ARNOLD JOURNALS

JOURNAL…STATISICAL MODELLING…ARTID…ST030OA…

TO: AUTHOR

AUTHOR QUERIES - TO BE ANSWERED BY THE AUTHOR

The following queries have arisen during the editing of your manuscript. Please answer the queries by marking the required corrections at the appropriate positions in the text.

Q1	Please provide received, revised and accepted dates.	
Q2	Lykken 1978 not in ref list, please add or delete citation.	
Q3	Wienke et al2002 not in ref list please add or delete citation.	
Q4	Initial for Hasman?	
Q5	“Let their individual hazards are” sense?	
Q6	Reworded as Fig. 1 will not fit here. OK as set.	

Arnold Journals, 338 Euston Road, London NW1 3BH
Tel: +44 20 7873 6338 Fax: +44 20 7873 6325

Statistical Modelling 2002: **2**: 1–14

Genetic analysis of cause of death in a mixture model of bivariate lifetime data

Andreas Wienke[1], Kaare Christensen[2], Axel Skytthe[2] and Anatoli I. Yashin[1]
[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] Danish Center for Demographic Research, and the Danish Twin Registry, University of Southern Denmark, Odense, Denmark

Abstract: A mixture model in multivariate survival analysis is presented, whereby heterogeneity among subjects creates divergent paths for the individual's risk of experiencing an event (i.e., disease), as well as for the associated length of survival. Dependence among competing risks is included and rendered testable. This method is an extension of the bivariate correlated gamma-frailty model. It is applied to a data set on Danish twins, for whom cause-specific mortality is known. The use of multivariate data solves the identifiability problem which is inherent in the competing risk model of univariate lifetimes. We analyse the influence of genetic and environmental factors on frailty. Using a sample of 1470 monozygotic (MZ) and 2730 dizygotic (DZ) female twin pairs, we apply five genetic models to the associated mortality data, focusing particularly on death from coronary heart disease (CHD). Using the best fitting model, the inheritance risk of death from CHD was 0.39 (standard error 0.13). The results from this model are compared with the results from earlier analysis that used the restricted model, where the independence of competing risks was assumed. Comparing both cases, it turns out, that heritability of frailty on mortality due to CHD change substantially. Despite the inclusion of dependence, analysis confirms the significant genetic component to an individual's risk of immortality from CHD. Whether dependence or independence is assumed, the best model for analysis with regard to CHD mortality risks is an AE model, implying that additive factors are responsible for heritability in susceptibility to CHD. The paper ends with a discussion of limitations and possible further extensions to the model presented.

Key words: coronary heart disease; dependent competing risks; frailty; mixture models; survival analysis

① **Data and software link available from:** http://stat.uibk.ac.at/SMIJ
Received xxxxxx; **revised** xxxxxx; **accepted** xxxxxx.

1 Introduction

Many studies of genetic epidemiology focus on the analysis of binary phenotypic traits such as the presence or absence of a particular disease. There is often, however, additional data such as information concerning the interval of time before the onset of the disease that is not included in such studies. In order to successfully incorporate all such useful information, it is necessary to combine models of survival analysis with models of epidemiology. Survival analysis models enhance the researcher's ability to handle censored and truncated data. Recent papers treating genetic analysis of time periods of events have been divided in their conclusions. One camp suggests the use of 'liability' models of survival (Neale *et al.*, 1989;

Address for correspondence: Dr Andreas Wienke, Max Planck Institute for Demographic Research, 18057 Rostock, Doberaner Strasse 114, Germany. E-mail: wienke@demogr.mpg.de

© Arnold 2002 10.1191/1471082x02st030oa

Techset Composition Ltd, Salisbury Doc: i:/Hodder/stat mod/st030oa.3d Manu No. 0000
Printed: 14/6/02 Page#: 14 page(s) Opp: Task: 3B2 Version: 7.51a/W (May 2 2001)/TechRef:9.01/H (March 20 1997)

Meyer *et al.*, 1991) while another has focused on frailty models (Clayton and Cuzick, 1985; Hougaard *et al.*, 1992; Vaupel, 1988; Yashin and Iachine, 1995). Bivariate frailty models have provided an especially powerful analytic tool for managing identifiability problems within univariate approaches (Tsiatis, 1975).

All of the above-mentioned models are flawed in that they do not take into account specific causes of death. The problem with this is that the importance of genetic factors varies with each disease. Genetic epidemiology seeks to discover the association between genes and diseases. It might be useful to examine the genetic components of the susceptibility to specific diseases and death rather than to longevity. For this purpose we have extended the correlated gamma-frailty model of Yashin and Iachine (1995), which takes into account the life-spans of related individuals in order to better estimate the effect of genetic factors influencing frailty and morbidity. This approach, in our case using Danish twin females, allows us to combine lifespan data with morbidity data.

Recently, we analysed cause-specific mortality data using the correlated gamma-frailty model, assuming independence among causes of death in a 'competing risk' scenario (Wienke *et al.*, 2000, 2001). In this paper, we investigate the effect of removing this limitation. The model allows us to test the hypothesis on dependence between the competing risks. The class of multivariate distributions presented is characterized by the association parameters, using arbitrary marginal distributions. The multivariate distribution is specified in full by the association and variance parameters and the marginal distribution functions.

We can empirically demonstrate the advantages of this new model, having revisited the statistical analysis of the lifespan data previously explored in Wienke *et al.* (2000, 2001). In this analysis, we focused on the mortality rates of coronary heart disease (CHD). To simplify description, in this paper we consider models limited to two competing risks (death as a result of CHD and death arising from other causes). The model can be extended to the case of multiple competing risks or multivariate lifetimes. Results of a simulation study are included. Both limitations and future uses for this model are discussed further on.

2 Statistical model

Identifying correlations of durations is a requirement for successfully analysing genetic factors. In survival analysis there is a recurring problem of censored data, which complicates observations far more than does complete data. Using a survival model to estimate correlations among lifetimes can solve this problem. In this paper, instead of treating life spans directly, we wish to analyse both genetic and environmental factors acting on frailty for cause-specific mortality. The correlated gamma-frailty model can be used to fit the lifetime data and provide a specific parameter for the correlation among frailties.

