The Heritability of Breast Cancer: A Bayesian Correlated Frailty Model Applied to Swedish Twins Data

Isabella Locatelli[1], Paul Lichtenstein[2], and Anatoli I. Yashin[3]

[1]*University Luigi Bocconi, Milan, Italy*
[2]*Karolinska Institut, Stockholm, Sweden*
[3]*Max Planck Institute for Demographic Research, Rostock, Germany*

The aim of this study was to investigate the role of genes and environment in susceptibility to breast cancer and to give an estimate of heritability in the propensity to develop the disease. To do this we applied an interdisciplinary approach, merging models developed in the field of demography and survival analysis — so-called *frailty models* — and models coming from quantitative genetics and epidemiology, namely *genetic models*. In our study, the inferential problem was solved in a Bayesian framework and the numerical work was carried out using MCMC methods. We used the special information coming from twin data, particularly breast cancer data, from the Swedish Twin Register. The application of a correlated log-normal frailty model leads to a very large estimate of the population heterogeneity (σ = 6.7), and relatively small correlations between co-twins' frailties — around 0.3 for monozygotic and 0.1 for dizygotic twins. Comparing three different genetic models (an ACE, an AE and an ADE model), we furthermore concluded 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 effects would be primarily individual-specific, that is, non-shared effects. Finally, a model which includes *dominance genetic* effects (ADE model) is preferred for genetic and statistical reasons.

Frailty Models

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 toward mortality (overall or cause-specific mortality) and it is often impossible to include all the relevant covariates in the model. More frail individuals die earlier than 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 analyze the individual risk of mortality at different frailty levels (Vaupel & Yashin, 1985).

In the present study, we were dealing with multivariate frailty models, which were created with the aim of assessing mutual dependence between the life spans 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; Sahu et al., 1997; Vaupel et al., 1992). 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 particularly spread in the field of animal genetics (Ducrocq et al., 1988; Ducrocq & Casella, 1996). In this context, the shared frailty term — multiplicatively added to models describing the time of culling for groups of breeding animals — represents the effect of belonging to a particular sire. Thus, this random effect reflects genetic-specific features, which are "shared" by all animals coming from the same sire (Yazdi et al., 2000).

Shared frailty models are useful for explaining correlations within groups, but they have some limitations. Firstly, they deal with a definition of frailty which is not consistent with the definition given in the univariate framework (Vaupel et al., 1979). The

Received 23 July, 2003; accepted 25 November, 2003.

Address for correspondence: Isabella Locatelli, Via Mincio 30, 20139 MILANO, Italy. Email: isabella.locatelli@uni-bocconi.it; locatelli@demogr.mpg.de

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

frailty term represents a part of individual frailty, only capturing the components that 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 also allow for a negative correlation between life spans within the groups (Xue & 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 on the analysis of pairs of related individuals, for example twins. The correlated frailty assumption is more flexible than the shared frailty assumption 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 life spans 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 toward 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 merging of traditional approaches of quantitative genetics and epidemiology with survival analysis methods (Yashin & Iachine, 1995, 1997).

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 semiparametrically. The most adopted parametrical hypothesis is the Gompertz baseline hazard (Iachine et al., 1998; Vaupel et al., 1992; Wienke et al., 2001) but other shapes are also possible, for example, Weibull (Do et al., 2000; Sahu et al., 1997; Visscher et al., 2001) or (piecewise) exponential (Xue & Ding, 1999; Scurrah et al., 2000). 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 estimation of 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; Hougaard, 2000; Oakes, 1982; Vaupel et al., 1979; Wienke et al., 2001; Yashin & Iachine, 1994). The gamma choice is convenient from a mathematical point of view, because of the simplicity of the Laplace transformation, which allows for the use of traditional maximum likelihood procedures in parameter estimation. Another possibility is to assume that frailty is log-normal distributed (Do et al., 2000; Korsgaard et al., 1998; Ripatti & Palmgren, 2000; Scurrah et al., 2000; Spiegelhalter et al., 1996; Xue & Ding, 1999). The log-normal approach is much more flexible than the gamma one 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 Estimation Strategy section below).

