A Correlated Frailty Model with Long-term Survivors for Estimating the Heritability of Breast Cancer

Isabella Locatelli[1], Alessandro Rosina[2], Paul Lichtenstein[3], Anatoli I. Yashin[4]

[1] Department of Quantitative Methods, University Luigi Bocconi, Milan, Italy

[2] Institute of Population and Geographical studies, Catholic University, Milan, Italy

[3] Karolinska Institut, Stockholm, Sweden

[4] Max Planck Institute for Demographic Research, Rostock, Germany

Corresponding author: Isabella Locatelli
Tel: 0039 02 5391567
Fax: 0039 02 58365630
E-mail: isabella.locatelli@uni-bocconi.it (locatelli@demogr.mpg.de)

Abstract

The aim of this study is to investigate the role of genetics and environment in the susceptibility towards breast cancer. We adopt an interdisciplinary approach combining a bivariate survival model including non observed heterogeneity - a correlated frailty model - with genetic models. These ones enable to decompose the frailty variance into its genetic and environmental components. The methodology is applied to breast cancer data from the Swedish Twin Registry, including information about all the female monozygotic and dizygotic twin pairs born in Sweden between 1886 and 1967. The estimate of heritability in the propensity to develop a breast cancer is obtained taking into account the possibility that a fraction of the population is not susceptible to experience the event. The inferential problem is solved in a Bayesian framework and the numerical work is carried out using MCMC methods. Possible extensions, advantages and limitations of the proposed method are discussed.

1. Introduction

This study is concerned with an attempt to investigate the role played by genetic and environmental factors in determining the individual susceptibility towards breast cancer, and to derive an estimate of heritability in the propensity to develop the disease. This kind of questions have been addressed by different authors working in the fields of medicine, genetics and biostatistics, and answers have often arisen from interdisciplinary considerations (McGue et al. 1993, Yashin and Iachine 1995, Do et al. 2000, Scurrah et al. 2000). In our research we try to estimate heritability of breast cancer via application of a correlated frailty-mixture model to a set of data concerning the onset of a breast cancer in a population of identical (monozygotic) and fraternal (dizygotic) female twins. We furthermore use quantitative genetics techniques in order to provide a genetic interpretation of our estimates.

Frailty-mixture models (Aalen 1988, Hougaard 1994, Longini 1996, Price and Manatunga 2001) are models of survival analysis in which, on the one hand, the possibility of an unsusceptible fraction in the population is accounted for, and, on the other hand, susceptible individuals are allowed to be heterogeneous in their risks to experience the event of interest.

The introduction of an unsusceptible fraction leads to consider the population under study as a mixture of two populations: susceptibles and long-term survivors. The susceptibles will eventually develop the disease before the end of the complete period at risk. The long-term survivors (sometimes called ”cured”, ”immune”, ”stayers”) will survive until the end of the complete period at risk without experiencing the event of interest. We do not adopt here the ’relative’ interpretation of long-term survivors as the set of individuals who have not experienced the event after the end of the ’normal period of risk’ (Wang 1994). Even if the latter definition has the advantage of allowing to identify an a priori set of long-term survivors, we believe in the difficulty to define a ’normal period of risk’ in the case of the onset of a disease.

The individual heterogeneity among the fraction under risk is allowed by using frailty models (frailty-mixture models) where the frailty distribution is a mixture of a discrete and a continuous part. If p is the susceptible proportion of the population, the frailty distribution has point mass at zero with probability $(1 - p)$ and is modeled via a continuous distribution with probability p.

In our study we deal with bivariate (twin) data and, by consequence, we need to specify a dependence structure between the durations in each pair. We do this via application of a correlated frailty-mixture model (Wienke et al. 2003). The typical assumption of the correlated frailty model (Yashin et al. 1995) is that the frailties of the two twin partners are different but eventually correlated, with a correlation coefficient to be estimated by the model. In a correlated frailty-mixture model this parameter represents the correlation between cotwins' frailties in the subpopulation of susceptible individuals. We furthermore assume that the susceptible statuses of the two individuals in a twin pair are independent of each other (Wienke et al. 2003).

The model described above is applied to combined data for monozygotic and dizygotic twins, leading to a different estimate of the correlation coefficient for the two groups of twins. Monozygotic twins share all the genetic endowment, while dizygotic twins, like all siblings, share in average half of the genes. That is why a difference in the estimated correlation between cotwins' propensities to develop a breast cancer for the two groups can give important suggestions in the attempt to estimate the role of genes in the susceptibility towards the disease. Following an approach introduced by Yashin and Iachine (1995), we adopt here quantitative genetic equations (Falconer 1990) in order to obtain an estimate of heritability of breast cancer. We are especially interested to see if such estimate is sensitive to the introduction of the unsusceptible fraction into the model.

