A bivariate survival model with compound Poisson frailty

A. Wienke,^{a,*†} S. Ripatti,^{b,c} J. Palmgren^{d,e} and A. Yashin^f

A correlated frailty model is suggested for analysis of bivariate time-to-event data. The model is an extension of the correlated power variance function (PVF) frailty model (correlated three-parameter frailty model) (*J. Epidemiol. Biostat.* 1999; 4:53–60). It is based on a bivariate extension of the compound Poisson frailty model in univariate survival analysis (*Ann. Appl. Probab.* 1992; 4:951–972). It allows for a non-susceptible fraction (of zero frailty) in the population, overcoming the common assumption in survival analysis that all individuals are susceptible to the event under study. The model contains the correlated gamma frailty model and the correlated inverse Gaussian frailty model as special cases. A maximum likelihood estimation procedure for the parameters is presented and its properties are studied in a small simulation study. This model is applied to breast cancer incidence data of Swedish twins. The proportion of women susceptible to breast cancer is estimated to be 15 per cent. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: survival analysis; correlated frailty model; compound Poisson distribution; cure fraction; breast cancer

1. Introduction

Traditional analysis of time-to-event data is based on two main assumptions: first, independence of observed event times and second, all individuals are susceptible to the event of interest and will eventually experience this event if the follow-up is sufficiently long.

Various approaches have been investigated to overcome the first restriction of independence, known as multivariate survival models. Available statistical models fall into two broad classes—marginal and frailty models.

Marginal methods of analysis specify models for the effect of covariates on the hazards of the individual events (the margins), taking into account the fact that observed event times are correlated but without the need for explicitly modelling this correlation [1]. However, they provide no insight into the multivariate relationship among failure times. These types of questions are answered by frailty models, considering the association between various events explicitly, and we will focus on this approach in the present paper.

Frailty models have been used frequently for modelling dependence in multivariate time-to-event data [2–8]. The dependence usually arises because individuals in the same group (family, litter, study center) are related to each other or because of the multiple recurrence of an event for the same individual.

To overcome the second strong assumption mentioned above, cure models are introduced into survival analysis. Cure models allow for a fraction of individuals in the population who are not susceptible to the event under study. Univariate cure models are well established in the literature, but only a few papers exist dealing with bivariate cure models, combining the concepts of frailty and cure modelling. A shared frailty model with an extra latent variable was used to model a non-susceptible fraction [9].

^aInstitute of Medical Epidemiology, Biostatistics and Informatics, University Halle-Wittenberg, Germany

^bInstitute for Molecular Medicine Finland, FIMM, University of Helsinki, Finland

^cPublic Health Genomics, National Institute for Health and Welfare, Finland

^dDepartment of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden

^eDepartment of Mathematical Statistics, Stockholm University, Stockholm, Sweden

^fCenter for Population Health and Aging, Duke University, Durham, U.S.A.

*Correspondence to: A. Wienke, Institute of Medical Epidemiology, Biostatistics and Informatics, University Halle-Wittenberg, Magdeburger Str. 8, 06097 Halle, Germany.

†E-mail: andreas.wienke@medizin.uni-halle.de

Contract/grant sponsor: National Institute on Aging; contract/grant numbers: R01AG027019, R01AG028259, 5P01AG008761

Contract/grant sponsor: German Ministry for Education and Research; contract/grant number: 01ZZ0404

Contract/grant sponsor: German Research Council; contract/grant number: DFG WI 3288/1-1

Extending this approach a correlated gamma frailty model was applied [6]. Other authors suggest a compound Poisson frailty model with random scale [10].

In frailty models, distributional assumptions about the frailty are often made because of mathematical convenience. The most common frailty distribution is the gamma distribution. To check the appropriateness of the correlated gamma frailty model (without a cure fraction), the correlated PVF frailty model was proposed [11], which is a bivariate extension of the univariate PVF frailty model introduced in [12]. Zahl and Harris considered three important submodels of the correlated PVF frailty model (gamma, inverse Gaussian and positive stable model) and applied these models to cancer incidence data from the Swedish Twin Registry [13]. A model similar to that proposed here was recently published [10], but with a different correlation structure.

Often there are no biological reasons as to why one frailty distribution is preferable over another. Nearly all arguments in favor or against a distribution are mathematically based. For some diseases (e.g. testicular cancer [14, 15], breast cancer [6, 9, 16, 17], recurrence of leukaemia [18], schizophrenia [19]), there are theories suggesting the existence of a fraction of non-susceptible individuals and this speaks in favor of choosing a compound Poisson frailty model.

Correlated frailty models have been developed for the analysis of multivariate failure time data [4, 6, 7, 20–23] in which associated random variables are used to characterize the frailty effect for each cluster. Correlated frailty models provide not only variance parameters of the frailties as in shared frailty models, but they also contain additional parameters for modelling the correlation between frailties.

The correlated compound Poisson frailty model will be applied to data from the Swedish Twin Registry, the largest population-based twin registry in the world. Information about onset of cancer was obtained from the National Cancer Registry. Unfortunately, such kind of registers do not commonly contain detailed information about individual risk factors. To account for this missing information frailty models can be used. In these models the frailty variable represents varying levels of risk of different individuals in the study population.

