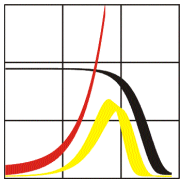

Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR WORKING PAPER WP 2003-025
JULY 2003

A Bayesian correlated frailty model applied to Swedish breast cancer data

Isabella Locatelli (isabella.locatelli@uni-bocconi.it)
Paul Lichtenstein (paul.lichtenstein@mep.ki.se)
Anatoli I. Yashin (yashin@demogr.mpg.de)

Abstract

Frailty was first introduced in survival analysis to control for unobserved heterogeneity. Frailty models represent an extension of the proportional hazards model in which both the frailty term and the covariate effects are assumed to act multiplicatively on the baseline hazard. Multivariate frailty models were then developed with the aim to introduce mutual dependence between the life spans of related individuals.

In this study we use a correlated log-normal frailty model in order to analyse breast cancer data from the Swedish Twin Registry. An estimate of the narrow sense heritability for the individual susceptibility towards breast cancer is given via the application of three genetic models.

We solve the inferential problem in a Bayesian framework and the numerical work is carried out using MCMC methods. Limitations and possible extensions of the model are discussed.

1. Introduction

Frailty was first introduced in survival analysis in order to assess unobserved heterogeneity (Vaupel et al. 1979). Frailty models represent an extension of the proportional hazards model (Cox 1972) in which both the frailty term and the covariate effects are assumed to act multiplicatively on the baseline hazard. The term including covariates allows for observed heterogeneity, while the frailty term captures that part of the individual heterogeneity that refers to unobserved risk factors. Individuals differ substantially in their susceptibility towards mortality (overall or cause specific mortality) and it is often impossible to include in the model all the relevant covariates. More frail individuals die earlier than the stronger ones and this leads to a systematic selection effect over time. When unobserved heterogeneity is introduced in the model, it is possible to identify the influence of selection on the observed hazard and to analyse the individual risk of mortality at different frailty levels (Vaupel and Yashin 1985).

In this work, we are dealing with multivariate frailty models, which were created with the aim to assess mutual dependence between the lifespans of related individuals. The first approach developed in the literature, and still much employed, is based on the concept of 'shared frailty' (Clayton 1978, Oakes 1982, Hougaard 1984, Vaupel et al. 1992, Sahu et al. 1997). Groups of individuals (family, litter, clinic or recurrent events from the same individual) share the

same frailty and their durations are assumed to be conditionally independent, given the frailty variable.

Shared frailty models are useful when we want to explain correlations within groups, but they have some limitations. First, they deal with a definition of frailty, which is not consistent with the definition given in the univariate framework (Vaupel et al. 1979). The frailty term in fact represents a part of individual frailty, only capturing the components, which are 'shared' by all individuals within a cluster. Second, they force all unobserved risk factors to be the same within a cluster, which is not always reasonable. For example, when one deals with pairs of twins there is no reason to assume that both partners in a pair share the same unobserved heterogeneity. Third, shared frailty will only induce positive association within a group. However, in some situations it could be useful to allow also for a negative correlation between lifespans within the groups (Xue and Ding 1999).

To overcome these limitations, a 'correlated frailty' approach has been developed. The importance of taking into account the dependence between heterogeneity variables describing different processes related to the same individual was first emphasised by Butler et al. (1986) and Lillard (1993). Yashin et al. (1995) introduced a correlated gamma frailty model to describe bivariate survival data, focusing their attention to the analysis of pairs of related individuals, for example twins. The correlated frailty assumption is more flexible than the shared frailty one in the sense that the model includes different - but correlated - frailties for the two individuals in a pair. It is of interest to estimate the correlation coefficient between these two variables, that is the degree of dependence between frailties in each pair. As in the shared frailty model, the two lifespans in a pair are assumed to be conditionally independent given the frailties.

In the correlated frailty model, unobserved risk factors are not forced to be the same in each group, the frailty term represents the entire susceptibility towards death exactly as in the univariate framework, and the possibility of a negative association between survival times is taken into account. In addition, the correlated frailty concept allows for the integration of survival data for related individuals with different levels of relationship, for example identical (monozygotic) and fraternal (dizygotic) twins, and to merge traditional approaches of quantitative genetics and epidemiology with survival analysis methods (Yashin and Iachine 1995, 97).

Two important assumptions in frailty models are related to the shape of the underlying hazard and the distribution of the frailty variables.

Shared and correlated frailty models have been estimated both parametrically and semi-parametrically. The most adopted parametrical hypothesis is the Gompertz baseline hazard (Vaupel et al. 1992, Iachine et al. 1998, Wienke et al. 2001) but other shapes are also possible, for example Weibull (Sahu et al. 1997) or exponential (Xue and Ding 1999). Yashin and Iachine (1994) derived a semiparametric representation for the correlated gamma frailty model, which opened new opportunities for the statistical analysis of bivariate data. This representation allows to estimate the model without making assumptions about the shape of the baseline hazard. The semiparametric approach was also adopted in a Bayesian framework to estimate different shared frailty models by Clayton (1991) and Spiegelhalter et al. (1996), among others.