The present study worked with correlated frailty models. The Model Description section below provides a general description of the theory of correlated frailty models. In the section titled Estimation Strategy, the estimation procedure is presented. An interdisciplinary approach based on quantitative genetics models is described in the section titled Genetic Models. The Data section introduces data from the Swedish Twin Register. The Results section presents results of the analysis. Some comments and suggestions for further research are presented in the Discussion.

Materials and Methods

The Model Description

As we have already described, frailty models represent a particular area of survival analysis. This discipline typically studies the behavior of a random variable X, describing the time since the origin of an observation period and the moment of the occurrence of an event of interest. The survival function is defined as the probability of the event occurring after a certain time:

$$S(x) = \Pr(X > x). \quad (1)$$

In the case of continuous time, another quantity is introduced, the so-called *hazard function*, which is defined as the probability of the event occurring in the interval $(x, x + \Delta x)$, given that it has not occurred

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

before x, divided by the length of the interval, and for $\Delta x \to 0$:

$$\mu(x) = \lim_{\Delta x \to 0} \frac{P(x \le X < x + \Delta x \mid X \ge x)}{\Delta x} \quad (2)$$

The hazard function characterizes the risk changing over time, specifying the instantaneous failure rate at time x, for an individual who is still at risk of experiencing the event at that time.

Being H(x) the *cumulative hazard function*, $H(x) = \int_0^x \mu(t)dt$, the following relations hold:

$$\begin{aligned} S(x) &= \exp(H(x)) \\ \mu(x) &= \frac{f(x)}{S(x)} \end{aligned} \quad (3)$$

where $f(x)$ is the density function of the random variable X.

Frailty models are typically based on the so-called *multiplicative assumption* (Cox, 1972), that is, the hazard function (2) is represented by the product of a *baseline hazard*, $\mu_0(x)$, and a *frailty* term (Z), the latter describing the role played by unobserved risk factors on the individual risk (Vaupel et al., 1979):

$$\mu(x, Z) = Z\mu_0(x). \quad (4)$$

In the current study we dealt with a particular class of frailty models, the so-called correlated frailty models, which are adopted in order to describe the correlation between frailties within pairs of individuals (namely twins) and, by consequence, between their duration times.

Let X_{i1},X_{i2} be the vector of life spans (duration times) for the two individuals from the pair i (i = 1,...,n). The typical assumption of correlated frailty models is that X_{i1} and X_{i2} are conditionally independent given the frailties Z_{i1} and Z_{i2}:

$$X_{i1} \mid Z_{i1}, Z_{i2} \perp X_{i2} \mid Z_{i1}, Z_{i2} \quad (5)$$

The conditional likelihood of the model is given by:

$$L(x \mid z) = \prod_{i=1}^{n} f_{X_{i1}, X_{i2} \mid Z_{i1}, Z_{i2}}(x_{i1}, x_{i1} \mid z_{i1}, z_{i2}) \quad (6)$$

where $x = (x_1,...,x_n)$, $x_i = (x_{i1}, x_{i2})$; $z = (z_1,...,z_n)$, $z_i = (z_{i1}, z_{i2})$ and $f_{X_{i1}, X_{i2} \mid Z_{i1}, Z_{i2}}$ represents the bivariate conditional density of the life spans for the pair i. The conditional independence of the life spans given the frailties (5) allows rewriting of (6) as follows:

$$L(x \mid z) = \prod_{i=1}^{n} \prod_{j=1}^{2} f_{X_{ij} \mid Z_{ij}}(x_{ij} \mid z_{ij}) \quad (7)$$

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

Given the relations (see equation [3]):

$$\begin{aligned} f_{X \mid Z}(x \mid z) &= \mu(x, z) S_{X \mid Z}(x \mid z) \\ S_{X \mid Z}(x \mid z) &= \exp(-zH_0(x)) \end{aligned} \quad (8)$$

where $S_{X \mid Z}$ is the conditional survival function given the frailty variable, the expression for the conditional likelihood becomes:

$$L(x, \delta \mid z) = \prod_{i=1}^{n} \prod_{j=1}^{2} [z_{ij} \mu_0(x_{ij})]^{\delta_{ij}} \exp(-z_{ij} H_0(x_{ij})) \quad (9)$$

where for each individual a censoring indicator δ_{ij} is introduced, taking value 1 if the subject experiences the event and 0 otherwise. When δ_{ij} = 0, the value of X_{ij} represents a censoring time, corresponding to the end of the observation period, instead of a failure time.