The paper is organized as follows. In Section 2.1, we introduce the general correlated frailty model. The correlated gamma frailty model is described in Section 2.2 and the new compound Poisson frailty model is introduced in Sections 2.3 and 2.4. A simulation study is presented in Section 3 and the results of an analysis of Swedish breast cancer data are given in Section 4. The paper concludes with a discussion in Section 5.

2. Statistical models

2.1. General correlated frailty model

Consider some bivariate observations, for example, cause-specific lifetimes of twins, or age at onset of a disease in spouses. Here we are dealing with single-event data (at most one event per individual). We assume that the frailties are acting multiplicatively on the baseline hazard function and that the observations in each pair are conditionally independent given the frailties. Hence, the hazard of individual $j(j=1, 2)$ in pair $i(i=1, \dots, n)$ is

$$\mu(t|Z_{ij}, X_{ij}) = Z_{ij}\mu_0(t) \exp(\beta^T X_{ij}) \quad (1)$$

where t denotes age or time, X_{ij} is a vector of observable covariates and β a vector of unknown regression coefficients describing the effect of the covariates X_{ij} , $\mu_0(t)$ is some baseline hazard function and Z_{ij} is a frailty (random effect).

Bivariate correlated frailty models are characterized by the joint distribution of a two-dimensional vector of frailties (Z_{i1}, Z_{i2}) . The form of the baseline hazard is important for the analysis because all methods described below are parametric. In principle, any parametric formula for a hazard rate is possible (e.g. Gompertz, Gompertz–Makeham, Weibull, exponential, piecewise exponential). A vast literature on human mortality and time to onset of specific diseases suggests using the Gompertz hazard rate. Correlated frailty models with the Gompertz baseline hazard have been used quite frequently [4, 6, 7, 24]. For that reason and to save space, we investigate only bivariate frailty models that have the Gompertz baseline hazard rate $\mu_0(t) = a \exp(bt)$, and the cumulative baseline hazard function is of the form $M_0(t) = (a/b) \{\exp(bt) - 1\}$ with $a, b > 0$. Denote by θ the vector of all parameters describing the distribution of the i.i.d. frailty vectors (Z_{i1}, Z_{i2}) . Let δ_{ij} be the event indicator for an individual $j(j=1, 2)$ in pair $i(i=1, \dots, n)$. Indicator δ_{ij} is 1 if the individual has experienced the event of interest, and it is 0 otherwise. According to (1), the conditional survival function of the j th individual in the i th pair is

$$S(t|Z_{ij}, X_{ij}) = \exp\{-Z_{ij}M_0(t) \exp(\beta^T X_{ij})\}$$

with $M_0(t)$ denoting the cumulative baseline hazard function. Here and below, S is used as a generic symbol for a survival function. The contribution of the j th individual in the i th pair of the conditional likelihood is given by

$$L_i(\beta, a, b|Z_{ij}) = \{Z_{ij}\mu_0(t_{ij}) \exp(\beta^T X_{ij})\}^{\delta_{ij}} \exp\{-Z_{ij}M_0(t_{ij}) \exp(\beta^T X_{ij})\}$$

where t_{ij} stands for age/time at onset or the censoring time of the individual. Then, assuming the conditional independence of life spans given frailty and integrating out the random effects, we obtain the marginal likelihood function:

$$L(\beta, a, b, \theta) = \prod_{i=1}^n \int \int \prod_{j=1}^2 L_i(\beta, a, b|z_{ij}) f_Z(z_{i1}, z_{i2}, \theta) dz_{i1} dz_{i2}$$

with $t = (t_1, \dots, t_n)$, $t_i = (t_{i1}, t_{i2})$, $\delta = (\delta_1, \dots, \delta_n)$, $\delta_i = (\delta_{i1}, \delta_{i2})$, $X = (X_1, \dots, X_n)$, $X_i = (X_{i1}, X_{i2})$ and $f_Z(\cdot, \cdot | \theta)$ is the p.d.f. of the corresponding frailty distribution. If there are explicit forms of $S(t_{i1}, t_{i2} | X_{i1}, X_{i2})$ ($i = 1, \dots, n$) and their partial derivatives, the likelihood function can be given in the following way:

$$L(\beta, a, b, \theta) = \prod_{i=1}^n \{ \delta_{i1} \delta_{i2} S_{12}(t_{i1}, t_{i2} | X_{i1}, X_{i2}) + \delta_{i1} (1 - \delta_{i2}) S_1(t_{i1}, t_{i2} | X_{i1}, X_{i2}) \\ + (1 - \delta_{i1}) \delta_{i2} S_2(t_{i1}, t_{i2} | X_{i1}, X_{i2}) + (1 - \delta_{i1})(1 - \delta_{i2}) S(t_{i1}, t_{i2} | X_{i1}, X_{i2}) \}$$

with partial derivatives

$$S_j(t_1, t_2 | X_1, X_2) = \frac{\partial S(t_1, t_2 | X_1, X_2)}{\partial t_j} \quad (j = 1, 2), \quad S_{12}(t_1, t_2 | X_1, X_2) = \frac{\partial^2 S(t_1, t_2 | X_1, X_2)}{\partial t_1 \partial t_2}$$

2.2. Correlated gamma frailty model

In the following we drop the covariates because our data example contains no covariates. The gamma distribution $\Gamma(k, \lambda)$ with density $f(t) = \lambda^k t^{k-1} e^{-\lambda t} / \Gamma(k)$ is the most popular frailty distribution. Here $\Gamma(k)$ denotes the Gamma function $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt$. Gamma-distributed frailty yields some useful mathematical results, for example a simple form for the Laplace transform. This implies an explicit form of the likelihood function. To make sure that the model can be identified, we use the parameter restriction $\mathbf{E}Z = 1$, which results in $k = \lambda$ for the gamma distribution. Denoting the variance of the frailty by $\sigma^2 := 1/\lambda$, the marginal survival function is represented by

$$S(t) = \mathbf{L}_Z(M_0(t)) = \{1 + \sigma^2 M_0(t)\}^{-1/\sigma^2}$$

where $M_0(t)$ denotes the cumulative baseline hazard function and $\mathbf{L}_Z$ is the Laplace transform of the random variable Z .

