Mathematical Population Studies
1995, Vol. 5(2), pp. 145–159
Reprints available directly from the publisher
Photocopying permitted by license only

CORRELATED INDIVIDUAL FRAILTY: AN ADVANTAGEOUS APPROACH TO SURVIVAL ANALYSIS OF BIVARIATE DATA

ANATOLI I. YASHIN, JAMES W. VAUPEL and IVAN A. IACHINE

Odense University Medical School, Winslowparken 17,1,
DK 5000 Odense C, Denmark

Duke University, Center for Demographic Studies, Box 90408,
Durham, NC 27708-0408, USA

March 30, 1993

Frailty models have been developed for the analysis of bivariate survival data. The notion of shared frailty in these models, however, is different from the notion of individual frailty in demographic models. Shared frailty models have important shortcomings. We develop a new model of bivariate survival based on the notion of correlated individual frailty. We analyze the properties of this model and suggest a new approach to the analysis of bivariate data that does not require a parametric specification—but permits estimation—of the form of the hazard function for individuals. We empirically demonstrate the advantages of the model in the statistical analysis of bivariate data.

KEY WORDS: Random effect, frailty, bivariate data, twins, survival analysis, aging, mortality model.
Communicated by Ken Wachter.

1. INTRODUCTION

What role do genetic vs. environmental factors play in influencing mortality? Galton (1875) first pointed out that bivariate data on twins are uniquely suited to address questions of this kind; numerous researchers since have used such data as well as data on other pairs of related individuals to analyze the determinants of longevity (Cohen 1964, Vaupel 1988, McGue et al. 1993). Traditionally the data have been studied by analysis-of-variance methods developed by quantitative geneticists (Bulmer 1980, Falconer 1990). A new approach has been opened up by the extension of the methods of survival analysis (Kalbfleish and Prentice 1980, Cox and Oakes 1984) to bivariate survival data. The survival analysis approach is superior to the analysis-of-variance approach when the data to be analyzed are censored (e.g., when for some individuals, it is known that the individuals were alive at some age but it is not known when the individuals died). Furthermore, survival analysis is advantageous when the mechanisms of mortality, including the effect of covariates, can be appropriately captured by a hazards model.

The use of bivariate survival data (e.g., twins data) to analyze processes of aging and survival is motivated by the idea that the association between the life spans of related individuals can help distinguish genetic vs. environmental influences. In particular, a difference in the associations between the life spans of MZ (monozygotic, identical) and DZ (dizygotic, fraternal) twins may have a genetic origin: MZ twins can be considered to be genetically identical whereas DZ twins share only half their genes, like ordinary siblings. If this genetic difference is represented in an appropriate bivariate survival model, then statistical analysis can help clarify the role of genetic factors in the aging process.

In this article we review the shared-frailty models generally used in bivariate survival analysis and suggest a new model for the analysis of bivariate data which exploits the idea of correlated individual frailty. Despite the fact that this model is not a general correlated-frailty model it is comprehensive enough to include the widely-used univariate proportional hazards individual-frailty and shared-frailty models as particular cases. The model has several properties worth highlighting.

Note that a crucial assumption used in the analysis of proportional-hazards frailty models concerns the parametric structure of the underlying hazard. It is often difficult to find a theoretical reason for such an assumption. So in particular cases some simple functions are used for the sake of computational simplicity. The main advantage of the new approach is that it does not require the assumptions about the parametric structure of the underlying hazard. The semiparametric estimate of this hazard, calculated from bivariate data, can then be used in the univariate (demographic) models of mortality.

Another important advantage of the new model is that data concerning MZ and DZ twins can be analyzed together, increasing the precision of parameter estimates. In addition, the correlated-frailty model permits two kinds of diagnostic checks. First, if the correlated-frailty model approximates reality, then the estimated variance in frailty and baseline hazard function should be the same for MZ and DZ individuals: if they are significantly different then this is a diagnostic signal that the model is inappropriate. Second, if the model approximates reality, then the estimated variance in frailty should be consistent with the estimate produced by a univariate analysis of the data as if it pertained to individuals: again, if the two estimates are not similar, then this is a diagnostic signal that the model is inappropriate. We also derive a useful formula which establishes the correspondence between the correlation coefficient of lifespans and the parameters of the bivariate frailty distribution.

The terminology and examples used in the article stem from our main motivation —to use data on twins to analyze genetic and environmental factors influencing mortality. Such a focus also simplifies the exposition. Readers should note, however, that the methods developed in the article can be extended to a broader range of applications. Instead of age at death, the duration of interest could be age at first heart attack, time from marriage to divorce, length of an unemployment spell, etc. Instead of pertaining to twins, the data could pertain to: (1) other relatives, such as mothers and daughters; (2) individuals who are "related" not genetically but because they share some attributes, such as matched pairs in clinical trials; or (3) a single individual who experiences two durations, such as waiting times to conception of first and second child or durations of first and second bouts of some illness. Furthermore, most of the concepts and formulas presented in this article can be extended to survival data on n individuals rather than 2, such as groups of individuals who share a common value of some covariate. As this last example suggests, the methods developed here for the analysis of data on the longevity of twins are relevant for a much wider class of problems, namely, the analysis of duration data for individuals with observed covariates.

2. BIVARIATE SURVIVAL MODELS

Research on bivariate survival models has burgeoned over the last decade and a half. Clayton's (1978) random-effect model of bivariate survival was a key innovation: he introduced the notion of shared relative-risk. This model was further developed by Oakes (1982) to analyze the association between two non-negative random variables. Clayton and Cuzick (1985) added observed covariates to the bivariate survival model with shared relative-risk. Crowder (1985) and Hougaard (1986) proposed random-effect models of bivariate Weibull distributions.

