STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2511–2526 (DOI: 10.1002/sim.1197)

A multistate model for the genetic analysis of the ageing process[‡]

Nicole Giard[1,*,†], Paul Lichtenstein[2] and Anatoli I. Yashin[1]

[1] *Max Planck Institute for Demographic Research, Doberaner Str. 114, 18057 Rostock, Germany*
[2] *Department of Medical Epidemiology, Box 281, 17177 Stockholm, Sweden*

SUMMARY

In this paper a multivariate frailty model is suggested that can be used in the genetic analysis of the ageing process as a whole, simplified to consisting of the states 'healthy', 'disabled' and 'deceased'. The model allows us to evaluate simultaneously the relative magnitude of genetic and environmental influences on frailty variables corresponding to the period of good health and to the life span. The frailty variables can be interpreted as susceptibility to illness or death. The model can be applied to data on groups of related individuals (twins, siblings, a litter). One of the major advantages of this model is that it allows one to include groups of individuals where some or all members of the group are already deceased at the time of observation. The current health status of the living individuals and the exact life span of individuals who are already deceased is the only information necessary for the application of the model. Questions concerning the identifiability of the model based on current health status data and estimation strategies are discussed in the context of specifying the model for twins. Finally, the results of a sample analysis of twin data on prostate cancer are presented. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: frailty model; heritability in lifespan; heterogeneous population; ageing

1. INTRODUCTION

Ageing is a very complex process which is influenced by a lot of factors such as health and living circumstances. There is much variation among individuals in this process. To analyse the relative importance of different factors such as genes and the environment, an interdisciplinary approach can be helpful.

On the one hand, demographers investigate empirically the mortality pattern of populations, recent trends in mortality, and life expectancy. On the other hand, genetic epidemiologists analyse statistically the association between genes and certain diseases.

[*] Correspondence to: Nicole Giard, Max Planck Institute for Demographic Research, Doberaner Str. 114, 18057 Rostock, Germany

[†] E-mail: giard@demogr.mpg.de

[‡] Presented at the International Society of Clinical Biostatistics, Twentieth International Meeting, Heidelberg, Germany, September 1999.

Received October 1999
Accepted October 2001

However, to study the role of genes in the ageing process one needs to deal with both mortality data and health information, since one can look at the process from different points of view. One can be interested in a quantitative description of ageing, which means studying its duration or life expectancy, or one can ask about the quality of this process and thus treat questions about health and disability.

An approach is needed that merges the methods of demography with those of genetic epidemiology. Moreover, this approach must take into account the structure of the data that are available. There is both demographic information about life span and information about different diseases and/or health status. To address questions about the influence of genes one usually uses data on relatives, such as twins, for example.

From a mathematical point of view it follows that one must deal with:

(i) methods of survival analysis, since life span information is usually censored;
(ii) methods of quantitative genetics to estimate genetic parameters such as heritability;
(iii) a model of the ageing process of related individuals.

For the genetic analysis of life span a bivariate correlated frailty model was successfully applied on Danish twin survival data [1]. This model combines the ideas of survival analysis and demography with those of genetic epidemiology. In this approach it is not the genetic parameters of life span that are estimated but those of an unobserved variable called frailty. Frailty represents the individual susceptibility to disease and death. The estimated heritability of frailty was 50 per cent. The properties of this model are discussed in Yashin *et al.* [2].

In order to generalize the bivariate correlated frailty model for a description of the entire ageing process, we introduce in this paper an additional frailty variable for each individual, which represents susceptibility to disease or disability. This leads in the case of twins to a four-dimensional correlated frailty model, thereby simplifying the ageing process to a process consisting of the three states healthy, ill or disabled and deceased.

The model for the ageing process as a whole that we introduce in this paper allows for a simultaneous consideration of genetic and environmental influences on 'disease susceptibility' and 'longevity'. The advantage over a separate analysis is the type of data one can use in the new model. It is possible to include groups of individuals with some or all members deceased at the time of observation. In an analysis that only examines disease status such groups would be excluded.

We illustrate our methods with an example in Section 4. It uses data on prostate cancer incidence in male Swedish twins. We wish to establish the relative importance of genetic and environmental influences on the development of prostate cancer and on life span in general.

2. FRAILTY MODELS AND THEIR APPLICATIONS

The notion of frailty as a measure of general susceptibility to all causes of death was introduced to describe mortality in heterogeneous populations [3]. If T is the life span then the conditional survival function of T given frailty Z is

$$S(x|Z) = \exp\left\{-\int_0^x \mu(u,Z)\,\mathrm{d}u\right\}$$

where μ is the conditional hazard function. Usually a proportional hazard model $\mu(x,Z) = Z\mu_0(x)$ is used. Then the unconditional survival function of T is

$$S(x) = p\left(\int_0^x \mu_0(u)\,\mathrm{d}u\right) = p(H(x))$$

where $H(x)$ is the cumulative hazard function and p the Laplace transform of Z.

To make such a univariate frailty model identifiable one has to make assumptions concerning the parametric structure of the underlying hazard rate μ_0. A bivariate correlated gamma frailty model does not have this disadvantage [2].

If T_1 and T_2 are the life spans of two related persons and $H_i(x) = \int_0^x \mu_i(u)\,\mathrm{d}u$ and $Z_i,\ i = 1,2$, are the corresponding cumulative hazard rates and frailty variables, respectively, then one gets for the conditional bivariate survival function of (T_1, T_2) given (Z_1, Z_2)

$$S(x_1, x_2 | Z_1, Z_2) = \exp\{-Z_1 H_1(x_1) - Z_2 H_2(x_2)\}$$

if one assumes a proportional hazard model and conditional independence of the life spans given the frailty variables.