The correlated gamma frailty model [4, 20, 21] has been developed for the analysis of bivariate failure time data, in which two associated random variables are used to characterize the frailty effect for each cluster. To be more specific, let k_0, k_1 be some real positive variables. Set $\lambda = k_0 + k_1$ and let Y_0, Y_1, Y_2 be independent gamma-distributed random variables with $Y_0 \sim \Gamma(k_0, \lambda)$, $Y_1 \sim \Gamma(k_1, \lambda)$, $Y_2 \sim \Gamma(k_1, \lambda)$. Consequently,

$$Z_1 = Y_0 + Y_1 \sim \Gamma(k_0 + k_1, \lambda) = \Gamma(\lambda, \lambda) \\ Z_2 = Y_0 + Y_2 \sim \Gamma(k_0 + k_1, \lambda) = \Gamma(\lambda, \lambda) \quad (2)$$

are the frailties of individuals 1 and 2 in a pair. The survival function is given by

$$S(t_1, t_2) = S(t_1)^{1-\rho} S(t_2)^{1-\rho} \{S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} - 1\}^{-\rho/\sigma^2} \quad (3)$$

where $S(t)$ denotes the marginal univariate survival function, assumed to be equal for both partners and $0 \leq \rho \leq 1$ holds. Furthermore, $\rho = \text{corr}(Z_1, Z_2)$ and $\sigma^2 = \mathbf{V}(Z_i)$ ($i = 1, 2$). The model is symmetric. This is because the example data in the present paper are symmetric twin data. Extensions to non-symmetric situations (e.g. lifetimes of fathers and sons) are straightforward. Obviously, the popular shared gamma frailty model [2] is a special case of (3) when $\rho = 1$.

2.3. Compound Poisson distribution

The compound Poisson distribution was introduced as a frailty distribution in univariate models [25]. An interesting property is that it yields a subgroup of zero frailty, which never experiences the event under study. Despite the fact that the density of the continuous part is an infinite series, which has to be calculated numerically, the distribution is mathematically convenient because of the simple Laplace transform. It may also be seen as a natural choice. The distribution can be constructed as the sum of a Poisson-distributed number of independent and identical gamma distributed random variables.

$$Z = \begin{cases} V_1 + V_2 + \dots + V_N & \text{if } N > 0 \\ 0 & \text{if } N = 0 \end{cases} \quad (4)$$

where N is a Poisson-distributed random variable with expectation v and Laplace transform $\mathbf{L}_N(s) = \exp(-v + v e^{-s})$, while $V_1, V_2, \dots$ are independent and identically gamma distributed with $V_i \sim \Gamma(k, \lambda)$, and Laplace transform $\mathbf{L}_V(s) = (1 + s/\lambda)^{-k}$. An appealing interpretation of this model is that each individual suffers from several hits causing damage. The effect of these hits cumulates over time and increases individual frailty. It should be noted that this process takes place before follow-up starts, maybe in early life. With the beginning of follow-up frailties are assumed to be fixed. For models with time-dependent frailty see [26]. Following [25] we assume independence between the random variables N and $V_1, V_2, \dots$ and get

$$\mathbf{L}(s) = \mathbf{E} \exp(-sZ) = \mathbf{E} \exp\{-s(V_1 + \dots + V_N)\} = \mathbf{E} \mathbf{L}_V(s)^N = \mathbf{L}_N[-\ln\{\mathbf{L}_V(s)\}]$$

Inserting the previous expressions gives the following Laplace transform of Z :

$$\mathbf{L}_Z(s) = \exp \left\{ -v + v \left(1 + \frac{s}{\lambda} \right)^{-k} \right\}$$

In the following, we use another parameterization

$$v = -\frac{k\lambda^\gamma}{\gamma}, \quad \lambda = \lambda, \quad k = -\gamma \quad (5)$$

and write the Laplace transform of the compound Poisson distribution as

$$\mathbf{L}(s) = \exp \left[-\frac{k}{\gamma} \{ (\lambda + s)^\gamma - \lambda^\gamma \} \right] \quad (6)$$

Expectation and variance of a compound Poisson-distributed random variable Z are

$$\mathbf{E}Z = k\lambda^{\gamma-1} \quad \text{and} \quad \mathbf{V}(Z) = k(1-\gamma)\lambda^{\gamma-2} \quad (7)$$

