On the conditional probability for assessing time dependence of association in shared frailty models with bivariate current status data

Steffen Unkel*

Department of Medical Statistics
University Medical Center Göttingen, Germany

May 5, 2016

Abstract

Shared frailty models are frequently used for inducing dependence between survival times. In this paper, we consider bivariate current status data that are reasonable to model by shared frailty models. A time-dependent association measure that has a conditional probability interpretation is revisited for its potential application to such data. We propose a method of estimation and derive asymptotic standard errors for this measure. Its small sample performance and its performance in assessing the temporal variation in the strength of association in realistic scenarios is investigated by means of experiments. We show that the measure based on the conditional probability can vary with time even in the absence of any time-dependent effects. Furthermore, we give evidence that it lacks interpretability in suggesting appropriate frailty models. We provide an illustration with multivariate current status data arising from a community-based study of cardiovascular diseases in Taiwan. We compare the observed patterns of association with the ones obtained by employing a fairly new time-varying association measure that is relevant for shared frailty models, owing to its connection to the cross-ratio function, and which serves as a diagnostic tool for suggesting classes of frailty distributions with constant, increasing or decreasing association over time.

Key words: Conditional probability; Current status data; Cross-sectional data; Frailty models; Heterogeneity; Time-varying association.

* Correspondence should be addressed to: Dr. Steffen Unkel, Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany (e-mail: steffen.unkel@med.uni-goettingen.de).

1 Introduction

Bivariate current status data can be formally represented as $\{X, \delta_1 = I(T_1 \leq X), \delta_2 = I(T_2 \leq X)\}$, where I denotes the indicator function, T_1 and T_2 are the failure times of interest and X is the monitoring time at which T_1 and T_2 are measured from the same observational units and that is independent of the failure times (Jewell et al. 2005; Sun 2006). In this paper, we consider bivariate current status data that are reasonable to model by shared frailty models (Duchateau and Janssen 2008; Hougaard 2000; Wienke 2011), with the frailty solely generating the association structure between T_1 and T_2 and the variability between the observational units, also referred to as the heterogeneity in the data, being represented by the variance of the frailty. Such bivariate current status data arise in various fields (Jewell et al. 2005). Consider for example tumorigenicity experiments on a single non-lethal tumor at two different sites, e.g. liver and brain, to investigate whether the environment accelerates the time until tumor onset in animals. In these experiments, the time to tumor onset in the animals is only known to be less than or greater than the observed time of death or sacrifice. A shared frailty model is natural in this setting, the shared latent frailty representing environmental exposures relevant to the progression of disease at different sites.

Another examples arises in twin pair studies in genetics, where the phenotypes of interest are the ages at onset of a specific disease. For neurological disorders such as Alzheimer's disease, the exact age of onset is usually not known even when a definitive diagnosis is available. If T_j $(j = 1, 2)$ is the unknown age of onset for the jth twin, then in such cases only bivariate current status information (δ_1, δ_2) is observed instead of (T_1, T_2). In other words, it is only known at the monitoring time whether the jth twin has the disease or not. Interest may focus on the strength of association between T_1 and T_2 for both monozygotic and dizygotic twins. Again, a shared frailty model is natural, the latent frailty variable representing genetic characteristics that may have a bearing on the onset of the disease of interest.

In some circumstances, it is also of interest to assess the time dependence of association (Anderson et al. 1992; Oakes 1989), for example to investigate the age-varying influence of

genetic factors on the disease-free life expectancy of individuals by comparing the disease-free life spans of monozygotic and dizygotic twins.

In shared frailty models for bivariate survival data, the frailty distribution is identifiable through Clayton's local cross-ratio function (Clayton 1978), which describes how the heterogeneity of the hazard functions in the survivor population evolves over time. Hence, the cross-ratio function may serve as a diagnostic tool in an exploratory analysis for suggesting appropriate frailty distributions and assessing the temporal variation in the strength of association in bivariate survival data (Farrington et al. 2012; Viswanathan and Manatunga 2001). This association measure is unavailable for current status data, though.

The odds ratio is the most obvious and popular association measure for binary data and is widely used in many fields, such as in epidemiological studies as a measure of association between the occurrence of a particular disease state or condition and an exposure factor (Jewell 2003). It can easily be estimated after fitting a linear logistic model to dichotomous data (Collett 2002). The odds ratio can be computed from current status data. However, the odds ratio suffers the disadvantages that it can vary with time even in the absence of any time-dependent effects and that it does not reliably suggest appropriate frailty distributions (Unkel and Farrington 2012).

Anderson et al. (1992) introduced the following time-dependent measure for association based on the conditional probability:

$$\psi(t_1, t_2) = \frac{\mathrm{P}(T_1 > t_1 | T_2 > t_2)}{\mathrm{P}(T_1 > t_1)} = \frac{S(t_1, t_2)}{S_1(t_1) S_2(t_2)} \ , \tag{1}$$

where $S_j(t_j) = \mathrm{P}(T_j > t_j)$ denotes the marginal survivor function for T_j $(j = 1, 2)$ and $S(t_1, t_2) = \mathrm{P}(T_1 > t_1, T_2 > t_2)$ is the joint survivor function. Large values of $\psi(t_1, t_2)$ indicate positive dependence between T_1 and T_2. For independent events $T_1 > t_1$ and $T_2 > t_2$, $\psi(t_1, t_2) = 1$. If $S(t_1, t_2) < S(t_1)S(t_2)$, then there is negative dependence between T_1 and T_2. In Anderson et al. (1992) the measure (1), indexed by age, is applied to right-censored data from the Danish Twin Registry to describe the differences in the strength of association between monozygotic and dizygotic twins with respect to their life spans and to investigate how these associations depend on the age of the twins.

Unfortunately, for current status data, the joint survivor function $S(t_1, t_2)$ is unobservable;

only $S(x, x)$, where $X = x$ denotes the observed monitoring (censoring) time, is available along with the marginals $S_1(x) = S(x, 0)$ and $S_2(x) = S(0, x)$. This means that for current status data, one can assess the association between two survival variables by means of

$$\psi(x) = \frac{S(x, x)}{S_1(x)S_2(x)} \ . \tag{2}$$

In this paper, we investigate the measure (2) and pay particular attention to the evaluation of its usefulness to govern the temporal variation in the strength of association inherent in bivariate current status data and in serving as an exploratory tool for suggesting frailty distributions. We propose a method of estimation and introduce asymptotic standard errors for (2). We also apply an association measure introduced in Unkel and Farrington (2012), which is based on the association parameter derived from Clayton's copula (Clayton 1978) for quantifying time-dependent association. This measure tracks the variation of the cross-ratio function with time. Therefore, it serves as a diagnostic tool for suggesting classes of frailty distributions with constant increasing or decreasing association over time. The shape of the observed time-varying association aids identification of a suitable frailty model, which then could be fitted to the data set at hand. Up to the author's knowledge, the association measure by Unkel and Farrington (2012) has been only applied so far to bivariate serological survey data on pairs of infections with similar and different transmission routes, where the time-varying association is likely to represent heterogeneities in activity levels and/or susceptibility to infection (see also Unkel et al. (2014) and Farrington et al. (2013)).