Yashin *et al.* [2] use an additive decomposition of the frailty variables into the sum of independent gamma distributed variables to construct a bivariate frailty distribution. The resulting bivariate unconditional survival function can be represented in two different ways:

$$\begin{aligned} S(x,y) &= [1+\sigma_1^2 H_1(x)]^{\frac{\varrho}{\sigma_1\sigma_2}-\frac{1}{\sigma_1^2}}[1+\sigma_2^2 H_2(y)]^{\frac{\varrho}{\sigma_1\sigma_2}-\frac{1}{\sigma_2^2}}[1+\sigma_1^2 H_1(x)+\sigma_2^2 H_2(y)]^{-\frac{\varrho}{\sigma_1\sigma_2}} \\ &= [S_1(x)]^{1-\frac{\sigma_1}{\sigma_2}\varrho}[S_2(y)]^{1-\frac{\sigma_2}{\sigma_1}\varrho}[S_1^{-\sigma_1^2}(x)+S_2^{-\sigma_2^2}(y)-1]^{-\frac{\varrho}{\sigma_1\sigma_2}} \end{aligned}$$

where S_1 and S_2 are the marginal univariate survival functions. σ_i^2 is the variance of Z_i and ϱ the correlation between the frailty variables.

The second representation of the survival function makes possible a semi-parametric approach to the estimation of the parameters of the bivariate frailty distribution if one uses a non-parametric estimate for the marginal univariate survival functions [2].

One of the main advantages of the correlated frailty model is that it allows for a joint analysis of data on monozygotic and dizygotic twins and therefore an estimation of the heritability of frailty. For such an analysis one has to assume that the twins in a pair and monozygotic and dizygotic twin individuals all have the same mortality pattern, that is, equal frailty variances and univariate survival functions. This assumption can be tested, for example, with a likelihood ratio test. It is also possible to compare different genetic models for frailty [4].

According to the Akaike information criterion (AIC), a genetic model consisting only of additive genetic and non-shared environmental components was the best fitting in the application of such an analysis to Danish twin data [1].

3. THE FOUR-DIMENSIONAL MODEL

3.1. Construction of the model

Let us now look at the ageing process as a whole and simplify it to a process consisting of the three states ‘healthy’, ‘ill’, and ‘deceased’. The ageing of an individual can then be

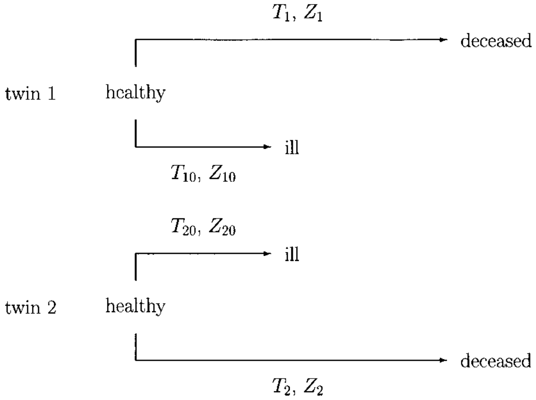

Figure 1. Four-dimensional correlated frailty model for a twin pair.

described by the time he stays in a 'healthy' state and the life span. If there is an underlying unobservable frailty variable corresponding to each of these times, then one represents factors that influence susceptibility to disease and the other represents factors that change individual chances of survival.

The ageing of a twin pair can be described in this way by a four-dimensional correlated frailty model. Figure 1 illustrates such a model. T_{i0}, $i=1,2$, is the time spent in a healthy state and T_i, $i=1,2$, the life span of twin i. The frailty variables Z_{i0} and Z_i, $i=1,2$, correspond to these time periods, respectively. We assumed that it makes no difference whether a twin is the first or the second in a pair. Therefore the bivariate distributions of (T_{10},T_1) and (T_{20},T_2) are equal.

Let S_0 and S_L denote the univariate survival functions of T_{i0}, $i=1,2$, and T_i, $i=1,2$, respectively. $S_L(x)$ is the probability of surviving to age x and $S_0(x)$ is the probability of still being healthy at age x. A proportional hazard model with underlying hazard rate μ_0 and cumulative hazard rate H_0 is assumed for the conditional survival function of T_{i0} given Z_{i0}. The conditional survival function of T_i given Z_i is defined similarly, with underlying hazard rate μ and cumulative hazard rate H:

$$P(T_{i0}>x|Z_{i0}) = \exp\left\{-Z_{i0}\int_0^x \mu_0(u)\,\mathrm{d}u\right\} = \exp\{-Z_{i0}H_0(x)\}$$

$$P(T_i>x|Z_i) = \exp\left\{-Z_i\int_0^x \mu(u)\,\mathrm{d}u\right\} = \exp\{-Z_iH(x)\}$$

It is assumed that the times T_{i0}, $i=1,2$, and T_i, $i=1,2$, are conditionally independent given the frailties:

$$\begin{aligned}&P(T_{10}>x_1, T_{20}>x_2, T_1>y_1, T_2>y_2|Z_{10},Z_{20},Z_1,Z_2)\\ &\quad =P(T_{10}>x_1|Z_{10})P(T_{20}>x_2|Z_{20})P(T_1>y_1|Z_1)P(T_2>y_2|Z_2)\end{aligned}$$

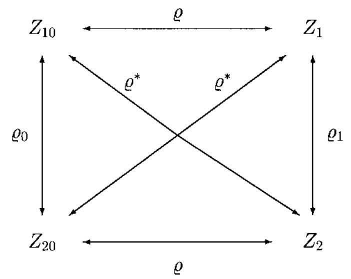

Figure 2. Correlation structure of the four-dimensional frailty distribution.

This assumption means that all genetic influence on the times T_{i0} and T_i is expressed by the frailty variables.

