A new measure of time-varying association for shared frailty models with bivariate current status data

Steffen Unkel* and C. Paddy Farrington

Department of Mathematics and Statistics
The Open University
Milton Keynes, UK

September 26, 2011

Abstract

In this paper, a new measure for assessing the temporal variation in the strength of association in bivariate current status data is proposed. This novel measure is relevant for shared frailty models. We show that this measure is particularly convenient, owing to its connection with the relative frailty variance and its interpretability in suggesting appropriate frailty models. We introduce a method of estimation and standard errors for this measure. We discuss its properties and compare it to two existing measures of association applicable to current status data. Small sample performance of the measure in realistic scenarios is investigated using simulations. The methods are illustrated with bivariate serological survey data on different infections, where the time-varying association is likely to represent heterogeneities in activity levels and/or susceptibility to infection.

Key words: Current status data, Cross-ratio function, Frailty models, Heterogeneity, Measures of association, Serological survey, Temporal variation.

* Correspondence should be addressed to: Steffen Unkel, Department of Mathematics and Statistics, Faculty of Mathematics, Computing & Technology, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom (e-mail:S.Unkel@open.ac.uk).
This research was supported by a grant from the UK Medical Research Council.

1 Setting the scene

Current status data, also known as case I interval-censored data, arise in survival analysis when the exact timing of an event is unobserved and it is only known at a given point in time whether or not the event has occurred (Sun 2006). Such data occur in various fields including demographical studies (Diamond and McDonald 1992), tumorigenicity experiments (Dunson and Dinse 2002) and epidemiology (Ding and Wang 2004).
Consider a bivariate setting and let T_j $(j = 1, 2)$ be the two failure times of interest. Let X denote the univariate monitoring time at which T_1 and T_2 are measured from the same observational units (e.g. individuals) and assume that (T_1, T_2) are independent of X. A concise representation of the observed information is $\{X, \delta_1, \delta_2\}$, where

$$\delta_j = \begin{cases} 1 & \text{if} \quad T_j \leq X \ , \\ 0 & \text{if} \quad T_j > X \ , \end{cases}$$

for $j = 1, 2$. In this paper, we are interested in current status data that are reasonable to model by shared frailty models (Duchateau and Janssen 2008; Hougaard 2000; Wienke 2011). Such bivariate data occur naturally in infectious disease epidemiology, for instance, when T_1 and T_2 represent the ages at the onset of infection by two distinct infectious agents whose onset can only be determined to lie below or above X. In this context, the time scale is age and the defining time point from which times are measured is birth. The association between the ages T_1 and T_2 may carry information about relevant infection processes and can be examined using paired serological survey data on two infections (Farrington and Whitaker 2005). Serological data, which provide the main motivating example for this paper, are a key resource in infectious disease epidemiology and are obtained by testing blood serum residues for the presence of antibodies to one or more infections. A positive (negative) result indicates prior infection (susceptibility to infection), giving rise to current status data. Suppose that paired data are available on two infections and let $X = x$ be the age at the monitoring time. For infection j $(j = 1, 2)$, the hazard rate (force of infection) at age x for an individual with a positive random effect Z is assumed to be of the form

$$\lambda_j(x, Z) = Z\lambda_{0j}(x) \ , \tag{1}$$

where the baseline hazards $\lambda_{0j}(x)$ are independent of Z and describe the age effect. The random variation in Z induces association between the two failure times T_1 and T_2; T_1 and T_2 are conditionally independent given $Z = z$. The individual latent effects may be viewed as individual frailties, yielding shared frailty models for the hazard rates. A shared frailty model is natural in this setting, the latent frailty variable representing individual characteristics, such as strength of the immune system or propensity to make contact, which may have a bearing on several distinct infection processes.

Another example for bivariate current status data arise from 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. Again, a shared frailty model is natural, the latent frailty representing environmental exposures relevant to the progression of disease at different sites.

In some circumstances, it is also of interest to assess the time dependence of association. Farrington et al. (2001) showed how bivariate serological survey data on two infections could be used to estimate heterogeneity using shared frailty models. Such heterogeneity can reflect individual variation in susceptibility and effective contact rates (Coutinho et al. 1999) and has implications for infection control. However, this work does not use the available information on how the strength of association, and hence the degree of heterogeneity, varies over time. Such information is important as it can suggest pointers to the source of the heterogeneity – for example if the association is sustained in adulthood it may reflect a common source of transmission for the two infections.

In a shared frailty model such as (1) the frailty Z solely generates the association structure between the two variables T_1 and T_2. Therefore, a time-dependent association measure should be free from the influence of the baseline hazards

$$\Lambda_{0j}(x) = \int_0^x \lambda_{0j}(t)\mathrm{d}t \ ,$$

where $\Lambda_{0j}(x)$ is the cumulative baseline hazard rate to age x for infection j. Beyond that, a useful time-dependent measure should reflect the variation in the strength of association in survivors over time. Such a requirement is particularly relevant to shared frailty

models, where the pattern of association can be used to select an appropriate frailty distribution (Viswanathan and Manatunga 2001).

Several measures to quantify the association in bivariate survival data are available (Drouet Mari and Kotz 2001), including the odds ratio, Kendall's τ (Kendall 1938), and association parameters derived from copula models. A variety of parametric and semi-parametric estimation methods have been proposed (Dale 1986; Hougaard 2000; Plackett 1965; Wang and Ding 2000). Various local indices to assess the time dependence of association and applicable in survival analysis have been introduced (Drouet Mari and Kotz 2001, pp. 178-190), describing the degree of dependence at a single time point.

The measure of choice in survival analysis for assessing time-varying dependence is Clayton's local cross-ratio function, originally introduced by Clayton (1978) and studied by Oakes (1989) and Anderson et al. (1992). Let T_j $(j = 1, 2)$ have marginal survivor functions $S_j(t_j) = \mathrm{P}(T_j > t_j)$ and joint survival function $S(t_1, t_2) = \mathrm{P}(T_1 > t_1, T_2 > t_2)$. The cross-ratio function (CRF) at (t_1, t_2) is defined as

$$\theta^*(t_1, t_2) = \frac{S(t_1, t_2)\mathrm{D}_1\mathrm{D}_2 S(t_1, t_2)}{[\mathrm{D}_1 S(t_1, t_2)]\,[\mathrm{D}_2 S(t_1, t_2)]} \ , \tag{2}$$

where D_j denotes the derivative operator $\partial/\partial t_j$. Unlike global measures, such as Kendall's τ, the CRF is a local dependence function, related to the hazard of events. It is the ratio of the hazard of T_1 given T_2 has taken place at time t_2 over the hazard of T_1 given T_2 has not yet taken place at t_2 (Oakes 1989). A CRF greater than one (< 1) corresponds to a positive (negative) association between T_1 and T_2 and $\theta^*(t_1, t_2) < 1$ if and only if $\mathrm{D}_1\mathrm{D}_2 \ln S(t_1, t_2) < 0$ (Gupta 2003). If T_1 and T_2 are independent, then $\theta^*(t_1, t_2) = 1$.

It is well-established that the CRF is constant in case of the gamma frailty distribution, decreases with time e.g. for the inverse Gaussian, and increases with time e.g. for the compound Poisson distribution (Duchateau and Janssen 2008). The CRF is a frequently used local measure of association for both right censored and interval censored survival data (Bogaerts and Lesaffre 2008; Chen and Bandeen-Roche 2005). However, 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. This implies that the CRF (2) cannot be evaluated directly from current status data.

Anderson et al. (1992) showed that the CRF has a local odds ratio (OR) interpretation.

The CRF can also be interpreted as a local version of Kendall's τ (Oakes 1989). However, as previously noted, the CRF (2) cannot be evaluated directly from current status data and hence neither can the corresponding local OR. Nevertheless, a non-local OR at (t_1, t_2) can be. Let $\pi_{00} = \mathrm{P}(T_1 > t_1, T_2 > t_2) = S(t_1, t_2)$, $\pi_{10} = \mathrm{P}(T_1 \leq t_1, T_2 > t_2)$, $\pi_{01} = \mathrm{P}(T_1 > t_1, T_2 \leq t_2)$, $\pi_{11} = \mathrm{P}(T_1 \leq t_1, T_2 \leq t_2)$ and define

$$\mathrm{OR}(t_1, t_2) = \frac{\pi_{00}\pi_{11}}{\pi_{10}\pi_{01}} = \frac{\mathrm{odds}(T_1 \leq t_1 \mid T_2 \leq t_2)}{\mathrm{odds}(T_1 \leq t_1 \mid T_2 > t_2)} \ . \tag{3}$$

In terms of paired current status data on two infections, let $\pi_{00}(x)$ be the probability that an individual of age x has been infected by neither infection and $\pi_{10}(x)$ the probability that an individual of age x has been infected by infection 1 but not infection 2, and similarly define $\pi_{01}(x)$ and $\pi_{11}(x)$. One can then assess the association between the two infections by means of $\mathrm{OR}(x) = \frac{\pi_{00}(x)\pi_{11}(x)}{\pi_{10}(x)\pi_{01}(x)}$. If $\mathrm{OR}(x)$ is estimated at each time point available (in the context of paired serological data at each age x), one can assess the temporal strength in the association in bivariate current status data.