2. Statistical methods

2.1 Frailty models

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 till the moment of occurring of an event of interest. The survival function is defined as the probability of the event occurring after a certain time:

$$S(x) = \Pr(X > x) . \tag{1}$$

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

$$\mu(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x} \tag{2}$$

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:

$$S(x) = \exp(H(x)) \tag{3}$$

$$\mu(x) = \frac{f(x)}{S(x)} \tag{4}$$

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), i.e. 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):

$$\mu(x, Z) = Z\mu_0(x). \tag{5}$$

In our study, we are dealing with multivariate frailty models, which were created with the aim to assess for 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. In the case of pairs of individuals, and being (X_{i1}, X_{i2}) the vector of life spans (duration times) for the two individuals from the pair i ($i = 1, ..., n$), the shared frailty model is thus characterized by the two main assumptions:

$$\text{(i)} \quad \mu(x_{ij}, Z_i) = Z_i\mu_0(x_{ij}) \qquad i = 1, ..., n \qquad j = 1, 2 \tag{6}$$

$$\text{(ii)} \quad X_{i1} \mid Z_i \perp X_{i2} \mid Z_i \qquad\qquad i = 1, ..., n \tag{7}$$

where Z_i represents the unobserved heterogeneity term, which is supposed to be 'shared' by the two individuals in the pair i.

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). In a shared frailty model, the 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 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 (Butler et al. 1986, Lillard 1993, Yashin et al. 1995). 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. The correlation coefficient between the two frailties is one of the parameters to be estimated by the model.

A correlated frailty model is thus characterized by:

$$\text{(i)} \quad \mu(x_{ij}, Z_{ij}) = Z_{ij}\mu_0(x_{ij}) \qquad i = 1, ..., n \qquad j = 1, 2 \tag{8}$$

$$\text{(ii)} \quad X_{i1} \mid Z_{i1}, Z_{i2} \perp X_{i2} \mid Z_{i1}, Z_{i2} \qquad\qquad i = 1, ..., n \tag{9}$$

where now we have a specific frailty variable for each individual in the population.

The conditional likelihood of the model is given by:

$$L(x \mid z) = \prod_{i=1} f_{X_{i1}, X_{i2} \mid Z_{i1}, Z_{i2}}(x_{i1}, x_{i2} \mid z_{i1}, z_{i2}) \tag{10}$$

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 (9) allows to rewrite (10) as follows:

$$L(x|z) = \prod_{i=1}\prod_{j=1} f_{X_{ij}|Z_{ij}}(x_{ij}|z_{ij}) \tag{11}$$

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

If some individuals in the population are censored, then their contribution to the likelihood is only given by the survival function. By consequence, the conditional likelihood takes the form:

$$L(x, \delta|z) = \prod_{i=1}\prod_{j=1} \left[f_{X_{ij}|Z_{ij}}(x_{ij}|z_{ij}) \right]^{\delta_{ij}} \left[S_{X_{ij}|Z_{ij}}(x_{ij}|z_{ij}) \right]^{1-\delta_{ij}}, \tag{12}$$

where δ_{ij} is the censoring indicator for the j-th individual in the i-th pair, with $\delta_{ij} = 1$ if the individual experiences the event before the end of the observation period, and $\delta_{ij} = 0$ otherwise.

Considering the relations (see equations (3) and (4)):

$$f_{X|Z}(x|z) = \mu(x, z)\, S_{X|Z}(x|z) \tag{13}$$

$$S_{X|Z}(x|z) = \exp(-zH_0(x)), \tag{14}$$

where $S_{X|Z}$ is the conditional survival function given the frailty variable, and taking into account the multiplicative assumption (8), the expression for the conditional likelihood becomes:

$$L(x, \delta|z) = \prod_{i=1}\prod_{j=1} \left[z_{ij}\mu_0(x_{ij}) \exp(-z_{ij}H_0(x_{ij})) \right]^{\delta_{ij}} \left[\exp(-z_{ij}H_0(x_{ij})) \right]^{1-\delta_{ij}}. \tag{15}$$

To complete the model, it is necessary to make assumptions about the shape of the baseline hazard $\mu_0(x)$ and the form of the bivariate distribution of the vector of frailties $f_{Z_{i1}, Z_{i2}}$.