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

$$\begin{aligned} L(x, \delta) = \prod_{i=1}^{n} \iint \prod_{j=1}^{2} [z_{ij} \mu_0(x_{ij})]^{\delta_{ij}} \\ \exp(-z_{ij} H_0(x_{ij})) f_{Z_{i1}, Z_{i2}}(z_{i1}, z_{i2}) dz_{i1} dz_{i2} \end{aligned} \quad (10)$$

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

In the current study, we adopted a Gompertz baseline hazard, $\mu_0(x) = ae^{bx}$, and we did not take into account observed covariate effects.

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

For identifiability reasons, a restriction was placed 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 was constrained to be equal to one, $E(Z_{ij})$ = 1, for i = 1,...,n and j = 1, 2. In that way, it was assumed that the hazard function of a "standard" individual corresponded to the baseline hazard function, and the hazard rate of any individual in the population was multiplicatively distorted by their frailty value z_{ij}. This assumption differs from the one generally made in the context of correlated log-normal frailty models. Usually the restriction is placed on the logarithm of the frailty variable, whose mean is assumed to be equal to zero (Do et al., 2000; Korsgaard et al., 1998; Ripatti & Palmgren, 2000; Scurrah et al., 2000; Spiegelhalter et al., 1996; Xue & Ding, 1999). This hypothesis does not imply that the average frailty in the population is equal to 1 [$E(\log Z) \neq \log E(Z)$], as originally assumed in the first formulations of frailty models (Clayton, 1978; Vaupel et al., 1979). Thus, in the present study, the estimated variance and correlation refer to the frailty variable itself, instead of to its logarithm. The same can be said of the genetic decomposition of the frailty variance and the estimate of heritability (see Genetic Models section below).

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

Finally, it was assumed that the two frailties in each pair had the same variance σ^2, because of the symmetry of twin data, which were the object of application in the present paper.

Hence, the study dealt with the following distribution of the vector of frailties:

$$\begin{bmatrix} Z_{i1} \\ Z_{i2} \end{bmatrix} \sim \log N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right) \quad i=1,\ldots,n \qquad (11)$$

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 & Lai, 1991):

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} -\frac{1}{2}\log(\sigma^2+1) \\ -\frac{1}{2}\log(\sigma^2+1) \end{bmatrix}, \begin{bmatrix} \log(\sigma^2+1) & \log(\rho\sigma^2+1) \\ \log(\rho\sigma^2+1) & \log(\sigma^2+1) \end{bmatrix} \right) \quad i=1,\ldots,n \qquad (12)$$

with *N* denoting the bivariate normal distribution.

Estimation Strategy

Methods that have been adopted for parameter estimation in frailty models can be approximately classified into two categories: (1) maximum likelihood, and (2) Markov chain Monte Carlo (MCMC) methods.

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

Bayesian MCMC methods have also been applied as estimation procedures especially in shared frailty models (Clayton, 1991; Sahu et al., 1997; Sinha & Dey, 1997; Spiegelhalter et al., 1996) but also in correlated frailty models (Do et al., 2000; Scurrah et al., 2000; Xue & Ding, 1999). The Bayesian framework is natural when dealing with conditionally independent observations and 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 generates a set of Markov chains whose joint stationary distribution corresponds to the joint posterior of the model, this one being 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 determine the marginal posterior of each random parameter. The MCMC methods enable circumvention of 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 (for a review see Gilks et al., 1996). Gibbs Sampling (Geman & 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 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 the current study, Bayesian MCMC methods were adopted to estimate the correlated log-normal frailty model described above. Calculations were performed within the software WinBUGS 1.4 (Spiegelhalter et al., 1999). WinBUGS 1.4 is a package which enables solution of 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-level) model in the following way:

1. Likelihood function:

$$L(x,\delta \mid y,a,b) = \prod_{i=1}^{n}\prod_{j=1}^{2}\left[\exp(y_{ij})a\exp(bx_{ij})\right]^{\delta_{ij}} \exp\left(-\exp(y_{ij})\frac{a}{b}\left[\exp(bx_{ij})-1\right]\right) \qquad (13)$$

2. Priors:

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

(ii) $a \sim \Gamma(0.01, 0.01)$

(iii) $b \sim \Gamma(0.01, 0.01)$

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, $y = (y_1,\ldots, y_n)$, $y_i = (y_{i1}, y_{i2})$, and Γ and U denote the gamma and uniform distribution, respectively. Non-informative priors are assigned on the parameters of the Gompertz curve and

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

on the frailty parameters (hyperparameters). A prior distribution is called non-informative when it covers, with a large variance, the reasonable interval of values of a parameter.

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 the current study, the joint distribution took the form:

$$\pi(x,\delta,y,a,b,\sigma^2,\rho) = L(x,\delta \mid y,a,b) \prod_{i=1}^{n}\left[\prod_{j=1}^{2}\pi(Y_{ij} \mid \sigma^2,\rho)\right]\pi(a)\pi(b)\pi(\sigma^2)\pi(\rho) \quad (14)$$

where $\pi(\cdot)$ indicates the density function of the corresponding 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 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 (Besag & Green, 1993; Hastings, 1970; Metropolis et al., 1953). In this case, the adaptive phase is of 4,000 iterations. The Metropolis within Gibbs procedure is applied in the log-normal case.

Different models proposed for the same set of data (even if they are not nested) can be compared with the help of a Bayesian criterion, the Deviance Information Criterion (DIC), recently introduced by Spiegelhalter et al. (2002). This criterion allows comparison of different Bayesian hierarchical models in terms of adequacy and complexity. The DIC statistic is defined as:

$$DIC = \overline{D(\theta)} + p_D \quad (15)$$

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, and 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 $-2\log p(y \mid \theta)$ where y comprises all stochastic nodes giving values (that is, 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.

Genetic Models

Typical models of quantitative genetics can easily be incorporated into the correlated frailty model described above. Quantitative genetics models (Falconer, 1990) are based on the decomposition of a phenotypic trait into 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 that is related to genetic factors. This proportion is defined as the "heritability" of the phenotypic trait. In particular, a heritability estimate can be calculated for human longevity by identifying the phenotype with the life span variable (McGue et al., 1993).

The definition of heritability given by Yashin and Iachine (1995) was used in the current study. To study the role of genetic and environmental factors on longevity, they suggest an approach based on the frailty variable Z instead of the life span (duration time) X. The phenotype is thus identified with the unobserved heterogeneity term. With this approach, the problem of censoring in the estimate of heritability does not arise because heritability is calculated as a function of the correlation coefficient between co-twins' frailties — estimated via application of a correlated frailty model — instead of the correlation between observed duration times. In this context, heritability is defined as the proportion of the total variability of frailty explained by genetic factors and it is thus obtained via decomposition of the frailty variance (Do et al., 2000; Scurrah et al., 2000). 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 (Yashin & Iachine, 1995).

As mentioned above (in The Model Description section), in the literature of correlated log-normal frailty models the decomposition is usually made with respect to the variance of the logarithm of frailty (Do et al., 2000; Scurrah et al., 2000). The current study referred to the variance of the frailty itself. We believe this interpretation is more consistent with the multiplicative assumption than the usual one, which is based on a definition of frailty as a term acting additively on the logarithm of the baseline hazard, $\log\mu(t, Z) = Z + \log\mu_0(t)$.