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

$$\psi(t_1, t_2) = \frac{\mathrm{P}(T_1 > t_1 | T_2 > t_2)}{\mathrm{P}(T_1 > t_1)} = \frac{S(t_1, t_2)}{S_1(t_1)S_2(t_2)} \ . \tag{4}$$

Large values of $\psi(t_1, t_2)$ indicate positive dependence between T_1 and T_2. It holds that $\theta^*(t_1, t_2) > 1 \Rightarrow \psi(t_1, t_2) > 1$ (Gupta 2003). 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 terms of paired current status data on two infections, the conditional probability measure provides insights into the time-dependent nature of association by estimating $\psi(x) = \frac{\pi_{00}(x)}{\pi_{0+}(x)\,\pi_{+0}(x)}$, where $\pi_{0+}(x) = \pi_{00}(x) + \pi_{01}(x)$ and $\pi_{+0}(x) = \pi_{00}(x) + \pi_{10}(x)$.

However, as will be demonstrated in the paper, the odds ratio and the conditional probability measure suffer the disadvantage that they can vary with time even in the absence of any time-dependent effects. Furthermore, they lack interpretability in suggesting appropriate frailty models.

The main aim of this paper is to propose a new method for studying the temporal variation in the strength of association found in bivariate current status data. The proposed

measure of association is relevant for shared frailty models and is based on the association parameter derived from Clayton's copula (Clayton 1978; Nelson 2006) for quantifying time-dependent association. We will show that this new measure is particularly convenient, owing to its connection with the relative frailty variance that describes the heterogeneity of the hazard functions in the survivor population, and its interpretability in suggesting appropriate frailty models. Maximum likelihood estimates of the time-varying association parameter are obtained at each time point and their standard errors are calculated by means of the Delta method. We make use of scatterplot smoothers to improve the interpretability of the dependency pattern and to capture trends with age.

We emphasize that the methods developed here are entirely exploratory. At no stage do we model the data. Our aim is simply to provide a useful representation of bivariate current status data, to facilitate the choice of a frailty model.

The remainder of the paper is organized as follows. In Section 2, we present a brief motivating example. In Section 3, the new association measure relevant for shared frailty models is introduced along with a method of estimation and standard errors. A simulation study is carried out to evaluate bias and variance of the new measure in small samples under realistic scenarios. An evaluation of how the proposed measure as well as existing association measures perform with respect to identifying time-varying effects in shared frailty models with bivariate current status data is given in Section 4. In Section 5, the methods developed in this paper are applied to paired serological survey data on a range of infections. Concluding comments are given in Section 6. Computations in this paper were carried out using the software package R version 2.13.1 (R Development Core Team 2011). All computer code used is available upon request.

2 A motivating example

We provide a motivating example using a set of serological survey data collected in 1994 on Epstein-Barr virus (EBV) and herpes simplex virus type 1 (HSV1) infections (cf. Section 6 for a more detailed description of the data). Residual blood samples from individuals of 1-30 years of age were tested for antibodies to both infections. A positive result for

either infection indicates that the individual is immune; a negative result indicates that the individual is susceptible. For each blood sample, paired results for EBV and HSV1 are available.

Serological survey data for the presence or absence of antibodies to one or more infections are a key resource for understanding the epidemiology of infectious diseases and for designing and monitoring the implementation of control strategies. Serological data are of importance in any quantitative assessment of disease transmission. For communicable infectious diseases, the variability between individuals in the rate at which they make contact with others, where a contact between two individuals is an opportunity for transmission of infection, is often of primary public health interest (Farrington and Whitaker 2005). Individual heterogeneity in contact rates can have a large impact on the transmission of infection (Farrington et al. 2001). Individuals who make many contacts will tend to acquire more infections transmitted via the same route.

Age-related heterogeneities in contact rates are particularly important. For example, contact rates in measles transmission are believed to be highest between children, yet there is substantial heterogeneity owing for example to variation in family environment, nursery attendance, individual behaviour and susceptibility. Thus, one needs to adequately represent the age-related heterogeneities in contact rates that might be relevant to the transmission of infection. Such heterogeneities can seldom be measured directly. However, for two infections with a common route of transmission, the variability between individuals can naturally be modelled by a shared frailty that is induced on the force of infection. In this context, the frailty represents variation in activity levels in interacting with other individuals relevant to the transmission route. The higher the value of the frailty, the higher the contact rate for that individual. The variation in contact rates will induce associations between the two infections and can be examined using paired serological survey data. Thus, the pattern of association by age can be used to investigate the presence and degree of heterogeneity, and to suggest appropriate models.

Paired serological survey data on n_x fixed individuals of age x give rise to a multinomial observation $(n_{00x}, n_{10x}, n_{01x}, n_{11x})$, where n_{00x} is the number of individuals of age x in the sample that are uninfected by either infection, n_{10x} is the number of individuals

Table 1: Dichotomous contingency table of a 4-tuple $(n_{00x}, n_{10x}, n_{01x}, n_{11x})$ at age x

Infection 1 \ Infection 2	$\delta_2 = 0$	$\delta_2 = 1$	$\sum$
$\delta_1 = 0$	n_{00x}	n_{01x}	$n_{0+x} = n_{00x} + n_{01x}$
$\delta_1 = 1$	n_{10x}	n_{11x}	$n_{1+x} = n_{10x} + n_{11x}$
$\sum$	$n_{+0x} = n_{00x} + n_{10x}$	$n_{+1x} = n_{01x} + n_{11x}$	n_x

that are uninfected by infection 2 but have been infected by infection 1, and so on (see also Table 1). When bivariate data are thought to arise from a shared frailty model with right-censored survival data, diagnostic plots based on the CRF are used to suggest an appropriate frailty distribution (Viswanathan and Manatunga 2001; Duchateau and Janssen 2008). The CRF is unavailable for current status data, though. Instead, one can easily obtain maximum likelihood estimates (MLEs) of $\mathrm{OR}(x)$ and $\psi(x)$, respectively, as

$$\widehat{\mathrm{OR}}(x) = \frac{n_{00x} n_{11x}}{n_{01x} n_{10x}} \qquad \text{and} \qquad \hat{\psi}(x) = \frac{n_x n_{00x}}{n_{0+x} n_{+0x}} \ ,$$

where n_{0+x} and n_{+0x} are defined in Table 1. It is customary to work with the log of an association measure than with the association measure itself, so we shall do so in the sequel. The plots of Figure 1 display estimates of the log odds ratio and the log conditional probability versus age, respectively, for the data on EBV and HSV1 infections. In the plots the sizes of the points are made proportional to the precision (reciprocal of the variance) of the estimates, that is, small points correspond to relatively large standard errors. A LOESS (locally weighted scatterplot smoothing) curve (Cleveland 1979) is fitted to the set of points to capture trends with age. Weights for the cases are chosen according to the precision of the estimates. This is done to ensure that the curve is less influenced by estimates with relatively high standard error. Asymptotic standard errors for the log odds ratio are well-established (e.g., Agresti 2002, p. 71). The derivation of the asymptotic standard error of $\ln(\hat{\psi})$ is given in the supplementary material.

In Figure 1, the association patterns for the log odds ratio and the log conditional probability are rather different and it is not obvious which one (if any) provides guidance about the degree of association in the population and how it evolves over time (cf. Subsection

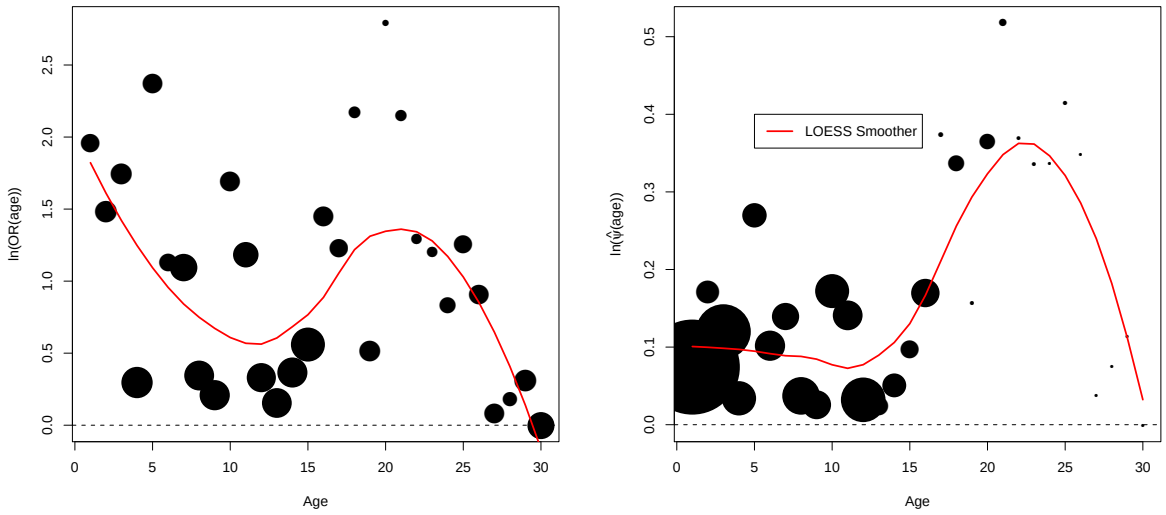

Figure 1: Plot of $\ln(\widehat{\mathrm{OR}})$ (left) and $\ln(\hat{\psi})$ (right) by age for EBV and HSV1 data (dashed line: no association). For $\ln(\hat{\psi})$, a minimum point size was defined to make all points visible.