Let $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ be independent and identically distributed (i.i.d.) non-negative two-dimensional random vectors (pairs of lifetimes). The lifetimes (X_{1j}, X_{2j}) $(j = 1, \ldots, n)$ are censored from the right by i.i.d. pairs of non-negative random variables $(C_{11}, C_{21}), \ldots, (C_{1n}, C_{2n})$ independent of the (X_{1j}, X_{2j}). Thus, instead of (X_{1j}, X_{2j}) we observe $(T_{1j}, T_{2j}, \Delta_{1j}, \Delta_{2j})$ with $T_{ij} + \min\{X_{ij}, C_{ij}\}$, $\Delta_{ij} = 1(X_{ij} \leq C_{ij})$ $(i = 1, 2; j = 1, \ldots, n)$ where $1(\cdot)$ denotes the indicator function of the event in the brackets.

Let us assume that the lifetimes follow a distribution given by the survival function $S(x_1, x_2) = P(X_{1j} > x_1, X_{2j} > x_2)$ and denote by $C(c_1, c_2) = P(C_{1j} > c_1, C_{2j} > c_2)$ the survival function of censoring times. Hence, the survival function of the four-dimensional non-observable data is

$$P(X_{1j} > x_1, X_{2j} > x_2, C_{1j} > c_1, C_{2j} > c_2) = S(x_1, x_2)C(c_1, c_2). \tag{2.1}$$

This form is a consequence of the independence between lifetimes and censoring times. Furthermore, due to the structure of the data we will be using as an example, let us assume that both lifetimes in each pair are left truncated at the same time w_j, which is common in twin studies. (Note that it is not a problem to deal with different truncation times in other pairs of relatives.) Consequently, observable data $(T^*_{1j}, T^*_{2j}), \Delta^*_{1j}, \Delta^*_{2j}, w_j)$ have a conditional distribution of the form

$$\mathfrak{L}(T^*_{1j}, T^*_{2j}, \Delta^*_{1j}, \Delta^*_{2j}, w_j) = \mathfrak{L}(T_{1j}, T_{2j}, \Delta_{1j}, \Delta_{2j} | T_{1j} > w_j, T_{2j} > w_j). \tag{2.2}$$

Here $\mathfrak{L}(X)$ denotes the distribution of the random variable X. With this model we derive the likelihood function of the bivariate left truncated and bivariate right censored data in (2.2), which is given by

$$\begin{aligned} L(t_1, t_2, \delta_1, \delta_2, w) = (&\delta_1\delta_2 S_{t_1t_2}(t_1, t_2) - \delta_1(1-\delta_2)S_{t_1}(t_1, t_2) \\ &- (1-\delta_1)\delta_2 S_{t_2}(t_1, t_2) + (1-\delta_1)(1-\delta_2)S(t_1, t_2))/S(w, w). \end{aligned} \tag{2.3}$$

Here $(t_1, t_2, \delta_1, \delta_2, w)$ denotes a realisation of the random vector $(T^*_{1j}, T^*_{2j}, \Delta^*_{1j}, \Delta^*_{2j}, w_j)$. Partial derivatives of the marginal survival functions are given by $S_{t_i}(t_1, t_2) = \dfrac{\partial S(t_1, t_2)}{\partial t_i}$ $(i = 1, 2)$ and $S_{t_1t_2}(t_1, t_2) = \dfrac{\partial S(t_1, t_2)}{\partial t_1 \partial t_2}$. Because of the independence of lifetimes (X_{1j}, X_{2j}) and censoring times (C_{1j}, C_{2j}) $(j = 1, \ldots, n)$ the distribution of the censoring times does not enter the likelihood function. Assuming a correlated gamma-frailty model for the survival times such that:

$$S(x_1, x_2) = \frac{S(x_1)^{1-\rho} S(x_2)^{1-\rho}}{(S(x_1)^{-\sigma^2} + S(x_2)^{-\sigma^2} - 1)^{\frac{\rho}{\sigma^2}}}, \tag{2.4}$$

this model was used to describe total mortality in twins by Yashin and Iachine (1995) and cause-specific mortality in twins under the assumption of independence between competing risks by Wienke *et al.* (2000, 2001). Here, $S(x) = S(x, 0) = S(0, x)$ denotes the marginal survival functions, which are assumed to be equal for twins. However, the assumption of independence between competing risks is questionable. Typically, in clinical and epidemiological studies two different types of censoring occur. The observation of certain individuals are censored due to the fact that they are still alive at the end of the study. Other individuals drop from follow-up for reasons not associated with the disease under study, but through life-events beyond the control of the researcher, such as migration.

If censoring can be assumed to be non-informative with regard to all different causes, then the model above may be applied with the censoring times (C_{1j}, C_{2j}) taken as the minimum of the hypothetical censoring times arising from the different causes of censoring. For estimating the marginal survival function S in (2.4) the Kaplan–Meier estimator is appropriate. However, the situation becomes much more difficult if the censoring arising from at least one of the different causes can be assumed to be informative.

In the following, we consider a case where two types of censoring occur, one non-informative and the other informative. Let $(X_{1j}, Y_{1j}, C_{1j}, X_{2j}, Y_{2j}, C_{2j})$ $(j = 1, \ldots, n)$ be i.i.d. vectors of non-negative random variables. The variables (X_{1j}, X_{2j}) denote the (usually non-observable) lifetimes (with respect to the cause of death of interest) of pairs of individuals. The (Y_{1j}, Y_{2j}) are informative censoring times (which may be lifetimes with respect to causes of death not under study) and the (C_{1j}, C_{2j}) are non-informative censoring times (for example caused by end of study). Again, for $j = 1, \ldots, n$ and $i = 1, 2$ we observe $T_{ij} = \min\{X_{ij}, Y_{ij}, C_{ij}\}$ and

$$\Delta_{ij} = \begin{cases} 1: & \text{if } X_{ij} \leq \min\{C_{ij}, Y_{ij}\} \\ 0: & \text{if } C_{ij} < \min\{X_{ij}, Y_{ij}\} \\ -1: & \text{if } Y_{ij} < \min\{X_{ij}, C_{ij}\} \end{cases} \tag{2.5}$$

where $\Delta_{ij} = 1$ means no censoring, $\Delta_{ij} = 0$ is non-informative censoring and $\Delta_{ij} = -1$ is informative censoring. Now we derive the six-dimensional survival function of the data. Suppose that we use (X_1, Y_1, X_2, Y_2) as a shorthand for $(X_{1j}, Y_{1j}, X_{2j}Y_{2j})$ $(j = 1, \ldots, n)$. Let (X_1, Y_1, X_2, Y_2) and (Z_1, Z_2, Z_3, Z_4) be the survival times of life- and (informative) censoring
(5) times and the frailties of the two individuals with respect to two different causes of death; let their individual hazards are represented by the proportional hazards model

