors, assumed to be independent for all. The second component models the genetic effects, and can be independent between father and mother, whereas parent–child and child–child pairs share half of their genes,

and hence have correlation 0.5. To make the genetic inheritance as realistic as possible, the children do not inherit the same half of their genes from their parents. On the third level, one may have a component for common environmental effects, shared by all individuals in a family. Sun exposure may contribute on this level, as it could to some extent be shared within families.

We analyze a case-cohort sample of the data, and compare the fit of the new model to a similar additive model. The melanoma data are described in Sect. 2. In Sect. 3, the hierarchical Lévy frailty model is described for general survival data. It includes examples on models for different types of data, likelihood construction, properties of the expected value, variance and correlation structure within families, and methods for quantifying the dependence in times to events. We then turn back to the melanoma example, and the results of the analysis are presented in Sect. 4. A discussion is given in Sect. 5. We only consider parametric models in this paper, the extension to semi-parametric models is a topic for further research.

2 The melanoma data

The Swedish Multi-Generation Register contains information on first-degree relatives of all residents born in Sweden from 1932. The database encompassed around 11 million individuals in 2000. It includes unique national identification numbers for every individual, which can be used to link data from several registers. The Migration Register includes information on immigration and emigration years both of people born in Sweden and immigrants. Melanoma cancer cases are recorded by the Swedish Cancer Register since 1958, using the ICD-7 cancer codes and the histopathological type (PAD). Death years are found in the National Death Register. Registration of cancers and deaths improved greatly from 1961 and on, so we analyze data from January 1st 1961 to 31th December 2001. In addition, we have information on first immigration year and last emigration year. The outcome variable is age in years at first melanoma diagnosis recorded in the register after January 1961, and the (possibly) censored event time is defined as the minimum of emigration year, diagnosis year, death year and 2001, subtracted the year of birth. Since we do not have information on cancer before 1961, the data are left-truncated, and an adjustment has to be made in the likelihood. The age at truncation is defined as the maximum of immigration year and 1961, subtracted the year of birth for individuals born before 1961.

Due to the vast size of the Swedish Multi-Generation Register, we analyze a case-cohort sample of the data. This is expected to give efficient parameter estimates (Moger et al. 2008). All multiple births in the register are excluded as they are expected to share more of the genetics and environment than ordinary siblings, and in case of a mother having children with several partners, we restrict to the first partner in order to avoid half siblings. Define a nuclear family as two parents and their one (in case of only one child in the family) or two oldest children. In addition, define a case family as a family with at least one melanoma case, and a control family as a family with no melanoma cases. All case families in the register are included in the sample. In addition, we sample three control families, without replacement, for each case family. The sampling is matched on the birth cohort of the oldest child (one-year strata from

Table 1 Distribution of number of families according to which family members are affected and number of children in the family

	None	Mother	Father	Both parents
No. affected, families with one child				
None	24,052	3,021	2,814	17
1st child	2,045	26	23	2
No. affected, families with two children				
None	70,690	8,104	8,202	56
1st child	4,002	64	68	1
2nd child	2,893	39	36	0
Both children	37	3	1	0

1932 to 2001) and on the size of the case family (one or two children). In order to avoid some suspiciously high ages, the 1,382 individuals in the sample born before 1900 and with no record of death or emigration until 2001 are omitted from the analysis. Individuals are also censored at 90 years of age, as very old individuals are likely to be a highly selected group. Hence, 95 individuals (and families) shift status from cases to controls. For the 18,459 individuals where age at event/censoring and age at truncation are equal, we adjust the age at event/censoring by adding one half to ensure that they give a contribution to the likelihood. In total, there are 126,196 families (471,402 individuals), 31,454 case families, and 94742 control families. Also, 32,000 families have one child, whereas 94,196 have two children. Table 1 shows the distribution of the families according to which family members are affected in the sample. The covariates included in the analysis are birth cohort and gender. The matching on birth cohort in the sampling should improve the precision of the regression coefficient for birth year in the analysis. Further details on the likelihood and estimation are given in Sect. 4.

3 Hierarchical Lévy frailty models

In this section we describe our hierarchical frailty model, give two examples of the general model formulation, and discuss certain properties of the model. Throughout the section we focus on general survival time data, leaving a discussion on the melanoma data to Sect. 4. An alternative description of the model and discussion of some simple examples are given in Chap. 7.4 of Aalen et al. (2008).

In the multiplicative frailty model the hazard for an individual is given as the product of a frailty variable Y and a basic rate $\lambda(t)$ common to all individuals. This paper only considers parametric choices for the baseline hazard $\lambda(t)$. Conditional on Y, the individual hazard is given by:

$$h(t|Y) = Y\lambda(t). \tag{1}$$

The survival function is given by $S(t) = L(\Lambda(t))$, where $\Lambda(t) = \int_0^t \lambda(s)ds$ is the cumulative baseline hazard and $L(s)$ is the Laplace transform of Y. For ease of notation

there are no covariates in model (1). However, the model may easily be extended to include covariates.