Every distribution of a positive random variable can be adopted to model frailty. The gamma distribution has been widely applied in the literature (Clayton 1978, Vaupel et al. 1979, Oakes 1982, Yashin and Iachine 1994, Hougaard 2000, Wienke et al. 2001). The gamma choice is convenient from a mathematical point of view, because of the simplicity of the Laplace transform, which allows for the use of traditional maximum likelihood procedures in parameter estimation. Another possibility is to assume that frailty is log-normal distributed (Korsgaard et al. 1998, Spiegelhalter et al. 1996, Xue and Ding 1999, Ripatti and Palmgren 2000). The log-normal approach is much more flexible than the gamma model in creating correlated but different frailties as required in the case of the correlated frailty model. Unfortunately, with a log-normal assumption it is impossible to derive the marginal likelihood function in an explicit form and parameter estimation has to be performed with the help of more sophisticated estimation strategies, such as numerical methods of integration or Bayesian MCMC methods (see Section 3).

In the present study, we work with correlated frailty models. Section 2 provides a general description of the theory of correlated frailty models. In Section 3, the estimation procedure is presented. An interdisciplinary approach based on quantitative genetics models is described in Section 4. Sections 5 and 6 show results of the analysis of Swedish breast cancer data. Some comments and suggestions for further research can be found in Section 7.

2. The model description

We assume that, given some bivariate observations, for example life spans of twins or age at the onset of some disease in pairs of related individuals, the hazard of individual j $(j = 1, 2)$ in the pair i $(i = 1, ..., n)$ takes the form:

$$\mu(x, Z_{ij}) = Z_{ij}\mu_0(x) \tag{1}$$

where $\mu_0(x)$ denotes some baseline hazard function and Z_{ij} are unobserved random effects or frailties. In this study, we adopt a Gompertz baseline hazard ($\mu_0(x) = ae^{bx}$). We are not taking into account covariate effects.

Let (X_{i1}, X_{i2}) be the vector of life spans for the two individuals from the pair i $(i = 1, ..., n)$. We are assuming that X_{i1} and X_{i2} are conditionally independent given the frailties Z_{i1} and Z_{i2}:

$$X_{i1}|Z_{i1}, Z_{i2} \perp X_{i2}|Z_{i1}, Z_{i2}. \tag{2}$$

The conditional likelihood of the model is given by:

$$L(x|Z) = \prod_{i=1}^{n} f_{X_{i1},X_{i2}|Z_{i1},Z_{i2}}(x_{i1}, x_{i2}|z_{i1}, z_{i2}) \tag{3}$$

where $x = (x_1, ..., x_n)$, $x_i = (x_{i1}, x_{i2})$; $Z = (Z_1, ..., Z_n)$, $Z_i = (Z_{i1}, Z_{i2})$ and $f_{X_{i1},X_{i2}|Z_{i1},Z_{i2}}$ represents the bivariate conditional density of the life spans for the pair i. The conditional independence of the life spans given the frailties (2) allows us to rewrite (3) as follows:

$$L(x|Z) = \prod_{i=1}^{n}\prod_{j=1}^{2} f_{X_{ij}|Z_{ij}}(x_{ij}|z_{ij}) \tag{4}$$

where now we deal with the univariate densities $f_{X_{ij}|Z_{ij}}$ $(j = 1, 2)$.

Given the relations:

$$\begin{aligned} f_{X|Z}(x|z) &= \mu(x, Z) S_{X|Z}(x|z) \\ S_{X|Z}(x|z) &= \exp(-ZH_0(x)), \end{aligned} \tag{5}$$

where $S_{X|Z}$ is the conditional survival function and $H_0(x) = \int_0^x \mu_0(u)\,du$ represents the cumulative baseline hazard function, we obtain the following expression for the conditional likelihood:

$$L(x, \delta|Z) = \prod_{i=1}^{n} \prod_{j=1}^{2} \left[Z_{ij}\mu_0(x_{ij})\right]^{\delta_{ij}} \exp\left(-Z_{ij}H_0(x_{ij})\right). \tag{6}$$

where $\delta = (\delta_1, ..., .\delta_n)$, $\delta_i = (\delta_{i1}, \delta_{i2})$, with δ_{ij} representing the censoring indicator for the individual j $(j = 1, 2)$ in the pair i $(i = 1, ..., n)$.

Integrating out the random effects, we obtain the marginal likelihood function:

$$L(x, \delta) = \prod_{i=1}^{n} \int\int \prod_{j=1}^{2} \left[z_{ij}\mu_0(x_{ij})\right]^{\delta_{ij}} \exp\left(-z_{ij}H_0(x_{ij})\right) f_{Z_{i1},Z_{i2}}(z_{i1}, z_{i2})\,dz_{i1}dz_{i2} \tag{7}$$

where $f_{Z_{i1},Z_{i2}}$ represents the joint density function of the vector of frailties (Z_{i1}, Z_{i2}).