6.1). One might ask, for example, if an overall decreasing $\ln(\widehat{\mathrm{OR}})$ and/or $\ln(\hat{\psi})$ would imply decreasing association in the survivor population, as a decreasing cross-ratio function would? We show in this paper that this is not the case and propose a new measure for current status data that has a useful interpretation in terms of shared frailty models and does not suffer from the shortcomings of the log odds ratio and the log conditional probability.

3 A new time-varying association measure relevant for shared frailty models

For shared frailty models, the CRF (2) can be expressed as

$$\theta^*(t_1, t_2) = 1 + a^*(t_1, t_2) \ , \tag{5}$$

where

$$a^*(t_1, t_2) = \frac{\mathrm{var}(Z|T_1 > t_1, T_2 > t_2)}{\mathrm{E}(Z|T_1 > t_1, T_2 > t_2)^2} \tag{6}$$

is the variance of the distribution of the relative frailty, $Z/\mathrm{E}(Z|T_1 > t_1, T_2 > t_2)$, in the population of survivors at (t_1, t_2) (Anderson et al. 1992). The properties of the relative

frailty were first studied by Hougaard (1984). The quantity (6), called the relative frailty variance (RFV), is thus a readily interpretable measure of how the heterogeneity of the hazard functions of survivors, as represented by a frailty model, evolves over time.

Oakes (1989) showed that in Archimedean copula models (Genest and MacKay 1986a,b) the CRF depends on (t_1, t_2) only through some function of $S(t_1, t_2)$. Hence, this also holds for $a^*(t_1, t_2)$. As previously noted, $\theta^*(t_1, t_2)$ cannot be calculated directly for current status data or indeed any other measure of local dependence. Instead, we propose a new measure with a simple interpretation and which can be calculated from current status data, which shares some of the properties of the cross-ratio function and the relative frailty variance and tracks their variation with age.

3.1 A scheme to track the cross-ratio function for current status data

As shown in (5), the CRF $\theta^*(x, x)$ can for shared frailty models be expressed in terms of the variance of the frailty in survivors at time x, $a^*(x, x)$. Since this cannot be identified from current status data, we instead use the variance of a gamma distributed frailty, which reproduces the observed bivariate distribution of survivors at age x, $S(x, x)$, and the marginal distributions $S_1(x)$ and $S_2(x)$ (a formal definition of our measure will follow). There are three reasons for choosing a gamma frailty. First, because the gamma frailty corresponds to the time-invariant association case, which serves as a reference in frailty models. Second, owing to the close link between the gamma frailty and the Clayton copula (Clayton 1978), this choice can represent negative association, if it is present – and thus can indicate that a shared frailty model is inappropriate, should this be the case. A third reason - to be discussed later - is that this choice produces a measure that tracks $\theta^*(x, x)$ and $a^*(x, x)$.

Suppose a shared gamma frailty model with the frailty having mean one and shape parameter $\theta > 0$. For convenience, the following reparameterization is proposed: $\phi = \ln\left(1 + \frac{1}{\theta}\right)$, where ln denotes the natural logarithm, hence $\theta = 1/(e^{\phi} - 1)$. Then, for an observed pattern $S(x, x)$, $S_1(x)$ and $S_2(x)$, it holds that (Clayton 1978)

$$S(x, x) = \max\left\{0, (S_1(x)^{1-e^{\phi}} + S_2(x)^{1-e^{\phi}} - 1)^{1/(1-e^{\phi})}\right\} \quad , \tag{7}$$

where the value $\phi = 0$ corresponds to independence between the two survival variables. When $\phi < 0$, (7) allows for negative dependence. Note that there is no frailty interpretation in this case. Let

$$f(\phi, S(x,x), S_1(x), S_2(x)) = \left(S_1(x)^{1-e^{\phi}} + S_2(x)^{1-e^{\phi}} - 1\right)^{\frac{1}{1-e^{\phi}}} - S(x,x) \ . \tag{8}$$

It is easily shown that, if $S(x,x)$ lies between $S_1(x)S_2(x)$ and $\min\{S_1(x), S_2(x)\}$, the equation

$$f(\phi, S(x,x), S_1(x), S_2(x)) = 0$$

considered as a function of ϕ, has a unique root $\phi_0(S(x,x), S_1(x), S_2(x))$. We define our new measure to be this root:

$$\phi(x) \equiv \phi_0(S(x,x), S_1(x), S_2(x)) \ . \tag{9}$$

For a shared gamma frailty model with frailty variance θ^{-1}, $\phi(x)$ is constant and $\phi(x) = \ln(1 + 1/\theta) = \ln(\theta^*(x,x))$. This follows from the definition of ϕ and the fact that if $\theta^*(t_1, t_2)$ is a positive constant, then the frailty is gamma.

In general, however, the frailty might not be gamma distributed, that is, the association in survivors might vary with time. If so, $\phi(x)$ will not be equal to the CRF $\theta^*(x,x)$. Nevertheless, according to the following results, $\phi(x)$ tracks the RFV $a^*(x,x)$ and hence the local CRF $\theta^*(x,x)$ for *all* shared frailty models with monotone CRF *regardless* of the frailty distribution, in the sense that it shows the same direction of travel.

Proposition 1. Consider a shared frailty model with cumulative baseline hazards $\Lambda_1(t)$ and $\Lambda_2(t)$, and suppose that the cross-ratio function is such that $\theta^*(t,t)$ is monotone. Let $a(t)$ be such that $\theta^*(t,t) = a(\Lambda_1(t) + \Lambda_2(t)) + 1$. Then,

(a) there exists a unique function $u(t)$ such that $\phi(t) = \ln\{a(u(t)) + 1\}$, with $u(t) \in [0, \Lambda_1(t) + \Lambda_2(t)]$, and $\phi(0) = \ln(\theta^*(0,0))$.

(b) under a weak identifiability condition, $u(t)$ is non-decreasing and hence $\phi(t)$ is monotone in the same direction as $\theta^*(t,t)$. Furthermore, if $\Lambda_1(t)$ and $\Lambda_2(t)$ are unbounded then, if $\ln(\theta^*(t,t))$ tends to a limit $c \geq 0$ as $t \to \infty$ where c can equal ∞, $\phi(t)$ also tends to c.

The proof of this proposition is in the supplementary material and the condition required in part (b) is stated in the proof. Note that the identifiability condition required for part (b) is sufficient but could perhaps be weakened. The condition fails, for example, when $\lambda_1(t) = \lambda_2(t) = 0$ on some interval. Proposition 1 implies that, when $\theta^*(t,t)$ is monotone, $\phi(t)$ is a lagged version of $\ln(\theta^*(t,t))$ and hence is also monotone. In addition $\phi(t)$ tends to the same limits as $\ln(\theta^*(t,t))$ when $t \to 0$ and $t \to \infty$.

Most standard frailties, such as from the power variance family (Aalen et al. 2008) have cross-ratio functions such that $\theta^*(t,t)$ is monotone. Farrington et al. (2010) describe some for which $\theta^*(t,t)$ is not monotone. In such cases the results of Proposition 1 apply to the initial section $[0, t_1)$, where t_1 is the first turning point of $\theta^*(t,t)$: on this interval, $\phi(t)$ is a lagged version of $\ln(\theta^*(t,t))$ with $\phi(0) = \ln(\theta^*(0,0))$.

One might query the choice of the Clayton copula and ask whether an association measure such as $\phi(x)$ could also be derived from Archimedean copulas other than the Clayton copula, for example using the copula representation of a shared inverse Gaussian frailty model. Such measures could indeed be defined. However, a measure derived from such a copula representation would not track the heterogeneity over time. For example, for the inverse Gaussian the RFV and the CRF are decreasing over time but the association measure $\phi(x)$ derived from such a copula would be constant if the inverse Gaussian frailty assumption holds.

3.2 Estimation and standard errors

An estimate of $\phi(x)$ is obtained by finding the root of the implicit function

$$f\left(\phi(x), \hat{S}(x,x), \hat{S}_1(x), \hat{S}_2(x)\right) = \left(\hat{S}_1(x)^{1-e^{\phi(x)}} + \hat{S}_2(x)^{1-e^{\phi(x)}} - 1\right)^{1/(1-e^{\phi(x)})} - \hat{S}(x,x) \ , \quad (10)$$

where $\hat{S}(x,x)$, $\hat{S}_1(x)$ and $\hat{S}_2(x)$ are calculated by the observed proportions, that is,

$$\hat{S}(x,x) = \frac{n_{00x}}{n_x} , \tag{11}$$

$$\hat{S}_1(x) = \frac{n_{0+x}}{n_x} , \tag{12}$$

$$\hat{S}_2(x) = \frac{n_{+0x}}{n_x} . \tag{13}$$

Since the estimates (11)–(13) are multinomial MLEs obtained from $(n_{00x}, n_{10x}, n_{01x}, n_{11x})$, $\hat{\phi}(x)$ is itself an MLE. The bisection algorithm as implemented in the function `uniroot` in R version 2.12.1 (R Development Core Team 2011) is used for finding the value $\hat{\phi}(x)$ such that $f\left(\hat{\phi}(x), \hat{S}(x,x), \hat{S}_1(x), \hat{S}_2(x)\right)$ is equal to zero.

Estimated asymptotic standard errors for $\hat{\phi}(x)$ were computed by means of the delta method (Benichou and Gail 1989). For the derivation of the asymptotic standard error of $\hat{\phi}$, the reader is referred to the supplementary material.