As described in the Introduction, we assume that the combined frailty Y for an individual is composed of two or more frailty variables in a hierarchical way. We denote the frailty variable on the first level by Z_1, the one on the second level by Z_2, etc., and let f_{Z_i} be the probability density of Z_i. We use the notation Z_i for random variables, and z_i for the value given all Z_j's on higher levels (z_i is then constant). Throughout, the Z_i's are assumed to follow distributions defined by non-negative Lévy processes, either single distributions or sums of distributions as in Sect. 3.2. However, the theory on Lévy processes is not important here, it is mainly introduced due to simple notation and properties when calculating the expected value and variance of the models in Sect. 3.3. For more information on frailty models derived from Lévy processes, see Aalen and Hjort (2002), Gjessing et al. (2003) and Chap. 11 in Aalen et al. (2008). For any such frailty Z_i, the Laplace transform can be written as

$$L_{Z_i}(s) = \exp\left[-z_{i+1}\Phi_i(s)\right]. \tag{2}$$

Here, z_{i+1} corresponds to the time parameter in a Lévy process, s is the argument of the Laplace transform, and $\Phi_i(s)$ is the Laplace exponent or cumulant generating function. The constant z_{i+1} can also be seen as a scale transformation of a parameter in the distribution of Z_i. For instance, if Z_i is gamma distributed, the Laplace transform is

$$L_{Z_i}(s) = \frac{\theta_i^{z_{i+1}\rho_i}}{(\theta_i + s)^{z_{i+1}\rho_i}} = \exp(-z_{i+1}\rho_i\,[\ln(\theta_i + s) - \ln\theta_i]),$$

yielding the Laplace exponent $\Phi_i(s) = \rho_i\,[\ln(\theta_i + s) - \ln\theta_i]$. Hence, z_{i+1} is a scale transformation of ρ_i, and $\mathrm{E}(Z_i) = z_{i+1}\rho_i/\theta_i$ and $\mathrm{Var}(Z_i) = z_{i+1}\rho_i/\theta_i^2$ (Gjessing et al. 2003).

Let Z_1 be the frailty variable at the first level, and let the frailty on all higher levels be given. The variable Z_1 may have independent values for all individuals, for instance to model individual heterogeneity not captured by a parametric baseline hazard $\lambda(t)$, or it may be shared for individuals within clusters. If there are no higher level frailties and Z_1 is shared within clusters, one gets the standard shared frailty model. Conditional on Z_1, the individual hazard is then $h(t|Z_1) = Z_1\lambda(t)$ from (1), and the Laplace transform is $L_{Z_1}(s) = \exp(-z_2\Phi_1(s))$ from (2). One may now randomize the parameter z_2 by a second frailty Z_2 (with distribution f_{Z_2}) to get a two-level model. At this level, Z_2 will typically be independent for some individuals but shared for others, thus creating dependence between individuals in a cluster. The marginal Laplace transform $L(s)$ of the combined frailty Y for each individual will now be

$$L(s) = \mathrm{E}\left(L_{Z_1}(s)|Z_2\right) = \int \exp\left(-z_2\Phi_1(s)\right) f_{Z_2}(z_2)dz_2 = \exp\left[-z_3\Phi_2(\Phi_1(s))\right].$$

One may further randomize z_3 by using a third frailty Z_3 to create a model with yet another level of dependence. The marginal Laplace transform of the combined frailty Y will then be $L(s) = \exp\left[-z_4\Phi_3\left(\Phi_2(\Phi_1(s))\right)\right]$, and so on for further levels. Hence,

new levels of frailty are introduced by randomizing the Lévy time parameter, or a scale transformation of the Laplace exponent. An interesting special case for the three-level model applies when positive stable distributions are used for all Z_i, that is, when $\Phi_i(s) = \rho_i s^{\alpha_i}$. The marginal Laplace transform of Y for an individual then becomes

$$L(s) = \exp\left[-z_4\rho_3\rho_2^{\alpha_3}\rho_1^{\alpha_3\alpha_2}s^{\alpha_3\alpha_2\alpha_1}\right],$$

which again is the Laplace transform of a stable distribution. This is identical to the Laplace transform of a multiplicative stable frailty model, as presented in Hougaard (2000, pp. 354–362).

3.1 A frailty model for family data in different neighborhoods

Consider a model for data with two levels of dependence, consisting e.g. of individuals organized into families living in different neighborhoods. The frailty variable Z_1 could model individual environmental factors which are independent for all, while Z_2 models environmental/genetic factors which are shared within families, but independent between families. The variable Z_3 could describe common environmental factors shared by all individuals in the neighborhood. Let different superscripts denote independent values of the Z_i's. As a simple example, a neighborhood with lifetimes (t_{11}, t_{12}) in family 1 and (t_{21}, t_{22}) in family 2 gets the joint Laplace transform $L(s_{11}, s_{12}, s_{21}, s_{22})$ found as

$$\begin{aligned}
&\iint \exp[-z_2^{(1)}\Phi_1(s_{11}) - z_2^{(1)}\Phi_1(s_{12}) - z_2^{(2)}\Phi_1(s_{21}) - z_2^{(2)}\Phi_1(s_{22})] f_{Z_2}(z_2)dz_2 f_{Z_3}(z_3)dz_3 \\
&\quad = \int \exp[-z_3\Phi_2(\Phi_1(s_{11}) + \Phi_1(s_{12})) - z_3\Phi_2(\Phi_1(s_{21}) + \Phi_1(s_{22}))] f_{Z_3}(z_3)dz_3 \\
&\quad = \exp\left[-z_4\Phi_3(\Phi_2(\Phi_1(s_{11}) + \Phi_1(s_{12})) + \Phi_2(\Phi_1(s_{21}) + \Phi_1(s_{22})))\right].
\end{aligned}$$

Generally, assume there are m_j individuals in family j and n families in the neighborhood. The survival function for the neighborhood is then

$$S(t_{11}, \ldots, t_{nm_n}) = \exp\left[-z_4\Phi_3\left(\sum_{j=1}^{n}\Phi_2\left(\sum_{l=1}^{m_j}\Phi_1(\Lambda(t_{jl}))\right)\right)\right].$$

The survival function may be used directly in the likelihood, see Sect. 3.5.