In more detail, let the frailty be represented by:

$$Z = A + D + I + C + E \quad (16)$$

where A represents additive genetic effects, D corresponds to dominance genetic effects, I denotes epistatic genetic effects, and C and E stand for shared and non-

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

shared 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:

$$1 = a^2 + d^2 + i^2 + c^2 + e^2 \quad (17)$$

$$\rho = \rho_1 a^2 + \rho_2 d^2 + \rho_3 i^2 + \rho_4 c^2 + \rho_5 e^2 \quad (18)$$

where lowercase letters a^2, d^2, i^2, c^2, e^2 indicate the proportions of the total variability associated with the corresponding 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 $0 \le m \le 0.25$ is an unknown parameter. Not all parameters of the genetic decomposition of frailty can be estimated simultaneously. The model reduces to three equations (two relationships (18) for monozygotic and dizygotic twins and one constraint (17)) allowing estimation of 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:

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

This system can be integrated into the correlated frailty model described above in the section titled The Model Description (see equation (12)) giving place to a reparameterization of the original model. The only difference is that when interested in estimating parameters of a genetic model, data for monozygotic and dizygotic twins have to be analyzed 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 & Iachine, 1995). In this paper we compare three different genetic models (ACE, AE and ADE).

The Data

In the current analysis we used 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 is now a unique source. Since its establishment, the Registry has been expanded and updated on several occasions, and the focus has similarly broadened to include 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. Both cohorts were included in the current analysis making a total of 12,568 pairs of female twins. The data are described in Table 1, categorized according to the censoring status. The event under study was the onset of breast cancer. If a woman did not develop breast cancer or she was deceased at follow-up, the corresponding observation was censored.

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

Results

Results of the 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 parallel chains were run from different starting points and 25,000 iterations were generated per chain. Estimates were calculated after discarding a burn-in of 4,000 iterations for each chain. Two estimates for each parameter were calculated in terms of the mean and the median of the corresponding Markov chain. In all cases, the two values were very close to each other. This means that empirical estimates of the marginal posteriors densities (Kernel density estimates) are approximately symmetric (Figure 1). The symmetry of the posterior distribution of all the parameters of interest around the posterior mean (which is thus equal to the posterior median) allows us to be sufficiently confident in our estimates, even if in some cases the sample standard deviation is quite large. For each parameter, in addition to the sample standard deviation, an estimate of the standard error of the mean is also given. This was obtained following the batch means method outlined by Roberts (1996). The value of the Corrected Scale Reduction Factor (CSRF) for each parameter is reported in the final row of Table 2. This value corresponds to the Gelman-Rubin convergence statistic (Gelman & 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 1, the sample can be considered to have arisen from the stationary distribution and descriptive statistics can be seen as valid estimates of unknown parameters.

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

Table 1

Composition of the Dataset by Zygosity and Censoring Status. Swedish Twin Registry

	Both censored	One censored	None censored	Total	% of Individual affected
MZ	4304	335	33	4672	0.0429
DZ	7236	625	35	7896	0.0432
Total	11540	960	68	12568	0.0431

Table 2

Results of a Correlated Log-normal Frailty Model Applied to Swedish Breast Cancer Data. Convergence Achieved After 50,000 Iterations

	a	b	σ^2	ρ_{MZ}	ρ_{DZ}
Mean	2.54E-5	0.0715	45.190	0.3107	0.1044
Median	2.52E-5	0.0715	41.500	0.2991	0.0967
Standard deviation	3.24E-6	0.0025	17.040	0.0456	0.1084
MC error	7.92E-8	8.94E-5	0.824	0.0051	0.0021
CSRF	1.0021	1.0063	1.055	1.0082	1.0048

Table 3

Results of Three Genetic Models Applied to Swedish Breast Cancer Data. Convergence Achieved After 50,000 Iterations

	a	b	σ^2	a^2	d^2	c^2	e^2	DIC
ACE	2.55E–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	2.50E–5	0.0721	47.31	0.2304			0.7696	15102.3
	(3.16E–6)	(0.003)	(18.3)	(0.091)			(0.091)	
ADE	2.52E–5	0.0719	48.30	0.1273	0.1491		0.7239	15091.8
	(3.16E–6)	(0.002)	(16.7)	(0.086)	(0.100)		(0.084)	