Shared and correlated frailty models have been estimated both parametrically and semiparametrically. The most adopted parametrical hypothesis is a 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, Do et al. 2000, Visscher et al. 2001) or (piecewise) exponential (Xue and 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 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 lognormal distributed (Korsgaard et al. 1998, Spiegelhalter et al. 1996, Xue and Ding 1999, Ripatti and Palmgren 2000, Do et al. 2000, Scurrah et al. 2000). The lognormal 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 lognormal 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.

2.2 Mixture and frailty-mixture models

Most approaches to the analysis of duration data implicitly assume that all individuals in the study population will eventually experience the event if followed-up for a sufficiently long time. This means that all individuals are susceptible to the event. However, in many situations, it is more reasonable to allow for the possibility that a fraction of the population will never experience the event. These situations can arise when one is interested in the onset or in the recurrence of a disease (Farewell et al. 1977, Langlands et al. 1979, Maller and Zhou 1995, Price and Manatunga 2001, Wienke et al. 2003) or in the case of toxicological experiments (Farewell 1982, Kuk and Chen 1992). More in general, the existence of an unsusceptible fraction in the study population should be taken into account in the case of the presence of a big number of censored observations and when the empirical survival function seems to level off far from the zero line. This can also happen for many socio-demographic phenomena like the contraceptive use (Wang 1994, Wang and Murphy 1997) and the birth of a child (Yamaguchi and Ferguson 1995, Li and Choe 1997, McDonald and Rosina 2001).

Mixture models have been created in order to allow for the existence of an unsusceptible fraction in the study population. The first formulation is due to Farewell (1977) who introduced into the model a binary variable Y_i taking value $Y_i = 1$ if the i-th individual in the population is susceptible to the event of interest and $Y_i = 0$ otherwise. The probability $p_i = \Pr(Y_i = 1)$ is assumed to be related to a set of individual characteristics C_i by means of a logistic relationship:

$$p_i = \frac{\exp\left(\beta^T C_i\right)}{1 + \exp\left(\beta^T C_i\right)}, \tag{16}$$

where β is a vector of coefficients. Conditionally on $Y_i = 1$, an exponential model is defined on the duration time X_i for the i-th individual:

$$S(x_i|Y_i = 1) = \exp(-\lambda x_i). \tag{17}$$

In this model the conditional hazard is constant and given by;

$$\mu(x_i|Y_i = 1) = \lambda. \tag{18}$$

Other shapes of the conditional hazard have been specified later (Farewell 1982, Kuk and Chen 1992, Wienke et al. 2003) and the effect on the time of failing of the individual caracteristics C_i has been introduced, following the multiplicative assumption:

$$\mu(x_i|Y_i = 1) = \mu_0(x_i|Y_i = 1)\exp\left(\gamma^T C_i\right), \tag{19}$$

where $\mu_0(x_i|Y_i = 1)$ is the baseline conditional hazard function, and γ are coefficients describing the effect of covariates on the risk of experiencing the event, for those individuals who are susceptible. In the simple case of the exponential model, $\mu_0(x_i|Y_i = 1) = \lambda$.

In order to give the likelihood function for a mixture model, let us start from the general expression of the likelihood function in a survival model, with censored observations (see also equation (12)):

$$L(x,\delta) = \prod_{i=1} f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}. \tag{20}$$

Indicating with $f(x|Y=1)$ and $f(x|Y=0)$ the density functions for the susceptible and the unsusceptible fraction, the marginal density is the result of the following mixture:

$$f(x) = pf(x|Y=1) + (1-p) f(x|Y=0). \tag{21}$$

Equivalently, the marginal survival function is given by:

$$S(x) = pS(x|Y=1) + (1-p) S(x|Y=0). \tag{22}$$

Giving that the long-term survivors will never experience the event, their conditional density function is equal to zero and their conditional survival function is equal to one. Thus equations (21) and (22) can be simplified as follows:

$$f(x) = pf(x|Y=1) \tag{23}$$

$$S(x) = pS(x|Y=1) + (1-p) \tag{24}$$

Substituting (23) and (24) in equation (20), and considering that:

$$f(x_i|Y_i=1) = \mu(x_i|Y_i=1) S(x_i|Y_i=1),$$

where $S(x_i|Y_i=1) = \exp\left[-\int_0^{x_i} \mu(t|Y_i=1)\,dt\right]$ and $\mu(t|Y_i=1)$ is given by equation (19), we obtain the following expression for the likelihood function:

$$\begin{aligned} L(x,\delta) &= \prod_{i=1} [p_i f(x_i|Y_i=1)]^{\delta_i} [(1-p_i) + p_i S(x_i|Y_i=1)]^{1-\delta_i} = \\ &= \prod_{i=1} \left[p_i \mu_0(x_i|Y_i=1) \exp\left(\gamma^T C_i\right) \exp\left(-\exp\left(\gamma^T C_i\right) \int_0^{x_i} \mu_0(t|Y_i=1)\,dt \right) \right]^{\delta_i} \\ &\quad \left[(1-p_i) + p_i \exp\left(-\exp\left(\gamma^T C_i\right) \int_0^{x_i} \mu_0(t|Y_i=1)\,dt \right) \right]^{1-\delta_i}. \end{aligned} \tag{25}$$