To calculate the unconditional survival function S of $(T_{10}, T_{20}, T_1, T_2)$ one has to specify the distribution of $(Z_{10}, Z_{20}, Z_1, Z_2)$, or more precisely, the Laplace transform of this random vector.

The frailty variables are assumed to be gamma distributed with mean value 1, variances $\sigma_0^2 = \text{var}(Z_{i0})$, $\sigma^2 = \text{var}(Z_i)$ and correlations $\varrho_0 = \text{corr}(Z_{10}, Z_{20})$, $\varrho_1 = \text{corr}(Z_1, Z_2)$, $\varrho = \text{corr}(Z_{i0}, Z_i)$, $\varrho^* = \text{corr}(Z_{10}, Z_2) = \text{corr}(Z_{20}, Z_1)$. Figure 2 illustrates the resulting correlation structure.

To construct the joint distribution of the four frailty variables one can represent them as the sum of independent gamma-distributed random variables. The aim of such a construction is to achieve a wide range of possible dependence structures, that is, a large permissible range for the correlations. The correlations ϱ_0 and ϱ_1 are of special interest for the analysis since with their values for monozygotic and dizygotic twins one can calculate the heritabilities of the frailty in becoming ill and of the frailty in life span.

Each frailty is decomposed into the sum of four independent gamma-distributed random variables. There is one common summand for all four frailty variables, one common summand for each state and one for each twin, and an individual summand for each frailty. Thus

$$Z_{10} = \sigma_0^2(Y_0 + Y_1 + Y_3 + Y_{10}) \quad Z_1 = \sigma^2(Y_0 + Y_1 + Y_4 + Y_{11})$$

$$Z_{20} = \sigma_0^2(Y_0 + Y_2 + Y_3 + Y_{20}) \quad Z_2 = \sigma^2(Y_0 + Y_2 + Y_4 + Y_{21})$$

where the Y_i and Y_{ij} are pairwise independent gamma-distributed random variables with the scale parameter 1 and different shape parameters:

$$Y_1 \sim \Gamma(k_1, 1), \quad Y_2 \sim \Gamma(k_1, 1), \quad Y_i \sim \Gamma(k_i, 1),\ i = 0, 3, 4$$

$$Y_{i0} \sim \Gamma(k_{10}, 1),\ i = 1, 2, \quad Y_{i1} \sim \Gamma(k_{11}, 1),\ i = 1, 2$$

There is a one-to-one correspondence between the shape parameters k_i, k_{ij} and the variances and correlations of the frailty variables. Since the shape parameters must be positive one gets the following constraints on the variances and correlations:

$$
\begin{aligned}
0 < \varrho^* < \varrho < \min\left(\frac{\sigma_0}{\sigma}, \frac{\sigma}{\sigma_0}\right) \\
\varrho^* \frac{\sigma_0}{\sigma} < \varrho_0 < 1 - (\varrho - \varrho^*)\frac{\sigma_0}{\sigma} \\
\varrho^* \frac{\sigma}{\sigma_0} < \varrho_1 < 1 - (\varrho - \varrho^*)\frac{\sigma}{\sigma_0}
\end{aligned}
$$

These constraints can be serious limitations if the variances of the frailties σ_0 and σ differ significantly or if the correlations between the frailties of one person (ϱ) or between the frailties for different states of different persons (ϱ^*) are large. If σ_0 and σ are of the same magnitude and ϱ and ϱ^* are small, then one gets acceptable permissible ranges for the correlations between the frailty variables of the same health state for different twins (ϱ_0, ϱ_1).

Similarly to the bivariate correlated frailty model, the unconditional survival function of $(T_{10}, T_{20}, T_1, T_2)$ can be expressed in two different ways:

$$
\begin{aligned}
S(x_1, x_2, y_1, y_2) &= [S_0(x_1)S_0(x_2)]^{1-\varrho_0-(\varrho-\varrho^*)\frac{\sigma_0}{\sigma}} [S_L(y_1)S_L(y_2)]^{1-\varrho_1-(\varrho-\varrho^*)\frac{\sigma}{\sigma_0}} \\
&\quad \times [(S_0^{-\sigma_0^2}(x_1) + S_L^{-\sigma^2}(y_1) - 1)(S_0^{-\sigma_0^2}(x_2) + S_L^{-\sigma^2}(y_2) - 1)]^{-\frac{\varrho-\varrho^*}{\sigma_0\sigma}} \\
&\quad \times [S_0^{-\sigma_0^2}(x_1) + S_0^{-\sigma_0^2}(x_2) - 1]^{-\frac{\varrho_0}{\sigma_0^2}+\frac{\varrho^*}{\sigma_0\sigma}} [S_L^{-\sigma^2}(y_1) + S_L^{-\sigma^2}(y_2) - 1]^{-\frac{\varrho_1}{\sigma^2}+\frac{\varrho^*}{\sigma_0\sigma}} \\
&\quad \times [S_0^{-\sigma_0^2}(x_1) + S_0^{-\sigma_0^2}(x_2) + S_L^{-\sigma^2}(y_1) + S_L^{-\sigma^2}(y_2) - 3]^{-\frac{\varrho^*}{\sigma_0\sigma}} \\
&= [(1 + \sigma_0^2 H_0(x_1))(1 + \sigma_0^2 H_0(x_2))]^{-\frac{1-\varrho_0}{\sigma_0^2}+\frac{\varrho-\varrho^*}{\sigma_0\sigma}} \\
&\quad \times [(1 + \sigma^2 H(y_1))(1 + \sigma^2 H(y_2))]^{-\frac{1-\varrho_1}{\sigma^2}+\frac{\varrho-\varrho^*}{\sigma_0\sigma}} \\
&\quad \times [(1 + \sigma_0^2 H_0(x_1) + \sigma^2 H(y_1))(1 + \sigma_0^2 H_0(x_2) + \sigma^2 H(y_2))]^{-\frac{\varrho-\varrho^*}{\sigma_0\sigma}} \\
&\quad \times [1 + \sigma_0^2 H_0(x_1) + \sigma_0^2 H_0(x_2)]^{-\frac{\varrho_0}{\sigma_0^2}+\frac{\varrho^*}{\sigma_0\sigma}} [1 + \sigma^2 H(y_1) + \sigma^2 H(y_2)]^{-\frac{\varrho_1}{\sigma^2}+\frac{\varrho^*}{\sigma_0\sigma}} \\
&\quad \times [1 + \sigma_0 H_0(x_1) + \sigma_0^2 H_0(x_2) + \sigma^2 H(y_1) + \sigma^2 H(y_2)]^{-\frac{\varrho^*}{\sigma_0\sigma}} \qquad (1)
\end{aligned}
$$