The Laplace transform given above implies the marginal survival and hazard function in case of a compound Poisson frailty model:

$$S(t) = \exp \left(-\frac{k}{\gamma} [\{ \lambda + M_0(t) \}^\gamma - \lambda^\gamma] \right)$$

and

$$\mu(t) = k\mu_0(t) \{ \lambda + M_0(t) \}^{\gamma-1}$$

Using the constraint $\mathbf{E}Z=1$ and relation $\sigma^2 = (1-\gamma)/\lambda$ (see (7)) it holds

$$S(t) = \exp \left(-\frac{1-\gamma}{\gamma\sigma^2} \left[\left\{ 1 + \frac{\sigma^2}{1-\gamma} M_0(t) \right\}^\gamma - 1 \right] \right)$$

and

$$\mu(t) = \mu_0(t) \left\{ 1 + \frac{\sigma^2}{1-\gamma} M_0(t) \right\}^{\gamma-1}$$

By assuming $\mathbf{E}Z=1$, the positive stable frailty distribution is not included in the family considered here. It should be noted that the integral of $\mu(t)$ over $[0, \infty)$ is finite when $\gamma < 0$. Consequently, the survival function is incomplete because a fraction of individuals have zero frailty, who could never experience the event under study.

2.4. Correlated compound Poisson frailty model

The notation $cP(\gamma, k, \lambda)$ is used for a compound Poisson distribution, given by (6). To introduce the bivariate model, let Y_0, Y_1, Y_2 be independent compound Poisson-distributed random variables with $Y_0 \sim cP(\gamma, k_0, \lambda)$, $Y_1 \sim cP(\gamma, k_1, \lambda)$, $Y_2 \sim cP(\gamma, k_1, \lambda)$. Consequently, using the same idea of additive composition as in the correlated gamma frailty model,

$$\begin{aligned} Z_1 &= Y_0 + Y_1 \sim cP(\gamma, k_0 + k_1, \lambda) \\ Z_2 &= Y_0 + Y_2 \sim cP(\gamma, k_0 + k_1, \lambda) \end{aligned} \quad (8)$$

are the frailties of individuals one and two in a pair. The bivariate survival function of this model can explicitly be calculated (see Appendix) and takes the following simple form:

$$S(t_1, t_2) = S(t_1)^{1-\rho} S(t_2)^{1-\rho} e^{\rho(1-\gamma)/(\gamma\sigma^2)(1 - \{1 - (\gamma\sigma^2/(1-\gamma)) \ln S(t_1)\}^{1/\gamma} + \{1 - (\gamma\sigma^2/(1-\gamma)) \ln S(t_2)\}^{1/\gamma} - 1)^\gamma} \quad (9)$$

with $S(t)$ denoting the marginal univariate survival function, assumed to be equal for both twin partners. Furthermore, it holds that $\rho = \text{corr}(Z_1, Z_2)$ with $0 \leq \rho \leq 1$ and $\sigma^2 = \mathbf{V}(Z_i)$, $i=1, 2$. The size of the susceptible fraction (in the case of $\gamma < 0$) is given by $1 - \exp((1-\gamma)/(\gamma\sigma^2))$. This model has an interesting feature: it is possible to have two related individuals where one has zero frailty (is non-susceptible) and the other has a positive frailty (is susceptible). This is an aspect that makes the model fit better than a shared frailty model. Of course, the shared compound Poisson frailty model considered in [5] is a special case of the proposed correlated compound Poisson frailty model when $\rho=1$. By construction, because of the additive structure of the frailty variables, the correlation has to be non-negative, so that the model cannot handle negative dependencies.

Up to now γ is negative by construction (see (5)). It turns out that the model makes sense up to the case $\gamma \leq 1$. The survival function (9) continues to be a survival function as $\gamma \in [0, 1]$. Consequently, parameter γ divides the class of distributions in two major subfamilies: For $\gamma \geq 0$, the model was already considered in [11]. The inverse Gaussian model is given by $\gamma=0.5$. The extension

Table I. Parameter estimation in the simulation study. ρ_{MZ} and ρ_{DZ} are the correlations between frailties of monozygotic and dizygotic twins, respectively.		Mean (s.e.)	Mean (s.e.)	Mean (s.e.)	Mean (s.e.)	Mean (s.e.)
γ (true)	True value	0.000	−0.050	−0.100	−0.300	−0.600
γ		−0.012 (0.046)	−0.063 (0.045)	−0.116 (0.066)	−0.326 (0.105)	−0.671 (0.202)
σ	7.000	6.861 (0.730)	6.876 (0.676)	6.910 (0.718)	6.942 (0.615)	6.894 (0.569)
ρ_{MZ}	0.120	0.123 (0.054)	0.121 (0.055)	0.122 (0.054)	0.122 (0.054)	0.122 (0.064)
ρ_{DZ}	0.100	0.102 (0.050)	0.101 (0.053)	0.100 (0.051)	0.101 (0.054)	0.101 (0.055)
a	5.00e−7	6.80e−7 (4.75e−7)	6.80e−7 (4.10e−7)	6.80e−7 (4.76e−7)	6.40e−7 (4.28e−7)	7.00e−7 (4.13e−7)
b	0.150	0.148 (0.013)	0.147 (0.013)	0.148 (0.014)	0.149 (0.013)	0.147 (0.013)

to $\gamma < 0$ in the univariate case was suggested in [25] and yields the compound Poisson distribution. The gamma distribution is given by $\gamma = 0$.