In the above model the interest is not in estimating the size of the unsusceptible fraction but in estimating the separate effect of the covariates on the overal risk of experiencing the event (the probability of being susceptible) and on the time of failing for the susceptible fraction of the population (Farewell et al. 1977).

The unobserved heterogeneity has been introduced in traditional mixture models during the last fifteen years, giving place to so-called frailty-mixture models. Only a portion of heterogeneity is explainable in terms of observed covariates; there remains a degree of heterogeneity induced by unobserved risk factors. Failing to account for unobserved heterogeneity between individuals may lead to distorted results. In a frailty-mixture model the frailty distribution is a mixture of a discrete and a continuous part: the frailty distribution for the i-th individual in the population has point mass at zero with probability $(1 - p_i)$ and is a continuous distribution with probability p_i. For those individuals who are susceptible, frailty acts multiplicatively on the baseline hazard, giving place to the following expression for the conditional hazard function (see equation(19)):

$$\mu\left(x_i|Y_i = 1, Z_i\right) = Z_i\mu_0\left(x_i|Y_i = 1\right)\exp\left(\gamma^T C_i\right), \tag{26}$$

where with Z_i we indicate the unobserved heterogeneity term. The likelihood function, which is now conditional to the frailty variables, takes the form:

$$L\left(x, \delta|z\right) = \prod_{i=1} \left[p_i z_i\mu_0\left(x_i|Y_i = 1\right)\exp\left(\gamma^T C_i\right)\exp\left(-z_i\exp\left(\gamma^T C_i\right)\int_0^{x_i}\mu_0\left(t|Y_i = 1\right)dt\right)\right]^{\delta_i}$$
$$\cdot\left[(1 - p_i) + p_i\exp\left(-z_i\exp\left(\gamma^T C_i\right)\int_0^{x_i}\mu_0\left(t|Y_i = 1\right)dt\right)\right]^{1-\delta_i}. \tag{27}$$

All considerations made in the Section 2.1 about the shape of the baseline hazard and the distribution of the frailty variable in a frailty model are still valid here.

Price and Mantunga (2001) gave a good introduction into this area and applied leukemia remission data to different mixture, frailty and frailty-mixture models. They conclude that frailty models are useful in modeling data with an unsusceptible fraction. McDonald and Rosina (2001) propose a mixture model that combines a discrete-time survival model (with constant baseline hazard) with a logistic regression model for the probability of never experiencing the event of interest. They also introduce a non-observed heterogeneity term (frailty) in their discrete-time event-history model. Chatterjee and Shih (2001) give an extension in the bivariate setting, estimating a shared frailty model with a cure fraction. Wienke et al. (2003) provide a further generalization with a correlated gamma frailty model for those who are susceptible to experience the event.

2.3 A lognormal correlated frailty-mixture model

As we have pointed out before, in our study we deal with bivariate (twin) data and, by consequence, we need to specify a dependence structure between the durations in each pair. We do this via application of a correlated frailty-mixture model. The typical assumptions of the correlated frailty model (Yashin et al. 1995) is that the frailties of the two twin partners are different but eventually correlated, and the durations of the two individuals in a pair are conditionally independent given the frailty variables (see equations (8) and (9)). We furthermore assume that the susceptible statuses of the two individuals in a twin pair are independent of each other. Wienke et al. (2003) also considered a model relaxing the restriction of independence, but they showed that the more complicated mixture model without the independence assumption does not introduce a significant improvement. Finally, being especially interested in estimating the size of the unsusceptible fraction and the correlation between the frailty variables of twin partners, we do not introduce covariate information into the model.