$$\begin{aligned} X_1 &\sim \mu_1(x_1, Z_1) = Z_1\mu_1(x_1) \qquad & X_2 &\sim \mu_1(x_2, Z_3) = Z_3\mu_1(x_2) \\ Y_1 &\sim \mu_2(y_1, Z_2) = Z_2\mu_2(y_1) \qquad & Y_2 &\sim \mu_2(y_2, Z_4) = Z_4\mu_2(y_2) \end{aligned} \tag{2.6}$$

where $X \sim \mu$ means, that μ denotes the hazard function of X. Hence, the conditional distribution of the lifetime of the first (X_1) and second twin (X_2) with respect to the first cause of death are assumed equal (given by μ_1). The same is true for the lifetime of the first (Y_1) and second twin (Y_2) with respect to the second cause of death (μ_2). Note, that $\mu_i(x, Z) = Z\mu_i(x)$ implies the relation $S_i(x|Z) = S_{0i}(x)^Z$ $(i = 1, 2)$ and S_{0i} are the survival functions related to the baseline hazard functions μ_i. We assume that X_1, Y_1, X_2, Y_2 are independent given the vector of frailties (Z_1, Z_2, Z_3, Z_4). Let $V_1, V_8 \sim \Gamma(k_1, \lambda_0)$, $V_2 \sim \Gamma(k_2, \lambda_1)$, $V_3 \sim \Gamma(k_3, \lambda_2)$, $V_4, V_7 \sim \Gamma(k_4, \lambda_2)$, $V_5, V_6 \sim \Gamma(k_5, \lambda_1)$ independent gamma distributed random variables with parameters $k_1 + k_2 + k_5 := \lambda_1 = 1/\sigma_1^2$ and $k_1 + k_3 + k_4 := \lambda_2 = 1/\sigma_2^2$. Now the frailties
(6) are given in Figure 1.

Here Z_1, Z_3 denote frailties with respect to the first cause of death (cause under study) and Z_2, Z_4 denote frailties with respect to the second cause of death of both individuals. Furthermore, ρ describes variously the correlations between the frailties: $\rho_1 = \mathbf{corr}(Z_1, Z_3)$, $\rho_2 = \mathbf{corr}(Z_2, Z_4)$ and $\rho = \mathbf{corr}(Z_1, Z_2) = \mathbf{corr}(Z_3, Z_4)$. Now the six-dimensional survival function can be derived by averaging over the conditional lifetimes, using relation

(2.6) and applying the laplace transform of gamma distributed random variables (for more detailed calculations see Appendix):

$$\begin{aligned} S(x_1, y_1, c_1, x_2, y_2, c_2) &= \mathrm{E} S_1(x_1)^{Z_1} S_2(y_1)^{Z_2} S_1(x_2)^{Z_3} S_2(y_2)^{Z_4} C(c_1, c_2) \\ &= (S_1(x_1)^{-\sigma_1^2} + S_1(x_2)^{-\sigma_1^2} - 1)^{-\frac{\rho_1}{\sigma_1^2}} * (S_2(y_1)^{-\sigma_2^2} + S_2(y_2)^{-\sigma_2^2} - 1)^{-\frac{\rho_2}{\sigma_2^2}} \\ &* (S_1(x_1)^{-\sigma_1^2} + S_2(y_1)^{-\sigma_2^2} - 1)^{-\frac{\rho}{\sigma_1\sigma_2}} * (S_1(x_2)^{-\sigma_1^2} + S_2(y_2)^{-\sigma_2^2} - 1)^{-\frac{\rho}{\sigma_1\sigma_2}} \\ &* S_1(x_1)^{1-\rho_1-\frac{\sigma_1}{\sigma_2}\rho} S_1(y_1)^{1-\rho_1-\frac{\sigma_1}{\sigma_2}\rho} * S_2(x_2)^{1-\rho_2-\frac{\sigma_2}{\sigma_1}\rho} S_2(y_2)^{1-\rho_2-\frac{\sigma_2}{\sigma_1}\rho} C(c_1, c_2) \end{aligned} \tag{2.7}$$

with $0 \leq \rho \leq \min\{\sigma_2/\sigma_1(1-\rho_1), \sigma_1/\sigma_2(1-\rho_2)\}$. In this model ρ_1 denotes the correlation between Z_1 and Z_3, the main parameter of interest. This parameter describes the correlation of frailties of individuals in a pair with respect to the cause of death under study and is the key figure for genetic analysis of susceptibility to death from cause-specific mortality. The second parameter ρ_2 models the correlation between frailties with respect to all other causes of death (combined to the second cause of death or informative censoring). The parameter ρ is responsible for the association between causes of death in each individual. With this parameter. it is possible to test the hypothesis of dependence between competing risks in the above model. S_1 and S_2 are the marginal survival functions with respect to the first and second cause of death. Please note that it is impossible to use the Kaplan–Meier estimator to get non-parametric estimates of the marginal survival functions because of the assumed dependence between the two competing risks. To overcome this problem we used a parametric approach by fitting a Gamma–Gompertz model to the data, e.g.