About this time, the word "frailty" started replacing the corresponding phrases "relative risk" or "random effects." A shared-frailty model with a positive stable distribution of frailty was suggested by Hougaard (1987); he also discussed several other bivariate distributions with biomedical and reliability applications. Oakes (1989) developed a bivariate shared-frailty model related to the "archimedean distributions" studied by Genest and MacKay (1986); he also proposed a local time-dependent association measure between bivariate life spans and discussed its use for a large class of bivariate survival functions. This and other time dependent association measures were studied by Anderson et al. (1991). Applications of shared-frailty models to twins data were discussed by Hougaard (1990, 1992a, 1992b). Sorensen et al. (1988), Thomas et al. (1990), Vaupel (1991a,b), Vaupel et al. (1991, 1992), Nielsen et al. (1992), and Guo (1993) studied genetic and environmental influences on longevity using bivariate survival models. Approaches to non-parametric bivariate survival analysis were examined by Pruitt (1990) and Dabrowska (1990). Gill (1990) provided an interesting discussion of problems in multivariate survival analysis.

Yashin et al. (1993a, b) noted that the shared-frailty model does not satisfy some natural conditions and suggested a correlated-frailty model for the analysis of twins data. This model, which is developed below in this article, is the natural bivariate generalization of the frailty model used in demographical applications (Vaupel et al. 1979, Vaupel and Yashin 1985a, b). At the first glance the correlated-frailty model is closely related to the shared-frailty model widely used in the literature on bivariate survival analysis. This impression, however, is wrong and misleading: these models are based on different concepts.

3. SHARED-FRAILTY MODELS

Let T_i, $i = 1,2$ be the survival times for two related individuals, e.g., a pair of twins. In a shared-frailty model, frailty is defined as a measure of the relative risk which twins in a pair share. Thus the frailty variable is associated with pairs of twins, rather than individuals. The hazard model for each individual twin, however, looks exactly the same as in the standard demographic frailty model: $\mu(Z,x) = Z\mu_0(x)$. Here Z represents shared frailty and $\mu_0(x)$ is the underlying hazard. The conditional survival function $P(T_i > x \mid Z) = S(x \mid Z)$ for each twin with shared frailty Z is:

$$S(x \mid Z) = \exp\left\{-Z \int_0^x \mu_0(u)du\right\}. \tag{1}$$

It is convenient to assume that Z is gamma-distributed with mean 1 and variance σ_z^2. Under the assumption that the life spans of twins in a pair are conditionally independent given Z, the bivariate conditional survival function is:

$$S(x_1,x_2 \mid Z) = \exp\{-Z(H(x_1)+H(x_2))\}. \tag{2}$$

Averaging (2) with respect to Z produces the marginal bivariate survival function which has been used to derive the likelihood function for bivariate data (Vaupel et al. 1991, 1992). Application of the maximum likelihood approach to a Danish twin data set indicated that the estimated variance of the distribution of frailty for MZ twins is higher than for DZ twins (Vaupel et al., 1991, 1992; see also Hougaard et al. 1992a), suggesting that genetic factors influence longevity.

The notion of shared-frailty is different from the definition of individual frailty given by Vaupel et al. (1979) and used in Heckman and Singer (1984), Vaupel and Yashin (1985a, b), Trussell and Rodriguez (1990), and other demographic and econometric analyses of univariate duration data. This difference has gone largely unrecognized, perhaps because of the superficial similarity of the individual hazards in the two approaches. The "frailty" in the shared-frailty model of bivariate survival is only a part of the individual "frailty" used in demographical models, capturing only the components of frailty that a pair of twins share. The unshared components of frailty, which differ for MZ and DZ twins, are implicitly represented by differences in the underlying hazard functions. For instance, Vaupel et al.'s (1991, 1992) shared-frailty analysis of Danish twin data indicated that the underlying hazard function is steeper for MZ and DZ twins.

It seems natural to develop a bivariate survival model consistent with the univariate definition of frailty. In such a model, the underlying hazard function should be the same for MZ and DZ persons and individuals with the same value of frailty Z should experience the same risk of death. Such model should have at least one parameter that characterizes the bivariate model alone and is not included in the marginal (univariate) survival function. Since parametric descriptions of the forces of mortality are the same for MZ and DZ persons such a model should be useful for combining MZ and DZ survival data (Yashin et al. 1993a). The model also should be helpful in taking advantage of the integration of these data sets with survival data for unrelated individuals, for example, with demographic data. The correlated-frailty model of bivariate survival that is developed below meets these criteria.

4. THE CORRELATED-FRAILTY MODEL

In the correlated-frailty model the frailty of each twin in a pair is defined by a measure of relative risk, i.e., exactly as it was defined in demographical applications (Vaupel et al. 1979). For two twins in a pair, frailties are not necessarily the same, as they are in the shared-frailty model. The specific feature of twins—association—is captured by the correlation between the individual frailties. Thus, the correlated-frailty model requires specification of the bivariate distribution of frailty. We construct this distribution using three independent gamma-distributed random variables. Note that because of this construction the model is not a general correlated-frailty model or even general correlated-gamma-distributed-frailty model. Its properties, however, are good enough to establish a hierarchical correspondence with univariate frailty models and with shared-frailty models. Gamma distributions are widely used in both univariate and bivariate frailty models; the correlated-frailty model developed here follows that tradition.

Let T_i and Z_i, $i = 1,2$ be the life spans and the frailties for two individuals who are twins: their individual hazards are represented by the proportional hazards model $\mu(Z_i,t) = Z_i\mu_0(t)$, $i = 1,2$. We assume that $Z_1 = Y_0 + Y_1$ and $Z_2 = Y_0 + Y_2$, where Y_0, Y_1, Y_2 are independent non-negative gamma-distributed random variables with parameters (k_i, λ_i), $i = 0,1,2$, respectively.