The likelihood function of the correlated frailty-mixture model can thus been seen as a generalization of (15), taking into account the unsusceptible fraction, or an extension in the bivariate setting of equation (27), which describes the likelihood in the case of an univariate frailty-mixture model:

$$L(x,\delta|z) = \prod_{i=1}\prod_{j=1}^{2}\left[p z_{ij}\mu_0(x_{ij}|Y_{ij}=1)\exp\left(-z_{ij}\int_0^{x_{ij}}\mu_0(t|Y_{ij}=1)\,dt\right)\right]^{\delta_{ij}} \left[(1-p)+p\exp\left(-z_{ij}\int_0^{x_{ij}}\mu_0(t|Y_{ij}=1)\,dt\right)\right]^{1-\delta_{ij}} \qquad (28)$$

where now the probability of being susceptible p does not depend on individual caracteristics.

In our study we specified a Gompertz conditional baseline hazard:

$$\mu_0(x|Y_{ij}=1) = a\exp(bx), \qquad (29)$$

and the vector of frailties is assumed to follow a bivariate log-normal distribution. This one is adopted because of its large flexibility in multivariate modeling, 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 here 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}. This assumption differs from the one generally made in the context of correlated log-normal frailty models. Usually in fact the restriction is on the logarithm of the frailty variable, whose mean is assumed to be equal to zero (Korsgaard et al. 1998, Spiegelhalter et al. 1996, Xue and Ding 1999, Ripatti and Palmgren 2000, Do et al. 2000, Scurrah et al. 2000). This hypothesis does not imply that the average frailty in the population is equal to one ($E(\log(Z)) \neq \log(E(Z))$), as originally assumed in the first formulations of frailty models. Thus, in our case, the estimated variance and correlation refer to the frailty variable itself, instead of to its logarithm.

Finally, we 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.

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

$$\begin{bmatrix} Z_{i1} \\ Z_{i2} \end{bmatrix} \sim LogN\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right) \qquad i = 1,...,n \tag{30}$$

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

$$\begin{bmatrix} W_{i1} \\ W_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} -\frac{1}{2}\log\left(\sigma^2+1\right) \\ -\frac{1}{2}\log\left(\sigma^2+1\right) \end{bmatrix}, \begin{bmatrix} \log\left(\sigma^2+1\right) & \log\left[\rho\sigma^2+1\right] \\ \log\left[\rho\sigma^2+1\right] & \log\left(\sigma^2+1\right) \end{bmatrix}\right) \qquad i = 1,...,n \tag{31}$$

with N denoting the bivariate normal distribution.

2.4 Quantitative genetics models

The lognormal correlated frailty-mixture model described in Section 2.3 will be applied (see Section 4) to combined data for monozygotic and dizygotic twins, leading to distinct estimates of the correlation coefficient ρ for the two groups of twins. We will refer to these two correlation estimates with ρ_M and ρ_D, respectively for monozygotic and dizygotic twins. As we have already pointed out in Section 1, a difference in the estimated correlation between cotwins' frailties for the two groups of twins can give important suggestions in the attempt to estimate the role of genes in the susceptibility towards the disease. In particular, when $\rho_M > \rho_D$, one can say that, according to the model, individuals who are more similar from a genetic point of view - monozygotic twins - also present a larger correlation in their propensities to develop a breast cancer. This result is generally interpreted as an evidence of the role played by genetic factors in determining the individual susceptibility towards a disease. Following an approach introduced by Yashin and Iachine (1995), we adopt here quantitative genetic equations (Falconer 1990) in order to quantify the role of genetics and environment in determining the propensity to develop a breast cancer, and in order to give an estimate of heritability of the disease.

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. The interdisciplinary approach introduced by Yashin and Iachine (1995) consists in identifying the phenotype with the frailty variable (Z).

Let the frailty be represented by:

$$Z = A + D + I + C + E \tag{32}$$

where A represents 'additive genetic' effects, D corresponds to 'dominance genetic' effects, I denotes 'epistatic genetic' effects, C and E stand for 'common environmental' and 'uncommon environmental' effects, respectively. All factors are assumed to be independent. Equation (32) and the independence assumption lead to an additive decomposition of the frailty variance and of the correlation coefficient between cotwins' frailty:

$$1 = a^2 + d^2 + i^2 + c^2 + e^2 \tag{33}$$

$$\rho = \rho_A a^2 + \rho_D d^2 + \rho_I i^2 + \rho_C c^2 + \rho_E e^2 \tag{34}$$

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 (A, D, I, C and E), and ρ_k ($k = A, D, I, C, E$) represent correlations between respective components within a twin pair.