To complete the model, it is necessary to make assumptions about the form of $f_{Z_{i1},Z_{i2}}$. In this study, the vector of frailties is assumed to follow a log-normal distribution. This one is adopted because of its large flexibility in multivariate modelling, especially when we are interested in introducing a correlation between frailties, as in the case of the correlated frailty model.

For identifiability reasons, we have to make a restriction on the parameters of the frailty distribution. Following the usual definition of frailty used in demography (Clayton 1978, Vaupel et al. 1979), the expected value of frailty is constrained to be equal to one ($E(Z_{ij}) = 1$, for $i = 1, ..., n$ and $j = 1, 2$). In that way, one is assuming that the hazard function of a 'standard' individual corresponds to the baseline hazard function, and any individual in the population has the hazard rate multiplicatively distorted by his frailty value z_{ij}. We also assume that the two frailties in each pair have the same variance σ^2, because of the symmetry of twin data, which are the object of applications in the present paper (see Sections 5 and 6).

Hence, we deal with the following distribution of the vector of frailties:

$$\begin{bmatrix} Z_{i1} \\ Z_{i2} \end{bmatrix} \sim LogN\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right) \quad i = 1, ..., n \tag{10}$$

with *logN* denoting the bivariate log-normal distribution. This can be obtained by assuming a

bivariate normal distribution on the logarithm of the frailty vector $\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} = \log \begin{bmatrix} Z_{i1} \\ Z_{i2} \end{bmatrix}$ whose parameters are some functions of the frailty parameters σ^2 and ρ (see for example Hutchinson and Lai 1991):

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} -\frac{1}{2}\log\left(\sigma^2+1\right) \\ -\frac{1}{2}\log\left(\sigma^2+1\right) \end{bmatrix}, \begin{bmatrix} \log\left(\sigma^2+1\right) & \log\left[\rho\sigma^2+1\right] \\ \log\left[\rho\sigma^2+1\right] & \log\left(\sigma^2+1\right) \end{bmatrix} \right) \qquad i = 1, ..., n \qquad (11)$$

with N denoting the bivariate normal distribution.

3. Estimation strategy

Methods that have been adopted for parameter estimation in frailty models can be approximately classified into the two categories of maximum likelihood and Markov chain Monte Carlo (MCMC) methods.

Procedures based on the maximum likelihood have been applied in the gamma context, where an explicit representation of the likelihood function is always available (Yashin and Iachine 1994, Yashin et al. 1995, Wienke et al. 2001). The maximum likelihood method has also been adopted in the lognormal framework with the help of different numerical algorithms (McGilchrist and Aisbett 1991, McGilchrist 1993, Lillard 1993, Lillard et al. 1995, Sastry 1997, Ripatti and Palmgren 2000, Arbeev et al. 2003). These methods are also implemented in the aML software package (aML version 1, see Lillard and Panis 2000).

Bayesian MCMC methods have also been applied as estimation procedures especially in shared frailty models (Clayton 1991, Spiegelhalter et al. 1996, Sahu et al. 1997, Sinha and Dey 1997) but also in correlated frailty models (Xue and Ding 1999). The Bayesian framework is in fact natural when we are dealing with conditionally independent observations and we are working with hierarchical models, with the frailty variables at an intermediate stage between the observations and the so-called hyperparameters. In the Bayesian context the frailty distribution represents a 'prior' of the model and its parameters (hyperparameters) are also considered as random variables following some non-informative distribution.

An MCMC method consists in generating a set of Markov chains whose joint stationary distribution corresponds to the joint posterior of the model, this one being in the Bayesian

framework the distribution of random parameters given observed data. In a hierarchical model, the posterior distribution is often very difficult to work with and almost always impossible to integrate out in order to find the marginal posterior of each random parameter. The MCMC methods enable us to circumvent this problem. The posterior of each parameter is approximated by the empirical distribution of the values of the corresponding Markov chain and empirical summary statistics calculated along each chain can be used to make inferences about the true value of the corresponding parameter (see for a review Gilks et al. 1996). The Gibbs Sampling (Geman and Geman 1984) is one of the algorithms that have been created in order to obtain Markov chains with the desired stationary distribution. The basic idea behind the Gibbs Sampling is to successively sample from the conditional distribution of each random node, whether parameter or observable, given all the others in the model. These distributions are known as 'full conditional distributions'. It can be shown that, under broad conditions, this process eventually provides samples from the joint posterior distribution of the unknown quantities.

In this study, Bayesian MCMC methods have been adopted to estimate the correlated log-normal frailty model described in Section 2. Calculations are performed within the software WinBUGS 1.4 (Spiegelhalter et al. 1999). This is a package, which enables us to solve Bayesian hierarchical models, essentially using the Gibbs Sampling algorithm.