3.2 Combined additive and hierarchical genetic model for family data

One may also construct a model combining genetic and environmental effects, for instance for data on nuclear families consisting of parents and up to two children. This model is applied to the melanoma data in Sect. 4. On the bottom level, we have a frailty for individual environment, Z_1, which is assumed independent for all. The Laplace transform is $L_{Z_1}(s) = \exp(-z_2\Phi_1(s))$, as usual. On the middle level, we randomize

z_2 by an additive genetic component, which looks like the following for a family of a mother (M), a father (F) and two children (C_1, C_2):

$$\begin{aligned} Z_{2M} &= M_1 + M_2 + M_3 + M_4 \\ Z_{2F} &= F_1 + F_2 + F_3 + F_4 \\ Z_{2C_1} &= M_2 + M_3 + F_2 + F_3 \\ Z_{2C_2} &= M_3 + M_4 + F_3 + F_4. \end{aligned}$$

All random effects M_i, F_i are assumed i.i.d., with Laplace transform $L(s) = \exp(-z_3\Phi_2(s))$. Hence, the parents have independent genetic frailties, but the children share half of the genetic frailty with their parents and each other. Moreover, they do not share the same half of the genes. This is the same method as used for standard additive frailty models (Korsgaard and Andersen 1998). Let m and f be the argument of the Laplace transform for the parents, and c_1, c_2 the arguments for the children. By using function iteration on the Laplace exponent, we get the joint Laplace transform $L(m, f, c_1, c_2)$ for the family:

$$\begin{aligned} &\mathrm{E}(\exp\left[-Z_{2M}\Phi_1(m) - Z_{2F}\Phi_1(f) - Z_{2C_1}\Phi_1(c_1) - Z_{2C_2}\Phi_1(c_2)\right] | M_i, F_i, i=1,\ldots,4) \\ &\quad = \mathrm{E}(\exp\{-M_1\Phi_1(m) - F_1\Phi_1(f) - M_2\left[\Phi_1(m) + \Phi_1(c_1)\right] \\ &\qquad -F_2\left[\Phi_1(f) + \Phi_1(c_1)\right] - M_4\left[\Phi_1(m) + \Phi_1(c_2)\right] \\ &\qquad -F_4\left[\Phi_1(f) + \Phi_1(c_2)\right] - M_3\left[\Phi_1(m) + \Phi_1(c_1) + \Phi_1(c_2)\right] \\ &\qquad -F_3\left[\Phi_1(f) + \Phi_1(c_1) + \Phi_1(c_2)\right]\} | M_i, F_i, i = 1,\ldots,4) \\ &\quad = \exp\{-z_3[\Phi_2\left(\Phi_1(m)\right) + \Phi_2\left(\Phi_1(f)\right) + \Phi_2\left(\Phi_1(m) + \Phi_1(c_1)\right) \\ &\qquad +\Phi_2(\Phi_1(f) + \Phi_1(c_1)) + \Phi_2(\Phi_1(m) + \Phi_1(c_2)) \\ &\qquad +\Phi_2(\Phi_1(f) + \Phi_1(c_2)) + \Phi_2(\Phi_1(m) + \Phi_1(c_1) + \Phi_1(c_2)) \\ &\qquad +\Phi_2(\Phi_1(f) + \Phi_1(c_1) + \Phi_1(c_2))]\}. \end{aligned}$$

One may further expand the model by introducing a common environmental term. This is done by randomizing z_3 by Z_3, which is shared by all. Hence, Z_3 has the Laplace transform $L_{Z_3}(s) = \exp(-z_4\Phi_3(s))$. The joint Laplace transform then becomes

$$\begin{aligned} &\exp\{-z_4\Phi_3[\Phi_2(\Phi_1(m)) + \Phi_2(\Phi_1(f)) + \Phi_2(\Phi_1(m) + \Phi_1(c_1)) \\ &\quad +\Phi_2(\Phi_1(f) + \Phi_1(c_1)) + \Phi_2(\Phi_1(m) + \Phi_1(c_2)) \\ &\quad +\Phi_2(\Phi_1(f) + \Phi_1(c_2)) + \Phi_2(\Phi_1(m) + \Phi_1(c_1) + \Phi_1(c_2)) \\ &\quad +\Phi_2(\Phi_1(f) + \Phi_1(c_1) + \Phi_1(c_2))]\}, \end{aligned} \tag{3}$$

and one gets the joint survival function by substituting the Laplace arguments with the $\Lambda(t_i)$'s. Expanding these models to an arbitrary number of children is straightforward, by adding more components to the additive genetic frailty Z_2. However, the expressions quickly become complicated. For only one child in every family, you need two additive components in Z_2, for two children you need four components, for three children you need eight components, and so on. If the data consist of a mixture of

families with one and two children, the model is still identifiable when using four additive terms in the genetic component.