In the present paper, our methods are illustrated with multivariate current status data arising from a community-based study on three cardiovascular diseases in Taiwan. These data were originally analyzed by Wang and Ding (2000) under the assumption of constant pairwise association over time. We explore the possible time dependence of association in the data. The remainder of the paper is organised as follows. In Section 2 we present maximum likelihood estimators for the association measure based on the conditional probability and derive asymptotic standard errors. An evaluation of how the conditional probability measure performs with respect to identifying time-varying effects in shared frailty models with bivariate current status data is given in Section 3. We also investigate the finite sample performance of the association measure in realistic scenarios

using simulations. In Section 4 the methodology discussed in this paper is applied to the aforementioned data set. Concluding comments are given in Section 5. Computations for this paper were carried out using the software package R version 3.2.1 (R Core Team 2015). All computer code used is available upon request.

2 Maximum likelihood estimation and asymptotic standard errors

Let $\pi_{00} = \mathrm{P}(T_1 > t_1, T_2 > t_2) = S(t_1, t_2)$, $\pi_{10} = \mathrm{P}(T_1 \leq t_1, T_2 > t_2)$, $\pi_{01} = \mathrm{P}(T_1 > t_1, T_2 \leq t_2)$ and $\pi_{11} = \mathrm{P}(T_1 \leq t_1, T_2 \leq t_2)$. In the sequel, the time scale is age and $X = x$ is the age at which individuals are monitored. In terms of bivariate current status data, let $\pi_{00}(x)$ be the probability that an individual of age x has experienced neither of the two events and $\pi_{10}(x)$ the probability that an individual of age x has experienced event 1 but not event 2, and similarly define $\pi_{01}(x)$ and $\pi_{11}(x)$. The conditional probability measure provides insights into the time-dependent nature of association by estimating $\psi(x) = \frac{\pi_{00}(x)}{\pi_{0+}(x)\,\pi_{+0}(x)}$ at each age x for which data are available, where $\pi_{0+}(x) = \pi_{00}(x) + \pi_{01}(x)$ and $\pi_{+0}(x) = \pi_{00}(x) + \pi_{10}(x)$.

Bivariate current status data on n_x fixed individuals of age x give rise to a multinomial observation $(n_{00x}, n_{10x}, n_{01x}, n_{11x})$, where $\sum_{i,j=0,1} n_{ijx}$ and n_{00x} is the number of individuals of age x in the sample for whom neither event has occurred, n_{10x} is the number of individuals that have not experienced event 2 but have experienced event 1, and so on. From the log-likelihood contribution $l_x = \sum_{i,j=0,1} n_{ijx} \ln(\pi_{ij}(x))$, one can obtain a maximum likelihood estimate (MLE) of $\psi(x)$ as

$$\hat{\psi}(x) = \frac{n_x n_{00x}}{n_{0+x} n_{+0x}} \ , \tag{3}$$

where $n_{0+x} = n_{00x} + n_{01x}$ and $n_{+0x} = n_{00x} + n_{10x}$. It is customary to work with $\ln(\hat{\psi})$ than with $\hat{\psi}$ itself, so we shall do so in the sequel.

To obtain asymptotic standard errors for $\ln(\hat{\psi})$, the delta method for a function of multinomial counts is applied (Agresti 2013, Section 16.1). Consider a 2×2 contingency table with multinomial cell counts $(n_{00}, n_{01}, n_{10}, n_{11})$. Let $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})^\top$ be the vector of cell probabilities with sample proportions $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{00}, \hat{\pi}_{01}, \hat{\pi}_{10}, \hat{\pi}_{11})^\top$, where $\hat{\pi}_{ij} = n_{ij}/n$

for $i, j = 0, 1$ and sample size $n = \sum_{i,j=0,1} n_{ij}$. The multivariate central limit theorem implies

$$\sqrt{n}\,[\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top) \ ,$$

where $\mathrm{diag}(\boldsymbol{\pi})$ is a 4×4 diagonal matrix with the elements of $\boldsymbol{\pi}$ on its main diagonal. Suppose that $g(u_{00}, u_{01}, u_{10}, u_{11})$ is a differentiable function that has a nonzero differential $\boldsymbol{\zeta} = (\zeta_{00}, \zeta_{01}, \zeta_{10}, \zeta_{11})^\top$ at $\boldsymbol{\pi}$, where

$$\zeta_{ij} = \frac{\partial g}{\partial \pi_{ij}} \quad (i, j = 0, 1)$$

denote $\partial g / \partial u_{ij}$ evaluated at $\mathbf{u} = \boldsymbol{\pi}$ with $\mathbf{u} = (u_{00}, u_{01}, u_{10}, u_{11})^\top$. By the delta method (Agresti 2013, Section 16.1),

$$\sqrt{n}\,[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})] \xrightarrow{d} \mathcal{N}(0, \sigma^2) \ ,$$

where the asymptotic variance σ^2 equals

$$\sigma^2 = \boldsymbol{\zeta}^\top \left[\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top\right] \boldsymbol{\zeta} = \sum_i \sum_j \pi_{ij} \zeta_{ij}^2 - \left(\sum_i \sum_j \pi_{ij} \zeta_{ij} \right)^2 \ .$$

We apply the delta method to the log conditional probability, taking

$$\begin{aligned} g(\boldsymbol{\pi}) = \ln(\psi) &= \ln\left(\frac{\pi_{00}}{\pi_{0+}\pi_{+0}}\right) = \ln(\pi_{00}) - \ln(\pi_{0+}\pi_{+0}) \\ &= \ln(\pi_{00}) - (\ln(\pi_{00} + \pi_{01}) + \ln(\pi_{00} + \pi_{10})) \\ &= \ln(\pi_{00}) - \ln(1 - \pi_{11} - \pi_{10}) - \ln(1 - \pi_{11} - \pi_{01}) \ . \end{aligned}$$

The partial derivatives $\zeta_{ij} = \frac{\partial(\ln(\psi))}{\partial \pi_{ij}}$ are

$$\begin{aligned} \zeta_{00} &= \frac{1}{\pi_{00}} \ , \\ \zeta_{01} &= \frac{1}{(1 - \pi_{11} - \pi_{01})} = \frac{1}{\pi_{+0}} \ , \\ \zeta_{10} &= \frac{1}{(1 - \pi_{11} - \pi_{10})} = \frac{1}{\pi_{0+}} \ , \\ \zeta_{11} &= \frac{1}{(1 - \pi_{11} - \pi_{10})} + \frac{1}{(1 - \pi_{11} - \pi_{01})} = \frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}} \ . \end{aligned}$$