The correlated log-normal frailty model applied here can be represented as a Bayesian hierarchical (3 - levels) model in the following way:

1. Likelihood function:

$$L\left(x,\delta|Y,a,b\right)=\prod_{i=1}^{n}\prod_{j=1}^{2}\left[\exp\left(Y_{ij}\right)a\exp(bx_{ij})\right]^{\delta_{ij}}\exp\left(-\exp\left(Y_{ij}\right)\frac{a}{b}\left[\exp(bx_{ij})-1\right]\right) \quad (12)$$

2. Priors:

$$\begin{aligned}
(i)\quad \begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} &\sim N\left(\begin{bmatrix} -\frac{1}{2}\log\left(\sigma^2+1\right) \\ -\frac{1}{2}\log\left(\sigma^2+1\right) \end{bmatrix}, \begin{bmatrix} \log\left(\sigma^2+1\right) & \log\left[\rho\sigma^2+1\right] \\ \log\left[\rho\sigma^2+1\right] & \log\left(\sigma^2+1\right) \end{bmatrix}\right) \quad i=1,...,n \\
(ii)\quad a &\sim \Gamma(0.01,0.01) \\
(iii)\quad b &\sim \Gamma(0.01,0.01)
\end{aligned}$$

3. Hyperpriors:

$$(i)\ \ \sigma^2 \sim \Gamma(0.01, 0.01)$$
$$(ii)\ \ \rho \sim U(-1, 1)$$

where $H(x) = (a/b) \cdot [\exp(bx) - 1]$ is the Gompertz cumulative hazard function; $Y = (Y_1, ..., Y_n)$, $Y_i = (Y_{i1}, Y_{i2})$; Γ and U denote the gamma and uniform distribution, respectively. Non-informative priors are assigned to the parameters of the Gompertz curve and on the frailty parameters (hyperparameters).

The full conditional distributions can be obtained considering that they are proportional to the joint distribution of all the random quantities of the model. In our case, this joint distribution takes the form:

$$\pi\left(x, \delta, Y, a, b, \sigma^2, \rho\right) = L(x, \delta | Y, a, b) \prod_{i=1}^{n} \left[\prod_{j=1}^{2} \pi\left(Y_{ij} | \sigma^2, \rho\right) \right] \pi(a)\, \pi(b)\, \pi\left(\sigma^2\right) \pi(\rho) \qquad (13)$$

where $\pi(\cdot)$ indicates the density function of the correspondent argument.

Often, the full conditional distributions have a complicated form, which makes it impossible to sample directly from them. In such cases, different modifications of the Gibbs Sampling algorithm originally proposed by Geman and Geman (1984) are available in the version 1.4 of the software WinBUGS. In particular, a slice-sampler algorithm is used for non log-concave densities defined on a restricted range (Neal 1997). This has an adaptive phase of 500 iterations, which are discarded from all summary statistics. A Metropolis within Gibbs algorithm based on a symmetric normal proposal distribution is applied in the case of non log-concave densities defined on an unrestricted range (Metropolis et al. 1953, Hastings 1970, Besag and Green 1993). In this case, the adaptive phase is of 5000 iterations. The Metropolis within Gibbs procedure is applied in the log-normal case.

4. Genetic models

Typical models of quantitative genetics can easily be incorporated into the correlated frailty model described in Section 2. Quantitative genetics models (Falconer 1990) are based on the decomposition of a phenotypic trait in a sum of different components, which are supposed to

be independent. Using this approach, it is possible to estimate the proportion of the total variability of the phenotype which is related to genetic factors. In particular, a heritability estimate can be calculated for human longevity by identifying the phenotype with the life span variable (McGue et al. 1993).

Yashin and Iachine (1995) suggested an approach based on the frailty variable Z instead of the life span X. It is now of interest to find out the relative importance of genes and environment in determining the individual susceptibility towards mortality (overall or cause specific). An advantage of this approach is that, through the additive decomposition of frailty into a genetic and an environmental component, one can obtain a competing risk structure for the respective survival model. That is, observed mortality is represented as a sum of two terms: one depends on genetic and another on environmental parameters, both estimated from bivariate data.

More in details, let the frailty be represented by:

$$Z = A + D + I + C + E \tag{16}$$

where A represents additive genetic effects, D corresponds to dominance genetic effects, I denotes epistatic genetic effects, C and E stand for shared and nonshared environmental effects, respectively. All factors are assumed to be independent. The following additive decomposition of the frailty variance and of the correlation coefficient between co-twins' frailty holds:

$$1 = a^2 + d^2 + i^2 + c^2 + e^2 \tag{17}$$

$$\rho = \rho_1 a^2 + \rho_2 d^2 + \rho_3 i^2 + \rho_4 c^2 + \rho_5 e^2 \tag{18}$$