To ensure that random variables Z_1 and Z_2 are gamma distributed we make an assumption that the scale parameters λ_0, λ_1 and λ_2 of gamma-distributed random variables Y_0, Y_1, Y_2 are the same, i.e. $\lambda_0 = \lambda_1 = \lambda_2 = \lambda$. Note that this assumption is not a restriction for population of unrelated individuals since gamma-distributed variables Z_i, $i = 1,2$, can always be decomposed this way. It does, however, influence survival process for bivariate populations of related individuals. It can be shown, for instance, that under this assumption the conditional correlation coefficient $\rho_z(x)$ between the frailties of twins who survived to age x is a declining function of x.

To force Z_1 and Z_2 to have the same distribution we assume that shape parameters k_1 and k_2 for the distributions of Y_1 and Y_2 are the same, $k_1 = k_2$. This condition is relevant for twin studies, when there is no reason to argue different distributions of frailty for two twins, and can be omitted for other applications. It can be easily shown that under this assumptions frailties Z_1 and Z_2 are gamma-distributed correlated random variables. When the variances $V(Y_0)$ and $V(Z_i)$, $i = 1,2$ are known the correlation coefficient ρ_z can be calculated using

$$\rho_z = \frac{V(Y_0)}{\sqrt{V(Z_1)V(Z_2)}}.$$

Since $V(Y_0) = k_0/\lambda^2$ and $V(Z_i) = (k_0 + k_1)/\lambda^2 = \sigma_z^2$, $i = 1,2$, the correlation coefficient ρ_z is simply given by:

$$\rho_z = \frac{k_0}{k_0 + k_1}.$$

We use a standard assumption that the mean frailty of individuals is one. This condition, which is typical for demographic proportional-hazards-frailty models

(Vaupel et al. 1979), is equivalent to the condition $(k_0+k_1)/\lambda=1$. It seems that formulated assumptions restrict significantly the class of correlated-frailty models which we propose to use. However, this class is still wide enough in order to include individual-frailty models and shared-frailty models with gamma distributed random effects as particular cases.

We also assume that given frailties Z_1 and Z_2 the life spans T_1 and T_2 are conditionally independent. Under these assumptions the marginal bivariate survival function $S(x_1,x_2)$ for the two individuals in a twin pair can be calculated as:

$$\begin{aligned} S(x_1,x_2) &= E(S(x_1,x_2\mid Z_1,Z_2)) \\ &= \int_0^\infty\int_0^\infty\int_0^\infty e^{-(y_0+y_1)H(x_1)-(y_0+y_2)H(x_2)}g_1(y_0)g_2(y_1)g_2(y_2)\,dy_0\,dy_1\,dy_2, \end{aligned} \tag{3}$$

where

$$H(x_i)=\int_0^{x_i}\mu_0(u)\,du, \qquad i=1,2 \qquad \text{and} \qquad g_i(y)=\frac{\lambda^{k_i}y^{k_i-1}e^{-\lambda y}}{\Gamma(k_i)}, \qquad i=0,1.$$

After integration (3) reduces to:

$$\begin{aligned} S(x_1,x_2) &= (\sigma_z^2H(x_1)+1)^{-(1-\rho_z)/\sigma_z^2}(\sigma_z^2H(x_2)+1)^{-(1-\rho_z)/\sigma_z^2} \\ &\quad\times(\sigma_z^2(H(x_1)+H(x_2))+1)^{-\rho_z/\sigma_z^2}. \end{aligned} \tag{4}$$

By differentiation of (4), the marginal bivariate probability density function can be obtained:

$$\begin{aligned} f(x_1,x_2) &= \mu_0(x_1)\mu_0(x_2)(\sigma_z^2H(x_1)+1)^{-(1-\rho_z)/\sigma_z^2}(\sigma_z^2H(x_2)+1)^{-(1-\rho_z)/\sigma_z^2} \\ &\quad\times(\sigma_z^2(H(x_1)+H(x_2))+1)^{-\rho_z/\sigma_z^2} \\ &\quad\times\Bigg(\frac{\rho_z(\rho_z+\sigma_z^2)}{(\sigma_z^2(H(x_1)+H(x_2)+1)^2}+\frac{\rho_z(1-\rho_z)}{\sigma_z^2(H(x_1)+H(x_2))+1} \\ &\quad\times\left[\frac{1}{\sigma_z^2H(x_1)+1}+\frac{1}{\sigma_z^2H(x_2)+1}\right]+\frac{(1-\rho_z)^2}{(\sigma_z^2H(x_1)+1)(\sigma_z^2H(x_2)+1)}\Bigg). \end{aligned} \tag{5}$$

This formula has been applied to MZ and DZ twin survival data by Yashin et al. (1993a) in a maximum-likelihood procedure for parameter estimation.