We carried out a simulation study to investigate bias and variance of the proposed measure for various sample sizes. The design of the study along with the results are given in the supplementary material. Results show that reliable estimates are obtained for moderate sample sizes.

To reduce the effect of differences in the baseline hazards, we suggest that, like the cross-ratio function, $\hat{\phi}(x)$ can also usefully be plotted, not against x, but against a function of $\hat{\pi}_{00}(x) = \hat{S}(x,x)$ such as $1-\hat{\pi}_{00}(x)$ or $-\ln(\hat{\pi}_{00}(x))$ (see also Viswanathan and Manatunga 2001), suitably isotonized. A suitable isotonizing procedure is the method of greatest convex minorant described by Groeneboom and Wellner (1992).

4 Evaluation of time-varying association measures for shared frailty models

In this Section, the performance of the log odds ratio, the log conditional probability and our proposed association measure in mirroring the temporal variation in the strength of association shall be examined by means of experiments. In Subsection 4.1 we consider the important case of gamma frailties and evaluate to what extent the three measures provide useful diagnostics for identifying such models. We also investigate the impact of different cumulative baseline hazards. The suitability of these measures in serving as a diagnostic

tool to suggest appropriate frailty distributions other than the gamma is examined in Subsection 4.2.

4.1 Identifying gamma frailties

Cumulative baseline hazards are generated for ages $x = 0.05, 0.06, ..., 50.00$ 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 log odds ratios, log conditional probabilities and our proposed measure are calculated at each x for the three baseline hazards.

Figure 2 (i)-(vi) display the three tracings for the three association measures, where the latter are plotted both against x and -ln($\pi_{00}(x)$). 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 plots (i)-(iv) both ln(OR) and ln(ψ) (similarly) increase with age for all baseline models. Moreover, when plotted against age the shape of the temporal variation in the strength of association clearly depends on the baseline hazard chosen, see Figure 2 (i) and (iii) for ln(OR) and ln(ψ), respectively. Hence, there is evidence that the odds ratio and the conditional probability are both severely influenced by the cumulative baselines. When plotted against $-\ln(\pi_{00}(x))$, however, ln(OR) and ln(ψ) are largely free of the influence of the baseline hazards. By contrast, $\phi(x)$ is the same constant for all three baseline hazards, as illustrated by Figure 2 (v) and (vi). That is, the obtained association pattern reflects the absence of any time-varying association in the population, free from the influence of the cumulative baselines.

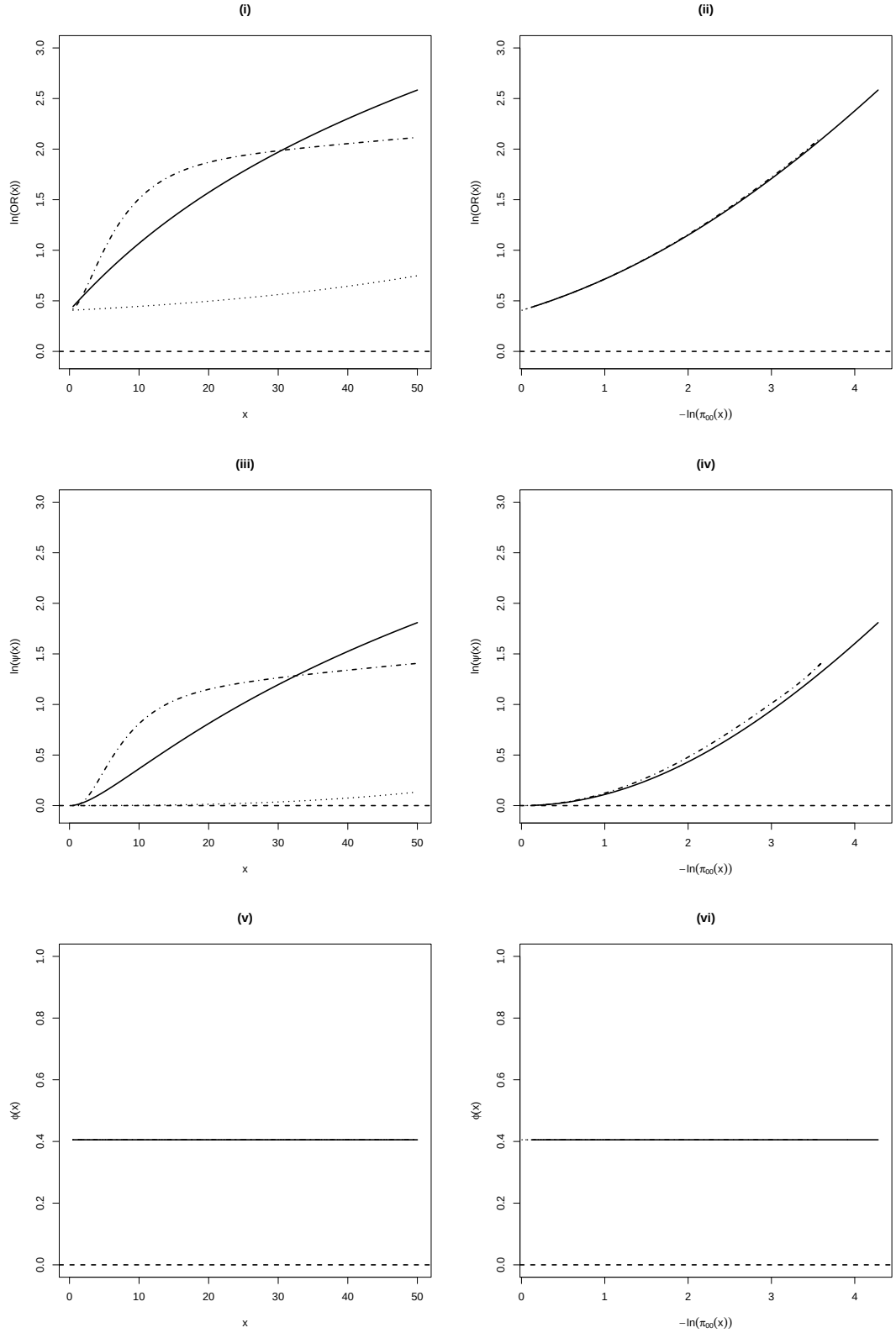

Figure 2: Shared gamma frailty model ($\theta = 2$): $\ln(\mathrm{OR})$, $\ln(\psi)$ and ϕ against x and $-\ln(\pi_{00}(x))$ for constant baseline (solid line), Gompertz baseline (dotted line) and EDL baseline (dot-dashed line) (horizontal dashed line: no association).

4.2 Tracking the time-varying association in survivors

It is well established that 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 the measures reflect these population effects, data are generated in the style described above and the three measures are calculated at each x for various shared inverse Gaussian ($Z \sim InvG(1, \theta)$) and compound Poisson frailty models ($Z \sim CP(1, \theta^{-1}, \nu)$ and $\nu = 1.5$), all with mean one and constant baseline hazards (with $c_1 = 0.2$ and $c_2 = 0.1$). The results are displayed in Figure 3 and Figure 4 for the inverse Gaussian and compound Poisson frailty models, respectively, for a range of values of $\text{Var}(Z) = \theta^{-1}$. Whereas the log conditional probability totally fails to mirror the declining heterogeneity induced by the inverse Gaussian distribution for the whole range of shape parameters, the log odds ratio reflects these population effects only for the case $\theta = 0.1$. For the compound Poisson frailty models the three measures increase, thus mirroring the increasing heterogeneity of the survivor population. However, neither ln(OR) nor $\ln(\psi)$ clearly differentiate between inverse Gaussian and compound Poisson frailties, and induce very different dependence patterns in survivors. Hence, there is evidence that neither ln(OR) nor $\ln(\psi)$ are suitable diagnostics for suggesting a frailty distribution. In contrast, for the whole variety of models tried our proposed measure adequately mirrors the decreasing (increasing) heterogeneity caused by the inverse Gaussian (compound Poisson) frailties (see Figure 3 and Figure 4 (v) and (vi), respectively).