Parameter values of $\gamma < 0$ imply the existence of a non-susceptible fraction in the population. In Section 4 the model is applied to breast cancer data of 5857 Swedish twin pairs to estimate the size of the susceptible fraction and the frailty parameters. The main question here is whether a correlated compound Poisson frailty model with cure fraction ($\gamma < 0$) shows a better fit to the data than a model without cure fraction ($\gamma \geq 0$) or not, which would speak in favor of the existence of a fraction of women, who are not susceptible to breast cancer either because of genetic or environmental predisposition. To check the behavior of the parameter estimates in a situation similar to the breast cancer example in Swedish twins, a simulation study was conducted in the following section.

3. Simulation study

All simulations involve generating correlated compound Poisson distributed frailties by formula (8), bivariate life and censoring times. We will try to mimic the characteristics of the Swedish twin data, which we analyze in the following section. Five thousand twin pairs are simulated:

- Let N be a Poisson-distributed random variable.
- Generate frailties using sums of N -independent gamma-distributed random variables.
- Generate lifetimes given the frailties by $S(t|Z) = \exp[-Z(a/b)\{\exp(bt) - 1\}]$.
- Birth years are generated by using a uniform distribution on the interval [1886,1925].
- Censored lifetimes are generated by using the year 2000 as the end of the study.

One thousand data sets were simulated for each parameter set. Censoring in the simulated data sets varied between 94 and 97 per cent. Because we are especially interested in the parameter γ and due to the long time that the simulations require, γ is the only parameter that is varied. The different true values of γ are given in the second line of Table I. The mean parameter estimates of the model are shown in comparison with the true values used for simulation in the following lines. It turns out that the parameter γ can be estimated with acceptable precision. There seems to be a slight underestimation of this parameter. There appears to be nearly no bias in the estimates of all other parameters, and the overall performance is almost accurate. Two different correlations are used in the model, ρ_{MZ} for MZ twins and ρ_{DZ} for DZ twins, respectively.

4. Swedish breast cancer twin data

The suggested model was applied to breast cancer incidence data of identical (MZ) and fraternal (DZ) female twins born 1886–1925 provided by the Swedish Twin Registry. This data set contains records of 5857 female twin pairs with both partners being alive in 1959–1961 (old cohort). Consequently, data are left truncated that were adjusted for in the analysis. Individuals were followed up from 1959/1961 to 27 October 2000. Altogether, we have 2003 monozygotic and 3854 dizygotic twin pairs, and 715 cases of breast cancer were identified during the follow-up. This results in 94 per cent censoring in the data. Owing to the design, age at onset of breast cancer ranges from 36 to 93 years. In the following analysis age is used as a time scale. More detailed information about the construction of the Swedish Twin Register can be found in [27].

Table II. Analysis of time to onset of breast cancer in 5857 Swedish twin pairs.			
	Gamma frailty	Inverse Gaussian frailty	Compound Poisson frailty
γ	0	0.5	-0.617 (0.908)
σ	5.726 (0.679)	4.178 (1.675)	3.993 (1.038)
ρ_{MZ}	0.154 (0.052)	0.344 (0.131)	0.154 (0.053)
ρ_{DZ}	0.126 (0.040)	0.300 (0.107)	0.130 (0.041)
Susceptible	1.000	1.000	0.152 (0.059)
a	$1.32e-5$ ($1.05e-5$)	$1.821e-4$ ($0.57e-4$)	$2.53e-5$ ($1.87e-05$)
b	0.099 (0.016)	0.047 (0.008)	0.080 (0.016)
Likelihood	-5122.3155	-5130.5856	-5121.2319

We consider three correlated frailty models. In the first case the frailty follows a gamma distribution ($\gamma=0$), in the second an inverse Gaussian distribution ($\gamma=0.5$) is assumed and in the third one the general compound Poisson/PVF distribution ($-\infty<\gamma\leq 1$) is used. The results are given in Table II. The estimate of parameter $\gamma=-0.617$ is negative, but with a large standard error (0.909). This indicates the existence of a subpopulation of individuals who are non-susceptible to breast cancer. Heterogeneity (σ^2) seems to be large and the difference between the intrapair correlations of MZ and DZ twins is small.

5. Discussion

The first model applied to the Swedish breast cancer twin data is the correlated gamma frailty model ($\gamma=0$). Parameter σ^2 is a measure of heterogeneity, which is large in this data, and all individuals are assumed to be susceptible to breast cancer.

In the inverse Gaussian frailty model ($\gamma=0.5$), parameters show smaller heterogeneity and larger correlations compared with the gamma case. All women are assumed to be susceptible.

The most interesting parameter in the compound Poisson model is parameter γ , which is negative ($\gamma=-0.617$) and indicates the existence of a fraction of individuals who are non-susceptible to breast cancer. The size of the susceptible fraction (in a univariate meaning) can be calculated from the parameters γ and σ^2 and is around 0.15—with a 95 per cent confidence interval not including one. This means that the new model shows a better fit to the data than the correlated gamma frailty model without a non-susceptible fraction. However, the estimate of the size of a susceptible fraction (due to breast cancer) is in a similar range compared with the estimate 0.22 found in a different study population [9] and the same value 0.22 obtained by applying a correlated gamma frailty cure model to the same data [6], especially with respect to the standard error. The compound Poisson model fits the data better than the two submodels, but at the cost of an additional parameter.