Using the relationship

$$H(x)=\frac{S(x)^{-\sigma_z^2}-1}{\sigma_z^2} \tag{6}$$

from the univariate individual-frailty model (Vaupel et al. 1979), the marginal bivariate survival function (4) can be transformed to a semiparametric form:

$$S(x_1,x_2)=\frac{S(x_1)^{1-\rho_z}S(x_2)^{1-\rho_z}}{(S(x_1)^{-\sigma_z^2}+S(x_2)^{-\sigma_z^2}-1)^{\rho_z/\sigma_z^2}}. \tag{7}$$

The semiparametric form for the bivariate marginal probability density function may be obtained by the differentiation of (7):

$$\begin{aligned} f(x_1,x_2) = & \frac{(1-\rho_z)^2 S(x_1)^{-\rho_z} S(x_2)^{-\rho_z} f(x_1) f(x_2)}{(S(x_1)^{-\sigma_z^2} + S(x_2)^{-\sigma_z^2} - 1)^{\rho_z/\sigma_z^2}} \\ & + \frac{(1-\rho_z)\rho_z S(x_1)^{-\rho_z} S(x_2)^{-\rho_z-\sigma_z^2} f(x_1) f(x_2)}{(S(x_1)^{-\sigma_z^2} + S(x_2)^{-\sigma_z^2} - 1)^{(\rho_z/\sigma_z^2)+1}} \\ & + \frac{(1-\rho_z)\rho_z S(x_1)^{-\rho_z-\sigma_z^2} S(x_2)^{-\rho_z} f(x_1) f(x_2)}{(S(x_1)^{-\sigma_z^2} + S(x_2)^{-\sigma_z^2} - 1)^{(\rho_z/\sigma_z^2)+1}} \\ & + \frac{\rho_z(\rho_z+\sigma_z^2) S(x_1)^{-\rho_z-\sigma_z^2} S(x_2)^{-\rho_z-\sigma_z^2} f(x_1) f(x_2)}{(S(x_1)^{-\sigma_z^2} + S(x_2)^{-\sigma_z^2} - 1)^{(\rho_z/\sigma_z^2)+2}}, \end{aligned} \qquad (8)$$

where $f(x)$ is the marginal univariate p.d.f. of the life span distribution. The semiparametric representations of these formulas play an important role in the development of estimation strategies and in the interpretation of the results of bivariate analysis.

When $\rho_z = 1$ formulas (4), (7) and (5), (8) represent survival functions and probability density functions for the shared-frailty model.

5. WHY THE ESTIMATES CAN BE BIASED

When the shared-frailty model is applied to data sets, parameter estimates usually differ for univariate vs. bivariate analyses. Our applications of the correlated-frailty model, on the other hand, have generally yielded similar estimates of the variance in frailty and the parameters of the underlying hazard function regardless of whether the data are treated as pertaining to pairs or to individuals. It turns out, however, that there is an important exception to this experience. When, in a bivariate analysis, the estimated value of the correlation of frailty, ρ_z, is 1—i.e., on the boundary of the possible range—then the estimated value of the variance in frailty, σ_z^2, tends to be higher than it is in a univariate analysis of the same data. The reason is suggested by the following approximate formula, derived in the Appendix, for the correlation coefficient of the life spans of twins:

$$\rho_x \approx \frac{\rho_z \sigma_z^2}{1+\sigma_z^2}. \qquad (9)$$

If ρ_z is on the boundary, then to satisfy (9) σ_z^2 may have to be inflated from its value as estimated from univariate data.

The basic, underlying problem is a familiar one: when a model does not fit the data, estimation procedures, attempting to squeeze the data onto the model's Procrustean bed, produce estimates that are a contorted compromise between reality and the constraints of the model. A major advantage of the correlated-frailty model is that when the univariate and bivariate estimates of σ_z^2 differ significantly, the discrepancy can be taken as a diagnostic signal that the model is inappropriate.

Hougaard, in his discussion of Clayton and Cuzick (1985), pointed out a weakness of shared-frailty models with gamma-distributed frailty: the parameter σ_z^2, which characterizes the measure of association between the life spans of twins, can be identified from the univariate data. This statement is correct, however, only if the mechanism generating the data is precisely captured by the model used in the parameter estimation procedure. In reality any model is just an approximation of the truth. Consequently the estimate of σ_z^2 calculated from univariate data can be far from the bivariate estimate of this parameter (as demonstrated in the Example 2 of the Applications section of this article). The situation can be explained as follows. In the univariate case the estimate of σ_z^2 is chosen to give the best fit to the univariate model. In the bivariate case this estimate should, in addition, preserve the observed correlation between the life spans of twins in accordance with formula (9), with $\rho_z = 1$. The resulting estimate can be far from the univariate one.

6. SEMIPARAMETRIC ESTIMATE OF THE UNDERLYING HAZARD

Is there any way to correct the erroneous model when such a discrepancy between univariate and bivariate analyses occurs? It turns out that the semiparametric representation of the bivariate survival function provides a solution. This representation, given in (7) and (8), was derived assuming some underlying hazard $\mu_0(x)$, i.e. assuming the proportionality of the hazard. The representation, however, could pertain directly to the survival function $S(x)$. Given either an empirical or a parametric description of $S(x)$, the underlying hazard $\mu_0(x)$ with $H(x) = \int_0^x \mu_0(u)du$ has to satisfy the equality

$$S(x) = \left(\frac{1}{1+\sigma_z^2 H(x)}\right)^{1/\sigma_z^2}. \tag{10}$$

When σ_z^2 is estimated from the bivariate data and $S(x)$ is either estimated nonparametrically or its parameters are estimated from univariate analysis, $\mu_0(x)$ can be easily calculated. Indeed, since $S(x)$ and σ_z^2 are known, the cumulative hazard $H(x)$ is simply given by (6). The appropriate estimate of the underlying hazard is

$$\hat{\mu}_0(x) = \overline{\mu}(x)\hat{S}(x)^{-\hat{\sigma}_z^2}, \tag{11}$$

where $\overline{\mu}(x)$ and $\hat{S}(x)$ are the observed mortality and estimated marginal survival functions calculated from univariate data, and $\hat{\sigma}_z^2$ is the estimate of the variance of the frailty distribution calculated from bivariate data.

In sum, the information provided by bivariate data permits construction of an appropriate underlying hazard function. The parameters σ_z^2 and ρ_z can be estimated from the bivariate data without specifying an underlying hazard function, using the semiparametric representation in (7) and (8). Given the value of σ_z^2, the appropriate underlying hazard function can be calculated using (11).