3.3 Expected values, covariance and correlation

Simple results are valid for the expectation and variance of the hierarchical Lévy frailty model, provided that they exist for the model in question (e.g. the stable distribution has no finite expected value or variance). If the baseline hazard $\lambda(t)$ in (1) includes a scale parameter, one often sets the expectation of the frailty distribution equal to one, to assure identifiability. Generally for distributions derived from non-negative Lévy processes, $\Phi_i(0) = 0$. Assume that the expectation of the Z_i's equals one, that is, $\Phi_i'(0) = 1$ for all i. For the genetic model in Sect. 3.2, this implies that E(M_i)=E(F_i)=1/4 for all i. The time parameter at the highest level, z_k, is set equal to one. By using the rules of double expectation, double variance and induction, one may show that

$$\mathrm{E}Y = 1, \quad \mathrm{Var}Y = \mathrm{Var}Z_1 + \mathrm{Var}Z_2 + \mathrm{Var}Z_3 \tag{4}$$

for three-level models, and similarly for higher level models. Hence, the variance of a hierarchical Lévy frailty variable is 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. These formulas are valid also when Z_2 is additive, as the expectation and variance on that level is the sum of the expectations and variances of each component (since they are i.i.d.). Even though the variance of Y can be written as a sum of the variances of the Z_i's, no simple general algebraic relation can be found between Y and (Z_1, Z_2, Z_3). If $\Phi_i'(0) \neq 1$, Var$Y$ is a function both of the expectation and the variance on each level, making the formula more complicated. When $\Phi_i(s)$ has two parameters, setting $\Phi_i'(0) = 1$ implies that only one parameter is left on that level (i.e. $\rho_i = \theta_i$ for gamma distributed Z_i or $\rho_i = \theta_i/4$ for all M_i, F_i in the genetic model).

For the genetic model in Sect. 3.2, a measure of the importance of the genetic component relative to the total variance of the components generating dependence in the model can be formulated as

$$h^2 = \frac{\mathrm{Var}Z_2}{\mathrm{Var}Z_2 + \mathrm{Var}Z_3}. \tag{5}$$

This is similar to the squared coefficient of heritability, but does not include the variance of the components creating individual variance (Z_1 and the baseline hazard $\lambda(t)$). Including the variability in the survival times due to $\lambda(t)$ seems very difficult, as the correlation structure is put on the latent frailty variables, not on the outcomes. This problem is present in all frailty models, as well as other models for latent variable modelling such as generalized linear mixed models (see e.g. Pawitan et al. 2004; Gjessing and Lie 2008). Including only the variance of Z_1 is also not satisfactory, as this value would greatly depend on the choice of $\lambda(t)$.

The covariance is derived in a similar manner as the variance, by using that

$$\mathrm{Cov}(Y_1, Y_2) = \mathrm{E}(\mathrm{Cov}(Z_1, Z_1|Z_2)) + \mathrm{Cov}(\mathrm{E}(Z_1|Z_2), \mathrm{E}(Z_1|Z_2)).$$

For a three-level model, this gives the following covariance between two individuals j and k:

$$\mathrm{Cov}(Y_j, Y_k) = \mathrm{Cov}(Z_{1j}, Z_{1k}) + \mathrm{Cov}(Z_{2j}, Z_{2k}) + \mathrm{Cov}(Z_{3j}, Z_{3k}).$$

The covariances in the sum are always either zero if the individuals are independent in a frailty component or equal to the variance of a shared frailty component. If the frailty on a level is additive, the covariance on that level can also be decomposed into a sum of variances of shared frailty terms. As an example, consider the genetic model in Sect. 3.2. Let the frailty on level i have variance σ_i^2 (the variance of all M_i and F_i on the additive second level is then $\sigma_2^2/4$), and let $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$ be the variance of the combined frailty for the family members. We get the following covariances between the frailties of the family members:

$$\begin{aligned}
\mathrm{Cov}(Y_M, Y_F) &= \mathrm{Var}Z_3 = \sigma_3^2 \\
\mathrm{Cov}(Y_M, Y_{C_1}) &= \mathrm{Var}M_2 + \mathrm{Var}M_3 + \mathrm{Var}Z_3 = \frac{1}{2}\sigma_2^2 + \sigma_3^2 \\
\mathrm{Cov}(Y_M, Y_{C_2}) &= \mathrm{Var}M_3 + \mathrm{Var}M_4 + \mathrm{Var}Z_3 = \frac{1}{2}\sigma_2^2 + \sigma_3^2 \\
\mathrm{Cov}(Y_F, Y_{C_1}) &= \mathrm{Var}F_2 + \mathrm{Var}F_3 + \mathrm{Var}Z_3 = \frac{1}{2}\sigma_2^2 + \sigma_3^2 \\
\mathrm{Cov}(Y_F, Y_{C_2}) &= \mathrm{Var}F_3 + \mathrm{Var}F_4 + \mathrm{Var}Z_3 = \frac{1}{2}\sigma_2^2 + \sigma_3^2 \\
\mathrm{Cov}(Y_{C_1}, Y_{C_2}) &= \mathrm{Var}M_3 + \mathrm{Var}F_3 + \mathrm{Var}Z_3 = \frac{1}{2}\sigma_2^2 + \sigma_3^2.
\end{aligned}$$

Thus the covariance between the children is equal to the covariance between a parent and a child. The covariance between the parents is smaller than the others. The correlations between the frailties of the family members will then be:

$$\mathrm{Corr}(Y_M, Y_F) = \frac{\sigma_3^2}{\sigma^2}, \quad \mathrm{Corr}\left(Y_j, Y_k\right) = \frac{\frac{1}{2}\sigma_2^2 + \sigma_3^2}{\sigma^2} \quad (j,k) \neq (M,F).$$