Additionally, the estimate of the size of the susceptible fraction in the compound Poisson frailty model is not far from the range of the figures obtained in [16] for different combinations of four risk factors. The authors found that if none of the risk factors is present, the susceptible fraction is around 0.015, and if all risk factors are present, the estimate increases to 0.272.

The size of the susceptible fraction has to be compared with the overall lifetime risk of breast cancer, which is around 8–12 per cent in current western populations [28–31]. The twin population considered is born between 1886 and 1925 when the lifetime risk for breast cancer was lower because of competing causes of death like infections, which are less important today. For comparison, the lifetime risk of women with a family history of breast cancer is around 40 per cent and for women with a BRCA1 mutation it is around 80 per cent [32].

In all models correlations in monozygotic pairs are higher than in dizygotic pairs, but the differences are small and not significant (for example $\rho=0.60$ in the correlated compound Poisson model). This is in line with the well-known fact that the influence of genetic factors on susceptibility to breast cancer is small [33]. Environmental factors contribute to the similarity between the twins regarding the age at the onset of breast cancer, because the correlations between frailties of twin partners are significantly different from zero.

The newly introduced correlated compound Poisson frailty model offers a very elegant approach to integrate the concept of cure models into multivariate frailty modelling. The likelihood function is explicitly available in a very simple form, which allows traditional maximum likelihood parameter estimation. Important frailty models like the correlated and shared gamma ($\gamma=0$) and inverse Gaussian model ($\gamma=0.5$) are included in this model family and provide great flexibility to the model. Because of the extension to negative values of γ , the gamma distribution ($\gamma=0$) as the most popular frailty distribution is no longer on the border of the parameter space. Consequently, standard tests can be applied to check hypotheses about the frailty distribution (for example $H_0:\gamma=0$ versus $\gamma\neq 0$). Simulation studies show a good performance of the parameter estimates in this model with nearly no bias. A disadvantage of the model is that it is not able to handle negative dependencies.

One question is related to the possible implications of the assumption of a Gompertz baseline hazard. An analysis with a Weibull baseline hazard reveals similar results (not shown). Of great interest would be a non-parametric version of the correlated compound Poisson frailty model, where the baseline hazard functions are not specified. But such models raise questions about the identifiability of the parameters and will be a part of future research.

Here we only consider the bivariate case. Extensions to arbitrary cluster size are possible, but require different estimation strategies, because the likelihood of a cluster needs derivatives of order p of the survival function where p is the number of events in the cluster.

Cure models suffer from an inherent identifiability problem with right-censored observations: The event under study has not occurred either because the person is insusceptible or the person is susceptible, but follow-up was not long enough to observe the event. The identifiability problem is growing with increasing censoring, but is reduced by the parametric modelling of the baseline hazard. As a consequence, the standard error of the estimate for parameter γ is relatively large in the simulations as well as in the example data. In cure models with fixed censoring times (caused by ending the study), censoring is no longer non-informative even in cases where the censoring and the survival times are independent. The proportion of censored observations contains important information about the parameters in the model. Of course, the size of the susceptible fraction is at least the proportion of non-censored observations. The large standard error of the parameter γ in the correlated compound Poisson frailty model points to this identification problem. It is hard to estimate latent trait parameters even with large sample size as in the present example. Maybe this problem becomes less pronounced in cases with observed covariates. This hypothesis needs further investigations.

Appendix A

A.1. Correlated compound Poisson frailty model

The following symmetry relations $\mathbf{E}Z_1 = \mathbf{E}Z_2 = 1, \mathbf{V}(Z_1) = \mathbf{V}(Z_2) = \sigma^2$ are assumed. This implies (see (7)) $(k_0 + k_1)\lambda^{\gamma-1} = 1$ and $(k_0 + k_1)(1 - \gamma)\lambda^{\gamma-2} = \sigma^2$. Consequently, $(k_0 + k_1)\lambda^{\gamma-2} = 1/\lambda$ and $\lambda = (1 - \gamma)/\sigma^2$, which results in

$$(k_0 + k_1)\lambda^\gamma = \lambda = \frac{1 - \gamma}{\sigma^2} \quad (\text{A1})$$

It holds that $\mathbf{E}Y_0^2 = \mathbf{V}(Y_0) + (\mathbf{E}Y_0)^2 = k_0(1 - \gamma)\lambda^{\gamma-2} + (k_0\lambda^{\gamma-1})^2$ and

$$\begin{aligned} \mathbf{E}Z_1 Z_2 &= \mathbf{E}(Y_0 + Y_1)(Y_0 + Y_2) = \mathbf{E}(Y_0^2 + Y_0 Y_1 + Y_0 Y_2 + Y_1 Y_2) \\ &= k_0(1 - \gamma)\lambda^{\gamma-2} + k_0^2 \lambda^{2\gamma-2} + k_0 k_1 \lambda^{2\gamma-2} + k_0 k_1 \lambda^{2\gamma-2} + k_1^2 \lambda^{2\gamma-2} \\ &= k_0(1 - \gamma)\lambda^{\gamma-2} + (k_0 + k_1)^2 \lambda^{2\gamma-2} \\ &= k_0(1 - \gamma)\lambda^{\gamma-2} + 1 \\ \text{cov}(Z_1, Z_2) &= \mathbf{E}Z_1 Z_2 - \mathbf{E}Z_1 \mathbf{E}Z_2 = k_0(1 - \gamma)\lambda^{\gamma-2} \end{aligned}$$