In Figure 5 (i)-(iv), $\phi(x)$ and $\ln(\theta^*(x,x)) = \ln(1+a^*(x,x))$ are plotted against $-\ln(\pi_{00}(x))$ and x for a shared inverse Gaussian and compound Poisson model (with $\nu = 1.5$) (both with mean one and variance 10), respectively, choosing the same three baseline hazards as described above. The upper three lines in plot (i) (in plot (iii)) correspond to $\phi(x)$ (to $\ln(\theta^*(x,x))$), whereas the lower three lines correspond to $\ln(\theta^*(x,x))$) (to $\phi(x)$). Figure 5 illustrates how, following Proposition 1, $\phi(x)$ is able to track the RFV $a^*(x,x)$ in case the frailty is not gamma distributed, that is, the association in survivors is not constant. Moreover, when plotted against $-\ln(\pi_{00}(x))$ the shape of the time-varying association does not depend on the baseline hazard chosen. When plotted against x, the shapes differ

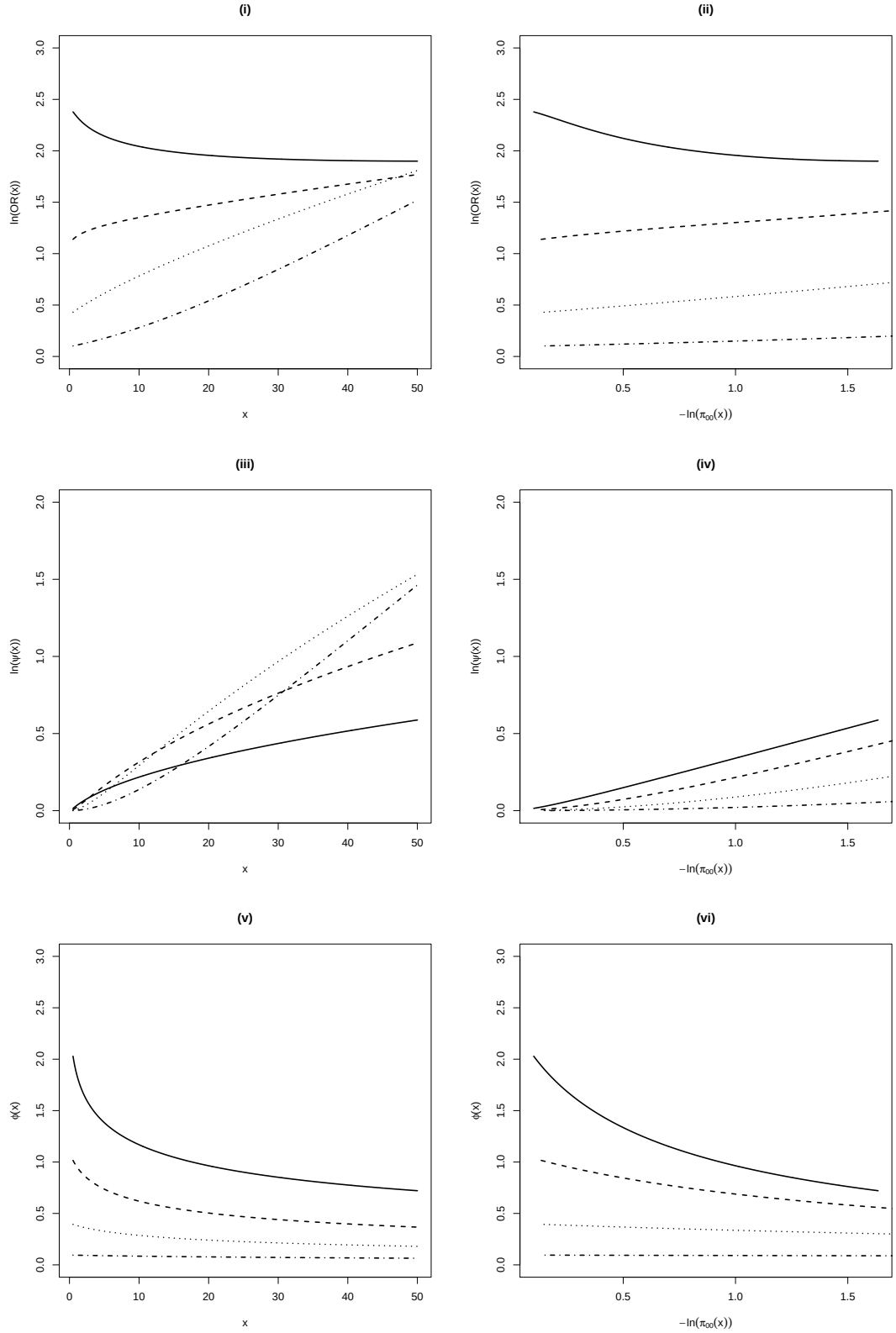

Figure 3: $\ln(\text{OR})$, $\ln(\psi)$ and ϕ against x and $-\ln(\pi_{00}(x))$ for $Z \sim InvG(1, 0.1)$ (solid line), $Z \sim InvG(1, 0.5)$ (dashed line) , $Z \sim InvG(1, 2)$ (dotted line), $Z \sim InvG(1, 10)$ (dot-dashed line) and constant baseline hazards.

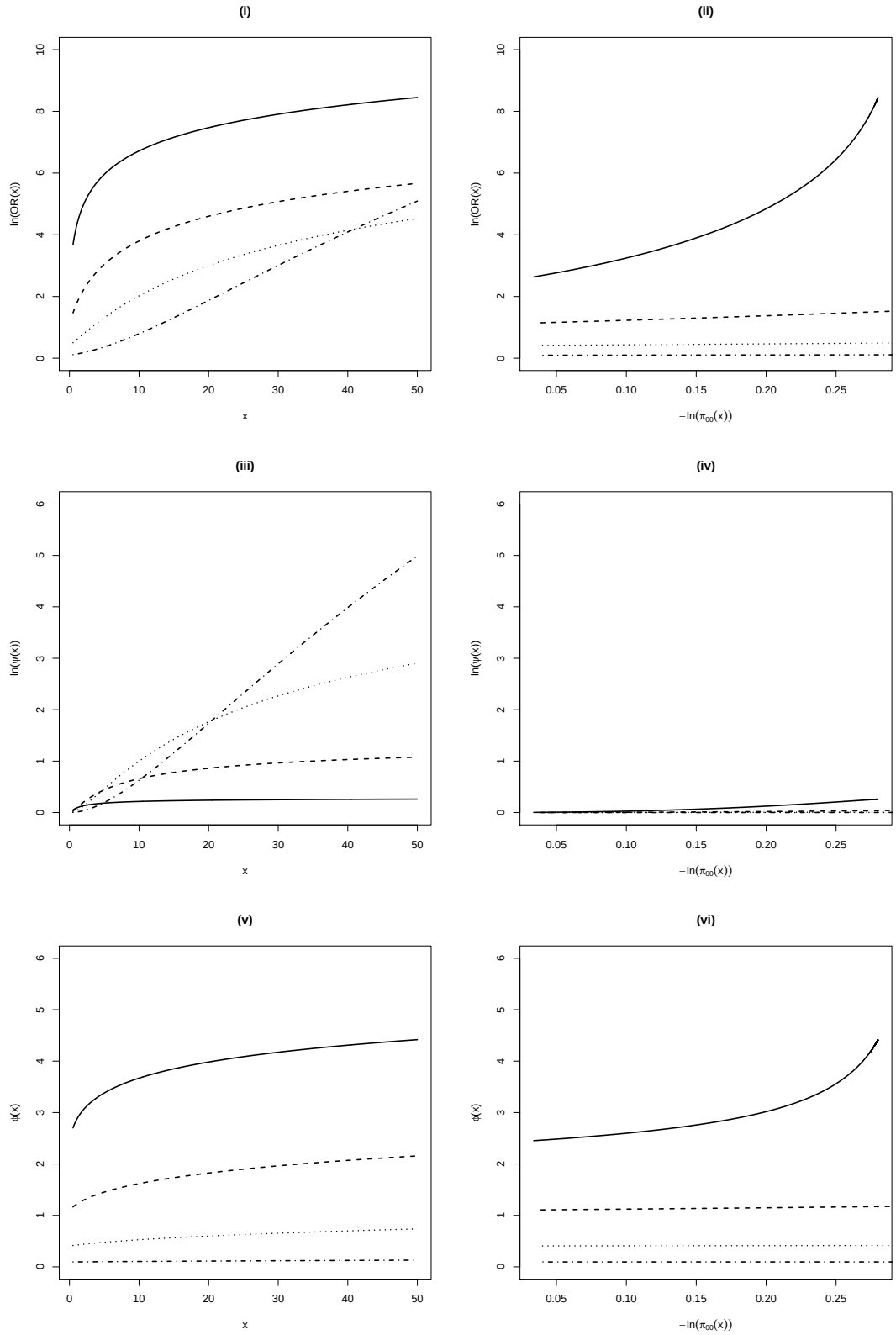

Figure 4: ln(OR), $\ln(\psi)$ and ϕ against x and $-\ln(\pi_{00}(x))$ for $Z \sim CP(1, 10, 1.5)$ (solid line), $Z \sim CP(1, 2, 1.5)$ (dashed line) , $Z \sim CP(1, 0.5, 1.5)$ (dotted line), $Z \sim CP(1, 0.1, 1.5)$ (dot-dashed line) and constant baseline hazards.