Since $\sum_i \sum_j \pi_{ij}\zeta_{ij} = \frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}} - 1$ and $\zeta_{00}^2 = 1/\pi_{00}^2$, $\zeta_{01}^2 = 1/\pi_{+0}^2$, $\zeta_{10}^2 = 1/\pi_{0+}^2$, $\zeta_{11}^2 = \frac{1}{\pi_{0+}^2} + \frac{1}{\pi_{+0}^2} + \frac{2}{\pi_{0+}\pi_{+0}}$, it holds that

$$\begin{aligned}
\sigma^2 &= \sum_i \sum_j \pi_{ij}\zeta_{ij}^2 - \left(\sum_i \sum_j \pi_{ij}\zeta_{ij}\right)^2 \\
&= \frac{1}{\pi_{00}} + \frac{\pi_{01}}{\pi_{+0}^2} + \frac{\pi_{10}}{\pi_{0+}^2} + \pi_{11}\left(\frac{1}{\pi_{0+}^2} + \frac{1}{\pi_{+0}^2} + \frac{2}{\pi_{0+}\pi_{+0}}\right) - \left(\frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}} - 1\right)^2 \\
&= \frac{1}{\pi_{00}} + \frac{\pi_{01} + \pi_{11} - 1}{\pi_{+0}^2} + \frac{\pi_{10} + \pi_{11} - 1}{\pi_{0+}^2} \\
&\quad + 2\,\frac{(\pi_{11} - 1)}{\pi_{0+}\pi_{+0}} + 2\left(\frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}}\right) - 1 \ .
\end{aligned}$$

Then, $\sigma/\sqrt{n}$ is an asymptotic standard error for $g(\hat{\boldsymbol{\pi}})$. Hence, the asymptotic standard error of $\ln(\hat{\psi})$ is

$$\begin{aligned}
\sigma(\ln(\hat{\psi})) &= \left(\frac{1}{n\pi_{00}} + \frac{\pi_{01} + \pi_{11} - 1}{n\pi_{+0}^2} + \frac{\pi_{10} + \pi_{11} - 1}{n\pi_{0+}^2}\right. \\
&\quad \left. + \frac{2}{n}\left(\frac{\pi_{11} - 1}{\pi_{0+}\pi_{+0}} + \frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}}\right) - \frac{1}{n}\right)^{1/2} \ . \qquad (4)
\end{aligned}$$

Replacing π_{ij} for $i, j = 0, 1$ in (4) by their estimates yields the estimated standard error

$$\begin{aligned}
\hat{\sigma}(\ln(\hat{\psi})) &= \left(\frac{1}{n\hat{\pi}_{00}} + \frac{\hat{\pi}_{01} + \hat{\pi}_{11} - 1}{n\hat{\pi}_{+0}^2} + \frac{\hat{\pi}_{10} + \hat{\pi}_{11} - 1}{n\hat{\pi}_{0+}^2}\right. \\
&\quad \left. + \frac{2}{n}\left(\frac{\hat{\pi}_{11} - 1}{\hat{\pi}_{0+}\hat{\pi}_{+0}} + \frac{1}{\hat{\pi}_{0+}} + \frac{1}{\hat{\pi}_{+0}}\right) - \frac{1}{n}\right)^{1/2} \ . \qquad (5)
\end{aligned}$$

If $\ln(\hat{\psi}(x))$ along with the estimated standard error $\hat{\sigma}(\ln(\hat{\psi}(x))$ is computed at each age x available, one can assess the time dependence of association in bivariate current status data. We also use the following summary measure of association across age groups $x = 1, 2, \ldots, M$:

$$\overline{\ln(\psi)} = \frac{\sum_{x=1}^{M} p_x \ln(\hat{\psi}(x))}{\sum_{x=1}^{M} p_x} \ , \quad \mathrm{Var}(\overline{\ln(\psi)}) = \frac{1}{\sum_{x=1}^{M} p_x} \ , \qquad (6)$$

where p_x is the (estimated) precision of $\ln(\hat{\psi}(x))$, that is, the reciprocal of its (estimated) variance $\hat{\sigma}^2(\ln(\hat{\psi}(x))$.

3 Experiments

It is assumed in the sequel that age x is the only measured attribute of an individual and that all unmeasured attributes are described by a random variable $Z > 0$ with density

$f(Z)$ and $\mathrm{E}(Z) = 1$. Consider the following shared frailty model (Duchateau and Janssen 2008; Hougaard 2000; Wienke 2011) for the hazard rate for an individual of age x and random effect (frailty) Z:

$$\lambda_j(x, Z) = Z\,\lambda_{0j}(x)\ ,$$

for $j = 1, 2$, where the baseline hazard rates $\lambda_{0j}(x)$ are independent of Z and describe the age effect. The random variation in Z induces the association between the failure times T_1 and T_2; T_1 and T_2 are conditionally independent given $Z = z$. In this Section we consider gamma distributed frailties ($Z \sim \Gamma(\theta, 1/\theta)$), inverse Gaussian frailties ($Z \sim InvG(1, \theta)$) and compound Poisson frailty models ($Z \sim CP(1, \theta^{-1}, \nu)$).

3.1 Evaluation of the time-varying association in shared frailty models

The heterogeneity at the population level or association in survivors is constant for the gamma distribution, decreases with time for the inverse Gaussian, and increases with time for the compound Poisson distribution (Aalen et al. 2008). To investigate whether $\ln(\psi(x))$ reflects these population effects, data are generated as follows. Cumulative baseline hazards $\Lambda_{0j}(x) = \sum_{\leq x} \lambda_{0j}(x)$ are generated for ages $x = 0.05, 0.06, \ldots, 50.00$ and the following three models for the baseline hazards λ_{0j} $(j = 1, 2)$: a constant baseline hazard, $\lambda_{0j}(x) = c_j$ (with $c_1 = 0.2$ and $c_2 = 0.1$), a Gompertz baseline of the form $\lambda_{0j}(x) = a_j \exp\{b_j x\}$ (with $a_1 = 0.006$, $b_1 = 0.02$, $a_2 = 0.008$ and $b_2 = 0.03$), and an exponentially damped linear (EDL) function of age (Farrington 1990), $\lambda_{0j} = (\alpha_j x - \gamma_j)\exp\{-\beta_j x\} + \gamma_j$ (with $\alpha_1 = 0.2$, $\gamma_1 = 0.02$, $\beta_1 = 0.2$, $\alpha_2 = 0.25$, $\gamma_2 = 0.03$, and $\beta_2 = 0.3$). A shared gamma frailty model with $Z \sim \Gamma(\theta, 1/\theta)$ and $\theta = 2$ is defined, so that $\mathrm{E}(Z) = 1$ and $\mathrm{Var}(Z) = 1/2$, and the log conditional probabilities are calculated at each x for the three baseline hazards. Figure 1 (i) displays the three tracings for $\ln(\psi(x))$ versus x.

[Figure 1 about here.]