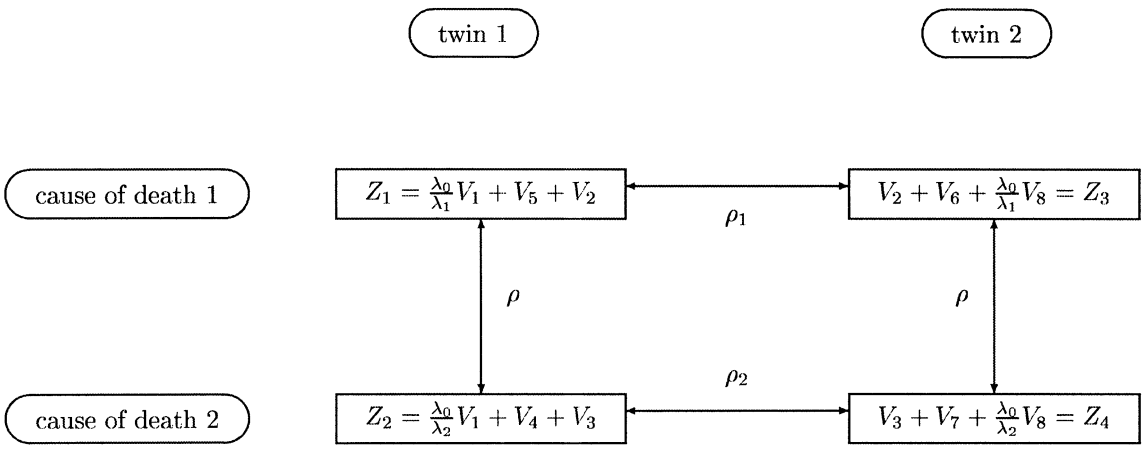

Figure 1 Cause-specific frailties and their correlations in a twin pair

$S_i(x) = (1 + s_i^2\alpha_i/\beta_i(e^{\beta_i x} - 1))^{-\frac{1}{s_i^2}}$, $(i = 1, 2)$, where α_i, β_i, s_i^2 are parameters to be estimated. Again, it is necessary to account for left truncation in the data. The likelihood function of this model is given in the Appendix. Due to the assumption concerning non-informative censoring with respect to the (C_{1j}, C_{2j}) the function C does not enter the likelihood function.

3 Quantitative genetics of frailty

In twin studies, the intrapair-correlations of traits found in monozygotic (MZ) and dizygotic (DZ) twin pairs (we use the notation $\rho_1(MZ), \rho_1(DZ)$) play a key role in the analysis of genetic and environmental factors.

Using these coefficients, we considered five genetic models of frailty that correspond to five different assumptions about structure. We followed the notation of Neale and Cardon (1992) for these models. Resemblance in twins are caused by three factors (completely for MZ twins and partly for DZ twins): additive genetic factors (A), genetic dominance factors (D) and shared environmental factors (C). Non-shared enviromnent (E) (including measurement errors) is (completely for MZ twins and partly for DZ twins) responsible for intrapair differences in twins. Additive genetic factors contribute twice as much to the correlation in MZ twins as DZ twins because MZ twins share all their identical genes by decent, while DZ twins (like non-twin siblings) share on average only half of their genes. Dominant genetic factors contribute four times as much to the correlation in MZ twins than DZ twins according to Mendelian theory. A shared environment (with environmental factors such as social class or parental behaviour, common familiar habits such as smoking, drinking, physical exercises and diet) contributes in equal measure to the correlation between MZ and DZ twins. Higher intra-pair correlations in MZ twins indicate how important a role genetic factors play. Readers unfamiliar with the use of latent variables in structural equation modelling may wonder how it is possible to reach conclusions about the role of genetic and environmental risk factors without actually measuring them directly. As in all latent variable models, the impact of genes and environment on the susceptibility to the disease of interest is inferred from the pattern of observed correlations in relatives, which are in turn predicted by Mendelian theory.

From the estimation point of view, only three parameters could be included in the model, because there are only data about two different groups of relatives (MZ and DZ twins). More complex models need data about additional groups of relatives such as parents or offspring. Each additional group of relatives in the study allow for an additional parameter, but this point is beyond the scope of the paper. The following biometric models were fitted to the data: AE, DE, ACE, ADE and CE, the ACE model refers to the decomposition of frailty $Z = A + C + E$; the CE model refers to the decomposition $Z = C + E$; ADE, AE and DE models are defined similarly. We use the small letters a^2, d^2, c^2, e^2, to refer to the respective proportions of variance. For example, the relation $1 = a^2 + c^2 + e^2$ corresponds to the decomposition of variance in the ACE model of frailty. Depending on the best fitting model, the proportion of variance in susceptibility due to additive (a^2) or dominant (d^2) genetic factors is termed heritability. Shared environmental factors and dominance factors cannot be estimated simultaneously, because they are completely confounded in the classical study

where twins are reared together (Heath *et al.*, 1989). Standard assumptions about the quantitative genetics yields in the following relations:

$$\begin{aligned} \rho_1(MZ) &= a^2 + d^2 + c^2 \\ \rho_1(DZ) &= 0.5a^2 + 0.25d^2 + c^2 \\ 1 &= a^2 + d^2 + c^2 + e^2 \end{aligned} \tag{3.1}$$

This includes the assumption that MZ and DZ twins have the same correlation in environments (equal environment assumption). This standard assumption of the classical twin method is necessary for the identifiability of parameters. To combine the approach of quantitative genetics with the methods of survival analysis, we used the extended correlated gamma-frailty model with genetic and environmental components of frailty. In this approach the genetic and environmental parameters of frailty decomposition are estimated directly by the maximum likelihood method. For more detailed information about this point see Yashin and Iachine (1994). The analysis was made using the statistical software package GAUSS.

4 Simulation

4.1 Simulation design

All simulations involve generating gamma-distributed frailties, bivariate lifetimes, dependent and independent censoring times and truncation times. We will try to mimic the characteristics of the Danish twin data, which we analyse in the next section. A total of 8000 twin pairs (3000 MZ and 5000 DZ pairs) are simulated, a number which is reduced by the truncation process to a final sample size of around 4200–4300 twin pairs. Samples are generated as follows:

- Generate frailty variables Z_1, Z_2, Z_3, Z_4 using independent gamma-distributed random variables $V_1, \ldots, V_8$.
- Generate lifetimes with respect to the first (X_1, X_2) and second disease (Y_1, Y_2) given the frailties using $S_i(x) = (1 + s_i^2 \alpha_i / \beta_i (e^{\beta_i x} - 1))^{-\frac{1}{s_i^2}}$, $(i = 1, 2)$
- The censored (bivariate) lifetimes $T_{ij} = \min\{X_{ij}, Y_{ij}, C_{ij}\}$ are generated by using the lifetimes with respect to the second cause of death as dependent censoring times and uniform distributed random variables on [40,100] as independent censoring times.
- Birth years are generated by using a uniform distribution on [1870,1930] to mimic the truncation pattern.
- Year of truncation is 1943.

The simulation program was written using GAUSS language. We simulated 1000 data sets.