3.4 Dependence measures for times to events

Instead of considering the frailty correlation, one may wish to study the dependence in the outcomes: The times to events. One may then calculate the frailty relative risk (FRR), introduced in Moger and Aalen (2005). It is defined as the probability of experiencing the event within a specific age t_1 given that a relative or another member of the cluster has experienced the event within age t_2, compared to the probability of getting the event within age t_1 if another member of the cluster has survived up to

age t_2 without the event. By using Bayes' theorem and that any pair of individuals are independent given all Z_i, $i > 1$, and then integrating out the frailties, one gets

$$\begin{aligned}\text{FRR} &= \frac{P(\text{Individual gets event within age } t_1 | \text{Relative gets event within age } t_2)}{P(\text{Individual gets event within age } t_1 | \text{Relative not event up to age } t_2)} \\ &= \frac{(1 - S(t_1) - S(t_2) + S(t_1, t_2))S(t_2)}{(1 - S(t_2))(S(t_2) - S(t_1, t_2))}. \end{aligned} \tag{6}$$

Hence, one needs the univariate and bivariate survival functions for different pairs within a cluster. These are found by setting the appropriate t's in (7) equal to zero. For instance, for the genetic model in Sect. 3.2, the bivariate survival function for mother and father is found by setting the survival times of the children equal to zero. Results may be presented for given choices of t_1 and t_2.

A frequently used measure for the dependence in times to events for frailty models, is Kendall's τ. No general analytic expression for the models presented here can be found, but Kendall's τ can be calculated using numerical integration of the bivariate survial functions and densities for different pairs within clusters (Hougaard 2000, Eq. 4.4).

3.5 Likelihood construction

It is difficult to give a general expression for the likelihood, since the formula for the likelihood contribution of each cluster depends on which individuals in the cluster experience the event. Let $\Lambda(t) = \int_0^t \lambda(s)ds$ be the cumulative baseline hazard, and let L_X denote the joint Laplace transform of the cluster, e.g. as given by (3) for the genetic model. Then

$$S(t_{j1}, \ldots, t_{jk}) = L_X(\Lambda(t_{j1}), \ldots, \Lambda(t_{jk})) \tag{7}$$

is the joint survival function of cluster j, $j = 1, \ldots, n$. Let δ_{jl} be an indicator on whether individual l, $l = 1, \ldots, k$, in cluster j has been censored ($\delta_{jl} = 0$) or not ($\delta_{jl} = 1$), and let $\delta_{j\cdot}$ be the number of cases in cluster j. The log-likelihood for n clusters can then be written as

$$\begin{aligned}\log L &= \sum_{j=1}^{n} \log \left[\frac{\partial^{\delta_{j\cdot}}}{(\partial t_{j1})^{\delta_{j1}} \cdots (\partial t_{jk})^{\delta_{jk}}} (-1)^{\delta_{j\cdot}} S(t_{j1}, \ldots, t_{jk}) \right] \\ &= \sum_{j=1}^{n} \log f_j(t_{j1}, \ldots, t_{jk}, \delta_{j1}, \ldots, \delta_{jk}). \end{aligned}$$

Hence, the contribution for a cluster where none have an event is equal to the joint survival function for that cluster. If one or more members has an event, one differentiates the joint survival function with respect to the t_{jl}'s of these members to find the joint density f_j for the cluster. For the model in Sect. 3.1, it may be possible to find a general expression for the $\delta_{j\cdot}$th derivative of the survival function, at least for the gamma

and stable distributions. For the genetic model in Sect. 3.2, one may use a software package like Mathematica to differentiate the survival function for all combinations of events in families observed in the data and paste the output into a function for the likelihood function in e.g. R. This is straightforward. Adjustment for left-truncation is handled by dividing the density f_j for a cluster with the joint probability that they have survived up to age at truncation r_{jl}, $S(r_{j1}, \ldots, r_{jk})$. For cohort data, the standard errors of the parameters are found by the inverse of the observed information matrix.

4 Analysis of the melanoma data

We apply the genetic model in Sect. 3.2 to the melanoma data. Likelihood maximization is done in R. To adjust for truncation, we use the joint probability that they have no melanoma before 1961, which will be elaborated upon in the Discussion. The baseline hazard $\lambda(t)$ in (1) is assumed to be of the Weibull form $\alpha\kappa t^{\kappa-1}$. The frailty on each level is gamma distributed. Using PVF distributions for the components give similar results. This yields three parameters for the frailty (individual environment θ_1, genetics θ_2 and common environment θ_3) with $\Phi_i(s) = \theta_i\,[\ln(\theta_i + s) - \ln\theta_i]$ for $i = 1, 2, 3$. A high value of θ_i yields low variance and correlation on that level. In addition, we have two regression coefficients for the covariates birth cohort and gender. The covariates are included as a Cox regression term, giving $\lambda(t) = \exp(\boldsymbol{\beta X})\alpha\kappa t^{\kappa-1}$. When using the model in Sect. 3.2 to analyze the data, age at melanoma diagnosis is the time variable of interest while death from other causes is treated as censoring.