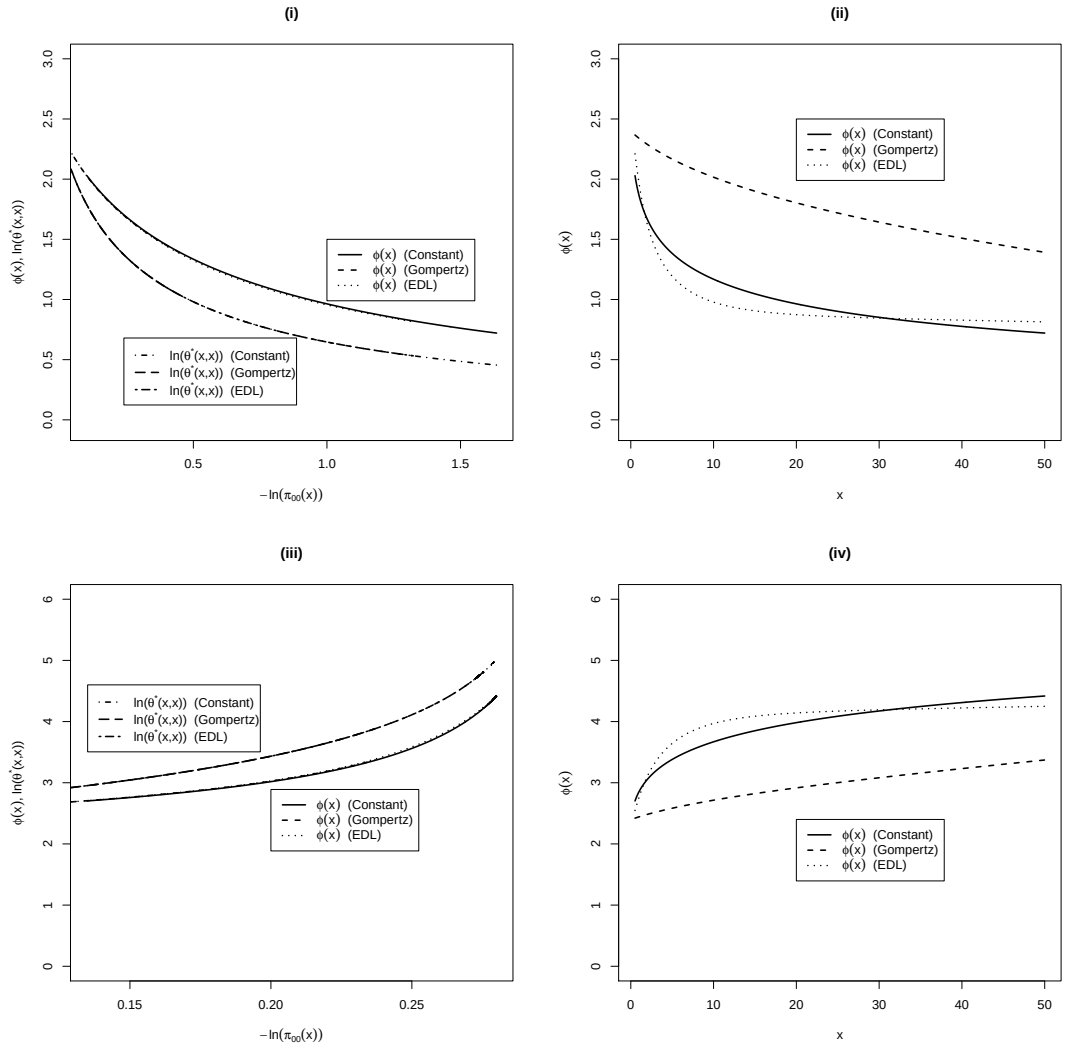

Figure 5: Tracking $\ln(\theta^*(x,x))$ by $\phi(x)$: inverse Gaussian (i)-(ii) and compound Poisson (iii)-(iv) frailty model (for three different baseline hazards).

(as expected), reflecting the strengths of the selection effects represented by the different baselines. We also varied the parameter values for θ in the inverse Gaussian distribution and for θ and ν in the compound Poisson distribution. Results (omitted) confirm these findings. In this sense, the new time-varying association measure helps to suggest a class of frailty distributions, based on how the heterogeneity in survivors varies with time.

5 Applications to serological survey data

In this Section, the methods developed in this paper are illustrated with two datasets, on Epstein-Barr virus (EBV) and herpes simplex virus type 1 (HSV1) infections (the motivating example of Section 2) and on Toxoplasma and *Helicobacter pylori* infections. The data have arisen from two large surveys undertaken in the United Kingdom (Data source: Health Protection Agency). Both are nationwide surveys of serum samples taken for diagnostic testing for conditions unconnected with the infections studied here. For each infection, a positive (negative) test result indicates prior infection (susceptibility to infection). Equivocal test results are recoded as being positive indicating prior exposure. Owing to ethical restrictions, the only information on each individual is locality of the testing laboratory, gender, age, and test results.

5.1 EBV and HSV1 infections

This survey was undertaken in 1994 and was reported by Morris et al. (2002). Serum samples from 2,803 individuals age 1-30 years were tested for evidence of prior infection by EBV and HSV1. There is no vaccination against EBV and HSV1. Both infections are transmitted through exchange of saliva. Two types of contacts are believed to be involved: general person-to-person contacts, which peak in childhood, and intimate contacts through kissing, which occur after puberty and peak in teenage years (Heymann, D. L. (ed.) 2008, pp. 300–304, 428–430). Since both infections share the same route of transmission, it is plausible that the association between them should be governed by a shared frailty. Farrington and Whitaker (2005) give odds ratios for the association between EBV and HSV1 infection within age groups. They found the association to be significant in all

groups. However, Farrington and Whitaker (2005) do not provide a diagnostic tool for shared frailty models with current status data, nor do they focus on how the strength of association varies with age.

Figure 6 (i), (ii) and (iii) are plots of estimates of ϕ, of the log conditional probability and of the log odds ratio versus age, respectively, for the data on EBV and HSV1 infections (cf. Section 2). For all of the 4-tuples $(n_{00x}, n_{01x}, n_{10x}, n_{11x})$ none of the counts are equal to zero. Note that in Figure 6 (i), $\hat{\phi}$ is plotted against x and not, as suggested in Subsection 3.2, against $-\ln(\hat{\pi}_{00}(x))$, where $\hat{\pi}_{00}(x)$ is isotonized by the greatest convex minorant method (for the latter plot see Figure 6 (iv)). This is because age is the key variable we are interested from the epidemiological point of view. By plotting $\hat{\phi}$ against $-\ln(\hat{\pi}_{00}(x))$ we would lose the age-related interpretation. However, both plots have their merits. Unlike in Figure 6 (i), $\hat{\phi}$ in Figure 6 (iv) reduces the dependence on the baseline hazards, therefore satisfying one of the desirable properties of time-varying association measures for shared frailty models (cf. Subsection 4.1). As such, a diagnostic plot like Figure 6 (iv) would allow us to compare association patterns in different datasets.

In the age-varying association pattern (i) for $\hat{\phi}$ two peaks can be identified, in early childhood and in late teen years, whereas the association is overall decreasing towards zero. The two peaks are due to a greater variation between children and between adolescents in contact patterns. The observed decreasing association in survivors is suggestive of a time-invariant frailty with decreasing relative frailty variance (e.g. an inverse Gaussian frailty model) or a frailty that varies over time. The plot (ii) for the log conditional probability tells a rather different story. At early ages, this might appear to suggest little association between the two infections. However, this is due to the fact that in the definition of $\hat{\psi}(x) = \frac{n_x\, n_{00x}}{n_{0+x} n_{+0x}}$ the cell n_{00x} determines the association if n_{01x} and n_{10x} are relatively small compared to n_{00x}. For paired serological survey data this is likely to be the case at early ages. The dominance of n_{00x} in the definition of the measure also has an effect on the standard errors; the precision of the estimates are much larger at early ages and hence lower at later ages compared to the ones obtained both for the log odds ratio and $\hat{\phi}$. It seems a serious shortcoming of the conditional probability measure that the degree of association is overly influenced by the single cell n_{00x}. The plot (iii) for the log odds