4.2 Simulation results

The mean parameter estimates of the model are shown in Table 1, in comparison with the true values used for simulation. Although there appears to be some bias in certain parameter

Table 1 Parameter estimation in the simulation study

Parameter	True value	Mean of estimates	Standard deviation
α_1	1.000	1.008	0.278
β_1	0.120	0.121	0.008
s_1	2.000	1.991	0.200
α_2	1.000	0.974	0.303
β_2	0.120	0.123	0.009
s_2	2.000	2.061	0.254
σ_1	2.000	1.964	0.288
σ_2	2.000	2.174	0.689
$\rho_1(MZ)$	0.400	0.408	0.073
$\rho_1(DZ)$	0.200	0.205	0.051
$\rho_2(MZ)$	0.100	0.107	0.064
$\rho_2(DZ)$	0.060	0.067	0.047
ρ	0.500	0.539	0.237

Table 2 Parameter estimation in a simulated data set

Parameter	True value	Estimates	Standard deviation
α_1	1.000	1.070	0.317
β_1	0.120	0.119	0.008
s_1	2.000	1.902	0.234
α_2	1.000	1.074	0.331
β_2	0.120	0.120	0.010
s_2	2.000	1.934	0.371
σ_1	2.000	2.056	0.375
σ_2	2.000	2.041	0.583
a^2	0.400	0.285	0.112
c^2	0.000	0.060	0.074
$\rho_2(MZ)$	0.100	0.098	0.044
$\rho_2(DZ)$	0.060	0.067	0.036
ρ	0.500	0.610	0.366
e^2	0.600	0.655	0.075

estimates, the magnitude does not appear to be of any practical significance and the overall performance is quite accurate.

To give the reader the possibility to reproduce the results presented in the article and to apply the software to their own problems we simulated a data set which is available on the website of the journal. Parameter estimates for this simulated data are given in Table 2.

5 Example

In our example, we investigated how well the method performed when used to analyse the respective influence of genetic and environmental factors affecting risk of mortality from coronary heart disease (CHD). In this example the second cause of death is all other causes combined. The data we use for our analysis are the survival times of MZ and DZ female twins sampled from the Danish Twin Registry, the first national twin registry world-wide (established in 1954 by Harvald and Hauge). This population-based registry includes all

twins born in Denmark during the period 1870–1910 and all same-sex pairs born between 1911 and 1930. For detailed information about the Danish Twin Registry see Hauge (1981). The data set contains records of female twin pairs born between 1 January 1870 and 31 December 1930 and both individuals were still alive on 1 January 1943. Consequently, the observations are left truncated. Pairs with at least one death before 31 December 1993 and incomplete ICD information or unknown zygosity were excluded. Individuals were followed up to 31 December 1993 and subjects identified as deceased after that date were classified here as alive. In total, we sampled 1470 MZ twin pairs and 2730 DZ twin pairs.

In addition to the lifetime data, there is documentation regarding the cause of death for all non-censored lifetimes. During the follow-up, 369 deaths due to CHD occured among MZ twins and 704 deaths among DZ twins. CHD was defined as ICD code 420 (revision 6 and 7) and 410–414 (revision 8). Death status, age at death, and cause of death were obtained from the Central Person Register, the Danish Cause-of-death Register, the Danish Cancer Register (founded in 1942), and other public registers in Denmark. The validity of the twin register has been checked by comparing information about year of death with the nationwide Danish Cancer Register. Both registers were independent, but 99% agreement was found (Holm, 1983). Further data corrections increased this part to almost 100%. Zygosity was determined by self reported similarities. Validations of this zygosity classification by comparing with laboratory methods (serological markers) show a misclassification rate of less than 5% (Lykken, 1978; Holm, 1993). For more detailed information about status, zygosity and cause of death in the population under study see Tables 3 and 4.

First, we compared the ADE and DE as well as the ACE and CE models. The likelihood ratio test prefers the DE and the ACE model (Table 5). The ACE model converges to the AE model. Standard errors are not given in the ACE model since 0 is the boundary of the parametric space. A comparison of the AE and the DE model is impossible with respect to the likelihood ratio test because the models are not nested. According to the Akaike Information Criterion (AIC), the AE model fit the data best and gives a inheritance estimate of 0.39, with standard error 0.13. Using the sub-model of independent causes of death ($\rho = 0$, model (2) and (4)), the inheritance estimate was 0.58 (0.14).

Table 3 Study population (number of pairs) by zygosity and life status

Status	Monozygotic twins	Dizygotic twins
Both twins dead	622	1072
One twin alive, co-twin dead	332	773
Both twins alive	516	885
All pairs together	1470	2730

Table 4 Study population (number of individuals) by zygosity and cause of death

Cause of death	Monozygotic twins	Dizygotic twins
Coronary heart disease	369	704
All causes together	1576	2917
Alive (censored)	1364	2543

Table 5 Genetic analysis of CHD

	σ	a^2	d^2	c^2	e^2	ρ	Log-L
ACE	1.70 (—)	0.39 (—)		0.00 (—)	0.61 (—)	0.45 (—)	22268.06
AE	1.70 (0.21)	0.39 (0.13)			0.61 (0.13)	0.45 (0.12)	22268.06
ADE	1.63 (0.23)	0.28 (0.24)	0.13 (0.26)		0.59 (0.12)	0.49 (0.06)	22267.81
DE	1.63 (0.24)		0.44 (0.12)		0.56 (0.12)	0.49 (0.07)	22268.61
CE	1.80 (0.26)			0.22 (0.00)	0.78 (0.07)	0.54 (0.06)	22271.93
AE*	1.87 (0.41)	0.58 (0.14)			0.42 (0.14)	0.00	22269.24

σ^2: variance of frailty; a^2: additive genetic effects; d^2: genetic effects due to dominance; c^2: shared environment; e^2: non-shared environment; ρ: correlation between frailties associated with competing risks; Log-L: value of the Log-Likelihood function divided by number of observations; DE*: DE model with $\rho = 0$ (independent model).

6 Discussion

Frailty models are mixture models within survival analysis. In survival analysis, one typically has to deal with censored observations. In most applications censoring is assumed to be simply non-informative. This assumption is realistic for example in clinical studies, where patients contribute censored observations because they are still alive at the preassigned termination point of study. Some others get lost during the time of follow-up for reasons that are not related to the event under study. In such cases censoring can be assumed to be non-informative. However, in some cases this assumption is questionable, especially in cases where there are competing causes of death. This paper has suggested using an extension of the bivariate correlated gamma-frailty model (Pickles *et al.*, 1994; Yashin and Iachine, 1995) in such cases, where only a part of the censored observations is assumed to be non-informatively censored. Because competing risks can also be correlated within families and may share unobserved dependencies with the cause of interest, the standard approach, which treats competing risks as independent, could lead to biased estimates of the variance components associated with the cause of interest. Here, the frailties are modelled in terms of standard variance components for additive and dominance genetic effects and shared and unique environmental effects. This thus provides a rich class of models for analysing this complex pattern of dependencies between family members and between causes of death. ③ Furthermore, frailty models are well suited for inclusion of observed covariates into the analysis (Wienke *et al.*, 2002).