where lowercase letters a^2, d^2, i^2, c^2, e^2 indicate the proportions of the total variance σ^2 associated with the correspondent components of frailty, and ρ_i $(i = 1, ..., 5)$ are correlations between respective components within a twin pair. Standard assumptions of quantitative genetics models specify different values of ρ_i $(i = 1, ..., 5)$ for monozygotic and dizygotic twins. In the case of monozygotic twins $\rho_i = 1$, $i = 1, ..., 4$ and $\rho_5 = 0$, while for dizygotic twins $\rho_1 = 0.5$, $\rho_2 = 0.25$, $\rho_3 = m$, $\rho_4 = 1$, $\rho_5 = 0$ and $0 \leq m \leq 0.25$ is an unknown parameter. Not all parameters of the genetic decomposition of frailty can be estimated simultaneously. The model in fact reduces to

three equations (two relationships (18) for monozygotic and dizygotic twins and one constraint (17)) allowing us to estimate no more than three parameters at the same time. One possibility is to consider an ACE (additive genetic - common environmental - uncommon environmental) model. In this case, equations (17) and (18) lead to the following:

$$\begin{cases} 1 = a^2 + c^2 + e^2 \\ \rho_{MZ} = a^2 + c^2 \\ \rho_{DZ} = 0.5a^2 + c^2 \end{cases} \tag{19}$$

This system can be integrated into the correlated frailty model described in Section 2 (see equation (11)) giving place to a reparameterisation of the original model. The only difference is that, when we are interested in estimating parameters of a genetic model, data for monozygotic and dizygotic twins have to be analysed simultaneously and a likelihood function for combined data has to be drawn.

Equivalently, other genetic models can be obtained combining no more than three components of frailty (Yashin and Iachine 1995). In this paper we compare three different genetic models (ACE, AE and ADE). Results are shown in Section 6.

5. The data

In this analysis we use breast cancer data from the Swedish Twin Registry. First established in the late 1950s to study the importance of smoking and alcohol consumption on cancer and cardiovascular diseases whilst controlling for genetic propensity to disease, it has today developed into a unique source. Since its establishment, the Registry has been expanded and updated on several occasions, and the focus has similarly broadened to most common complex diseases.

At present, the Swedish Twin Registry contains information about two cohorts of Swedish twins referred to as the'old' and the 'middle' cohort. The old cohort consists of all same-sexed pairs born between 1886 and 1925 where both members in a pair were living in Sweden in 1959. In 1970 a new cohort of twins born between 1926 and 1967, the middle cohort, was compiled. We have included both cohorts in our analysis and looked at a total of 12568 pairs of female twins. The data are described in Table 1, categorised according to the censoring status. The

	Both censored	One censored	None censored	*Total*
MZ	4304	335	33	4672
DZ	7236	625	35	7896
Total	11540	960	68	12568

Table 1: Composition of the dataset by zygosity and censoring status. Swedish Twin Registry.

	a	b	σ^2	ρ_{MZ}	ρ_{DZ}
Mean	2.54E-5	0.07155	45.19	0.3107	0.1044
Median	2.52E-5	0.07154	41.50	0.2991	0.0967
Standard dev.	3.239E-6	0.00251	17.05	0.0456	0.1084
MC error	7.92E-8	8.94E-5	0.824	0.0051	0.0021
CSRF	1.002	1.006	1.055	1.008	1.005

Table 2: Results of a correlated log-normal frailty model applied to Swedish breast cancer data. Convergence achieved after 50000 iterations.

event under study is the onset of breast cancer. If a woman did not develop breast cancer or she was died during the follow-up, the corresponding observation is censored.

For a comprehensive description of the Swedish Twin Registry database, with a focus on the recent data collection efforts and a review of the principal findings that have come from the Registry see Lichtenstein et al. (2002).

6. Results

The results of application of the correlated log-normal frailty model to the Swedish breast cancer data are presented in Table 2. Estimated values include the Gompertz parameters a and b, the variance of the frailty distribution σ^2, which can be seen as the extent of population heterogeneity with respect to breast cancer, and estimates of the correlation coefficient for both monozygotic twins (ρ_{MZ}) and dizygotic twins (ρ_{DZ}). Two estimates for each parameter are given in terms of the mean and the median of the correspondent Markov chain. In all cases, the two values are very close to each other. This means that empirical estimates of the marginal posteriors densities (Kernel density estimates) are approximately symmetric. For each parameter the sample standard deviation and an estimate of the standard error of the mean are

also given. This one is obtained following the batch means method outlined by Roberts (1996). In the last row, we reported the value of the *Corrected Scale Reduction Factor* (CSRF) for each parameter. This value corresponds to the Gelman-Rubin convergence statistic (Gelman and Rubin 1992), as modified by Brooks and Gelman (1998), and is based on a comparison of the within and between chain variance for each variable. When values of this diagnostic are approximately equal to one, the sample can be considered to have arisen from the stationary distribution. In this case, descriptive statistics can be seen as valid estimates of unknown parameters.

According to the model, the population under study would present a very large heterogeneity (σ^2) in terms of susceptibility towards breast cancer. The estimated correlation between frailties is larger for monozygotic than for dizygotic twins. This means that individuals who are more similar from a genetic point of view (MZ twins) also present a larger connection in terms of frailty towards breast cancer. This finding suggests that there is a genetic influence on breast cancer propensity. The extent of such an influence is estimated with the help of three different genetic models.