7. ESTIMATION STRATEGIES

To illustrate the ideas discussed above we generated two sets of bivariate survival data. Two main strategies were used for parameter estimation. One can be called the "parametric" strategy. It uses a parametric specification of the underlying hazard and bivariate frailty distribution in the model for the maximum likelihood estimation algorithm. Note that in this strategy the estimates of σ_z^2 can take on significantly different values for univariate and bivariate versions of the same data. Indeed, in the univariate case σ_z^2 is used to fit $S(x)$ to the data. In the bivariate case additional condition (9) can bias the estimate of σ_z^2. Thus, it is difficult to interpret this parameter as a variance of the frailty distribution. The real variance should have the same value in the univariate and bivariate analyses. That is why we use the notation s^2 instead of σ_z^2 to represent the results of parametric analysis as a model fitting procedure.

Another strategy can be called "semiparametric." It does not use a parametric specification of the underlying hazard. Two options can be used in this strategy: In the first option a parametric model of the univariate survival function is inserted in the semiparametric representation (7), (8). For example, one can use representation (10) with $\mu_0(x)$ defined by (12) below. It has been shown (Vaupel et al. 1979, Manton et al. 1986) that such model provides a good fit to univariate survival data. Note that since it is just a parametric representation of $S(x)$ the parameter σ_z^2 in (10) (which influences the shape of $S(x)$) does not necessarily take the same value as parameter σ_z^2 (outside $S(x)$) in (7), (8) which is responsible for association between the lifespans. To take this fact into account we use the notation s^2 again for the parameter in (10) and save the notation σ_z^2 for the parameter in (7), (8). In the second option non-parametric estimates of the univariate survival function $S(x)$ can be used in (7), (8). In this case only parameters ρ_z and σ_z^2 should be calculated in the maximum likelihood estimation procedure.

8. APPLICATIONS

In this section we use two examples of bivariate data to illustrate the properties of the estimation algorithms derived from different bivariate-survival models. Both examples are based on simulated data sets so that estimated values can be compared with the known values used to generate the data.

EXAMPLE 1 *The data for the first example were generated using a proportional-hazards correlated-frailty model with a Gompertz–Makeham baseline hazard function*

$$\mu_0(x) = ae^{bx} + c, \tag{12}$$

and bivariate frailty distribution with parameters ρ_z and σ_z^2 constructed above. There was no censoring. Data were generated for 780 *pairs of twins. Parameter values were estimated by maximizing likelihood functions derived from several bivariate probability density functions given by correlated-frailty model with parametric description of $S(x)$ and $\mu_0(x)$ given by* (12), *(CF-I); shared-frailty model with parametric description of $S(x)$ and $\mu_0(x)$ given by* (12), *(SF); correlated-frailty model with parametric description of $S(x)$ and semiparametric evaluation of $\mu_0(x)$, (CF-II); and semiparametric representation of correlated-frailty model with $S(x)$ estimated by Kaplan–Meier estimator from univariate data, (Semipar). Table* 1 *shows the results.*

The first row of the table displays the true value of the parameters, i.e., the values used to generate the data. A univariate analysis (*line* 2 *in Table* 1) *was carried out by setting the correlation coefficient* ρ_z *equal to zero in the likelihood function for the correlated-frailty model* (4), (5) *with the notation* s^2 *used for parameter* σ_z^2. *The parameter estimates are close to the true values. A parametric bivariate analysis based on* (4), (5) *produced essentially the same results as the univariate analysis, as well as well as a good estimate of* ρ_z (*line* 3 *in Table* 1). *The incorporation of* ρ_z *significantly improved the fit of the model* ($P = .003$) *using the likelihood ratio test.*

A shared-frailty analysis was carried out by setting the correlation coefficient in (4), (5) *equal to one* (*line* 4 *in Table* 1). *The parameter estimates for the shared-frailty model differ from both the univariate estimates and the parametric correlated-frailty model estimates. The correlated-frailty model provides a better fit to the data* ($P = .03$) *than the shared-frailty model. The difference in the estimates of the standard deviation of frailty can be explained by the fact that the frailty in shared-frailty models is only a part of the frailty in correlated-frailty models. In all three cases notation* s^2 *was used instead of* σ_z^2. *We save the notation* σ_z^2 *for the semiparametric model associated with representation* (7), (8).

Using the representation (7), (8) *with* $S(x)$ *described by* (10) *with the notation*s^2 *used instead of* σ_z^2 *in* (10), *leads to a model with* 6 *parameters*: a, b, c, *and* s *describe the univariate survival function* $S(x)$, *and* σ_z^2 *and* ρ_z *characterize bivariate life span distribution. Note that parameters* σ_z^2 *and* ρ_z *represent the variance and correlation coefficient of frailty distribution for some proportional hazards model with the underlying hazard defined by* (11). *Since data in this example were generated using a Gompertz–Makeham underlying hazard and gamma-distributed frailty, with variance* σ_z^2, *estimates of* s *and* σ_z *are actually estimates of the same quantity—the standard deviation of frailty. The estimates* (*line* 5 *in Table* 1) *turn out to be close to each other and to the true value, and the estimates of the other parameters are also appropriate. As indicated by the log likelihoods, the model fits almost precisely as well as the simpler correlated-frailty model in row* 3 *of the table.*

Finally, the last row of the table shows the values of σ_z^2 *and* ρ_z *estimated by the semiparametric representation* (7), (8) *using the empirical estimates of* $S(x)$ *and* $f(x)$. *These estimates are also similar to the true values, but there is no direct way of calculating standard errors for them.*