Since we are analyzing a case-cohort sample, we use the methods from Sect. 3 in Moger et al. (2008); stratified sampling without replacement. Sampling weights p_i enter the likelihood in Sect. 3.5, yielding a pseudo-likelihood, where p_i denotes the sampling probability for a nuclear family in stratum i. The stratum for case families has $i = 0$. The control families have $i = 1, \ldots, 140$, as there are 70 birth year strata $\times$ 2 family size strata. The log pseudo-likelihood function with (possibly) censored event times (the t_{jl}'s), indicators of censoring (the δ_{jl}'s) and truncation times (the r_{jl}'s) in bold vector form is then

$$\log L_{\text{pseudo}} = \sum_{i=0}^{140} \frac{1}{p_i} \sum_{j \in D_i} \log g_j(\boldsymbol{t}_j, \boldsymbol{\delta}_j, \boldsymbol{r}_j; \boldsymbol{\theta}),$$

where D_i is the set of families sampled from stratum i and $\boldsymbol{\theta}$ denotes all parameters in the model. The likelihood contribution for each family, g_j, is adjusted for truncation, $g_j(\boldsymbol{t}_j, \boldsymbol{\delta}_j, \boldsymbol{r}_j; \boldsymbol{\theta}) = f_j(\boldsymbol{t}_j, \boldsymbol{\delta}_j; \boldsymbol{\theta})/S_j(\boldsymbol{r}_j; \boldsymbol{\theta})$. As all case families in the Multi-Generation Register are included in the sample, $p_0 = 1$. With m_i families sampled out of the n_i families in each stratum, the control families get inverse weights $1/p_i = n_i/m_i$. The standard errors are estimated by a sandwich-type estimator $\boldsymbol{A}(\boldsymbol{\theta})^{-1} + \boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{B}_{\text{st}}(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})^{-1}$. Here, $\boldsymbol{A}(\boldsymbol{\theta})$ is estimated by

$$\widehat{\boldsymbol{A}}(\widehat{\boldsymbol{\theta}}) = \sum_{i=0}^{140} \frac{1}{p_i} \sum_{j \in D_i} \boldsymbol{I}_j(\widehat{\boldsymbol{\theta}}),$$

where $\boldsymbol{I}_j(\boldsymbol{\theta}) = -\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}' \log g_j(\boldsymbol{t}_j, \boldsymbol{\delta}_j, \boldsymbol{r}_j; \boldsymbol{\theta})$, the observed information matrix for family j. $\boldsymbol{B}_{\mathrm{st}}(\boldsymbol{\theta})$ is estimated by

$$\widehat{\boldsymbol{B}}_{\mathrm{st}}(\widehat{\boldsymbol{\theta}}) = \sum_{i=0}^{140} \frac{1-p_i}{p_i^2} \sum_{j\in D_i} (\boldsymbol{s}_j(\widehat{\boldsymbol{\theta}}) - \overline{\boldsymbol{s}}_i(\widehat{\boldsymbol{\theta}}))(\boldsymbol{s}_j(\widehat{\boldsymbol{\theta}}) - \overline{\boldsymbol{s}}_i(\widehat{\boldsymbol{\theta}}))',$$

where $\boldsymbol{s}_j(\boldsymbol{\theta}) = \partial/\partial\boldsymbol{\theta} \log g_j(\boldsymbol{t}_j, \boldsymbol{\delta}_j, \boldsymbol{r}_j; \boldsymbol{\theta})$, the score function for family j, and $\overline{\boldsymbol{s}}_i(\widehat{\boldsymbol{\theta}}) = m_i^{-1} \sum_{j\in D_i} \boldsymbol{s}_j(\widehat{\boldsymbol{\theta}})$, the estimated average value of the score function in stratum i.

First, we fit the model without covariates. For comparison, we also fit an additive gamma model with Weibull baseline and the following frailty structure:

$$\begin{aligned} Z_M &= I_1 + M_1 + M_2 + M_3 + M_4 + E \\ Z_F &= I_2 + F_1 + F_2 + F_3 + F_4 + E \\ Z_{C_1} &= I_3 + M_2 + M_3 + F_2 + F_3 + E \\ Z_{C_2} &= I_4 + M_3 + M_4 + F_3 + F_4 + E. \end{aligned}$$

Here, I_i denotes individual environmental frailty, M_i and F_i denote genetic frailty as before, and E is common environmental frailty, all gamma distributed with parameters θ_1 for I_i, θ_2 for M_i and F_i, and θ_3 for E. The marginal fit from the estimated multivariate models is compared to a Nelson–Aalen plot in Fig. 1. The hierarchical model has a better fit, especially around 60 years and after 80 years. The multivariate fit is also better for the hierarchical model, indicated by the log pseudo-likelihood values of −307050.7 for the additive gamma model, and −306835.1 for the hierarchical gamma model (even though the usual likelihood ratio test is not applicable for pseudo-likelihoods).

The results without covariates are shown in the upper part of Table 2. The variance $\widehat{\sigma}_2^2$ of the genetic component is $1/\widehat{\theta}_2 = 1.99$ and the common environmental variance $\widehat{\sigma}_3^2$ is 0.66. Hence, from (5), h^2 is around 75.1%. The lower part of Table 2 shows results with the covariates gender (0= female, 1=male) and birth cohort (continuous, per 10 years) included. The relative risk of birth year per 10 years is 1.49 (95% CI 1.48–1.51), hence a 10 year increase in birth year is associated with a 49% increase in risk of melanoma. The corresponding relative risk for gender is 0.97 (95% CI 0.94–0.99), meaning that males have a 3% lower risk of melanoma than females. The estimate of the gamma parameter in the genetic component is 0.32, indicating that some families have much higher values of the genetic frailty than others. The variance is 3.10, and the corresponding variance for common environment is 0.23. By calculating h^2, genetics now account for 93.1% of the total frailty variance of the components generating dependence in the model. Fitting a model where the individual component Z_1 is removed yields an equally good fit based on log pseudo-likelihood values (−304279.5, three-level model, vs. −304279.7, two-level model), indicating that the individual variation is captured by the Weibull baseline once covariates are included.