In Table 3, we compare an ACE, AE and ADE model. Estimates of each parameter are given in terms of the sample mean. Sample median values are omitted because they are very close to the mean as in Table 2. We chose to give the posterior standard deviation of each parameter in parenthesis. This quantity is a measure of the dispersion of the posterior density estimate, giving an idea of a parameter's significance.

A first observation can be made about the estimate of parameter c^2 in the ACE model. This value cannot be considered as being significantly different from zero. For this reason the ACE model, which is one the most wide spread in the literature, doesn't seem to be appropriate, and we therefore decided to compare it with two models which do not include the common environmental effect c^2, namely the AE and ADE model.

Moreover, the estimated value of the narrow sense heritability parameter resulting from the ACE model ($\widehat{a}^2 \simeq 0.18$) does not correspond to the one that could be obtained applying a 'two step procedure'. This procedure, which consists in substituting ρ_{MZ} and ρ_{DZ} estimates (Table 2) in ACE equations (19) would lead to a bigger estimate of the heritability parameter ($\widehat{a}^2_{2ST} \simeq 0.4$). The same procedure would also give a negative estimate of parameter c^2, which

may indicate the presence of non-additive genetic effects (Yashin and Iachine 1997).

Analogous considerations about the heritability estimate can be made in the case of the AE model.

The problem does not arise with the last estimated model, including dominance (non-additive) genetic effects (ADE model). The 'one step procedure' applied here, which consists in a re-parameterisation of the correlated frailty model to incorporate the ADE structure (Section 5), provides results, which are similar to the ones obtained with the procedure in two steps ($\widehat{a}^2 \simeq 0.13$ and $\widehat{d}^2 \simeq 0.15$ while $\widehat{a}^2_{2ST} \simeq 0.1$ and $\widehat{d}^2_{2ST} \simeq 0.2$).

The last column of Table 3 shows values of the Deviance Information Criterion (DIC) for the three models. This is a statistic introduced by Spiegelhalter et al. (2002) in order to compare Bayesian models in terms of adequacy and complexity. The DIC statistic is defined as:

$$DIC = \overline{D(\theta)} + p_D$$

where $\overline{D(\theta)}$ represents an estimate (in terms of posterior mean) of the deviance of the model and is suggested as a Bayesian measure of fit or adequacy; p_D is the difference between the posterior mean of the deviance and the deviance of the posterior mean of parameters of interest and is proposed as a measure of the effective number of parameters (complexity) of the model. The deviance $D(\theta)$ is defined as equal to $-2\log p(y|\theta)$ where y comprises all stochastic nodes giving values (i.e. data), and θ comprises the stochastic nodes upon which the distribution of y depends, when collapsing over all logical relationships. It can be shown (Spiegelhalter et al. 2002) that DIC is related to other information criteria and in particular, in models with negligible prior information, DIC is approximately equivalent to Akaike's criterion. The model with the smallest DIC is estimated to be the model that would best predict a replicate dataset of the same structure as that currently observed. In Table 3, the model which presents the lowest value of DIC is the ADE model.

Thus, from the comparison between the three models we can conclude that genetic effects would explain globally almost 30% of the total variability of propensity to breast cancer. Environmental effects would be predominant in determining breast cancer susceptibility and these ones would be primarily individual-specific, that is non-shared effects. Finally, a model which

	a	b	σ^2	a^2	d^2	c^2	e^2	DIC
ACE	2.546E-5	0.0715	45.21	0.1759		0.0529	0.7712	15138.6
	(3.36E-6)	(0.003)	(17.7)	(0.094)		(0.046)	(0.089)	
AE	2.502E-5	0.0721	47.31	0.2304			0.7696	15102.3
	(3.16E-6)	(0.003)	(18.3)	(0.091)			(0.091)	
ADE	2.522E-5	0.07188	48.30	0.127	0.1491		0.7239	15091.80
	(3.16E-6)	(0.002)	(16.7)	(0.086)	(0.100)		(0.084)	

Table 3: Results of three genetic models applied to Swedish breast cancer data. Convergence achieved after 50000 iterations

includes dominance genetic effects should preferably be used for genetic and statistical reasons.

7. Discussion

In the present paper, a Bayesian correlated frailty model has been adopted to analyse the onset of breast cancer in a population of female Swedish twins. A Gompertz assumption is made in order to model the baseline hazard function. The vector of frailties is assumed to follow a log-normal distribution, which is one of the most flexible in multivariate modelling and especially when we are interested in introducing a correlation between frailties, as in the case of the correlated frailty model. Estimates of the correlation coefficient between co-twins' frailties for the two groups of twins are given in Table 2, along with an estimate of the frailty variance. The latter, which measures the degree of heterogeneity in susceptibility towards breast cancer, is very large. Similar results in terms of the variance estimate have been obtained with a gamma assumption on the frailty distribution and using a ML estimation method (results are not shown here). It might be that these effects are partly due to the strong negative correlation between the estimates of σ^2 and ρ, which is typical of the correlated frailty model. Such correlation has been detected and discussed in a recent simulation study involving different assumptions on the frailty distribution and different estimation strategies (Wienke et al. 2003a). On the other hand, using a subset of the data analysed here (the old cohort of the Swedish Twin Registry), Wienke et al. (2003b) have shown that the heterogeneity estimate decreases when the possibility that a fraction of the study population is unsusceptible to experience the disease is accounted for.