Note that the frailty is independent of time and hence there is no time-varying association on an individual level, that is, the heterogeneity in individuals does not vary with age. Furthermore, since the frailty is gamma distributed, there is no time-varying association

on a population level either, that is, there is no time-varying association in survivors. Nevertheless, according to the plot (i), $\ln(\psi)$ increases with age for all baseline models. Moreover, the shape of the temporal variation in the strength of association clearly depends on the baseline hazard chosen. Hence, there is evidence that the conditional probability is severely influenced by the cumulative baselines. Recall that a time-dependent association measure should be free from the influence of the baseline hazards because in a shared frailty model the frailty Z solely generates the association structure between T_1 and T_2. To reduce the effect of differences in the baseline hazards, in Figure 1 (ii) $\ln(\psi(x))$ is plotted against -$\ln(\pi_{00}(x))$ (Viswanathan and Manatunga 2001). When plotted against $-\ln(\pi_{00}(x))$, $\ln(\psi)$ is largely free of the influence of the baseline hazards. Nevertheless, $\ln(\psi)$ still increases with age for all baseline models. Data are also generated from shared inverse Gaussian and compound Poisson frailty models and constant baseline hazards (with $c_1 = 0.2$ and $c_2 = 0.1$). The results are displayed in Figure 1 (iii)–(iv) and Figure 1 (v)–(vi) for the inverse Gaussian and compound Poisson frailty models, respectively, for a range of values of $\mathrm{Var}(Z) = \theta^{-1}$. The log conditional probability totally fails to mirror the declining heterogeneity induced by the inverse Gaussian distribution for the whole range of shape parameters. For the compound Poisson frailty models the measure increases, thus mirroring the increasing heterogeneity of the survivor population. However, $\ln(\psi)$ is not able to differentiate between inverse Gaussian and compound Poisson frailties, and induces very different dependence patterns in survivors. Hence, there is evidence that $\ln(\psi)$ is not a suitable diagnostic for suggesting a frailty distribution.

3.2 Simulation study

Cumulative baseline hazards are obtained for ages $x = 1, 2, \ldots, 40$ and constant baseline hazards $\lambda_{0j}(x) = c_j$ $(j = 1, 2)$ with $c_1 = c_2 = 0.05$. For each of the three frailty models $Z \sim \Gamma(0.1, 10)$, $Z \sim InvG(1, 0.1)$ and $Z \sim CP(1, 10, 1.5)$ the proportions $S_1(x), S_2(x)$ and $S(x, x)$ are calculated and $\ln(\psi(x))$ is obtained from these proportions for $x = 1, \ldots, 40$. The multinomial probabilities $\pi_{00}(x)$, $\pi_{01}(x)$, $\pi_{10}(x)$ and $\pi_{11}(x)$ are used to generate 10000 4-tuples of bivariate current status data $(n_{00x}, n_{10x}, n_{01x}, n_{11x})$ for each of the four fixed sample sizes $n_x = 50, 100, 200$ and 400. If one of the values in the 4-tuple of observations

is zero, 0.5 is added to all counts (Agresti 2013, Section 10.6). Estimates of the association measure along with its standard errors are obtained for the 10000 replications using the procedure described in Section 2. Figure 2 shows $\ln(\psi(x))$ along with the arithmetic mean of the $\ln(\hat{\psi}(x))$ for sample size $n_x = 50$ and $n_x = 400$.

[Figure 2 about here.]

For $n_x = 50$ and $n_x = 400$, there is virtually no bias for the three shared frailty models. For $x = 20$ and $x = 40$, Table 1 shows the bias of $\ln(\hat{\psi})$, its mean standard error (s.e.), the empirical s.e. obtained as the standard deviation of the 10000 simulated values, and the coverage probability of the 95% confidence intervals (P95), that is, the proportion of the 10000 confidence intervals containing $\ln(\psi(x))$.

[Table 1 about here.]

The relative bias is less than 3% even for $n_x = 50$. The empirical standard error matches the mean standard error of the asymptotic values. The coverage probabilities are close to the nominal level 0.95. As one would expect, as the sample size increases the bias and variance of $\ln(\hat{\psi}(20))$ and $\ln(\hat{\psi}(40))$ decrease.

4 Application

In this Section, the time dependence of association is analyzed using current status data arising from a community-based study in cardiovascular epidemiology conducted in Taiwan from 1991 to 1993. The sample consists of 6314 participants age 1-93 years. The data comprised measurements of the participants' current age at the time of study and the prevalence indicators of three diseases: diabetes mellitus, hypercholesterolemia and hypertension. The aim of this study was to investigate whether the onset ages of these diseases are correlated with one another. Because the natural history of these chronic diseases was difficult to trace precisely, the data contained only information about whether or not a subject under the study had already developed the diseases and about the subject's current age at the time of the study. Let T_1, T_2 and T_3 denote the onset age of diabetes mellitus, hypercholesterolemia and hypertension, respectively, and let x denote

the individual's age in years at the monitoring time. Then, for each individual the observed data are of the form $(x, \delta_1, \delta_2, \delta_3)$, where $\delta_j = I(T_j \leq x)$ $(j = 1, 2, 3)$. For a more detailed description of the data, the reader is referred to Wang and Ding (2000). Wang and Ding (2000) assumed that the pairwise dependence structures of the three diseases all follow a gamma frailty model and then estimated the Clayton copula association parameters. The Clayton copula association parameter is related to Kendall's tau, which is a measure of rank correlation for bivariate dependence that is invariant subject to both linear and nonlinear monotonic changes of scale of failure times (Kendall 1938). Kendall's tau between T_i and T_j, denoted by τ_{ij}, is defined as

$$\tau_{ij} = \mathrm{E}\left[\mathrm{sign}\left(\left(T_i^{(1)} - T_i^{(2)}\right)\left(T_j^{(1)} - T_j^{(2)}\right)\right)\right] \ ,$$

where $T^{(k)}$ $(k = 1, 2)$ are independent copies of T and $\mathrm{sign}(x) = -1$ for $x < 0$, $\mathrm{sign}(x) = 0$ for $x = 0$ and $\mathrm{sign}(x) = 1$ for $x > 0$. If T_i and T_j are independent, $\tau_{ij} = 0$. For each pair (T_1, T_2), (T_1, T_3) and (T_2, T_3), Wang and Ding (2000) tested H_0: $\tau = 0$ versus H_1: $\tau > 0$ and computed the p-value of the test. The estimated values of the corresponding Kendall's tau with p-values given in parentheses are $\hat{\tau}_{12} = 0.304$ $(p < 0.001)$, $\hat{\tau}_{13} = 0.128$ $(p = 0.028)$ and $\hat{\tau}_{23} = 0.082$ $(p = 0.052)$. Ding and Wang (2004) also analyzed the data and performed pairwise non-parametric independence tests of (T_1, T_2), (T_1, T_3) and (T_2, T_3). The associations between T_1 and T_2 as well as T_1 and T_3 were both very strong with p-values close to zero. The association between T_2 and T_3 was significant at the .05 level, but not at the .01 level. Note that this data example is used only for illustrative purposes here, because the prevalence of the three cardiovascular diseases were determined via participant interviews, health examinations or previous medical history, rather than based on a formal medical diagnosis.