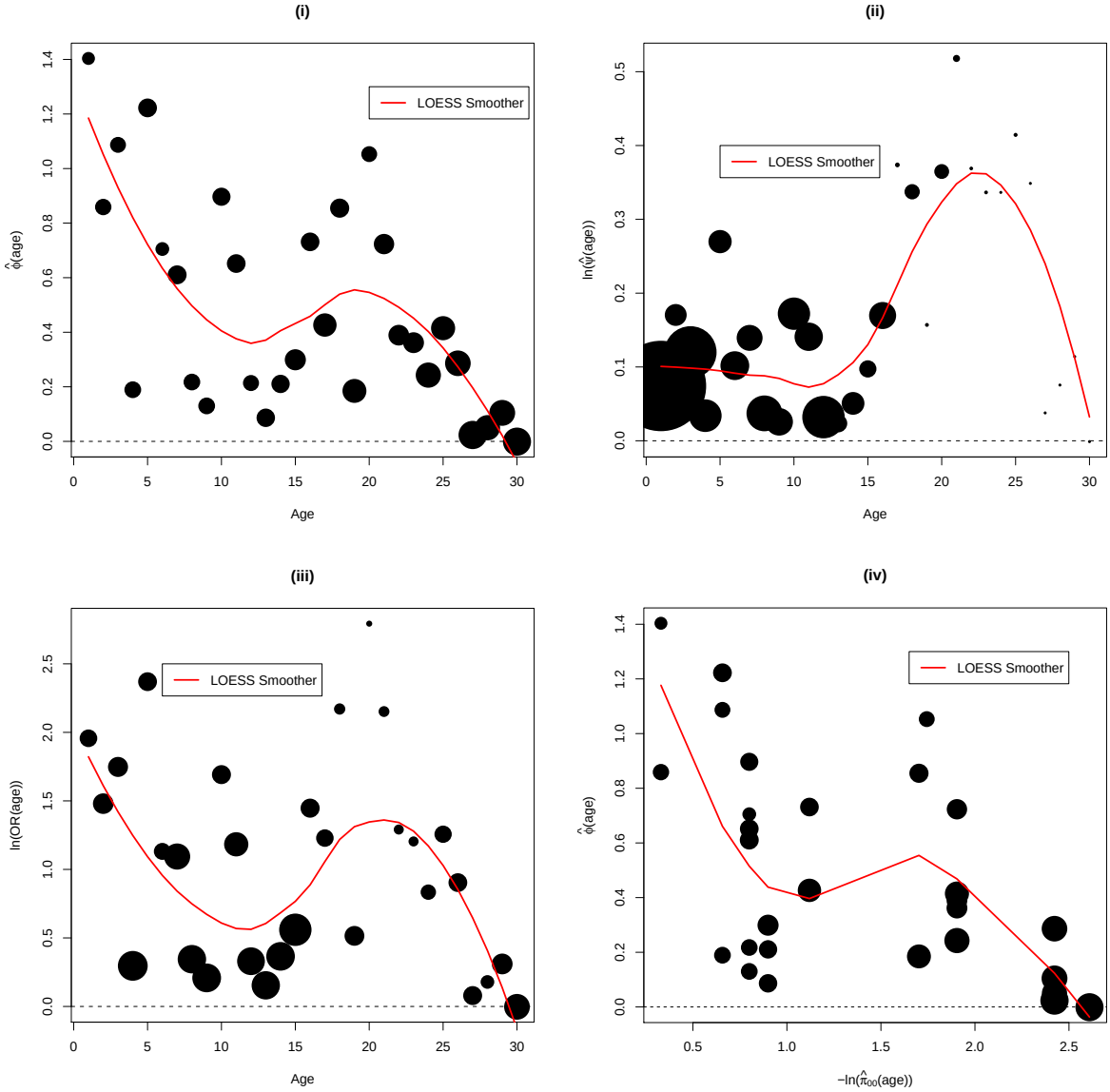

Figure 6: Plot of $\hat{\phi}$ (i), $\ln(\hat{\psi})$ (ii), and ln(OR) (iii) by age, as well as plot of $\hat{\phi}$ against $-\ln(\hat{\pi}_{00}(\text{age}))$ (iv), for EBV and HSV1 data (dashed line: no association). For (ii), a minimum point size was defined to make all points visible.

ratio is quite similar to plot (i). However, as it has been illustrated in Subsection 4.2, the odds ratio lacks a clear interpretation in terms of frailty models.

5.2 Toxoplasma and *Helicobacter pylori* infections

This survey was undertaken in 1996. Whereas the study of *Helicobacter pylori* was reported by Vyse et al. (2002), a study on the Toxoplasma data hasn't been published yet (Richard Pebody 2010, personal communication). A total of 3,632 individuals age 1-84 years were tested for antibodies to Toxoplasma and *Helicobacter pylori* infections. Toxoplasma is a protozoan zoonosis. Human infection in pregnancy at early stages may lead to death of the fetus or can seriously damage the baby later. *Helicobacter pylori* is a bacterial infection of humans causing acute and chronic gastritis and peptic ulcer disease. Both infections are transmitted by oral ingestion of contaminated matter (Heymann, D. L. (ed.) 2008, pp. 250-253, 613-617). There is no vaccination against either infectious agents for humans. Heterogeneity in hygiene is likely to result in association between the two infections. To the best of our knowledge, the two infections have not been studied together.

For this paired data set one or more of the counts within the 4-tuples $(n_{00x}, n_{01x}, n_{10x}, n_{11x})$ are zero. We propose to deal with zeroes as follows. If there is a single cell in the 4-tuple of counts, add 0.5 to all of the counts in the corresponding 4-tuple (Agresti 2002, Section 9.8). When there are two zeroes, but all 4 margins $n_{0+x}, n_{+0x}, n_{1+x}$, and n_{+1x} are greater than zero, add 0.5 to all of the cells. If there are two zeroes and at least one of the margin is zero or there are more than two zero cells, the point is not informative about association and should be combined with a neighbour, and plotted at the average age. For the data on Toxoplasma and *Helicobacter pylori* infections this was done for the points at $x = 2, 3, 6, 7, 13$, at which two cells and one margin were zero. Alternatively, the point could be deleted. Note that $\ln(\hat{\psi})$ can be calculated when all three cells other than n_{00} are zero, in which case $\ln(\hat{\psi}) = 0$ and so there is 'no association'. However, for the sake of comparability, we combined the same points for all measures.

Figure 7 (i), (ii) and (iii) are plots of estimates of ϕ, of the log conditional probability and of the log odds ratio versus age, respectively, for the data on Toxoplasma and *Helicobacter*

pylori infection. The plot (i) for $\hat{\phi}$ suggests that there is particularly strong heterogeneity among pre-school children and that the heterogeneity in the survivor population may be decreasing towards a positive asymptote. If so, Proposition 1 indicates that this is also true of the cross-ratio function and relative frailty variance. An equalisation effect caused by school socialisation may be responsible for the reduction in heterogeneity with age, whereas the remaining positive association in adulthood is likely to due to differences between individuals in behaviour such as differences in hygiene levels.

In terms of a modelling strategy, the decreasing heterogeneity in adulthood could be due to a selection effect caused by a time-invariant frailty model or to a temporal variation of the frailty itself. With respect to the former, Aalen et al. (2008) describes the Kummer family of densities for use in shared frailty models. This family includes distributions with relative frailty variance monotonically decreasing towards a positive asymptote (Farrington et al. 2010). Figure 7 (iv), in which $\hat{\phi}$ is plotted against a function of an isotonized version of $\hat{\pi}_{00}(x)$, is informative mainly about the heterogeneity between children at early ages, owing to the sparsity at higher ages. The plot (iii) for estimates of the log odds ratio closely resembles plot (i). For the log conditional probability, the observed association pattern (ii) indicates independence between the two infections. However, the observed association pattern is misleading; the absence of age-varying heterogeneity is solely a result of the dominant effect of the count of individuals being susceptible to both infections. As in the previous example, there are large differences in the standard errors for the estimates of $\ln(\hat{\psi})$. We conclude that this measure is seriously flawed and not appropriate to assess time-varying association in current status data.

6 Discussion

We introduced a new measure for assessing the temporal variation in the strength of association inherent in bivariate current status data. Owing to its connection with the relative frailty variance, the new measure ϕ serves as a diagnostic tool for suggesting classes of frailty distributions with constant, increasing or decreasing relative frailty variance. The shape of the observed time-varying association aids identification of a suitable frailty

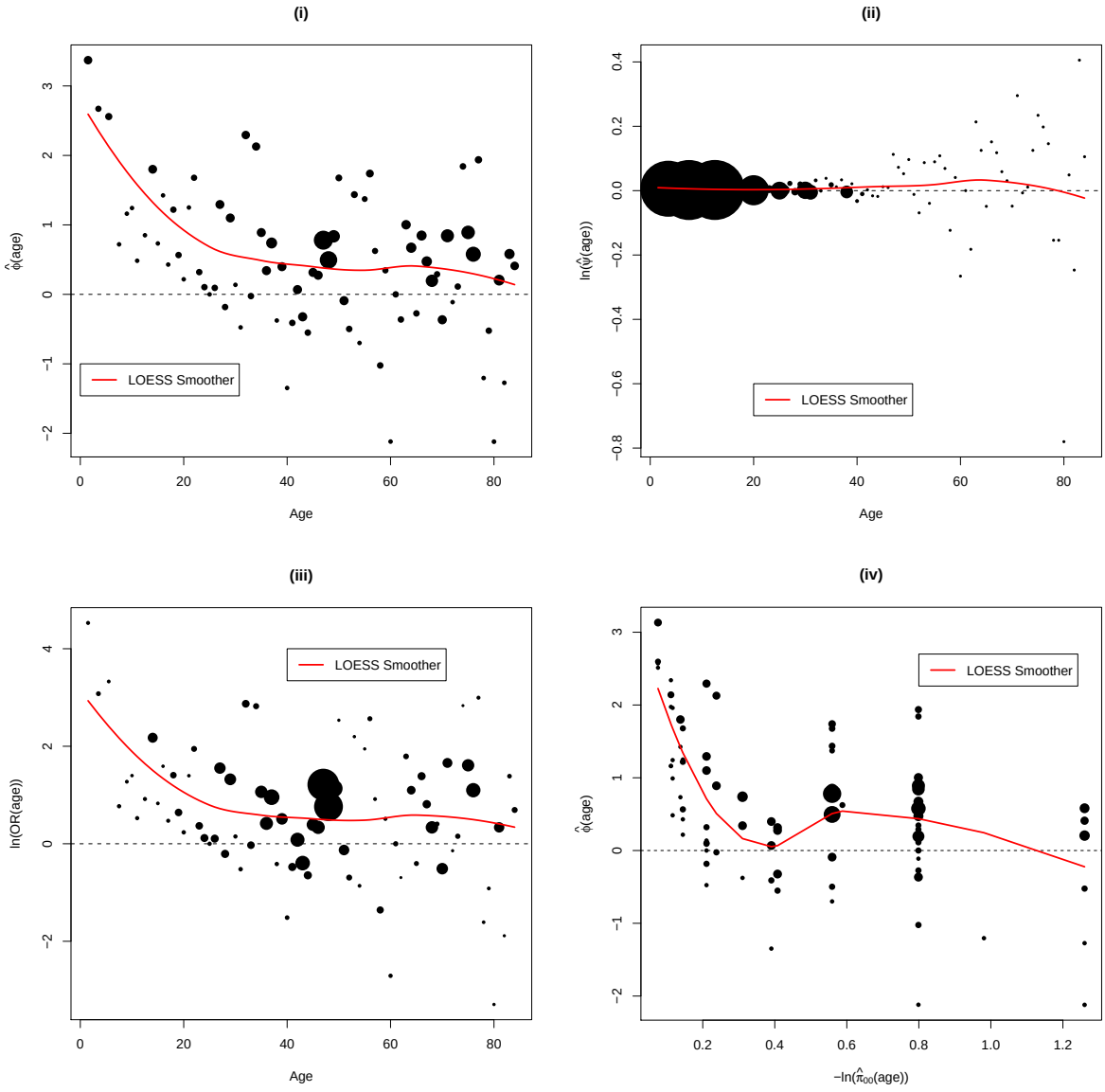

Figure 7: Plot of $\hat{\phi}$ (i), $\ln(\hat{\psi})$ (ii), and ln(OR) (iii) by age, as well as plot of $\hat{\phi}$ against $-\ln(\hat{\pi}_{00}(\text{age}))$ (iv), for Toxoplasma and *Helicobacter pylori* data (dashed line: no association). For (ii), a minimum point size was defined to make all points visible.

model, which then could be fitted to the data set at hand. The diagnostic plot of $\hat{\phi}$, or a smoothed version thereof, in which the time axis is suitably rescaled to remove the effect of the cumulative baselines would also allow us to compare different frailty models in different populations or different processes within the same population.