EXAMPLE 2 *The second example is also based on a simulated, uncensored data set, but with a different method of generation. We transformed* 592 *pairs of data points drawn from a bivariate normal distribution with correlation coefficient* 0.22 *so that the marginal univariate survival function was described by* (10) *with* $H(x)$ *given by a Gompertz–Makeham mortality curve with parameters* $a = .00005$, $b = .096$, *and* $c = 0$. *The notation* s *was used instead of* σ_z *in* (10) *and had a value of* .3. *Thus the data were generated, without any reference to proportional-hazards frailty models, so that* $S(x)$ *was described by a four-parameter logistic function. The value of* σ_z *is not part of the data-generation procedure and is not given in the table. As in the first example, parameter values were estimated by maximizing likelihood functions derived from bivariate probability density functions. Table* 2 *shows the results.*

TABLE 1

The parameter estimates and their true values for bivariate survival data corresponding to Example 1. *Univ.* denotes the univariate survival model; *CF-I* is the correlated frailty model; SF corresponds to the shared frailty model. In all three models s denotes the standard deviation of the frailty distribution. *CF-II* represents the correlated frailty model in which s is a parameter used to fit the univariate model and σ_z is the standard deviation of frailty distribution.

Model	a	b	c	s	ρ_z	σ_z	LogLik
True Values	9.900E-05	0.121	0.003	0.493	0.470	0.493	
Univ.	7.348E-06 (4.131E-06)	0.125 (0.008)	0.003 (0.001)	0.525 (0.074)	[0.000] (—)	— (—)	−6143.5526
CF-I	7.377E-06 (3.989E-06)	0.125 (0.008)	0.003 (0.001)	0.524 (0.071)	0.518 (0.192)	— (—)	−6139.3237
SF	1.566E-05 (5.340E-06)	0.113 (0.005)	0.003 (0.000)	0.403 (0.045)	[1.000] (—)	— (—)	−6141.6550
CF-II	7.410E-06 (4.571E-06)	0.125 (0.009)	0.003 (0.001)	0.524 (0.077)	0.500 (0.398)	0.538 (0.320)	−6139.3232
Semipar.					0.493	0.561	

Univariate analysis yields estimated parameter values that are close to their true values, as is to be expected given the way the data were generated. An estimated value is given for s rather than for σ_z in the table because it is s that actually describes the univariate survival function; if an analyst thought that the data could be described by a Gompertz–Makeham underlying hazard in a proportional-hazards frailty model, then the analyst would interpret the estimate of s as if it were an estimate of σ_z.

Bivariate analysis, based on the correlated-frailty model with Gompertz–Makeham underlying hazard yields inappropriate estimates for σ_z and for ρ_z, which is on the boundary of possible values for a correlation coefficient. Nonetheless, the values of the log-likelihood functions indicate that the bivariate model fits better than the univariate model ($P < .001$). *The significant difference between the estimated value of s in the univariate analysis and the corresponding parameter, σ_z, in the bivariate analysis implies that the model does not adequately describe the data. The empirical correlation coefficient between twins' life spans in the data set is* .24: *to satisfy* (9) *the estimated value of σ_z has to be close to .5 even if the estimated value of ρ_z is pushed to the boundary of* 1.

The shared-frailty model yields the same parameter estimates as the correlated-frailty model (which is understandable since in both cases $\rho_z = 1$).

A correlated-frailty model (7), (8) *in which both s and σ_z are represented, using σ_z in* (7), (8) *and using notation s instead of σ_z in the description of $S(x)$ given in* (10), *yields the estimates given in the fifth row of the table. The estimates of the four parameters of $S(x)$ are close to the true values. The estimate of σ_z is* 1.41. *Thus, the underlying hazard appropriate for the model is not a Gompertz–Makeham hazard but the steeper hazard given by* (11) *with* $\sigma = 1.41$. *This model fits the data better than the first correlated-frailty model* ($P = .002$). *As shown in the last row of the table, essentially the same estimates of s and σ_z are obtained from semiparametric estimation, when $S(x)$ and $f(x)$ in* (7) *and* (8) *are described empirically using the univariate data.*

TABLE 2
The parameter estimates and their true values for bivariate survival data corresponding to Example 2. *Univ.* denotes the univariate survival model; *CF-I* is the correlated frailty model; SF corresponds to the shared frailty model. In all three models s denotes the standard deviation of the frailty distribution. *CF-II* represents the correlated frailty model in which s is a parameter used to fit the univariate model and σ_z is the standard deviation of frailty distribution.

Model	a	b	c	s	ρ_z	σ_z	LogLik
True Values	5.000E-05	0.096	0.000	0.300	—	—	
Univ.	4.409E-05 (1.014E-05)	0.098 (0.001)	0.000 (0.006)	0.273 (0.026)	[0.000] (—)	— (—)	−4615.5312
CF-I	1.899E-05 (7.628E-06)	0.111 (0.006)	0.001 (0.000)	0.485 (0.056)	1.000 (0.000)	— (—)	−4602.3476
SF	1.899E-05 (6.702E-06)	0.111 (0.005)	0.001 (0.000)	0.485 (0.052)	[1.000] (—)	— (—)	−4602.3476
CF-II	4.518E-05 (1.315E-05)	0.097 (0.005)	0.000 (0.000)	0.276 (0.099)	0.383 (0.100)	1.415 (0.440)	−4597.7535
Semipar.					0.382	1.405	

8. DISCUSSION

To apply survival analysis to mortality data various assumptions have to be made. The assumptions constitute a compromise among such competing objectives as biological plausibility, flexibility, transparency, computational ease, implementability, and statistical power. All models are wrong, as Box (1976) wrote, but some models are useful. Useful models provide insights into reality.