The first representation could be called semi-parametric since it shows the functional dependence of the survival function S on the marginal univariate survival functions, which may be estimated parametrically or non-parametrically from univariate survival data. The second representation, on the other hand, could be called parametric, since it shows the dependence of the survival function on the underlying hazard rates. These rates cannot be estimated either parametrically or non-parametrically from univariate survival data. The first representation is to be preferred for the statistical analysis of data since it allows us to avoid unjustifiable assumptions about the parametric form of the underlying hazard.

A model for the description of the ageing and survival processes of more than two related individuals can be constructed in a completely analogous manner.

3.2. Identifiability of the model

If one has uncensored observations of the times $(T_{10}, T_{20}, T_1, T_2)$ then the identifiability of the model follows directly from that of the bivariate correlated frailty model, since all marginal bivariate survival functions correspond to this model. The identifiability of the bivariate model has been proven by Iachine and Yashin [5]. To assume that one can get such types of observation in reality is very unrealistic. It can be shown that a more realistic pattern of censored observations is sufficient for the identifiability of the model.

Let us assume that one has carried out a cross-sectional study where one examined the health state of twin pairs. Assume further that it is possible to obtain from a population registry the exact life span of twins of the same birth cohorts as the observed ones who have already died.

If A is the age of a twin pair at the time of observation, T_{i0} is the time spent in a healthy state and T_i is the life span of twin i, then one can observe for every twin in a pair the current health state

$$C_i = \begin{cases} 0 & \text{if the twin is healthy} \quad \Leftrightarrow \quad T_{i0} > A \text{ and } T_i > A \\ 1 & \text{if the twin is ill} \quad \Leftrightarrow \quad T_{i0} \leqslant A \text{ and } T_i > A \\ 2 & \text{if the twin is deceased} \quad \Leftrightarrow \quad T_i \leqslant A \end{cases}$$

and the censored life span

$$X_i = \min(T_i, A) \quad i = 1, 2$$

The observation for a twin pair is the vector (X_1, X_2, C_1, C_2). X_1 and X_2 are positive real numbers, C_1 and C_2 can take the value 0, 1 or 2.

The following proposition holds:

Proposition 3.1
Let A have a positive density function on an interval $[a, b]$ and let it be independent of T_{i0}, $i = 1, 2$, and T_i, $i = 1, 2$. If there exists at least one value x in $[a, b]$ for which it holds that $0 < S_L(x) < 1$, $S_L'(x) < 0$ and $0 < S_0(x) < 1$ and if for all c with $0 < c < 1$ there is at least one $x \in [a, b]$ such that

$$S_L^{-\sigma^2}(x) \neq \frac{S_0^{-\sigma_0^2}(x) - 1}{S_0^{-c\sigma_0^2}(x) - 1} \tag{2}$$

then the model corresponding to (1) is identifiable with the help of the observations (X_1, X_2, C_1, C_2).

This means that the parameters of the model ϱ_0, ϱ_1, ϱ, ϱ^*, σ_0 and σ and all values of the univariate survival functions on $[a, b]$ with $0 < S_0(x) < 1$, $0 < S_L(x) < 1$ and $S_L'(x) < 0$ are uniquely determined.

The following remark shows that condition (2) is not very restrictive. Furthermore, it contains only analytical causes in the proof of Proposition 3.1.

Remark 3.2
If $a=0$, that is, A is distributed on an interval $[0,b]$, and $S_0(x)<1$ for all $x>0$ then (2) is always fulfilled.

Remark 3.3
Since current health status data on two individuals are sufficient for the identifiability of the model, a model that describes the ageing of more than two individuals is obviously also identifiable with the help of current health status data.

A sketch of the proof of Proposition 3.1 is given in the Appendix.

3.3. Estimation strategies

For censored observations in the four-dimensional correlated gamma frailty model as described above, the log-likelihood function can be expressed as a function depending on the parameters of the four-dimensional frailty distribution σ_0, σ, ρ_0, ρ_1, ρ and ρ^* and the values of the univariate survival functions S_0 and S_L. The proof of identifiability shows that the univariate survival functions can be identified non-parametrically, that is, without any parametric specification. This property opens up different possibilities for the estimation of the parameters in the model and also provides us with the opportunity to evaluate the fit of the model.

On the one hand, one can choose a parametric specification for the univariate survival functions.

For the life span of every individual one has censored information where the age at the time of observation A is the censoring variable. On the other hand, therefore, one can calculate the Kaplan–Meier estimator for the univariate survival function S_L from the data. One can use the values of this non-parametric estimator in the log-likelihood function and choose a parametric specification for the function S_0. This approach could be called a partly semi-parametric estimation strategy for the model.