Moreover, the bigger estimate of ρ for monozygotic twins (Table 2) provides an evidence of a genetic influence on the propensity to develop breast cancer. To evaluate the amount of such an influence, three genetic models (ACE, AE and ADE) have been applied. All models are implemented in the software WinBUGS with the help of MCMC estimation procedures. We made comparisons between competing models using genetic arguments and with the help of a Bayesian criterion for model comparison (DIC) introduced by Spiegelhalter et al. (2002) and available in the last version (version 1.4) of the software WinBUGS.

Results of this study show that more than 70% of the variability in frailty towards breast cancer is due to environmental factors. These are essentially factors which are not shared by the two women in a pair (Table 3). The heritability estimate is then around 30% and genetic effects include both additive genetic and dominance genetic effects.

The WinBUGS package proved to be extremely useful and flexible enough to estimate correlated frailty models and to add to them equations typical of genetic models. Within the same software it is easy to modify the hypothesis on the frailty distribution, and it is also possible to follow a semiparametric strategy by assuming a prior process on the cumulative hazard function (the work on semiparametric methods is in progress). Different assumptions about the frailty distribution and the shape of the baseline hazard function can be compared within the same software (version 1.4) with the help of a Bayesian information criterion (DIC).

The disadvantage of using WinBUGS in the context described here is in the time required for estimation. In fact, we are working with models which include a very large number of parameters, especially when we deal with large data sets. This means that every MCMC algorithm which updates parameters one by one (like the Gibbs Sampling used in WinBUGS) will be very time consuming. To overcome this limitation, an algorithm which enables to update parameters all together (or groups of parameters at the same time) should be adopted.

Aknowledgment The authors wish to aknowledge Andreas Wienke and Konstantin Arbeev for the useful discussions. The authors whish also to say thank to the Max Planck Institute for Demographic Research of Rostock (Germany) for the opportunity to use all its technical facilities, during work on this paper.

Bibliography

[1] Arbeev, K.G., Vaupel, J.W., Yashin, A.I. (2003) Bivariate Lognormal Frailty Models: Estimation Methods, Simulation Studies and Application to Danish Twins Data, MPIDR Working Paper Series, to appear.

[2] Besag, J., Green, P.J. (1993) Spatial statistics and bayesian computation. Journal of the Royal Statistical Society, B 55, 25 - 37.

[3] Brooks S.P. and Gelman A. (1998) Alternative methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7, 434-455.

[4] Butler, J.S., Anderson, H.A., Burkhauser R.V. (1986) Testing the relationship between work and health. Economics Letters 20, 383 - 386.

[5] Clayton, D. (1978) A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence. Biometrika 65, 141 - 151.

[6] Clayton, D. (1991) A Monte Carlo Method for Bayesian Inference in Frailty Models. Biometrics 47, 467 - 485.

[7] Cox, D.R. (1972) Regression models and life tables (with discussion). Journal of the Royal Statistical Society, B 34, 187 - 220.

[8] Falconer, D.S. (1990) Introduction to Quantitative Genetics, Longman Group, New York.

[9] Gelman, A., Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7, 457 - 511.

[10] Geman, S., Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restauration of images. IEEE Trans. Pattern Analysis and Machine Intelligence 6, 721 - 741.

[11] Gilks, W.R., Richardson, S., Spiegelhalter, D.G. (1996) Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

[12] Hastings, W.K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57, 97 - 109.

[13] Hougaard, P. (1984) Life tables methods for heterogeneous populations: distributions describing the heterogeneity. Biometrika 71, 75 - 84.

[14] Hougaard, P. (2000) Analysis of Multivariate Survival Data. Springer, New York.

[15] Hutchinson, T.P., Lai, C.D. (1991) The engineering statistician's guide to continuous bivariate distributions. Rumsby, Adelaide.

[16] Korsgaard, I.R., Madsen, P., Jensen, J. (1998) Bayesian inference in semiparametric lognormal frailty model using Gibbs sampling. Genetics, Selection, Evolution 30, 241 - 256.

[17] Iachine, I.A., Holm, N.V., Harris, J.R., Begun, A.Z., Iachina, M.K., Laitinen, M., Kaprio, J., Yashin, A.I. (1998) How heritable is individual susceptibility to death? The results of an analysis of survival data on Danish, Swedish and Finnish twins. Twin Research 1, 196 - 205.