According to the model, the population under study would present a very large heterogeneity (σ^2) in terms of susceptibility toward breast cancer. The estimated correlation between frailties is larger for monozygotic than for dizygotic twins. A likely interpretation of this result is that individuals who are more similar from a genetic point of view (MZ twins) also present a larger connection in terms of frailty toward breast cancer. This finding provides evidence of a genetic influence on propensity to develop breast cancer. If genetic factors do influence individual susceptibility toward breast cancer, we would expect to see a higher correlation between frailties in MZ twins, who are genetically identical, than in DZ twins who, on the average, have just half of their genes in common. The extent of such a genetic influence was then estimated in the current study with the help of three different genetic models.

Table 3 compares an ACE, AE and ADE model. Estimates of each parameter are given in terms of the sample mean. Sample median values were omitted because they were very close to the mean in Table 2. The posterior standard deviation of each parameter is presented 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 0. For this reason the ACE model, which is the one most widely reported in the literature, does not seem to be appropriate, therefore, it was compared 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, $\hat{a}^2 \cong 0.18$, does not correspond to the one that could be obtained by applying a "two step-procedure". A two-step procedure, which simply consists of substituting ρ_{MZ} and ρ_{DZ} estimates (Table 2) in ACE equations (19), would lead to a bigger estimate of the heritability parameter, $\hat{a}^2_{2ST} \cong 0.4$. The same procedure would also give a negative estimate of parameter c^2, which may indicate the presence of non-additive genetic effects (Yashin & Iachine, 1997).

These problems do not arise with the other two models (AE and ADE). In particular, under the AE model, ρ_{MZ} and ρ_{DZ} are both estimates of a^2. The (unweighted) average of these two is around 0.25. This value is not too far from the 0.23 obtained with the "one-step procedure" adopted here, consisting of a reparameterization of the correlated frailty model in