It is important to keep in mind that the standard errors that one would get with traditional methods using this semi-parametric approach do not include the error that is created by using a non-parametric estimator for S_L. Therefore one has to use other methods such as, for example, a bootstrap approach to get standard errors of the parameter estimates.

Using a fully parametric approach has the advantage that there exist well-known theoretical results about the asymptotic distribution of the estimators. Thus standard errors can be derived directly. The disadvantage is the higher number of parameters that have to be estimated simultaneously and, of course, one makes restrictions on the form of the survival function. A semi-parametric approach does not make such restrictions, but it does require additional computational efforts to calculate the non-parametric estimate. To get standard errors for the estimators is not straightforward, and it is usually also computationally demanding.

If the parameter estimates from this semi-parametric approach and a fully parametric approach differ substantially, this can be a hint that the chosen parametric specification for S_L is inappropriate.

3.4. Properties and limitations

The four-dimensional correlated frailty model has several advantages and useful properties. One of its major advantages is that in a scenario like the one described above it is possible to include the information on broken pairs (these are twin pairs where one twin is deceased at the time of observation) and on pairs where both members are already deceased. This increases the sample size and therefore the accuracy of parameter estimates. Traditional methods would ignore the information about such pairs.

The possibility of using a non-parametric estimator for the survival function S_L, thus avoiding any parametric specification of this function, makes the model flexible.

The information that is needed to identify the model is quite easily available. Only the current health status of the individuals has to be known, not the age at onset of disease or disability. It is quite difficult to obtain the latter information for most diseases and sometimes even impossible to define it.

The model enables us to combine data on monozygotic and dizygotic twins in the analysis, thus allowing for the application of methods of quantitative genetics. That is, it is possible to estimate the heritability of the frailty in becoming ill and that of the frailty in dying simultaneously. In addition, one can compare different genetic models in order to explore the nature of genetic influences, that is, to compare additive and non-additive genetic factors or investigate the role of the shared environment.

The analytic form of the likelihood function is known, so traditional estimation methods (for example, maximum likelihood estimation) can be used. Censored and truncated data can be included in the analysis. Although this three-state model is still highly simplified, it provides more complex insight into the ageing process as a whole, since it combines different types of information – medical data about the health status and demographic data about mortality.

It is especially useful to investigate whether the influence of genes on the age at death is a confounder for the disease status relationship in a twin pair. This is an advantage over the traditional method of estimating the correlation in liability to the development of a certain disease.

Of course the model has also disadvantages which may present serious limitations in practical applications. First of all, there are constraints on the parameters of the four-dimensional frailty distribution, especially on the correlations between the frailty variables. This restricts the range of possible dependence structures between the times T_{i0} and T_i, which describe the ageing process. This could make the fitting of the model to real data nearly impossible.

In practice, the estimation of the model parameters can face some problems. The model is quite complex. In addition to the parameters of the four-dimensional frailty distribution, one has to choose a parametric specification for univariate survival functions or underlying hazard rates. The number of parameters is therefore quite large, which can create computational difficulties in the estimation process.

Moreover, the sample size in twin studies is usually relatively small, which will yield large standard errors for the parameter estimates, and the proportion of ill or disabled persons is also often small. This will lead to large standard errors for the parameters belonging to the transition from a healthy to an ill state, thus making the estimation of these parameters difficult. Nevertheless, the model was successfully applied to a data set on Swedish twins (see Section 4). A detailed simulation study could bring some knowledge about the sample size and the proportions of affected individuals that are necessary to get satisfactory estimation results in the four-dimensional correlated frailty model.

Table I. Proportion of pairs with no twin, one twin or two twins affected.

Zygosity	No twin with diagonis	One twin with diagonis	Both twins with diagonis	Total
MZ	1446 87.7%	177 10.7%	26 1.6%	1649 100%
DZ	2602 87.2%	367 12.3%	14 0.5%	2983 100%

4. APPLICATION TO PROSTATE CANCER DATA

4.1. Data

We tested the applicability of the four-dimensional correlated frailty model with data on prostate cancer in male Swedish twins. The data set was created by merging the Swedish Twin Registry with the Swedish Cancer Registry. The twin pairs come from the old Swedish Twin Registry, which includes the birth cohorts from 1886 to 1925. The data is left-truncated since both twins of a pair had to be alive in 1961 in order to be included [11]. The cancer follow-up extended from 1961 to 1995, so 1995 is the year of right censoring.

The year of birth and the zygosity is known for every twin pair. For individuals who died during the observation period (1961–1995) the year of death is included in the data, as is the year of diagnosis for individuals with prostate cancer.

There is information about 9264 individuals, that is, 4632 pairs (1649 monozygotic, 2983 dizygotic). For 624 individuals a cancer diagnosis was registered (6.7 per cent). The mean age at cancer diagnosis is 73.6 years (standard deviation 7.4 years).

Table I shows the proportions of pairs where for no twin, one twin or both twins a cancer diagnosis is registered for the different zygosities. The proportion of pairs with two affected twins is higher for monozygotic twins. This could be an indication of genetic influences.

Four different types of observation are possible for every individual:

1. A person died during the observation period and no cancer diagnosis is registered. Thus only the year of death is known for this person.
2. A person died during the observation period and a cancer diagnosis is registered. Then the year of diagnosis and the year of death are known.
3. A person is still alive at the end of the observation period and there is no cancer diagnosis. Then one only knows the censored life span.
4. A person is still alive at the end of the observation period and there is a cancer diagnosis. Then one knows the censored life span and the year of diagnosis.

Tables II and III show the cross-tabulation of these types of observation in a twin pair for monozygotic and dizygotic twins. In the majority of pairs both members are deceased and do not have a cancer diagnosis.