[18] Lichtenstein, P., de Faire, U., Floderus, B., Svartengren, M., Svedberg, P., Pedersen, N.L. (2002) The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. Journal of Internal Medicine 252, 184 - 205.

[19] Lillard, L.A. (1993) Simultaneous equations for hazards: marriage duration and fertility timing. Journal of Econometrics 56, 189 - 217.

[20] Lillard, L.A., Brian, M.J., Waite, M.J. (1995) Premarital Cohabitation and Subsequent Marital Dissolution: A Matter of Self-Selection? Demography 32, 437 - 457.

[21] Lillard, L.A., Panis, C.W.A. (2000) aML User's Guide and Reference Manual. Los Angeles: EconWare.

[22] McGilchrist, C.A. (1993) REML Estimation for Survival Models with Frailty. Biometrics 49, 221 - 225.

[23] McGilchrist, C.A., Aisbett, C.W. (1991) Regression with Frailty in Survival Analysis. Biometrics 47, 461 - 466.

[24] McGue, M., Vaupel, J.W., Holm, N., Harvald, B. (1993) Longevity is moderately heritable in a sample of Danish twins born 1870 - 1880. Journal of Gerontology: Biological Sciences, B 48: B237 - B244.

[25] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller A.H., Teller, E. (1953) Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics 21, 1087 - 1091.

[26] Neal R. (1997) Markov chain Monte Carlo methods based on 'slicing' the density function. Technical Report 9722, Department of Statistics, University of Toronto, Canada.

[27] Oakes, D. (1982) A Concordance Test for Independence in the Presence of Censoring. Biometrics 38, 451 - 455.

[28] Ripatti, S., Palmgren, J. (2000) Estimation of multivariate frailty models using penalised partial likelihood. Biometrics 56, 1016 - 1022.

[29] Roberts, G.O. (1996). Markov chain concepts related to sampling algorithms. In W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.) Markov chain Monte Carlo in Practice. Chapman and Hall, London.

[30] Sahu, K.S., Dey, D.K., Aslanidou, H., Sinha, D. (1997) A Weibull Regression Model with Gamma Frailties for Multivariate Survival Data. Lifetime Data Analysis 3, 123 - 137.

[31] Sastry, N. (1997) A Nested Frailty Model for Survival data, With an Application to the Study of Child Survival in Northeast Brazil. Journal of the American Statistical Association 92, 426 - 435.

[32] Sinha, D., Dey, K.D. (1997) Semiparametric Bayesian Analysis of Survival Data. Journal of the American Statistical Association 92, No 439.

[33] Spiegelhalter, D.J., Thomas, A., Best, N.G., Gilks, W.R. (1996). BUGS Examples Volume 1, Version 0.5, (version ii).

[34] Spiegelhalter, D.J., Thomas, A., Best, N.G. (1999) WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit.

[35] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002) bayesian measures of model complexity and fit. Journal of the Royal Statistical Society B 64, 583 - 639.

[36] Vaupel, J.W., Manton, K.G., Stallard, E. (1979) The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. Demography 16, 439 - 454.

[37] Vaupel, J.W., Yashin A.I. (1985) Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics. The American Statistician 39, 176 - 185.

[38] Vaupel, J.W., Harvald, B., Holm, N.V., Yashin, A.I., Xiu L. (1992) Survival Analysis in Genetics: Danish Twin Data Applied to a Gerontological Question. Kluwer Academic Publishers, Netherlands.

[39] Wienke, A., Holm, N., Skytthe, A., Yashin A.I. (2001) The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. Twin Research 4, 266 - 274.

[40] Wienke, A., Arbeev, K., Locatelli, I., Yashin, A.I. (2003) A comparison of different correlated frailty models and estimation strategies (submitted to Mathematical Biosciences)

[41] Wienke, A., Lichtenstein, P., Yashin, A.I. (2003) A biveriate frailty model with a cure fraction for modeling familial correlations in diseases (accepted by Biometrics)

[42] Xue, X., Ding, Y. (1999) Assessing heterogeneity and correlation of paired failure times with the bivariate frailty model. Statistics in Medicine 18, 907 - 918.

[43] Yashin, A.I., Iachine, I.A. (1994) Mortality models with application to twin survival data. In CISS-First Joint Conference of International Simulation Societies Proceedings (Halin J., Karplus W. and Rimane R. eds.). Zürich, Switzerland, pp. 567 - 571.

[44] Yashin, A.I., Iachine, I.A. (1995) Genetic Analysis of Durations: Correlated Frailty Models Applied to Survival of Danish Twins. Genetic Epidemiology 12, 529 - 538.

[45] Yashin, A.I., Iachine, I.A. (1997) How frailty models can be used for evaluating longevity limits: taking advantage of an interdisciplinary approach. Demography, vol. 34 , 31 - 48.

[46] Yashin, A.I., Vaupel, J.W., Iachine, I.A. (1995) Correlated Individual Frailty: an Advantageous Approach to Survival Analysis of Bivariate Data. Mathematical Population Studies 5, 145 - 159.