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

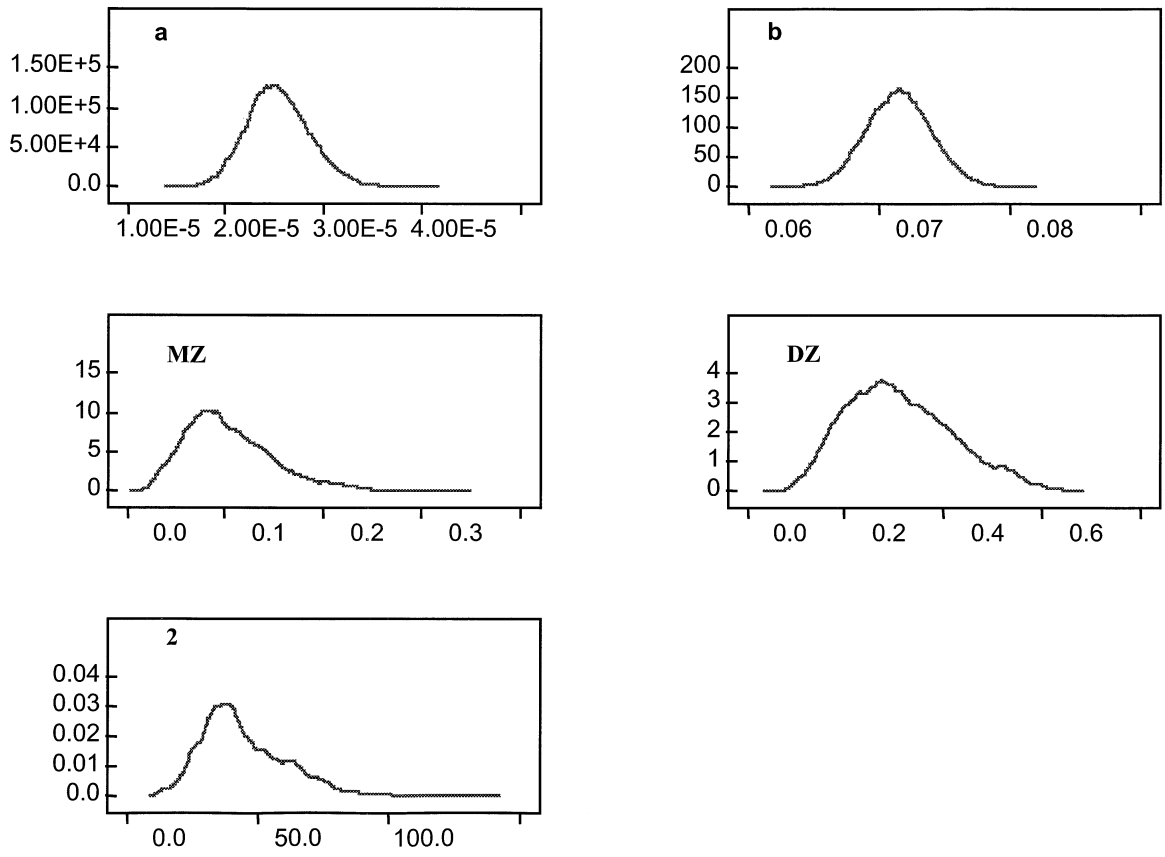

Figure 1

Posterior distributions of the parameters estimated via application of a correlated log-normal frailty model to Swedish breast cancer data. Convergence achieved after 50,000 iterations.

order to incorporate the ADE structure (see Genetic Models section above). Equivalently, under the ADE model, the one-step procedure provided results that were similar to those obtained with the procedure in two steps ($\hat{a}^2 \cong 0.13$ and $\hat{d}^2 \cong 0.15$ while $\hat{a}^2_{2ST} \cong 0.1$ and $\hat{d}^2_{2ST} \cong 0.2$).

Finally, the three models were compared using the Deviance Information Criterion (Spiegelhalter et al., 2002). As seen above (Estimation Strategy section), the DIC statistic allows comparison between different Bayesian models using the criteria of the best adequacy to the data and the lowest complexity. The model which presented the smallest value of DIC in the current study was the ADE model (Table 3). Therefore, the ADE model is the model that would best predict a replicate dataset of the same structure as the one currently observed.

Discussion

In the present paper, a Bayesian correlated frailty model was adopted to analyze the onset of breast cancer in a population of female Swedish twins. A Gompertz assumption was made in order to model the baseline hazard function. The vector of frailties was assumed to follow a log-normal distribution, which is one of the most flexible in multivariate modelling and especially when interested in introducing a correlation between frailties, as in the case of the correlated frailty model. Also, estimates of the frailty variance (Table 2), which measure the degree of heterogeneity in susceptibility toward breast cancer, were very large. This effect may be 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 analyzed 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.

The current study compared three different genetic models using the Deviance Information Criterion (see section titled Results). The ADE model, a model including dominance genetic effects, proved to be the best model in terms of adequacy and complexity. According to the parameter estimates of the ADE model, genetic effects would explain globally around

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

30% of the total variability of propensity to breast cancer. Environmental effects would be predominant in determining breast cancer susceptibility and would primarily be individual-specific, that is, non-shared 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 (BIC).

The disadvantage of using WinBUGS in the context described here is in the time required for estimation. Models being worked with include a very large number of parameters, especially when analyzing large data sets. Therefore, every MCMC algorithm which updates parameters one by one (like Gibbs Sampling used in WinBUGS) will be very time consuming. To overcome this limitation, an algorithm which enables updating of parameters all together (or groups of parameters at the same time) should be adopted.

Acknowledgments

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

References

Arbeev, K. G., Vaupel, J. W., & Yashin, A. I. (in press). Bivariate lognormal frailty models: Estimation methods, simulation studies and application to Danish twins Data. *MPIDR Working Paper Series*.

Besag, J., & Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, 25–37.

Brooks, S. P. & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

Butler, J. S., Anderson, H. A., & Burkhauser R. V. (1986). Testing the relationship between work and health. *Economics Letters, 20*, 383–386.

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.

Clayton, D. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics, 47*, 467–485.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, B34*, 187–220.