Using cause-specific mortality data of relatives (here twins) it is possible to overcome problems due to identifiability in univariate censored lifetimes as stated in Tsiatis (1975). The model we have evolved allows for dependencies among competing risks and makes it possible to test for such dependencies. Furthermore, combining methods from survival analysis (especially from frailty models) and genetic analysis as we did, improves the genetic analysis of time-to-event data in the case of informative and non-informative censoring together as well as accounting for heterogeneity in the population. Our example is an extension of the analysis in the case of independent causes of death (Wienke *et al.*, 2000, 2001), where deaths from other causes than the cause under study are treated as non-informative and collapsed with censored observations caused by end of study. In both cases (here called dependent and independent) the AE model is the best fitting model for CHD. This shows a certain degree of consistency in the model. Comparing both cases, it turns out,

that the heritability of frailty on mortality due to CHD change substantially. Fixing the correlation of 0.45 (0.12) between frailty on mortality from CHD and frailty on mortality from other causes to zero has a impact on the heritability estimate—changing it from 0.58 (0.14) to 0.39 (0.13). Both models detected the significant influence of genetic factors. The parameter ρ can be used to test the hypothesis of dependence between the competing risks. The likelihood ratio test indicates that the simpler independent model is sufficient to describe the data.

Mortality pattern in twins and in the general population are very similar (Christensen *et al.*, 1996, 2001), which is an important argument for generalising the results of twin studies to the general population.

The proof of consistency and asymptotic normality of the maximum likelihood estimators is still an open problem, but our simulation results seem to point to the asymptotic validity of the proposed method.

One important limitation of the presented model should be kept in mind, the correlation coefficient between the frailties are always non-negative by construction. This restriction makes sense when comparing the lifetimes of relatives, but it is not clear that the same holds for the competing risks in an individual. On one hand, many major diseases have risk factors in common and consequently, the presence of any one of these risk factors will increase the risk of death with respect to all diseases. On the other hand, everyone dies eventually, so logically, if the risk of death from one cause is decreased the risk from another cause must be increased, Furthermore, the parameter ρ is only identifiable in a 'real' multivariate case. Pairs of unrelated individuals (e.g. $\rho_1 = \rho_2 = 0$) implies the univariate case, which makes the parameter ρ non-identifiable. The nature of dependencies among competing risks deserves further study.

Classical twin studies are based on the important assumption that MZ and DZ twins have the same correlation in environments (equal environment assumption). This standard assumption is necessary for the identifiability of heritability i.e. so as to be able to interpret the difference in concordance between MZ and DZ twins as being explained in full by their difference in genetic concordance. However, without doubt, the assumption is also questionable: MZ twins are generally treated the same by their parents to a much greater extent than DZ twins by their parents. This implies an overestimation of heritability. The equal environment assumption seems to be acceptable with respect to environmental factors related to CHD.

The suggested model gives a clear illustration of how the methods of survival analysis and genetic epidemiology may be merged to improve the genetic investigation of time-to-event data. Further extensions of the model to multiple causes of death and/or multiple related lifetimes will be important in elucidating the properties of this strategy.

Acknowledgements

The authors wish to thank the Danish Twin Register for providing the data and Susann Backer for help in preparing the paper for publication. The research was partly supported by NIH/NIA grant 7PO1AG08761.

References

Christensen K, Vaupel JW, Holm NV, Yashin AI (1996) Mortality among twins after age 6: fetal origins hypothesis versus twin method. *British Medical Journal*, **310**, 432–36.

Christensen K, Wienke A, Skytthe A, Holm NV, Vaupel JW, Yashin AI (2001) Cardiovascular mortality in twins and the fetal origins hypothesis. *Twin Research*, **5**, 344–49.

Clayton DG, Cuzick J (1985) Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society*, *Series A*, **148**, 82–117.

Hauge M (1981) The Danish Twin Register. In Mednich SA, Baert AE, Bachmann BP, eds. *Prospective longitudinal research*. Oxford: Oxford Medical Publications, 217–22.

Heath AC, Neale MC, Hewitt JK, Eaves LJ, Fulker DW (1989) Testing structural equation models for twin data using LISREL. *Behavior Genetics*, **19**, 9–35.

Holm NV (1983) The use of twin studies to investigate causes of diseases with complex etiology with a focus on cancer. PhD thesis, University of Odense.

Hougaard P, Harvald B, Holm NV (1992) Measuring the similarities between the lifetimes of adult Danish twins born between 1881–1930. *Journal of the American Statistical Association*, **87**, 17–24.

Meyer JM, Eaves LJ, Heath AC, Martin NG (1991) Estimating genetic influences on the age at menarche: A survival analysis approach. *American Journal of Medical Genetics*, **39**, 148–54.

Neale MC, Cardon LR (1992) *Methodology for genetic studies of twins and families*. Dordrecht: Kluwer.

Neale MC, Eaves LJ, Hevitt JK, MacLean CJ, Meyer JM, Kendler KS (1989) Analysing the relationship between age at onset and risk to relatives. *American Journal of Human Genetics*, **45**, 226–39.

Pickles A, Crouchley R, Simonoff E, Eaves LJ, Meyer JM, Rutter M, Hevitt JK, Silberg J (1994) Survival models for developmental genetic data: age of onset of puberty and antisocial behavior in twins. *Genetic Epidemiology*, **11**, 155–70.

Tsiatis AA (1975) A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, **72**, 20–22.

Vaupel JW (1988) Inherited frailty and longevity. *Demography*, **25**, 277–87.

Wienke A, Christensen K, Holm NV, Yashin AI (2000) Heritability of death from respiratory diseases: an analysis of Danish twin survival data using a correlated frailty model. In Hasman *et al.*, eds. *Medical Infobahn for Europe*. Amsterdam: IOS Press, 407–11. ④

Wienke A, Holm N, Skytthe A, Yashin AI (2001) The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. *Twin Research*, **4**, 266–74.

Yashin AI, Iachine IA (1994) Environment determines 50% of variability in individual frailty: results from Danish twin study. Research Report, Population Studies of Aging 10, Odense University, Denmark.

Yashin AI, Iachine IA (1995) Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. *Genetic epidemiology*, **12**, 529–38.