In the sample, the prevalence indicator δ_1 is equal to 1 for 434 individuals, $\delta_2 = 1$ for 711 individuals and $\delta_3 = 1$ for 1111 individuals. Figure 3 is a cumulative sum diagram of the cumulative number of persons with the cardiovascular disease diabetes mellitus (DM), hypercholesterolemia (HC) and hypertension (HT), respectively, versus the cumulative number of persons in the sample. The greatest convex minorants of the three cumulative sums are superimposed. Underneath the x-axis in Figure 3 the ages in years are presented that correspond to the cumulative totals labelled above.

[Figure 3 about here.]

Across the age range the cumulative frequencies of hypercholesterolemia are greater than for diabetes mellitus but the slopes of the greatest convex minorants are virtually similar at higher ages. At age 57 with 75% of the total number of participants aged 57 or less, the cumulative sums of prevalence indicators for hypertension and hypercholesterolemia intersect. In the age range 57-93 hypertension is the cardiovascular disease that has the highest prevalence of the three diseases with the slope of greatest convex minorant being the steepest. Summary values for the association parameter $\overline{\ln(\psi)}$ shown in equation (6) are presented in Table 2. If one of the values in the 4-tuple of observations $(n_{00x}, n_{01x}, n_{10x}, n_{11x})$ is zero, 0.5 is added to all counts (Agresti 2013, Section 10.6).

[Table 2 about here.]

The values for $\overline{\ln(\psi)}$ are all close to zero and suggest that there is almost no correlation between (T_1, T_2), (T_1, T_3) and (T_2, T_3). We also use a fairly new association measure denoted $\phi(x)$ and introduced in Unkel and Farrington (2012), which is defined as the unique root $\phi_0(S(x,x), S_1(x), S_2(x))$ of the implicit function

$$f(\phi, S(x,x), S_1(x), S_2(x)) = \left(S_1(x)^{1-e^{\phi}} + S_2(x)^{1-e^{\phi}} - 1\right)^{\frac{1}{1-e^{\phi}}} - S(x,x) \ .$$

For a shared gamma fraily model, $\phi(x)$ is constant and equal to the logarithm of the cross-ratio function (CRF). If the frailty is not gamma distributed, $\phi(x)$ will not be equal to the logarithm of the CRF. However, Unkel and Farrington (2012) show that $\phi(x)$ tracks the CRF for all shared frailty models with monotone CRF regardless of the frailty distribution, in the sense that it shows the same variation with age. Hence, $\phi(x)$ can be used as a diagnostic tool for suggesting gamma frailty distributions with constant association over time or frailty models that lead to time-varying association. The value $\phi(x) = 0$ corresponds to independence; $\phi(x) > 0$ corresponds to positive association and $\phi(x) < 0$ corresponds to negative association although there is no frailty interpretation in this case. Summary values for this association measure, denoted $\overline{\phi}$, are computed analogously to equation (6), and are presented in Table 2. The values for $\overline{\phi}$ are similar to the association patterns found by Wang and Ding (2000). Associations are significantly

positive for all three pairs of diseases, while the association between (T_2, T_3) is notably lower than for (T_1, T_2) and (T_1, T_3). However, overall measures are crude, so we plotted the estimated values of the association parameters $\ln(\psi(x))$ and $\phi(x)$ at each age x in Figure 4. The areas of the points within each graph are proportional to the precisions; the smooth lines are nonparametric precision-weighted estimates of trend.

[Figure 4 about here.]

For the log conditional probability, the three tracings indicate independence between (T_1, T_2), (T_1, T_3) and (T_2, T_3). However, this is due to the fact that in the definition of ψ in (2) the numerator can only attain values between its lower Fréchet bound $\max\{0, S_1(x) + S_2(x) - 1\}$ and upper Fréchet bound $\min\{S_1(x), S_2(x)\}$. This implies that as $x \to 0$, $\ln(\psi(x)) \to 0$. In this sense, the observed association patterns in Figure 4 are misleading; the absence of positive association at early ages is solely a result of the definition of the conditional probability measure. It is a serious shortcoming of this measure that the degree of association at is influenced by its range restriction.

On the other hand, the plots for $\hat{\phi}$ suggest that there is a strong association at early ages for the three pairs of diseases and that the heterogeneity in the survivor population may be decreasing towards a positive asymptote for the pair DM/HC and towards zero for both DM/HT and HC/HT. The decreasing association in adulthood could be due to a selection effect caused by a time-invariant frailty model (e.g. an inverse Gaussian frailty model) or a temporal variation of the frailty itself. In any case, the observed association patterns of $\hat{\phi}$ that are displayed in Figure 4 raise doubts whether the assumption of gamma frailties with constant association over time is justified.

5 Discussion

The conditional probability measure $\psi(t_1, t_2)$ is available for case I interval-censored data, which are also named current status data. We presented its variant $\psi(x)$ for current status data and developed maximum likelihood estimation and standard errors for this measure. We also introduced a summary measure of association across age groups. Our proposed

estimator and standard errors have shown good finite sample performance in the simulation experiments. However, we showed that the conditional probability measure can vary with time even in the absence of any time-dependent effects. Furthermore, it lacks interpretability in suggesting appropriate frailty models. Moreover, because of its range restriction the conditional probability measure is flawed based on theoretical grounds.

The real data example illustrates that estimates of the conditional probability measure can lead to observed association patterns that are misleading. Contrary to the conditional probability measure $\psi(x)$, the shape of the time-varying association measure $\phi(x)$ aids identification of a suitable frailty model. Note that in the application, $\hat{\phi}$ is plotted against age in order to keep the age-related interpretation of the association pattern. To remove the dependence on the baseline hazards and to compare association patterns across different data sets, $\hat{\phi}$ can be plotted against $-\ln(\hat{\pi}_{00}(\text{age}))$, suitably isotonized. For the application that was considered in this paper, the estimated temporal patterns of $\phi(x)$ give evidence that the assumption of constant association over time is violated. One may argue that the observed pairwise dependence structures between the three cardiovascular diseases can be expected to be different from those obtained when associations are stratified according to levels of further covariates such as gender. However, age is the only available attribute for the set of individuals in the study and therefore we were not able to look further into such features.

Finally, recall that the methods developed in the present paper are entirely exploratory. At no stage do we model the data. We see ourselves in a stage prior to fitting current status data by means of frailty models. The aim is to investigate the observed association patterns, which then may serve as a guide for embarking on a suitable frailty model. Clearly, the conditional probability measure is not an appropriate tool for assessing the temporal variation in the strength of association in shared frailty models with bivariate current status data.

Acknowledgements

An Associate Editor, two reviewers and C. Paddy Farrington made valuable comments and suggestions on earlier drafts of this paper. The author is grateful to Weijing Wang for providing the data for the application.