This leads to the correlation

$$\rho = \frac{\text{cov}(Z_1, Z_2)}{\sqrt{\mathbf{V}(Z_1)\mathbf{V}(Z_2)}} = \frac{k_0(1 - \gamma)\lambda^{\gamma-2}}{(k_0 + k_1)(1 - \gamma)\lambda^{\gamma-2}} = \frac{k_0}{k_0 + k_1} \quad (\text{A2})$$

Consequently, because of (A1) and (A2)

$$k_0 \lambda^\gamma = \frac{k_0}{k_0 + k_1} (k_0 + k_1) \lambda^\gamma = \rho \frac{1 - \gamma}{\sigma^2} \quad (\text{A3})$$

Now we can derive the unconditional model, applying the Laplace transform of compound Poisson-distributed random variables (6):

$$\begin{aligned} S(t_1, t_2) &= \mathbf{E}S(t_1 | Z_1)S(t_2 | Z_2) \\ &= \mathbf{E}\exp\{-Z_1 M_0(t_1)\} \exp\{-Z_2 M_0(t_2)\} \\ &= \mathbf{E}\exp\{-Y_0(M_0(t_1) + M_0(t_2)) - Y_1 M_0(t_1) - Y_2 M_0(t_2)\} \\ &= \exp\left(-\frac{k_0}{\gamma} [\{\lambda + M_0(t_1) + M_0(t_2)\}^\gamma - \lambda^\gamma]\right) \exp\left(-\frac{k_1}{\gamma} [\{\lambda + M_0(t_1)\}^\gamma - \lambda^\gamma]\right) \\ &\quad \times \exp\left(-\frac{k_1}{\gamma} [\{\lambda + M_0(t_2)\}^\gamma - \lambda^\gamma]\right) \end{aligned} \quad (\text{A4})$$

The three terms are considered in detail. In the univariate case it holds that

$$S(t) = \exp\left(-\frac{k_0 + k_1}{\gamma} [\{\lambda + M_0(t)\}^\gamma - \lambda^\gamma]\right) \quad (\text{A5})$$

which implies

$$\lambda + M_0(t) = \left\{ \lambda^\gamma - \frac{\gamma}{k_0 + k_1} \ln S(t) \right\}^{1/\gamma} \quad (\text{A6})$$

Hence, using (A2) and (A5)

$$\begin{aligned} \exp\left(-\frac{k_1}{\gamma} [\{\lambda + M_0(t)\}^\gamma - \lambda^\gamma]\right) &= \exp\left(-\frac{k_1}{k_0 + k_1} \frac{k_0 + k_1}{\gamma} [\{\lambda + M_0(t)\}^\gamma - \lambda^\gamma]\right) \\ &= \exp\left(-(1-\rho) \frac{k_0 + k_1}{\gamma} [\{\lambda + M_0(t)\}^\gamma - \lambda^\gamma]\right) \\ &= \exp\left(-\frac{k_0 + k_1}{\gamma} [\{\lambda + M_0(t)\}^\gamma - \lambda^\gamma]\right)^{1-\rho} = S(t)^{1-\rho} \end{aligned}$$

The first term in (A4) holds because of (A6)

$$\begin{aligned} &\exp\left(-\frac{k_0}{\gamma} [\{\lambda + M_0(t_1) + M_0(t_2)\}^\gamma - \lambda^\gamma]\right) \\ &= \exp\left(-\frac{k_0}{\gamma} [\{\lambda + M_0(t_1) + \lambda + M_0(t_2) - \lambda\}^\gamma - \lambda^\gamma]\right) \\ &= \exp\left(\frac{k_0}{\gamma} \left(\left[\left\{ \lambda^\gamma - \frac{\gamma}{k_0 + k_1} \ln S(t_1) \right\}^{1/\gamma} + \left\{ \lambda^\gamma - \frac{\gamma}{k_0 + k_1} \ln S(t_2) \right\}^{1/\gamma} - \lambda \right]^\gamma - \lambda^\gamma \right) \right) \\ &= \exp\left(\frac{k_0 \lambda^\gamma}{\gamma} \left(1 - \left[\left\{ 1 - \frac{\gamma}{(k_0 + k_1) \lambda^\gamma} \ln S(t_1) \right\}^{1/\gamma} + \left\{ 1 - \frac{\gamma}{(k_0 + k_1) \lambda^\gamma} \ln S(t_2) \right\}^{1/\gamma} - 1 \right]^\gamma \right) \right) \end{aligned}$$

which results because of (A1) and (A3) in the representation (9) of the correlated compound Poisson frailty model.

Acknowledgements

The authors wish to thank the two reviewers and the associate editor for valuable comments and the Swedish Twin Register for providing the data. A. Wienke and A. I. Yashin would like to thank the Max Planck Institute for Demographic Research in Rostock (Germany) for providing excellent research facilities during several stays at the institute. A. I. Yashin's research was also supported by the National Institute on Aging grants R01AG027019, R01AG028259 and 5P01AG008761. A. Wienke was supported by a grant of the German Ministry for Education and Research (NBL3-programme, 01ZZ0404) and a grant of the German Research Council (DFG WI 3288/1-1).