The data provide more information than the type of data described in Section 3.2, since there is information about the year of diagnosis, which could be interpreted as year of onset of disease. For every individual the censored life span is used in the analysis and for people with a diagnosis of cancer the age at onset of disease is used. For individuals without a

Table II. Proportions of different types of observations for MZ twins in per cent (1 = deceased without cancer diagnosis, 2 = deceased with cancer diagnosis, 3 = alive without cancer diagnosis, 4 = alive with cancer diagnosis).

Twin 1\twin 2	1	2	3	4	Total
1	46.6	3.7	10.6	0.2	61.0
2	2.9	0.8	0.7	0.2	4.6
3	11.6	1.0	19.0	1.0	32.6
4	0.5	0.4	0.7	0.1	1.8
Total	61.6	6.0	30.9	1.5	100.0

Table III. Proportions of different types of observations for DZ twins in per cent (1 = deceased without cancer diagnosis, 2 = deceased with cancer diagnosis, 3 = alive without cancer diagnosis, 4 = alive with cancer diagnosis).

Twin 1\twin 2	1	2	3	4	Total
1	43.9	3.8	11.7	0.9	60.3
2	3.4	0.3	1.0	0.1	4.8
3	13.0	1.0	18.6	0.6	33.2
4	0.9	0.1	0.8	0.0	1.8
Total	61.2	5.1	32.1	1.6	100.0

diagnosis of cancer who are deceased at the end of the observation period, the fact that their age at diagnosis is greater than their age at death is included. The risk of death and the risk of getting cancer are two competing risks here. Finally, individuals who are still alive at the end of the observation period and who do not have a diagnosis of cancer contribute with the information that they can only have a diagnosis before the beginning or after the end of the study.

4.2. Methods

The data are pairwise left truncated since both twins in a pair had to survive the year 1961. For this reason the usual Kaplan–Meier estimator is not appropriate for the estimation of the survival function S_L. Thus we could not use the semi-parametric approach and a parametric specification had to be chosen for S_0 as well as for S_L.

First we did a separate analysis for monozygotic and dizygotic twins, fitting combinations of six different parametric specifications for the univariate survival functions S_0 and S_L to the data. All these specifications were submodels of

$$S(x)=\mathrm{e}^{-dx}[1+s^2H(x)]^{-1/s^2} \quad \text{with } H(x)=\frac{a}{b}(\mathrm{e}^{bx}-1)+cx \tag{3}$$

The survival function S in (3) is derived by using the conditional hazard rate $\mu(x,Z)=Z(a\mathrm{e}^{bx}+c)+d$. This model is a combination of the gamma–Makeham parameterization $\mu(x,Z)=Za\mathrm{e}^{bx}+d$ and the Gompertz–Makeham parameterization $\mu(x,Z)=Z(a\mathrm{e}^{bx}+c)$. Both parameterizations have been used in lifetime data analysis [6].

Let $S1$ denote the submodel of (3) with $c=0$, $d=0$ and $s^2=\sigma^2$ (variance of the corresponding frailty variable) and $S2$ denote the submodel with $c=0$ and $d=0$.

We determined maximum likelihood estimates of the parameters of the four-dimensional frailty distribution and of the univariate survival functions. The best fitting model was chosen according to the Akaike information criterion (AIC). The corresponding parametric specifications of S_0 and S_L were used in the following analysis.

We then carried out a joint analysis of all data, in which the same parameters were estimated under the assumption of equal marginal distributions for monozygotic and dizygotic twin individuals.

Finally, we did the genetic analysis. Six different genetic models were explored: ACE; AE; ADE; DE; DCE; CE; E. A refers to additive genetic effects, D to genetic effects due to dominance, C to shared environmental factors and E to non-shared environmental factors. Let a^2 denote the proportion of variance associated with additive genetic effects. This is called narrow sense heritability. d^2, c^2 and e^2 are defined similarly as proportions of the phenotypic variance associated with the corresponding genetic or non-genetic effects. In quantitative genetics one generally assumes that for a phenotypic trait the correlations between monozygotic and dizygotic twins can be expressed as

$$\varrho_{\mathrm{MZ}} = a^2 + d^2 + c^2$$

$$\varrho_{\mathrm{DZ}} = 1/2a^2 + 1/4d^2 + c^2$$

and that the normalizing equation

$$1 = a^2 + d^2 + c^2 + e^2$$

holds. Since only three of the parameters a^2, d^2, c^2 and e^2 are determined by these equations, one can only look at the seven genetic models mentioned above.

A detailed description of the methods of quantitative genetics can be found in McGue *et al.* [7] or Neale *et al.* [8].

The four-dimensional frailty model allows for the joint genetic analysis of the frailty in becoming ill and the frailty in dying. One can thus combine the seven genetic models and can carry out the estimation procedure for a total of 49 different models. Since these models are not nested, the best-fitting one was chosen using the Akaike information criterion.

4.3. Results

For monozygotic twins the parametric specifications $S1$ for S_0 and $S2$ for S_L fitted best according to AIC. For dizygotic twins a model with the parametric form $S2$ for S_0 and S_L provided the best fit. The parameter estimates and their standard errors are given in Table IV.

In the joint analysis we chose the parametric specification $S2$ for all univariate survival functions. Assuming equal values for σ_0, σ, ϱ, s_0, a_0, b_0, s_1, a_1 and b_1 for monozygotic and dizygotic twins, respectively, yields the estimation results given in the rows ‘MZ and DZ’ of Table IV.

The genetic analysis of the data remains as the last step. For the possible 49 combinations of the seven different genetic models, we calculated maximum likelihood estimates for σ_0, σ, ϱ, the parameters of S_0 and S_L and for the genetic parameters. Again, the AIC was used to determine the best-fitting model. It turned out that a DE model for the transition from

Table IV. Results of separate and joint analysis. Estimators of the variances and correlations of the frailty variables and of the parameters of the univariate survival functions S_0 and S_L together with their standard errors (in brackets).