Frailty models of univariate data have proven useful in various contexts, providing insights into why, for instance, mortality curves bend over at older ages and why improvements in mortality are slower at older ages (Vaupel and Yashin 1985, Manton et al. 1986). These models, however, have been strongly criticized because assumptions have to be made about both the shape of the underlying mortality trajectory and the distribution of frailty: different pairs of assumptions can result in equally good fits to the data (Trussell and Rodriguez 1992, Hoem 1992).

The correlated-frailty model, applied to bivariate data, is not limited by this weakness. Instead of making an assumption about the shape of the latent underlying hazard function, the observed survival function can be either used as an empirical function or captured by some appropriate parametric form. Then the variance in frailty and the correlation between twins' frailties can be estimated. Using the estimate of the variance in frailty, the trajectory of the underlying hazard function can be computed. Thus, by applying the correlated-frailty model, bivariate data on twin pairs can be analyzed to shed light on the shape of the mortality curve for *iridividuals*. Assumptions still have to be made—e.g., that hazards are proportional and that frailty is gamma distributed—but an assumption about the shape of the unobserved hazard function for individuals is not longer required.

The semiparametric representation of the shared-frailty model also permits computation of the underlying hazard function. This function, however, pertains not to individuals but to twin pairs, and different functions will be computed for MZ and DZ pairs. In contrast, the correlated-frailty model is an "inclusive" model that permits data for MZ and DZ twins to be analyzed in the same framework and that yields estimates that pertain to individuals. Consequently, compared with shared-frailty models, correlated-frailty models can contribute more to understanding of mechanisms of aging and survival.

ACKNOWLEDGMENT

This research was supported by NIH/NIA grant PO1 AG08791. Authors thank Alexei Ya. Chervonenkis for useful discussions of the concept of correlated frailty and Kenneth W. Wachter for valuable comments to the earlier version of this paper.

APPENDIX: APPROXIMATIONS FOR ρ_x

Let η and Z be two nonnegative independent random variables such that η is exponentially distributed, i.e.

$$P(\eta > x) = e^{-x}$$

and Z has an arbitrary distribution with $E(Z) = 1$. For any $Z > 0$ life span X can be defined by

$$X = H^{-1}\left(\frac{\eta}{Z}\right) \tag{13}$$

where $H^{-1}(\cdot)$ is the inverse function of $H(x) = \int_0^x \mu_0(u)du$. It is easy to check that

$$P(X > x \mid Z) = e^{-ZH(x)},$$

i.e., the distribution of X is described by a proportional hazards model with underlying hazard $\mu_0(x)$. Using the delta-method, (13) can be written as the approximation

$$X \approx H^{-1}(1) - (H^{-1}(y))'_y|_{y=1}(Z-1) + (H^{-1}(y))'_y|_{y=1}(\eta - 1). \tag{14}$$

Thus the variance σ_x^2 of X is approximately given by

$$\sigma_x^2 \approx ((H^{-1}(y))'_y)^2|_{y=1}(1 + \sigma_z^2).$$

To calculate $(H^{-1}(y))'_y$ note that

$$H(H^{-1}(y)) = y$$

and hence

$$H'_\beta(\beta)|_{\beta=H^{-1}(y)}(H^{-1}(y))'_y = 1$$

or

$$(H^{-1}(y))'_y = \frac{1}{\mu_0(H^{-1}(y))}.$$

Thus

$$\sigma_x^2 = \frac{1}{\mu_0^2(x^*)}(1+\sigma_z^2) \tag{15}$$

where x^* is defined by the equation $H(x^*) = 1$.

If the underlying hazard is a Gompertz hazard, then $\mu_0(x^*)$ can be easily calculated. For a Gompertz hazard, $H(x) = a/b(e^{bx} - 1)$. Hence $bH(x) = \mu_0(x) - a$ and $\mu(x^*) = b + a$.

If $a \ll b$, then σ_x^2 can be represented as

$$\sigma_x^2 \approx \frac{1}{b^2}(1+\sigma_z^2) \tag{16}$$

and we obtain simple relationship between the variance in life spans and the variance of frailty.

The correlation between twin's life spans can also be calculated using formula (14). Assuming that η_1 and η_2 for twins in a pair are independent and $Z_1 = Z_2$, as is the case in the shared-frailty model, we get (9) with $\rho_z = 1$. If Z_1 and Z_2 are not equal to each other but are correlated with correlation coefficient ρ_z then formula (14) yields (9). It is interesting to note that formulas (9) and (16) do not include the parameters of the underlying hazard. Because the parameter ρ_x is central to the analysis of variance methods used in quantitative genetics, these formulas provide a useful link between quantitative genetics and survival analysis.

REFERENCES

Anderson, J. E., Louis, T. A., and Holm, N. (1992) The dependent association measures for bivariate survival distributions. *JASA* **87:** 641–650.

Box, G. E. P. (1976) Science and statistics. *JASA* **71:** 791–802.

Bulmer, M. G. (1980) The mathematical theory of quantitative genetics. Clarendon Press, Oxford, 1980.

Clayton, D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65 (1):** 141–151.

Clayton, D. G., and Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model (with discussion). *Journal of Royal Statistical Society, Ser. A* **148:** 82–117.

Cox, D. R., and Oakes, D. (1984) *Analysis of Survival Data*, Chapman and Hall, New York.

Crowder, M. (1985) A distributional model for repeated failure time measurements. *Journal of Royal Statistical Society, Ser. B* **47:** 447–452.

Dabrowska, D. M. (1988) Kaplan–Meier estimate on the plane. *Annals of Statistics* **16:** 1475–1489.

Falconer, D. S. (1990) *Introduction to Quantitative Genetics*, Longman Scientific and Technical, New York.