Do, K-A., Broom, B. M., Kuhnert, P., Duffy, D. L., Todorov, A. A., Treloar, S. A., et al. (2000). Genetic analysis of the age of menopause by using estimating equations and Bayesian random effects models. *Statistics in Medicine, 19*, 1217–1235.

Ducrocq, V., & Casella, G. (1996). A Bayesian analysis of mixed survival models. *Genetics Selection Evolution, 28*, 505–529.

Ducrocq, V., Quaas, R. L., & Pollak, E. J. (1988). Length of productive life of dairy cows. variance component estimation and sire evaluation. *Journal of Dairy Science, 71*, 3171–3079.

Falconer, D. S. (1990). *Introduction to quantitative genetics*. New York: Longman Group.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–511.

Geman, S., & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restauration of images. *Institute of Electrical and Electronic Engineers Transactions Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. G. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97–109.

Hougaard, P. (1984). Life tables methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika, 71*, 75–84.

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

Hutchinson, T. P., & Lai, C. D. (1991). *The engineering statistician's guide to continuous bivariate distributions*. Adelaide: Rumsby.

Iachine, I. A., Holm, N. V., Harris, J. R., Begun, A. Z., Iachina, M. K., Laitinen, M., et al. (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.

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

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.

Lillard, L. A. (1993). Simultaneous equations for hazards: Marriage duration and fertility timing. *Journal of Econometrics, 56*, 189–217.

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press

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.

Lillard, L. A., & Panis, C. W. A. (2000). aML *User's guide and reference manual*. Los Angeles: Econ-Ware.

McGilchrist, C. A. (1993). REML estimation for survival models with frailty. *Biometrics, 49*, 221–225.

McGilchrist, C. A., & Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics, 47*, 461–466.

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, Series B 48*, B237–B244.

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.

Neal, R. (1997). *Markov chain Monte Carlo methods based on "slicing" the density function. Technical Report 9722*. Toronto, Canada: Department of Statistics, University of Toronto.

Oakes, D. (1982). A concordance test for independence in the presence of censoring. *Biometrics, 38*, 451–455.

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

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

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.

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.

Scurrah, K. J., Palmer, L. J., & Burton, P. R. (2000). Variance components analysis for pedegree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genetic Epidemiology, 19*, 127–148.

Sinha, D., & Dey, K. D. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association, 92*, 1195–1212.

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, Series B64*, 583–639.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). *WinBUGS Version 1.2 User Manual*. Cambridge, UK: MRC Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS examples Volume 1, Version 0.5 (version ii)*.

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*. Netherlands: Kluwer Academic Publishers.

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.

Vaupel, J. W., & Yashin A. I. (1985). Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *The American Statistician, 39*, 176–185.

Visscher, P. M., Yazdy, M. H., Jackson, A. D., Shalling, M., Lindblad, K., Yuan, Q-P., et al. (2001). Genetic survival analysis of age-at-onset of bipolar disorder: Evidence for anticipation of cohort effect in families. *Psychiatric Genetics, 11*, 129–137.

Wienke, A., Arbeev, K., Locatelli, I., & Yashin, A. I. (2003). *A comparison of different correlated frailty models and estimation strategies*. Submitted for publication.

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.

Wienke, A., Lichtenstein, P., & Yashin, A. I. (in press). A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. *Biometrics*.

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

Yashin, A. I., & Iachine, I. A. (1994). Mortality models with application to twin survival data. In J. Halin, W. Karplus, & R. Rimane (Eds.), *CISS — First joint conference of international simulation societies proceedings* (pp. 567–571). Zurich, Switzerland.

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.

Yashin, A. I., & Iachine, I. A. (1997). How frailty models can be used for evaluating longevity limits: Taking advantage of an interdisciplinary approach. *Demography, 34*, 31–48.

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.

Yazdy, M. H., Visscher, P. M., Ducrocq, V., & Thomson, R. (2000). Heritability, reliability of genetic evaluations and response to selection in proportional hazard models. *Journal of Dairy Science, 85*, 1563–1577.

https://doi.org/10.1375/twin.7.2.182 Published online by Cambridge University Press