Appendix

The following relations are valid: $\mathbf{E}Z_1 = \mathbf{E}Z_2 = \mathbf{E}Z_3 = \mathbf{E}Z_4 = 1$, $\mathbf{V}(Z_1) = \mathbf{V}(Z_3) = \dfrac{1}{k_1 + k_3 + k_5} = \sigma_1^2$, $\mathbf{V}(Z_2) = \mathbf{V}(Z_4) = \dfrac{1}{k_1 + k_3 + k_4} = \sigma_2^2$.

$$\mathbf{E}V_2^2 = \mathbf{V}(V_2) + (\mathbf{E}V_2)^2 = \frac{k_2}{\lambda_1^2} + \left(\frac{k_2}{\lambda_1}\right)^2 = \frac{k_2^2 + k_2}{\lambda_1^2}$$

$$\begin{aligned}
\mathbf{E}Z_1Z_3 &= \mathbf{E}\left(\frac{\lambda_0}{\lambda_1}V_1 + V_2 + V_5\right)\left(V_2 + V_6 + \frac{\lambda_0}{\lambda_1}V_8\right) \\
&= \mathbf{E}\left(\frac{\lambda_0}{\lambda_1}V_1V_2 + \frac{\lambda_0}{\lambda_1}V_1V_6 + \frac{\lambda_0^2}{\lambda_1^2}V_1V_8 + V_2^2 + V_2V_6 + \frac{\lambda_0}{\lambda_1}V_2V_8 + V_2V_5 + V_5V_6 + \frac{\lambda_0}{\lambda_1}V_5V_8\right) \\
&= \frac{k_1k_2}{\lambda_1^2} + \frac{k_1k_5}{\lambda_1^2} + \frac{k_1^2}{\lambda_1^2} + \frac{k_2^2 + k_2}{\lambda_1^2} + \frac{k_2k_5}{\lambda_1^2} + \frac{k_1k_2}{\lambda_1^2} + \frac{k_2k_5}{\lambda_1^2} + \frac{k_5^2}{\lambda_1^2} + \frac{k_1k_5}{\lambda_2^2} = \frac{k_2}{\lambda_1^2} + 1
\end{aligned}$$

$$\mathbf{cov}(Z_1, Z_3) = \mathbf{E}Z_1Z_2 - \mathbf{E}Z_1\mathbf{E}Z_2 = \frac{k_2}{\lambda_1^2}$$

$$\rho_1 = \frac{\mathbf{cov}(Z_1, Z_3)}{\sqrt{V(Z_1)V(Z_3)}} = \frac{k_2}{\lambda_1} = k_2\sigma_1^2 \tag{A.1}$$

Similar calculations imply $\rho_2 = k_3\sigma_2^2$ and $\rho = k_1\sigma_1\sigma_2$. Consequently, $k_1 + k_2 + k_5 = 1/\sigma_1^2$ and $k_1 + k_3 + k_4 = 1/\sigma_2^2$ imply the following relations:

$$k_5 = \frac{1}{\sigma_1^2} - k_2 - k_1 = \frac{1}{\sigma_1^2} - \frac{\rho_1}{\sigma_1^2} - \frac{\rho}{\sigma_1\sigma_2} \qquad \text{and} \qquad k_4 = \frac{1}{\sigma_2^2} - k_3 - k_4 = \frac{1}{\sigma_2^2} - \frac{\rho_2}{\sigma_2^2} - \frac{\rho}{\sigma_1\sigma_2}.$$

If $Y \sim \Gamma(k, \lambda)$, then $\mathbf{E}e^{-sY} = (1 + \frac{s}{\lambda})^{-k}$. Now we are able to derive the survival function in (2.7):

$$\begin{aligned}
&\mathbf{E}S_1(x_1)^{Z_1}S_2(y_1)^{Z_2}S_1(x_2)^{Z_3}S_2(y_2)^{Z_4} \\
&\quad = \mathbf{E}e^{-V_1(\frac{\lambda_0}{\lambda_1}H_1(x_1) + \frac{\lambda_0}{\lambda_2}H_2(y_1))}e^{-V_2(H_1(x_1) + H_1(x_2))} \\
&\quad * e^{-V_3(H_2(y_1) + H_2(y_2))}e^{-V_8(\frac{\lambda_0}{\lambda_1}H_1(x_2) + \frac{\lambda_0}{\lambda_2}H_2(y_2))}e^{-V_4H_2(y_1)}e^{-V_5H_1(x_1)}e^{-V_6H_1(x_2)}e^{-V_7H_2(y_2)} \\
&\quad = \left(1 + \frac{1}{\lambda_0}\left(\frac{\lambda_0}{\lambda_1}H_1(x_1) + \frac{\lambda_0}{\lambda_2}H_2(y_1)\right)\right)^{-k_1}\left(1 + \frac{1}{\lambda_1}H_1(x_1) + \frac{1}{\lambda_1}H_1(x_2)\right)^{-k_2} \\
&\quad * \left(1 + \frac{1}{\lambda_2}H_2(y_1) + \frac{1}{\lambda_2}H_2(y_2)\right)^{-k_3}\left(1 + \frac{1}{\lambda_0}\left(\frac{\lambda_0}{\lambda_1}H_1(x_2) + \frac{\lambda_0}{\lambda_2}H_2(y_2)\right)\right)^{-k_1} \\
&\quad * \left(1 + \frac{1}{\lambda_2}H_2(y_1)\right)^{-k_4}\left(1 + \frac{1}{\lambda_1}H_1(x_1)\right)^{-k_5}\left(1 + \frac{1}{\lambda_1}H_1(x_2)\right)^{-k_5}\left(1 + \frac{1}{\lambda_2}H_2(y_2)\right)^{-k_4} \\
&\quad = (S_1(x_1)^{-\sigma_1^2} + S_1(x_2)^{-\sigma_1^2} - 1)^{-\rho_1/\sigma_1^2}(S_2(y_1)^{-\sigma_2^2} + S_2(y_2)^{-\sigma_2^2} - 1)^{-\rho_2/\sigma_2^2} \\
&\quad * (S_1(x_1)^{-\sigma_1^2} + S_2(y_1)^{-\sigma_2^2} - 1)^{-\rho/\sigma_1\sigma_2}(S_1(x_2)^{-\sigma_1^2} + S_2(y_2)^{-\sigma_2^2} - 1)^{-\rho/\sigma_1\sigma_2} \\
&\quad * S_2(y_1)^{1-\rho_2-\frac{\sigma_2}{\sigma_1}\rho}S_1(x_1)^{1-\rho_1-\frac{\sigma_1}{\sigma_2}\rho}S_1(x_2)^{1-\rho_1-\frac{\sigma_1}{\sigma_2}\rho}S_2(y_2)^{1-\rho_2-\frac{\sigma_2}{\sigma_1}\rho}
\end{aligned}$$