Standard assumptions of quantitative genetics models specify different values of ρ_k ($k = A, D, I, C, E$) for monozygotic and dizygotic twins. In the case of monozygotic twins $\rho_k = 1$, $k = A, D, I, C$ and $\rho_E = 0$, while for dizygotic twins $\rho_A = 0.5$, $\rho_D = 0.25$, $\rho_I = m$, $\rho_C = 1$, $\rho_E = 0$ and $0 \le m \le 0.25$ is an unknown parameter (Falconer 1990).

Not all parameters of the genetic decomposition of frailty can be estimated simultaneously. The model in fact reduces to three equations (two relationships (34) for monozygotic and dizygotic twins and one constrain (33)) allowing us to estimate no more than three parameters at the same time (more components could be introduced if data about more then two family members were available). Thus, different genetic models can be considered according to different choices on the decomposition of frailty (see for example Yashin and Iachine, 1995).

In particular, the following systems of equations:

$$\begin{cases} \rho_M = a^2 + c^2 \\ \rho_D = 0.5a^2 + c^2 \\ 1 = a^2 + c^2 + e^2 \end{cases} \quad \begin{cases} \rho_M = a^2 \\ \rho_D = 0.5a^2 \\ 1 = a^2 + e^2 \end{cases} \quad \begin{cases} \rho_M = a^2 + d^2 \\ \rho_D = 0.5a^2 + 0.25d^2 \\ 1 = a^2 + d^2 + e^2 \end{cases} \tag{35}$$

correspond to the ACE, AE and ADE model, respectively. Inverting the systems, it is easy to see how, for each model, it is possible to calculate an estimate of genetic parameters from the estimated correlations ρ_M and ρ_D:

$$\begin{cases} a^2 = 2(\rho_M - \rho_D) \\ c^2 = \rho_M - a^2 \\ e^2 = 1 - a^2 - c^2 \end{cases} \quad \begin{cases} a^2 = \rho_M = 2\rho_D \\ e^2 = 1 - a^2 \end{cases} \quad \begin{cases} d^2 = 2(\rho_M - 2\rho_D) \\ a^2 = \rho_M - d^2 \\ e^2 = 1 - a^2 - d^2 \end{cases} \tag{36}$$

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

3 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, categorized according to the censoring status. The 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. As we can see, about 4,3% of the women involved in the study developed a breast cancer. Very similar proportions are registered for monoziygotic and dizygotic twins.

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).

4. Results

In this study, Bayesian Markov Chain Monte Carlo (MCMC) methods have been adopted in order to estimate the lognormal correlated frailty-mixture model described in Section 2.3. In the Bayesian framework the model is seen as a hierarchical model. The likelihood function (28) represents the first level; at the second level we have parameters characterizing the likelihood, i.e. the susceptible fraction p, the frailty variables Z_{ij} (i = $1, ..., n$; j = $1, 2$) and the Gompertz parameters a and b. Each one of these parameters is supposed to follow some 'prior' distribution; in particular, the distribution of the vector of frailties $[Z_{i1}, Z_{i2}]^T$ is a lognormal distribution with vector of means equal to one, correlation ρ and common variance σ^2 (equation (30)). The other parameters are supposed to follow a 'noninformative' distribution, that is a distribution which is flat in the reasonable range of values of each parameter. Finally, at the third level of the model we have the so-called 'hyperparameters', which are in this case parameters σ^2 and ρ of the frailty distribution. A Bayesian hierarchical model, as the one described above, can not be estimated using traditional Bayesian techniques, which would require the calculation of the expected value of the marginal distribution of each parameter, giving observed data (marginal 'posterior' distribution). Markov Chain Monte Carlo (MCMC) methods have recently been introduced in order to give a numerical solution to Bayesian models that can not be estimated analytically. They consist in generating a set of Markov chains whose joint stationary distribution corresponds to the joint posterior of the model. Synthetic values (mean, median) of the marginal Markov chain of each parameter can be considered as good approximations of the Bayesian estimate.

In our study calculations are performed within the software WinBugs 1.4 (Spiegelhalter et al. 1999). This one is a Bayesian software allowing to estimate hierarchical models with the help of MCMC algorithms.

Results of the lognormal correlated frailty-mixture model (MODEL 2) are summarized in Table 2, and compared with the ones obtained without taking into account the unsusceptible fraction, that is with the constraint p = 1 (MODEL 1). For more details about the lognormal correlated frailty model described in the first row of Table 2, see Locatelli et al. (2004).