References

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008), *Survival and Event History Analysis: A Process Point of View*, Springer: New York.

Agresti, A. (2013), *Categorical Data Analysis*, John Wiley & Sons: Hoboken, 3rd ed.

Anderson, J. E., Louis, T. A., Holm, N. V., and Harvald, B. (1992), "Time-dependent association measures for bivariate survival distributions," *Journal of the American Statistical Association*, 87, 641–650.

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

Collett, D. (2002), *Modelling Binary Data*, Chapman & Hall/CRC: Boca Raton, 2nd ed.

Ding, A. A. and Wang, W. (2004), "Testing independence for bivariate current status data," *Journal of the American Statistical Association*, 99, 145–155.

Duchateau, L. and Janssen, P. (2008), *The Frailty Model*, Springer: New York.

Farrington, C. P. (1990), "Modeling forces of infection for measles, mumps and rubella," *Statistics in Medicine*, 9, 953–967.

Farrington, C. P., Unkel, S., and Anaya-Izquierdo, K. (2012), "The relative frailty variance and shared frailty models," *Journal of the Royal Statistical Society Series B*, 74, 673–696.

Farrington, C. P., Whitaker, H. J., Unkel, S., and Pebody, R. (2013), "Correlated infections: quantifying individual heterogeneity in the spread of infectious diseases," *American Journal of Epidemiology*, 177, 474–486.

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

Jewell, N. P. (2003), *Statistics for Epidemiology*, Chapman & Hall/CRC: Boca Raton.

Jewell, N. P., Van der Laan, M., and Lei, X. (2005), "Bivariate current status data with univariate monitoring times," *Biometrika*, 92, 847–862.

Kendall, M. G. (1938), "A new measure of rank correlation," *Biometrika*, 30, 81–93.

Oakes, D. (1989), "Bivariate survival models induced by frailities," *Journal of the American Statistical Association*, 84, 487–493.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Sun, J. (2006), *The Statistical Analysis of Interval-censored Failure Time Data*, Springer: New York.

Unkel, S. and Farrington, C. P. (2012), "A new measure of time-varying association for shared frailty models with bivariate current status data," *Biostatistics*, 13, 665–679.

Unkel, S., Farrington, C. P., Whitaker, H. J., and Pebody, R. (2014), "Time-varying frailty models and the estimation of heterogeneities in transmission of infectious diseases," *Journal of the Royal Statistical Society Series C*, 63, 141–158.

Viswanathan, B. and Manatunga, A. K. (2001), "Diagnostic plots for assessing the frailty distribution in multivariate survival data," *Lifetime Data Analysis*, 7, 143–155.

Wang, W. and Ding, A. A. (2000), "On assessing the association for bivariate current status data," *Biometrika*, 87, 879–893.

Wienke, A. (2011), *Frailty Models in Survival Analysis*, Chapman & Hall/CRC: Boca Raton.

Table 1: Bias and variance of $\ln(\hat{\psi})$ for three shared frailty models and four different sample sizes evaluated at $x = 20$ and $x = 40$.

		$x = 20$				
Frailty model	$\ln(\psi(20))$		$n_x = 50$	$n_x = 100$	$n_x = 200$	$n_x = 400$
$Z \sim \Gamma(0.1, 10)$	0.2375	bias	0.0019	0.0014	0.0004	0.0003
		mean s.e.	0.0622	0.0443	0.0313	0.0222
		empirical s.e.	0.0633	0.0444	0.0317	0.0222
		P95	0.9199	0.9411	0.9423	0.9507
$Z \sim InvG(1, 0.1)$	0.1762	bias	0.0006	0.0007	0.0005	0.0006
		mean s.e.	0.0695	0.0491	0.0347	0.0246
		empirical s.e.	0.0698	0.0493	0.0344	0.0247
		P95	0.9358	0.9429	0.9487	0.9491
$Z \sim CP(1, 10, 1.5)$	0.2039	bias	0.0041	0.0030	0.0008	0.0008
		mean s.e.	0.0666	0.0476	0.0336	0.0238
		empirical s.e.	0.0669	0.0481	0.0338	0.0240
		P95	0.9452	0.9311	0.9492	0.9421
		$x = 40$				
Frailty model	$\ln(\psi(40))$		$n_x = 50$	$n_x = 100$	$n_x = 200$	$n_x = 400$
$Z \sim \Gamma(0.1, 10)$	0.1751	bias	0.0023	0.0013	0.0006	0.0005
		mean s.e.	0.0737	0.0523	0.0370	0.0262
		empirical s.e.	0.0531	0.0448	0.0374	0.0261
		P95	0.9279	0.9435	0.9455	0.9510
$Z \sim InvG(1, 0.1)$	0.2806	bias	0.0026	0.0009	0.0001	0.0006
		mean s.e.	0.0954	0.0668	0.0471	0.0333
		empirical s.e.	0.0972	0.0674	0.0474	0.0331
		P95	0.9436	0.9466	0.9453	0.9509
$Z \sim CP(1, 10, 1.5)$	0.2306	bias	0.0051	0.0009	0.0008	0.0002
		mean s.e.	0.0712	0.0506	0.0360	0.0255
		empirical s.e.	0.0724	0.0509	0.0363	0.0257
		P95	0.9381	0.9405	0.9449	0.9415

Table 2: Associations, $\overline{\ln(\psi)}$ and $\overline{\phi}$, for the three pairs of cardiovascular diseases along with 95% confidence intervals in parentheses.

Pair of diseases	$\overline{\ln(\psi)}$	$\overline{\phi}$
DM and HC	-0.0109 (-0.0135, -0.0084)	0.9055 (0.7301, 1.0810)
DM and HT	-0.0141 (-0.0165, -0.0117)	0.7375 (0.5983, 0.8768)
HC and HT	0.0221 (0.0193, 0.0248)	0.3838 (0.2523, 0.5152)