Freund, J. E. (1961) A bivariate extension of exponential distribution. *JASA* **56:** 971–977.

Galton, F. (1875) The history of twins as a criterion of the relative powers of nature and nurture. *Fraser's Magazine* **12:** 566–576.

Gill, R. D. (1990) Multivariate survival analysis. Invited paper at the Second World Congress of the Bernoulli Society and 53rd Annual, Uppsala 1990.

Genest, C., and MacKay, R. J. (1986) Copules archimediennes et familles de lois bidimensionelles dont les marges sont donnees. *Canadian Journal of Statistics* **14:** 145–159.

Guo, G. (1993) Use of siblings data to estimate family mortality effects in Guatemala. *Demography* **30 (1):** 15–32.

Heckman, J. J., and Singer B. (1984) Econometrics duration analysis. *Journal of Econometrics* **24.**

Hougaard, P. (1986) A class of multivariate failure time distributions. *Biometrika* **73:** 671–678.

Hougaard, P. (1987) Modelling multivariate survival. *Scandinavian Journal of Statistics* **14:** 291–304.

Hougaard, P., Harvald, B., and Holm, N. V. (1992a) Measuring similarities between the lifetimes of adult Danish twins born between 1881–1830. *JASA* **87:** N 417.

Hougaard, P., Harvald, B., and Holm, N. V. (1992b) Models for multivariate failure time data, with application to the survival of twins. *Statistical Modelling*, pp. 159–173, in P. G. M. van der Heijden, W. Jansen, B. Francis, and G. U. H. Seeber (eds.), Elsevier Science Publisher.

Kalbfleish, J. D., and Prentice, R. L. (1980) *The Statistical Analysis of Failure Type Data*, New York, Wiley.

Marshall, A. W., and Olkin, I. (1967) A multivariate exponential distribution. *JASA* **62:** 30–44.

McGue, M., Vaupel, J. W., Holm, N., and Harvald, B. (1993) Longevity is moderately heritable in a sample of Danish twins born 1870–1880. *Journal of Gerontology*, Biological Sciences, forthcoming.

Nair, N. U., and Nair, V. K. R. (1988) A characterization of the bivariate exponential distribution. *Biometrical Journal* **30 (1):** 107–112.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sorensen, T. I. A. (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19:** 25–43.

Oakes, D. (1982) A model for association in bivariate survival data. *Journal of Royal Statistical Society B* **44:** 414–422.

Oakes, D. (1986) Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73 (2):** 353–361.

Oakes, D. (1989) Bivariate survival models induced by frailties. *JASA* **84:** 487–493.

Prentice, R. L., and Cai, J. (1992) Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* **79 (3):** 495–512.

Pruitt, R. C. (1990) Bivariate survival curve estimation using non-parametric smoothing techniques. Tech. Report 543, School of Statistics, University of Minnesota.

Self, S. G., and Prentice, R. L. (1986) Incorporating random effects into multivariate relative risk regression models, pp. 167–177, in *Modern Statistical Methods in Chronic Disease Epidemiology*, in S. H. Molgavcar and R. L. Prentice (eds.), Wiley.

Shaked, M., and Shanthikumar, J. G. (1987) Multivariate hazard construction. *Stochastic Processes and Their Applications* **24:** 241–258.

Sorensen, T. I. A., Nielsen, G. G., Andersen, P. K., and Teasdale, T. W. (1988) Genetic and environmental influences on premature death in adult adoptees. *New England Journal of Medicine* **318:** 727–732.

Thomas, D. C., Langholz, B., Mack, W., and Floderus, B. (1990) Bivariate survival model for analysis of genetic and environmental effects in twins. *Genetic Epidemiology* **7:** 121–135.

Trussell, J., and Rodriguez, G. (1990) Heterogeneity in demograph research, pp. 111–132, in Adams et al. (ed.), *Convergent Questions in Genetics and Demography*, Oxford University Press.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16:** 439–454.

Vaupel, J. W. (1991a) Relatives' risks: Frailty models of life history data. *Theoretical Population Biology* **37 (1):** 220–234.

Vaupel, J. W. (1991b) Kindred lifetimes: Frailty models in population genetics, in Julian Adams et al. (ed.), *Convergent Questions in Genetics and Demography*, Oxford University Press.

Vaupel, J. W., and Yashin, A. I. (1985a) Deviant dynamics of death in heterogeneous populations, in N. Tuma (ed.), *Sociological Methodology*, 179–211, San Francisco, Jossey-Bass.

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

Vaupel, J. W., Yashin, A. I., Hauge, M., Harvald, B., Holm, N., and Liang Xue (1991) Survival analysis in genetics: Danish twins data applied to gerontological question, in J. P. Klein and P. K. Goel (eds.), *Survival Analysis: State of the Art*, Kluwer Academic Publisher.

Vaupel, J. W., Yashin, A. I., Hauge, M., Harvald, B., Holm, N., and Liang Xue (1992) Strategies of modelling genetics in survival analysis. *What Can We Learn From Twins Data?*. Paper presented on PAA meeting in Denver, CO, April 30–May 2, 1992.

Yashin, A. I., Vaupel, J. W., Chervonenkis, A. Y., Iachine, I. A., Harvald, B., and Holm, N. (1993a) *When Two Are Better Than One: Inclusive Survival Models for Combining Several Data Sets*. Paper presented on PAA meeting in Cincinnati, April 1-3, 1993, Cincinnati, OH.

Yashin, A. I., Vaupel, J. W., Chervonenkis, A. Y., Iachine, I. A., Harvald, B., and Holm, N. (1993b) *Twins Die Twice: No Evidence for Predetermined Life Span*. Paper, presented on the workshop on Oldest Old Mortality, March 4–7, Duke University, USA.