The likelihood function of the (left truncated data) is of the following form:

$$
\begin{aligned}
L(x_1, y_1, x_2, y_2, \delta_1, \delta_2) = {} & (\delta_1\delta_2 S_{x_1y_1}(x_1, y_1, x_2, y_2) + \delta_1(1+\delta_2) S_{x_1y_2}(x_1, y_1, x_2, y_2) \\
& + (1+\delta_1)\delta_2 S_{x_2y_1}(x_1, y_1, x_2, y_2) + (1+\delta_1)(1+\delta_2) S_{x_1y_1}(x_1, y_1, x_2, y_2) \\
& + \delta_1(1-\delta_2) S_{x_1}(x_1, y_1, x_2, y_2) + (1-\delta_1)\delta_2 S_{y1}(x_1, y_1, x_2, y_2) \\
& + (1-\delta_1)(1+\delta_2) S_{y_2}(x_1, y_1, x_2, y_2) + (1+\delta_1)(1-\delta_2) S_{x_2}(x_1, y_1, x_2, y_2) \\
& + (1-\delta_1)(1-\delta_2) S(x_1, y_1, x_2, y_2))/S(w, w, w, w)
\end{aligned}
$$

Journals Offprint Order Form

A member of the *Hodder Headline Group*

Arnold Journals
338 Euston Road
London NW1 3BH
Tel: +44 (0)20 7873 6000
Fax: +44 (0)20 7873 6325

This form should be returned at once to the above address

- Title of Journal: Stat Modelling 2(2)
- 1st Author:

A. FREE OFFPRINTS - 25 offprints of your article will be supplied free of charge

Please indicate opposite, the name and full postal address to whom they should be sent. In the case of multi-author articles, free offprints are only sent to the corresponding author.

_____

—

_____

—

_____

—

_____

—

B. PURCHASE OF ADDITIONAL OFFPRINTS

Please note that if an article is by more than one author, only one offprint form is sent and all offprints should be ordered on that form in consultation with the co-authors.

Offprint Price List (£ sterling UK and Europe; US$ Rest of World)

	25		50		100		150		200	
	£	$	£	$	£	$	£	$	£	$
1-4 pages	70	98	91	128	138	194	210	295	252	354
5-8 pages	92	130	122	171	185	259	255	357	332	466
9-16 pages	121	170	138	194	210	295	279	391	369	517
17-24 pages	138	194	160	224	231	324	330	463	420	588
Extra 8 pages	13	18	20	28	24	34	29	41	42	59

For larger quantities contact the publisher for a quotation.

Add 100% for any offprints including colour reproduction.

I wish to purchase .....additional offprints

ADDRESS FOR DELIVERY

(please print in capitals)

_____

_____

_____

_____

_____

IMPORTANT

1. Cheques drawn on a UK or US bank should be made payable to Hodder Headline Group. We are unable to accept credit or debit card payments.
2. Orders will not normally be mailed until the publisher is in receipt of either the appropriate payment or an official purchase order.
3. The above are prepublication prices and apply only to orders received before the publication goes to press.
4. All despatches are by surface mail, normally within four weeks of publication.

ADDRESS FOR INVOICE (please print in capitals)

_____

_____

_____

_____

_____

Payment enclosed

Please invoice

Official order follows

Official order attached

5. Claims cannot be considered more than three months after despatch. **VAT will be added to UK invoices. Members of the EU will be required to pay VAT unless a VAT number is provided with order.**	no. ..... Signed..... Date...../...../.....

TRANSFER OF COPYRIGHT

Please return completed form to:

Arnold Journals
338 Euston Road
London NW1 3BH
UK.

A member of the
Hodder Headline
PLC Group

STATISTICAL MODELLING

In consideration of the publication in the above journal *Statistical Modelling* the contribution entitled

……. ("the contribution")

by (all authors' names): .....

1. To be filled in if copyright belongs to you

I/we warrant that I am/we are the sole owner/s of the complete copyright in the Contribution and I/we hereby assign to Arnold (Publishers) Limited the complete copyright in the Contribution in all formats and media.

2. To be filled in if copyright does not belong to you

a) Name and address of copyright holder.....
.....
.....

b) I/we warrant that I am/we are the sole owner/s of the complete copyright in the Contribution and I/we hereby grant Arnold (Publishers) Limited the non-exclusive right to publish the Contribution throughout the world in all formats and media and to deal with requests from third parties in the manner specified in paragraphs 2 and 4 overleaf.

3. To be filled in if US Government exemption applies

I/we certify that the Contribution was written in the course of employment by the United States Government, and therefore copyright protection is not available.

4. I/we warrant that I/we have full power to enter into this Agreement, and that the Contribution does not infringe any existing copyright, or contain any scandalous, defamatory, libellous or unlawful matter.

Signed as (tick one)

- ❑ the sole author(s) of the Contribution
- ❑ one author authorised to execute this transfer on behalf of all the authors of the Contribution
- ❑ the copyright holder or authorised agent of the copyright holder of the Contribution

Name (block letters) .....

Address .....

Signature ..... Date .....

(Additional authors should provide this information on a separate sheet please)

Notes for Contributors

1. The Journal's policy is to acquire copyright in all contributions. There are two reasons for this: (a) ownership of copyright by one central organisation tends to make it easier to maintain effective international protection against unauthorised use; (b) it also allows for requests from third parties to reprint or reproduce a contribution, or part of it, to be handled in accordance with a general policy which is sensitive both to any relevant changes in international copyright legislation and to the general desirability of encouraging the dissemination of knowledge.

2. Arnold co-operates in various licensing schemes which allow organisations to copy material within agreed restrains (e.g. the CLA in the UK and the CCC in the USA).

3. All contributors retain the rights to reproduce their paper for their own purposes provided no sale is involved, and to reprint their paper in any volume of which they are editor or author. Permission will automatically be given to the publisher of such a volume, subject to the normal acknowledgement.

4. It is understood that in some cases copyrights will be held by the contributor's employer. If so, Arnold requires non-exclusive permission to deal with requests from third parties, on the understanding that any requests it receives will be handled in accordance with paragraph 3.

5. Arnold will provide each contributor with a complimentary copy of the issue of the Journal in which the Contribution appears.