References

1. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989; **84**:1065–1073.
2. Clayton D. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**:141–151. DOI: 10.1093/biomet/65.1.141.
3. Oakes D. A concordance test for independence in the presence of censoring. *Biometrics* 1982; **38**:451–455.
4. Yashin AI, Vaupel JW, Iachine IA. Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* 1995; **5**:145–159.
5. Hougaard P. *Analysis of Multivariate Survival Data*. Springer: New York, 2000.
6. Wienke A, Lichtenstein P, Yashin AI. A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. *Biometrics* 2003; **59**:1178–1183. DOI: 10.1111/j.0006-341X.2003.00135.x.
7. Wienke A, Holm N, Christensen K, Skytthe A, Vaupel J, Yashin AI. The heritability of cause-specific mortality: a correlated gamma-frailty model applied to mortality due to respiratory diseases in Danish twins born 1870–1930. *Statistics in Medicine* 2003; **22**:3873–3887. DOI: 10.1002/sim.1669.
8. Duchateau L, Janssen P. *The Frailty Model*. Springer: New York, 2008. DOI: 10.1007/978-0-387-72835-3.
9. Chatterjee N, Shih J. A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics* 2001; **57**:779–786. DOI: 10.1111/j.0006-341X.2001.00779.x.
10. Moger TA, Aalen OO. A distribution for multivariate frailty based on the compound Poisson distribution with random scale. *Lifetime Data Analysis* 2005; **11**:41–59. DOI: 10.1007/s10985-004-5639-z.
11. Yashin AI, Begun AZ, Iachine IA. Genetic factors in susceptibility to death: a comparative analysis of bivariate survival models. *Journal of Epidemiology and Biostatistics* 1999; **4**:53–60.

12. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 1986; **73**:387–396. DOI: 10.1093/biomet/73.2.387.
13. Zahl PH, Harris JR. Cancer incidence for Swedish twins studied by means of bivariate frailty models. *Genetic Epidemiology* 2000; **19**:354–365. DOI: 10.1002/1098-2272(200012)19:4<354::AID-GEPI7>3.0.CO;2-N.
14. Moger TA, Aalen OO, Halvorsen TO, Storm HH, Tretli S. Frailty modelling of testicular cancer incidence using Scandinavian data. *Biostatistics* 2004; **5**:1–14. DOI: 10.1093/biostatistics/5.1.1.
15. Moger TA, Aalen OO, Heimdal K, Gjessing HK. Analysis of testicular cancer data using a frailty model with familial dependence. *Statistics in Medicine* 2004; **23**:617–632. DOI: 10.1002/sim.1614.
16. Farewell VT, Math B, Math M. The combined effect of breast cancer risk factors. *Cancer* 1977; **40**:931–936.
17. Peto J, Mack TM. High constant incidence in twins and other relatives of women with breast cancer. *Nature Genetics* 2000; **26**:411–414. DOI: 10.1038/82533.
18. Price DL, Manatunga AK. Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine* 2001; **20**:1515–1527. DOI: 10.1002/sim.1832.
19. Haukka J, Suvisaari J, Lönngqvist J. Increasing age does not decrease risk of schizophrenia up to age 40. *Schizophrenia Research* 2003; **61**:105–110. DOI: 10.1016/S0920-9964(02)00233-5.
20. Pickles A, Crouchley R. Generalizations is applications of frailty models for survival and event data. *Statistical Methods in Medical Research* 2004; **3**:263–278. DOI: 10.1177/096228029400300305.
21. Petersen JH. An additive frailty model for correlated lifetimes. *Biometrics* 1998; **54**:646–661.
22. Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 2000; **56**:1016–1022. DOI: 10.1111/j.0006-341X.2000.01016.x.
23. Ripatti S, Larsen K, Palmgren J. Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. *Lifetime Data Analysis* 2002; **8**:349–360. DOI: 10.1023/A:1020566821163.
24. Iachine IA, Holm NV, Harris JR, Begun AZ, Iachine MK, Laitinen M, Kaprio J, Yashin AI. How heritable is individual susceptibility to death? The results of an analysis of survival data on Danish, Swedish and Finnish twins. *Twin Research* 1998; **1**:196–205. DOI: 10.1375/twin.1.4.196.
25. Aalen OO. Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability* 1992; **4**:951–972.
26. Gjessing HK, Aalen OO, Hjort NL. Frailty models based on Lévy processes. *Advances in Applied Probability* 2003; **35**:532–550.
27. Lichtenstein P, de Faire U, Floderus B, Svartengren M, Svedberg P, Pedersen NL. The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine* 2002; **252**:184–205. DOI: 10.1046/j.1365-2796.2002.01029.x.
28. Harris JR, Lippman ME, Veronesi U, Willett W. Breast cancer (i). *New England Journal of Medicine* 1992; **327**:319–328.
29. Feuer EJ, Wun LM, Boring CC, Flanders WD, Timmel MJ, Tong T. The lifetime risk of developing breast cancer. *Journal of the National Cancer Institute* 1993; **85**:892–897.
30. Rosenthal TC, Puck SM. Screening for genetic risk of breast cancer. *American Family Physician* 1999; **59**:99–104.
31. Ries LAG, Kosary CL, Hankey BF (eds). *SEER Cancer Statistics Review 1973–1999*. National Cancer Institute: Bethesda, MD, 1999.
32. Ponder B. Genetic testing for cancer risk. *Science* 1997; **278**:1050–1054.
33. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. Environmental is heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine* 2000; **343**:78–85. DOI: 10.1056/NEJM200007133430201.