To illustrate the dependence in time to melanoma, one may calculate the frailty relative risk of getting melanoma as a function of age given that a relative is affected by

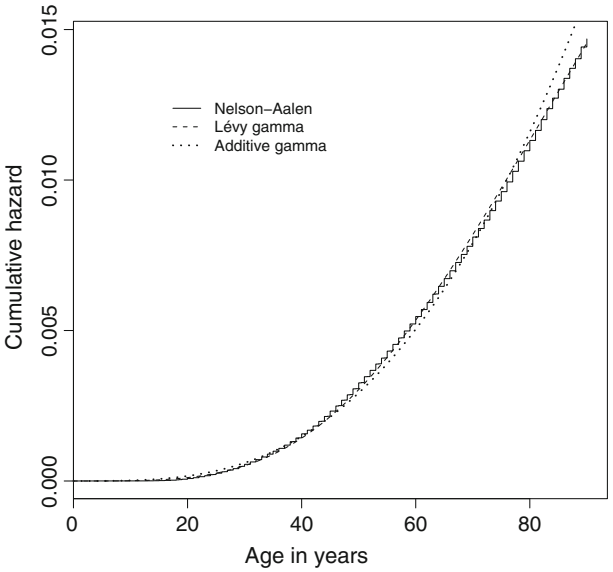

Fig. 1 Marginal fit of an additive gamma model and the hierarchical gamma model compared to Nelson–Aalen. Both models have Weibull baseline hazards and frailty components for genetics, common and individual environment, giving five parameters in total

Table 2 Parameter estimates from the hierarchical gamma model, analysis of the melanoma data

Parameter:	α	κ	θ_1	θ_2	θ_3	β gender	β birth year
Hierarchical gamma model without covariates							
Estimate	1.49×10^{-9}	3.75	1.08×10^{-2}	0.50	1.52	–	–
SE	2.1×10^{-10}	0.04	4.80×10^{-4}	0.13	0.44	–	–
Hierarchical gamma model with covariates							
Estimate	4.83×10^{-11}	4.55	7.02×10^{2}	0.32	4.34	−0.03	0.400
SE	5.51×10^{-12}	0.03	3.13×10^{4}	0.04	2.67	0.01	0.005

Females are the reference group for gender, and birth year is per 10 years
SE standard error

melanoma within 90 years (the upper limit for age), compared to the relative not being affected within that age. This is done by setting $t_2 = 90$ in (6), inserting the parameter estimates and some covariate values, as the frailty relative risk depends slightly on the covariates (we inserted the mean value of birth cohort, and appropriate genders). We computed the frailty relative risk for different values of t_1 but the estimates differed with only one tenth both for partners, siblings and parent–child pairs. If your partner gets melanoma, you have 1.23 times higher risk of getting it due to the shared environment for partners. If your parent, child or sibling gets melanoma, you have 2.70–2.80 times higher risk, due to shared environment and genetics. The dependence of the covariates results in slightly different estimates for e.g. mother–daughter and

brother–sister pairs. Similarly, with covariate values inserted, Kendall's τ is 0.10 for partners and 0.31–0.32 for sibling or parent–child pairs.

5 Discussion

This paper presents a new hierarchical frailty model for multivariate survival data. The model is especially well suited for analyzing data with hierarchical dependence structures, such as the neighborhood example in Sect. 3.1, as these structures naturally arise when adding levels to the model. However, by using additive levels one may also construct models with genetic components. The data example shows that a better marginal and multivariate fit can be achieved by using a hierarchical model compared to a corresponding additive one, indicating that the combined frailty distribution in the hierarchical model is more flexible. Including additional frailty terms for adult or childhood environment only, thus making the correlation between parent–child pairs and sibling pairs different, is easily achieved in an additive model. However, it seems difficult in the hierarchical frailty model. Once an additive level has been included, one may only put a frailty shared by all in the cluster on the level above. Both in the hierarchical and additive models, the correlations between family members are functions of the frailty variances. A log-normal model is more flexible in this regard, as covariance parameters may be estimated separately. However, for the log-normal model estimation is complicated by the fact that there is no analytical expression for the likelihood, making techniques like e.g. numerical integration necessary in the estimation. For large data sets such as ours, this might be a computational problem.

We have analyzed the melanoma data as left-truncated and right-censored failure time data with the age of melanoma diagnosis as the failure time of interest, treating death without melanoma as censoring. This way of analyzing the data is not without problems. The first problem is that we adjust for truncation using the joint probability of no melanoma in each family before 1961 (cf. Sect. 2), whereas the correct adjustment is to use the joint probability of survival up to 1961. Another problem is that we cannot know if there have been previous occurrences of melanoma for individuals alive in 1961, hence we implicitly assume in the truncation adjustment that any melanoma case for an individual is independent of possible previous occurrences, and this is the best we can do. For solving the first problem, it would have been more appropriate to use a competing risks model with occurrence of melanoma as the event of interest and death without melanoma as a competing cause. Formula (1) would then correspond to the conditional cause-specific hazard for the occurrence of melanoma (conditional on a frailty that is modeled in a hierarchical way as described in Sect. 3.2). In principle one could also assume a frailty model for death without melanoma, where a component of the frailty is common for the two causes; see Aalen et al. (2008, Sect. 6.6) for an example of such a model. To follow this approach is, however, outside the scope of the present paper. However, if the frailties for death without melanoma are independent of the frailties for occurrence of melanoma, and different sets of parameters are used to model the two competing causes, then the likelihood based on a competing risks model will factor into two components, one for each competing cause. Furthermore,

the component corresponding to occurrence of melanoma will be of the form described above, justifying the analysis we have presented.