As we can see from the table, when the unsusceptible fraction is taken into account, the estimate of the frailty variance gets smaller, while the estimates of correlation between cotwins' frailties get larger for both monozygotic and dizygotic twins. In the model with long-term survivors (MODEL 2) the susceptible fraction is estimated around 0.11, which means that only 10% of the population will experience the event before the end of the complete period of risk. Thus, according to the model, besides the proportion of women (4%) who did experience the event (a priori susceptible fraction), another 6% would have developed a breast cancer, if an early censoring had not occurred. Our estimate of 0.11 is smaller than the one obtained by Wienke et al. (2003) with the gamma assumption. Using a subset of the data that are the object of our application (the 'old cohort' of the Swedish Twin Register) they obtain an estimate of the susceptible fraction around 17%. Nevertheless an estimate of 0.11 is perfectly in the range of the estimated probabilities of developing a breast cancer, obtained by Farewell (1977) for different combinations of four risk factors.If none of the risk factors is present, such probability is estimated around 0.015; if all are present, the estimate increases to 0.272. The mean of the 16 estimated probabilities is around 0.1.

As we have pointed out before (Section 2.4), results in terms of correlation between frailties in pairs of monozygotic and dizygotic twins can be taken as starting point for further considerations about the influence of genetic factors on the susceptibility towards the event of interest. The larger estimate of ρ for monozygotic twins (Table 2) provides an evidence of the importance of genetics in determining the individual propensity to develop the disease. We consider here three genetic models presented in Section 2.4: the ACE, the AE and the ADE model. If we look to systems (36), it is possible to recognize that both the AE and the ADE models are not compatibles with our results in term of correlation estimates (Table 2, MODEL 2). In the AE model, parameter a^2 is at the same time equal to the correlation for monozygotic twins (ρ_M) and two times the correlation for dizygotic twins (ρ_D). But in our case $\rho_M = 0.78 \neq 2 \cdot 0.56 = 1.12$. If we consider the ADE model, we get $d^2 = 2(\rho_M - 2\rho_D) = 2(0.78 - 1.12) < 0$, which is of course impossible. The model that seems to better represent results of Table 2 is a model including additive genetic, common environmental and uncommon environmental effects (ACE). The estimates of parameters of the ACE model can be found in Table 3 (MODEL 2).

In the simpler model without the unsusceptible fraction (MODEL 1), analogous considerations lead to choose an ADE model (for details see Locatelli et al 2004). From Table 3 we can observe that, when the unsusceptible fraction is included, the estimate of heritability grows

	a	b	σ^2	ρ_M	ρ_D	p	
MODEL 1 (p = 1)	2.54E-5 (3.24E-6)	0.0725 (0.0025)	45.19 (17.04)	0.3107 (0.0456)	0.1044 (0.0967)		
MODEL 2	2.17E-5 (2.17E-5)	0.13 (0.005)	33.73 (10.42)	0.7859 (0.122)	0.5769 (0.1459)	0.1099 (0.0066)	

Table 2: Results of a correlated frailty model (MODEL 1) and a correlated frailty-mixture model (MODEL 2) applied to Swedish twin breast cancer data. Convergence achieved after 50,000 iterations

	a	b	σ^2	a^2	d^2	c^2	e^2	p
MODEL 1 (p = 1)	2.52E-5 (3.16E-6)	0.0719 (0.002)	48.30 (16.7)	0.1273 (0.086)	0.1491 (0.100)		0.7239 (0.084)	
MODEL 2	2.10E-5 (5.08E-6)	0.1364 (0.006)	38.07 (13.64)	0.5268 (0.22)		0.3039 (0.2179)	0.1693 (0.051)	0.11 (0.006)

Table 3: Results of three genetic models applied to Swedish twin breast cancer data. Convergence achieved after 50,000 iterations

from less than 0.3 (a^2 + d^2 in MODEL 1) to more than 0.5 (a^2 in MODEL 2).