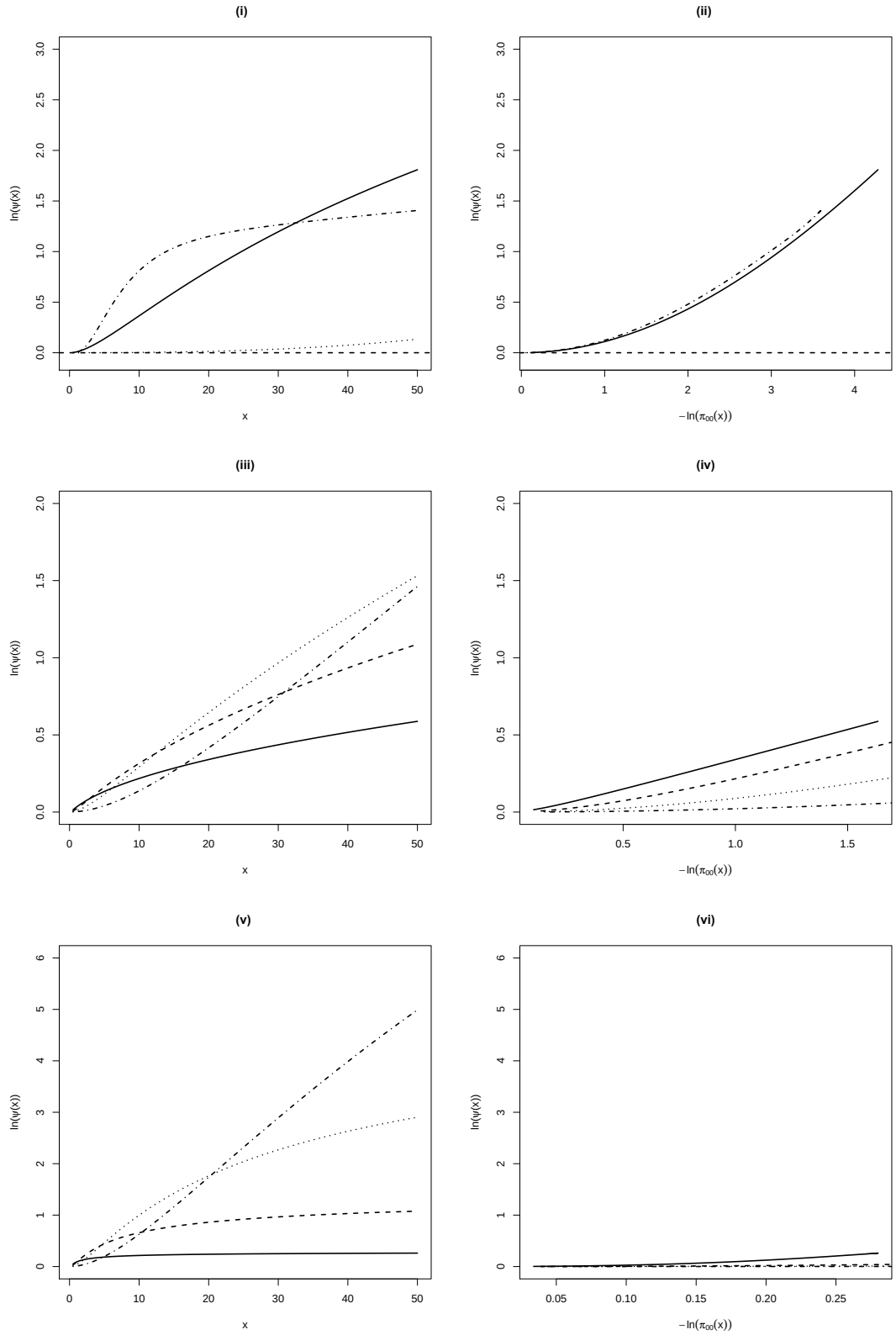

Figure 1: $\ln(\psi)$ against x and $-\ln(\pi_{00}(x))$ for various frailty models; upper panel: $Z \sim \Gamma(2, 1/2)$ for constant baseline (solid line), Gompertz baseline (dotted line) and EDL baseline (dot-dashed line) (horizontal dashed line: no association); middle panel: $Z \sim InvG(1, 0.1)$ (solid line), $Z \sim InvG(1, 0.5)$ (dashed line), $Z \sim InvG(1, 2)$ (dotted line), $Z \sim InvG(1, 10)$ (dot-dashed line) and constant baseline hazards; lower panel: $Z \sim CP(1, 10, 1.5)$ (solid line), $Z \sim CP(1, 2, 1.5)$ (dashed line), $Z \sim CP(1, 0.5, 1.5)$ (dotted line), $Z \sim CP(1, 0.1, 1.5)$ (dot-dashed line) and constant baseline hazards.

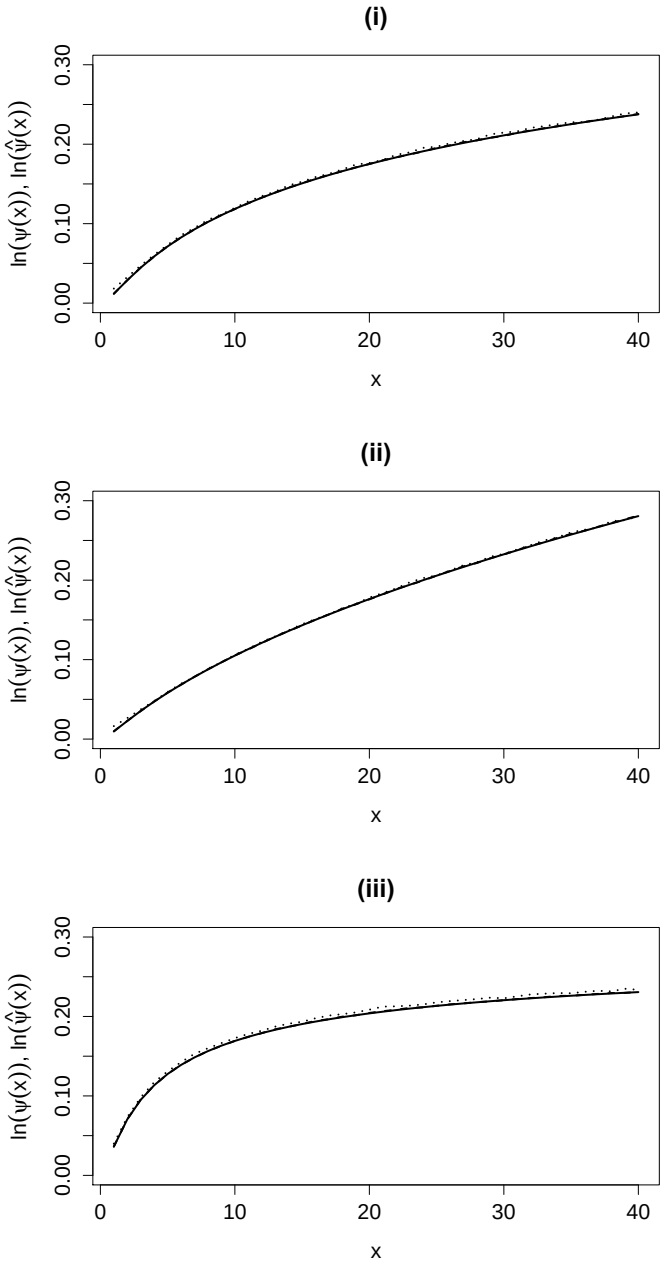

Figure 2: $\ln(\psi(x))$ (solid line) and (mean of) $\ln(\hat{\psi}(x))$ for $n_x = 50$ (dotted line) and $n_x = 400$ (dashed line) for the Gamma (i), inverse Gaussian (ii) and compound Poisson (iii) frailty model.