	σ_0	σ	ϱ_0	ϱ_1	ϱ	ϱ^*
MZ	3.904 (0.673)	1.687 (0.112)	0.321 (0.085)	0.496 (0.041)	0.379 (0.062)	0.093 (0.045)
DZ	3.011 (0.540)	2.546 (0.213)	0.037 (0.063)	0.147 (0.026)	0.762 (0.143)	0.031 (0.054)
MZ and DZ	2.741 (0.436)	2.023 (0.127)	0.231^{MZ} (0.092) 0.042^{DZ} (0.066)	0.429^{MZ} (0.036) 0.184^{DZ} (0.029)	0.656 (0.111)	0.120^{MZ} (0.069) 0.031^{DZ} (0.049)
	s_0	$a_0 \times 10^9$	b_0	s_1	$a_1 \times 10^5$	b_1
MZ	— —	0.288 (0.474)	0.252 (0.026)	0.260 (0.133)	2.032 (0.614)	0.105 (0.005)
DZ	1.592 (0.285)	1.832 (2.050)	0.227 (0.018)	0.321 (0.080)	1.925 (0.412)	0.106 (0.003)
MZ and DZ	1.886 (0.335)	2.783 (2.323)	0.216 (0.013)	0.298 (0.069)	1.988 (0.350)	0.106 (0.003)

healthy to ill ($d^2 = 0.22\,(0.09)$) and an AE model for the transition from healthy to deceased ($a^2 = 0.42\,(0.03)$) was the best-fitting combination.

4.4. Discussion

The analysis of the Swedish twin data on prostate cancer suggests that there is genetic influence on frailty in the development of the disease, which is due to dominance effects. The genetic influence on the frailty corresponding to the life span is moderate and results from additive genetic effects.

The latter result confirms that derived by Yashin *et al.* for Danish twins [1].

The findings from the prostate data must be judged critically, however. First, using the year of diagnosis to calculate the age at onset of disease is a questionable procedure. Second, only the years of birth and death are known in the data, which introduces a certain degree of inaccuracy. Third, the number of affected people, that is, the number of people with a cancer diagnosis, is small. This means that the estimators of the parameters for the transition from healthy to ill have large errors, so one should be careful with the interpretation of them.

Ahlbom *et al.* [9] applied traditional methods of quantitative genetics on the same data set about prostate cancer. They estimated correlations of liability and the relative risk of cancer for twins with an affected co-twin compared with twins with a non-affected co-twin. They found that 'prostate cancer displays a clear familial effect that is almost accounted for by heritable effects' and an indication of non-additive heritable effects. In the paper the limitations of the data set are discussed thoroughly.

Another study of prostate cancer in Swedish twins by Grönberg *et al.* [10] used other statistical methods (calculation of concordance rates and correlation of liability) but it also found that 'genetic factors might be important in prostate cancer development' on the basis of differences in proband concordance rates and correlations in liability between monozygotic and dizygotic twins. They calculated correlations of liability of 0.40 and −0.05 for monozygotic and dizygotic twins, respectively. The corresponding estimates of the correlation between the frailty variables influencing the transition from healthy to ill are 0.23 and 0.04, respectively (see Table IV, estimates for ϱ_0). These values are of the same magnitude.

Applying the traditional method of calculating tetrachoric correlations in liability on the data described in Section 4.1 would mean that one only uses the information given in Table I. In doing this one ignores, on the one hand, the fact that some of the living twins might still receive a diagnosis of cancer after the end of the study and, on the other hand, one does not use the additional information about the age at onset of the disease. All these facts are accounted for in the model that we use.

5. CONCLUSIONS

The present paper introduces a new model for the genetic analysis of the ageing process. It allows for a more sophisticated view of this process since it combines information about life span with information about health.

The model is complex, which means that it places higher demands on the quality of data. To carry out a meaningful analysis one needs relatively large sample sizes and a relatively high proportion of individuals in an ill state.

The main advantages of the model are that the health information (current health status) that is needed for identifiability is relatively easily available and that it is possible to include groups of individuals with some or all members deceased.

The use of a non-parametric estimate of the survival function corresponding to the life span allows for the avoidance of parametric specifications, which normally cannot be justified from a biological or medical point of view. In situations where it is not appropriate to use a non-parametric estimate of this survival function due to the mechanism of data ascertainment, one has to use a parametric specification. Such a parameterization could be chosen according to practical experience in lifetime data analysis, and it should be general in the sense that one can compare the fit of different submodels while fixing some of the parameters at zero.

A sample analysis of prostate cancer data using this model confirmed the results of other studies that examined either life span or the disease itself.

A future task for the further investigation of the model could be the development of a non-parametric estimator for the survival function corresponding to the transition from healthy to ill that can be calculated on the basis of the current health status data.

APPENDIX: PROOF OF PROPOSITION 3.1

If one only looks on the censored bivariate life span information of every twin pair then one can apply the identifiability in the bivariate correlated frailty model and get that of σ, ϱ_1 and $S_L(x)$ for $a \leqslant x \leqslant b$.