5. Discussion

In the present paper, a correlated frailty-mixture model (Section 2.3) has been adopted to analyze 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 modeling and especially when we are interested in introducing a correlation between frailties, as in the case of the correlated frailty model. Mixture models (Section 2.2) allow to introduce in traditional survival models the possibility that a fraction of the study population is not susceptible to the event of interest (long-term survivors). With the frailty component, susceptible individuals are allowed to be heterogeneous in their propensity to experience the event. When the model is applied to twin data (Section 3), very interesting interpretations can be given to the results, by applying quantitative genetics equations (Section 2.4). One of the most important concerns of the paper is to verify if the introduction of long-term survivors in the model leads to different estimates of the correlation between the frailty variables within a twin pair and to different heritability estimates. We found that the individual heterogeneity in the susceptibility towards breast cancer is extremely high and the correlation between frailties in a twin pair is larger for monozygotic than for dizygotic twins. Individuals who are more similar from a genetic point of view, monozygotic twins, also present a larger connection in terms of frailty towards breast cancer. This finding provides an evidence of a genetic influence on the breast cancer propensity. In fact, if genetic factors do influence the individual susceptibility towards breast cancer, we 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. When the unsusceptible fraction is taken into account, the heterogeneity estimate goes down while the estimated correlations between cotwins' frailties increase for both monozygotic and dizygotic twins, Table 2). Thus, when we include the possibility that a fraction of the population is not susceptible to experience the event, remaining individuals are less heterogeneous in their propensity to develop the disease and, within a pair, the two twins show a higher correlation in their susceptibility towards breast cancer. Wienke et al. (2003) obtain very similar results, with a different assumption on the frailty distribution (gamma instead of lognormal), working on a subset of the data used in our application (the 'old' cohort of the Swedish Twin Register). They also adopt a maximum likelihood, instead of a MCMC estimation procedure.

In our study we also investigate the effect of introducing the unsusceptible fraction on the genetic decomposition of frailty. As we have pointed out before (Section 2.4), with the help of quantitative genetics equations, it is possible to estimate the extent of the genetics-related part of the individual propensity to experience the event. Such estimate, which is often referred to as the 'heritability' estimate, is around 30% when the unsusceptible fraction is not taken into account. It increases to 50% including long-term survivors into the model.

References

Aalen, O.O. (1988) Heterogeneity in Survival Analysis. Statistics in Medicine 7, 1121 - 1137.

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

Chatterjee, N., Shih, J. (2001) A Bivariate Cure-Mixture Approach for Modeling Familial Association in Diseases. Biometrics 57, 779 - 786.

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, B 34, 187 - 220.

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

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

Farewell, V.T. (1977) A model for a binary variable with time-censored observations. Biometrika 64, 43 - 46.

Farewell, V.T. (1982) The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. Biometrics 38, 1041 - 1046.

Farewell, V.T., Math, B., Math, M. (1977) The combined effects of breast cancer risk factors. Cancer 40, 931 - 936.

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

Hougaard, P., Myglegard, P., Borch-Johnsen, K. (1994) Heterogeneity models of disease susceptibility, with application in diabetic nephropathy. Biometrics, 50:1178 - 1188.

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

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

Kuk, A.Y.C., Chen, C-H (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika 79, 531 - 41.

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.

Langlands, A.O., Pocock, S.J., Kerr, G.R., Gore, S.M. (1979) Long-term survival of patients with breast cancer: a study of the curability of the disease. Medical Journal 2, 1247 - 1251.

Li, L., Choe, M.K. (1997) A mixture model for duration data: analysis of second births in China. Demography 24, 189-197.

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.

Locatelli, I., Lichtenstein, P., Yashin, I.A. (2004) The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data (accepted by Twin Research).

Longini, I.M., Halloran, M.E. (1996) A Frailty Mixture Model for Estimating Vaccine E¢cacy. Applied Statistics 45, 165 - 173.

Testing for the Presence of Immune or Cured Individuals in Censored Survival Data (1995) Biometrics 51, 1197 - 1205.

McDonald, J.W., Rosina, A. (2001) Mixture modeling of recurrent event times with long-term survivors: Analysis of Hutterite birth intervals. Statistical Methods & Applications 10, 257 - 272.

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.

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

Price, D.L., Manatunga, A.K. (2001) Modeling survival data with a cured fraction using frailty models. Statistics in Medicine 20, 1515 - 1527.

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

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.

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.

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

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

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., 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.

Visscher, P.M., Yazdy, M.H., Jackson, A.D., Shalling, M., Lindblad, K., Yuan, Q-P., Porteous, D., Muir, W.J., Blackwood, D.H.R. (2001) Genetic survival analysis of age-at-onset of bipolar disorder: evidence for anticipation of cohort effect in families. Psychiatric Genetics 11, 129 - 137.

Wang, D. (1994) Contraceptive use in China. PhD Thesis, University of Southampton, Faculty of Social Sciences.

Wang, D., Murphy, M. (1997) The Use of Mixture Model for the Analysis of Contraceptive Use Duration with Long-term Users, London School of Economics.

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. (2003) A bivariate frailty model with a cure fraction for modeling familial correlations in diseases (accepted by 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.

Yamaguchi, K., Ferguson, L. (1995) The Stopping and Spacing of Childbirths and Their Birth-History Predictors: Rational-Choice Theory and Event-History Analysis. American Sociological Review 60, 272-298.

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.

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., 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.