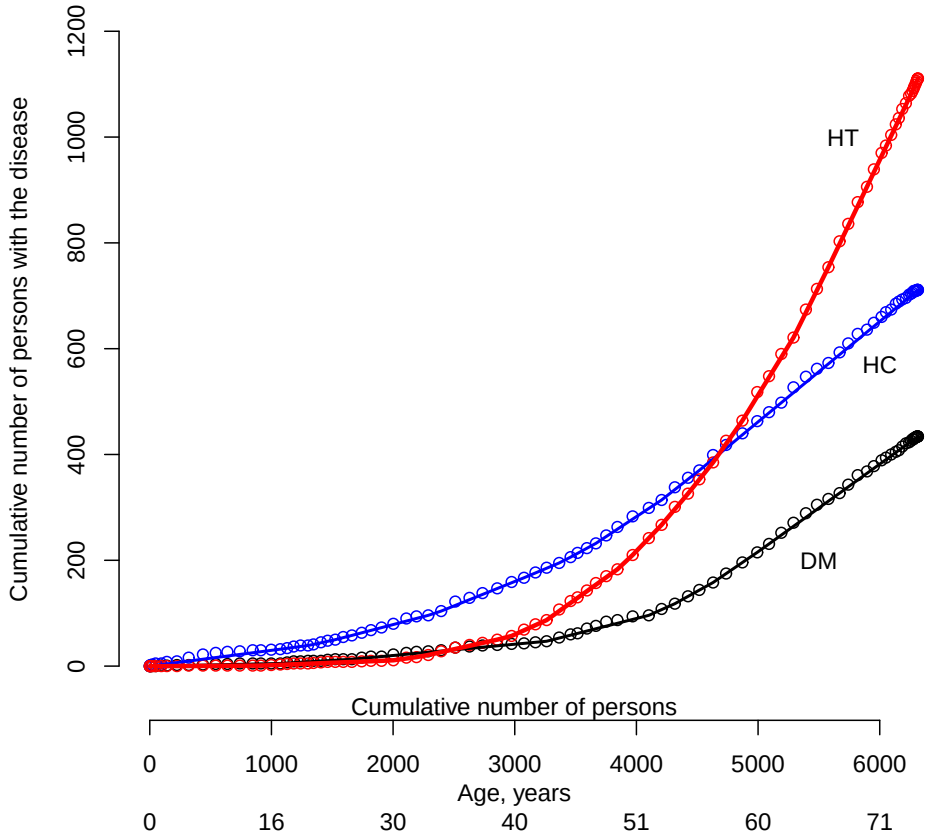

Figure 3: Cumulative number of persons with the disease versus cumulative number of persons in the sample (dots) and greatest convex minorant (line): diabetes mellitus (black), hypercholesterolemia (blue) and hypertension (red).

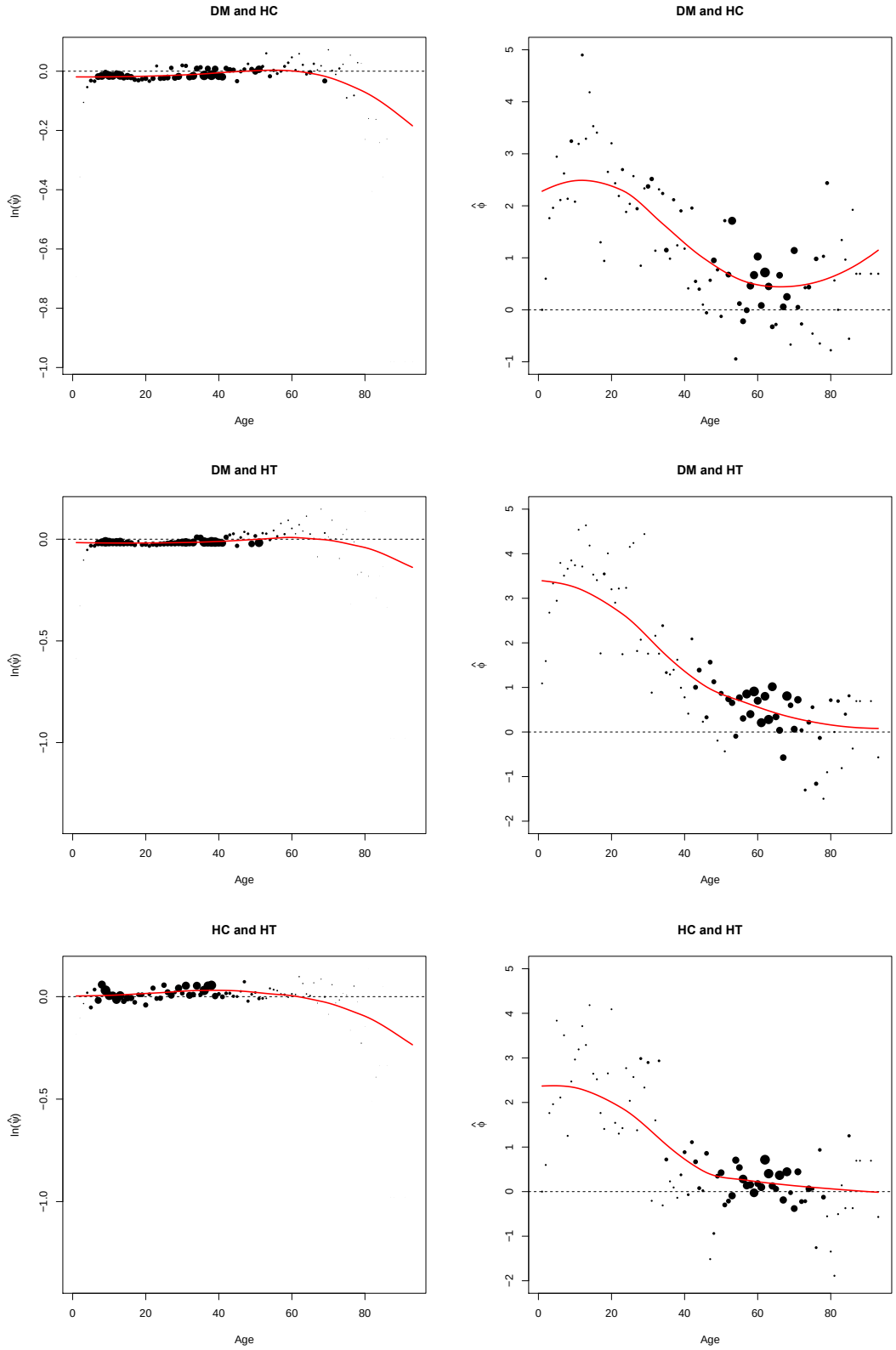

Figure 4: Association, measured by $\ln(\hat{\psi})$ (left) and $\hat{\phi}$ (right), between ages at disease onset for the three pairs of cardiovascular diseases. Top: diabetes mellitus (DM) and hypercholesterolemia (HC); center: DM and hypertension (HT); bottom: HC and HT. The dots represent empirical values and the lines show smoothed trends (horizontal dashed line: no association).