The information about every individual (censored life span+health status) involves the uniqueness of the function

$$S_T(x,x)=P(T_{i0}>x,T_i>x) \quad \text{for } a\leqslant x\leqslant b \tag{A1}$$

The calculation of the likelihood function for the described type of censored data shows that the functions

$$S(x,x,x,x) \quad \text{for } a\leqslant x\leqslant b \tag{A2}$$

$$S(0,x,x,x) \quad \text{for } a\leqslant x\leqslant b \tag{A3}$$

and

$$\left.\frac{\partial S(x_1,x_2,y_1,y_2)}{\partial y_1}\right|_{\substack{x_1=0,\\ x_2=y_2=x}} \qquad \text{for } a\leqslant y_1\leqslant x\leqslant b \tag{A4}$$

are uniquely determined. The uniqueness of (A2) from that of the sub-density of the pairs where both twins are healthy, that of (A3) from the identifiability of the sub-density of the pairs where one twin is healthy and one twin is ill and finally follows the uniqueness of (A4) from that of the sub-density of the pairs where one twin is healthy and the other is deceased. Since for every function f

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x}=f(x)\frac{\mathrm{d}\ln f(x)}{\mathrm{d}x} \quad \text{and} \quad \frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}=f(x)\left(\left[\frac{\mathrm{d}\ln f(x)}{\mathrm{d}x}\right]^2+\frac{\mathrm{d}^2\ln f(x)}{\mathrm{d}x^2}\right)$$

applying (A2), (A3) and (A4) yields that the functions

$$g_1(x)=\left.\frac{\partial}{\partial y_1}\ln S(x_1,x_2,y_1,y_2)\right|_{\substack{x_1=0,\\ x_2=y_1=y_2=x}}$$

and

$$g_2(x)=\left.\frac{\partial^2}{\partial y_1^2}\ln S(x_1,x_2,y_1,y_2)\right|_{\substack{x_1=0,\\ x_2=y_1=y_2=x}}$$

are identifiable for $a\leqslant x\leqslant b$. Using the uniqueness of S_L and its derivatives one can conclude from the explicit representation of functions g_1 and g_2 that the functions

$$h_1(x)=\varrho^*\frac{\sigma}{\sigma_0}\{[S_0^{-\sigma_0^2}(x)+2S_L^{-\sigma^2}(x)-2]^{-1}-[2S_L^{-\sigma^2}(x)-1]^{-1}\}$$

and

$$h_2(x)=\varrho^*\frac{\sigma}{\sigma_0}\{[S_0^{-\sigma_0^2}(x)+2S_L^{-\sigma^2}(x)-2]^{-2}-[2S_L^{-\sigma^2}(x)-1]^{-2}\}$$

are identifiable for all x with $a\leqslant x\leqslant b$, $S_L(x)\neq 0$ and $S_L'(x)\neq 0$. $S_0^{-\sigma_0^2}(x)$ and $\frac{\varrho^*}{\sigma_0}$ can be uniquely calculated as solutions of this system of non-linear equations for all $x\in[a,b]$ with $0<S_L(x)<1$,

$S'_L(x) < 0$ and $0 < S_0(x) < 1$. Taking into account the so far derived results and calculating the unique function $\ln S(x,x,x,x) - 2\ln S_T(x,x)$ yields the identifiability of $\frac{\varrho_0}{\sigma_0^2}$.

The known function $\ln S_T(x,x) - \ln S_L(x)$ yields for two different $x_1, x_2 \in [a,b]$ a system of linear equations with unique coefficients for $\frac{1}{\sigma_0^2}$ and $\frac{\varrho}{\sigma_0\sigma}$. If the determinant of the matrix of coefficients is unequal to zero for at least one pair of distinct numbers in $[a,b]$ then $\frac{1}{\sigma_0^2}$ and $\frac{\varrho}{\sigma_0\sigma}$ and therefore also σ_0 and ϱ are uniquely determined. The identifiability of ϱ_0, ϱ^* and $S_0(x)$ for all $x \in [a,b]$ is then obvious.

The determinant can only be zero for all x_1, x_2, $a \leqslant x_1 < x_2 \leqslant b$, if there exists a constant c, $0 < c < 1$, such that

$$S_L^{-\sigma^2}(x) = \frac{S_0^{-\sigma_0^2}(x) - 1}{S_0^{-c\sigma_0^2}(x) - 1} \qquad \text{(A5)}$$

for all $x \in [a,b]$ with $0 < S_0(x) < 1$.

REFERENCES

1. Yashin AI, Iachine IA. Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. *Genetic Epidemiology* 1995; **12**:529–538.
2. Yashin AI, Vaupel JW, Iachine IA. Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* 1995; **5**:145–159.
3. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**:439–454.
4. Yashin AI, Iachine IA. How long can humans live? Lower bound for biological limit of human longevity calculated from Danish twin data using correlated frailty model. *Mechanisms of Ageing and Development* 1995; **80**:147–169.
5. Iachine IA, Yashin AI. Identifiability of bivariate frailty models based on additive independent components. Research Report 8, Department of Statistics and Demography, Odense University, 1998.
6. Yashin AI, Vaupel JW, Iachine IA. A duality in aging: the equivalence of mortality models based on radically different concepts. *Mechanisms of Ageing and Development* 1994; **74**:1–14.
7. McGue M, Vaupel JW, Holm N, Harvald B. Longevity is moderately heritable in a sample of danish twins born 1870–1880. *Journal of Gerontology* 1993; **48**:B237–B244.
8. Neale MC, Cardon LR. *Methodology for Genetic Studies of Twins and Families*. Kluwer Academic Publishers: Dodrecht/Boston/London, 1992.
9. Ahlbom A, Lichtenstein P, Malmström H, Feychting M, Hemminki K, Pedersen NL. Cancer in twins: genetic and nongenetic familial risk factors. *Journal of the National Cancer Institute* 1997; **89**:287–293.
10. Grönberg H, Damber L, Damber J-E. Studies of genetic factors in prostate cancer in a twin population. *Journal of Urology* 1994; **152**:1484–1489.
11. Cederlöf R, Lorich U. The Swedish Twin Registry. In *Twin Research*: *Biology and Epidemiology*, Nance WE, Allan G, Parisi P (eds). Alan R. Liss: New York, 1978; 189–195.