Regarding the goodness-of-fit, several approaches can be considered. When comparing different models for cohort data, one may use the likelihood ratio test. One can remove a level to see how much it affects the likelihood value. It is also interesting to see how the multivariate model fits the marginal data by comparing fitted curves to Kaplan–Meier or Nelson–Aalen plots as in Fig. 1. The parameters of the model may also be estimated from marginal data. Ideally, one would like the total frailty variance to be similar in both cases, but as long as the marginal fit of the multivariate model is acceptable in the plots, this should not be an issue. Also, it is questionable whether the parameters are well identified when fitting the model to marginal data. We assumed that the dependence was the same for parent–child pairs as for sibling pairs. This may be checked by fitting the model to bivariate data, and comparing the total genetic and shared environmental variance to the results in Sect. 4 (as the components generating dependence cannot be properly identified from bivariate data, only total variance is relevant). This gives a variance of 3.45 for sibling pairs and 1.77 for mother–child pairs without covariates. The estimate from Sect. 4 of 2.65 is in-between. With covariates, we get 2.58 for sibling pairs and 2.40 for mother–child pairs, indicating that the correlation is similar after adjusting for birth cohort and gender. The estimate from Sect. 4 is 3.33, and hence a bit higher. Again, if the results are different, the multivariate fit of the model might still be acceptable. This could in principle be checked e.g. by comparing the estimated bivariate survival functions from the model to bivariate Kaplan–Meier plots of survival times of pairs of relatives, although this approach may not be very practical.

Family data gives the opportunity to study the heritability of some specific trait, here frailty (susceptibility) to melanoma. The trait is influenced by genetic as well as by environmental factors, assumed to be independent and that is why we assume an additive structure for the variances (4). The variance of the trait (total frailty variance) describes variability of genetic as well as environmental factors not included into the model without covariates. If covariates are included, we expect that the total frailty variance decreases, because these covariates explain some of the variability. The question is whether the included covariates are pure environmental - then only the environmental part of the frailty variance should decline - or if the covariates included have themselves some genetic background (e.g. BMI, Hypertension, etc.). Then we will also see a reduction in the genetic frailty variance. The inclusion of genetic marker information should only reduce the genetic variance, but not the environmental variance.

Confidence intervals for h^2 and the frailty relative risk are complex functions of the parameters in the model and not presented. Normally, one could have found confidence intervals by using the bootstrap method, but this is complicated by the fact that we are analyzing a case-cohort sample of families and that the data set is very large. Fitting the model takes around 4 h in R on a computer with 2 GHz processor and 2 GB RAM. Several studies show familial aggregation of melanoma. A meta-analysis (Ford et al. 1995) showed a relative risk of 2.24 for melanoma in individuals with affected first-degree relatives, somewhat lower than our estimate of 2.70–2.80. A previous study (Hemminki et al. 2003) on data from the Swedish Multi-Generation Register,

yielded a relative risk of 2.40 for parent–child pairs and 2.98 for sib–sib pairs, and our estimate is in-between these.

An extension to a 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 (Korsgaard and Andersen 1998) and multivariate log-normal models (Yau 2001). Also, in the absence of covariates, a non-parametric hazard would not allow for determining the distribution of the individual frailty component.

Acknowledgements We are grateful to Professor Yudi Pawitan at Karolinska Institutet in Stockholm, Sweden, for getting access to the melanoma data and discussions on the paper. We also wish to thank the associate editor and the referee for valuable comments. Marion Haugen was supported by Statistics for Innovation (sfi)^2, project number 460739.

References

Aalen OO, Hjort NL (2002) Frailty models that yield proportional hazards. Stat Probab Lett 58:335–342

Aalen OO, Borgan Ø, Gjessing HK (2008) Survival and event history analysis. A process point of view. Springer, New York

Ford D, Bliss JM, Swerdlow AJ et al (1995) Risk of cutaneous melanoma associated with a family history of the disease: the International Melanoma Analysis Group (IMAGE). Int J Cancer 62:377–381

Gjessing HK, Lie RT (2008) Biometrical modelling in genetics: are complex traits too complex? Stat Methods Med Res 17:75–96

Gjessing HK, Aalen OO, Hjort NL (2003) Frailty models based on Lévy processes. Adv Appl Probab 35:532–550

Hemminki K, Zhang H, Czene K (2003) Familial and attributable risks in cutaneous melanoma: effects of proband and age. J Invest Dermatol 120:217–223

Hougaard P (2000) Analysis of multivariate survival data. Springer, New York

Korsgaard IR, Andersen AH (1998) The additive genetic gamma frailty model. Scand J Stat 25:255–269

Moger TA, Aalen OO (2005) A distribution for multivariate frailty based on the compound Poisson distribution with random scale. Lifetime Data Anal 11:41–59

Moger TA, Pawitan Y, Borgan Ø (2008) Case-cohort methods for survival data on families from routine registers. Stat Med 27:1062–1074

Pawitan Y, Reilly M, Nilsson E et al (2004) Estimation of genetic and environmental factors for binary traits using family data. Stat Med 23:449–465

Petersen JH (1998) An additive frailty model for correlated life times. Biometrics 54:646–661

Ripatti S, Palmgren J (2000) Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56:1016–1022

Thörn M, Pontén F, Johansson AM et al (1998) Rapid increase in diagnosis of cutaneous melanoma in situ in Sweden, 1968–1992. Cancer Detect Prev 22:430–437

Yau KKW (2001) Multilevel models for survival analysis with random effects. Biometrics 57:96–102