A notable merit of ϕ is that it tracks Clayton's local cross-ratio function and hence the relative frailty variance in the sense that it has the same direction of travel. In contrast, existing global measures of association applicable to current status data such as the non-local version of the odds ratio or the conditional probability lack any connection to local dependence functions. Moreover, we illustrated that the odds ratio and the conditional probability may not reliably suggest appropriate frailty distributions.

The methods developed in this paper were applied to bivariate serological survey data. Based on our analysis of two data sets on pairs of infections, we conclude that this new measure is a fruitful one that can provide insights in representing heterogeneities between individuals in the acquisition and transmission of infectious diseases.

There are two broad limitations to our approach. The first limitation is due to data imperfections. With current status data the event of interest could have occurred at any time during the interval $(0, x]$. Therefore, the association observed at time x will not be truly local but 'averaged' in some sense over $(0, x]$. The second, which is shared with other measures of association, is due to identifiability issues inevitably associated with frailty models. In shared frailty models, the temporal pattern in the population association could be due to a time-varying frailty or to selection effects stemming from a time-invariant frailty, and there is no way of distinguishing between them. Nevertheless, our exploratory approach provides a new way of investigating the association structure in current status data before fitting models. In this sense, the exploratory tools presented in this paper could be viewed as the initial step of a comprehensive model selection procedure for analyzing current status data by means of frailty models.

Supplementary material

Asymptotic standard error for $\ln(\hat{\psi})$; Proof of Proposition 1; Asymptotic standard error for $\hat{\phi}$; Simulation study.

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. (2002), *Categorical Data Analysis*, John Wiley & Sons: Hoboken, New Jersey, 2nd 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.

Benichou, J. and Gail, M. H. (1989), "A delta method for implicitly defined random variables," *The American Statistcian*, 43, 41–44.

Bogaerts, K. and Lesaffre, E. (2008), "Estimating local and global measures of association for interval censored data with a smooth estimate of the density," *Statistics in Medicine*, 27, 5941–5955.

Chen, M.-C. and Bandeen-Roche, K. (2005), "A diagnostic for association in bivariate survival models," *Lifetime Data Analysis*, 11, 245–264.

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.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Coutinho, F., Massad, E., Lopez, L. F., Burattini, M. N., Struchiner, C. J., and Azevedo-Neto, R. S. (1999), "Modelling heterogeneities in individual frailties in epidemic models," *Mathematical and Computer Modelling*, 30, 97–115.

Dale, J. R. (1986), "Global cross-ratio models for bivariate, discrete, ordered responses," *Biometrics*, 42, 721–727.

Diamond, I. D. and McDonald, J. W. (1992), "Analysis of current status data," in *Demographic Applications of Event History Analysis*, eds. Trussell, J., Hankinson, J., and Tilton, R., Clarendon Press: Oxford, pp. 231–252.

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

Drouet Mari, D. and Kotz, S. (2001), *Correlation and Dependence*, Imperial College Press: London.

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

Dunson, D. B. and Dinse, G. E. (2002), "Bayesian models for multivariate current status data with informative censoring," *Biometrics*, 58, 79–88.

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

Farrington, C. P., Kanaan, M., and Gay, N. J. (2001), "Estimation of the basic reproduction number for infectious diseases from age stratified serological survey data (with discussion)," *Applied Statistics*, 50, 251–292.

Farrington, C. P., Unkel, S., and Anaya-Izquierdo, K. (2010), "The relative frailty variance and shared frailty models," The Open University Statistics Group Technical Report 10/13, available from `http://statistics.open.ac.uk/technical_reports`.

Farrington, C. P. and Whitaker, H. J. (2005), "Contact surface models for infectious diseases: estimation from serologic survey data," *Journal of the American Statistical Assoictaion*, 100, 370–379.

Genest, C. and MacKay, R. J. (1986a), "Copules Archimediennes et Familles de Lois Bidimensionelles Dont les Marges Sont Donnees," *Canadian Journal of Statistics*, 14, 145–159.

— (1986b), "The joy of copulas: Bivariate distributions with given marginals," *The American Statistician*, 40, 280–283.

Groeneboom, P. and Wellner, J. A. (1992), *Information Bounds and Non-Parametric Maximum Likelihood Estimation*, Birkhäuser: Boston.

Gupta, R. C. (2003), "On some association measures in bivariate distributions and their relationships," *Journal of Statistical Planning*, 117, 83–98.

Heymann, D. L. (ed.) (2008), *Control of Communicable Diseases Manual*, American Public Health Association: Washington D.C., 19th ed.

Hougaard, P. (1984), "Life table methods for heterogeneous populations: Distributions describing the heterogeneity," *Biometrika*, 71, 75–83.

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

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

Morris, M. C., Edmunds, W. J., Hesketh, L. M., Vyse, A. J., Miller, E., Morgan-Capner, P., and Brown, D. W. G. (2002), "Sero-epidemiological patterns of Epstein-Barr and Herpes Simplex (HSV-1 and HSV-2) Viruses in England and Wales," *Journal of Medical Virology*, 67, 522–527.

Nelson, R. B. (2006), *An Introduction to Copulas*, Springer: Berlin, 2nd ed.

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

Plackett, R. L. (1965), "A class of bivariate distributions," *Journal of the American Statistical Association*, 60, 516–522.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

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

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

Vyse, A. J., Gay, N. J., Hesketh, L. M., Andrews, N. J., Marshall, B., Thomas, H. I. J., Morgan-Capner, P., and Miller, E. (2002), "The burden of *Helicobacter pylori* infection in England and Wales," *Epidemiology and Infection*, 128, 411–417.

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 Press: Boca Raton, FL.

Supplementary material to "A new measure of time-varying association for shared frailty models with bivariate current status data"

Steffen Unkel and C. Paddy Farrington

Department of Mathematics and Statistics
The Open University
Milton Keynes, UK

September 26, 2011

Asymptotic standard error for $\ln(\hat{\psi})$

Let $g(\boldsymbol{\pi})$ denote a differentiable function of $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})^\top$ with sample value $g(\hat{\boldsymbol{\pi}})$, where $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{00}, \hat{\pi}_{01}, \hat{\pi}_{10}, \hat{\pi}_{11})^\top$. The delta method implies that (e.g., Agresti 2002, Section 14.1):

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

where

$$\sigma^2 = \sum_i \sum_j \pi_{ij} \zeta_{ij}^2 - \left(\sum_i \sum_j \pi_{ij} \zeta_{ij} \right)^2$$

is the asymptotic variance and

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

Then $\sigma/\sqrt{n}$ is an asymptotic standard error for $g(\hat{\boldsymbol{\pi}})$. We apply the delta method to the log conditional probability, taking

$$\begin{aligned} g(\boldsymbol{\pi}) = \ln(\psi) &= \ln\left(\frac{\pi_{00}}{\pi_{0+}\pi_{+0}}\right) = \ln(\pi_{00}) - \ln(\pi_{0+}\pi_{+0}) \\ &= \ln(\pi_{00}) - (\ln(\pi_{00} + \pi_{01}) + \ln(\pi_{00} + \pi_{10})) \\ &= \ln(\pi_{00}) - \ln(1 - \pi_{11} - \pi_{10}) - \ln(1 - \pi_{11} - \pi_{01}) \end{aligned} \tag{14}$$

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

$$\zeta_{00} = \frac{1}{\pi_{00}} , \tag{15}$$

$$\zeta_{01} = \frac{1}{(1-\pi_{11}-\pi_{01})} = \frac{1}{\pi_{+0}} , \tag{16}$$

$$\zeta_{10} = \frac{1}{(1-\pi_{11}-\pi_{10})} = \frac{1}{\pi_{0+}} , \tag{17}$$

$$\zeta_{11} = \frac{1}{(1-\pi_{11}-\pi_{10})} + \frac{1}{(1-\pi_{11}-\pi_{01})} = \frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}} . \tag{18}$$

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

$$\begin{aligned}
\sigma^2 &= \sum_i \sum_j \pi_{ij}\zeta_{ij}^2 - \left(\sum_i \sum_j \pi_{ij}\zeta_{ij}\right)^2 \\
&= \frac{1}{\pi_{00}} + \frac{\pi_{01}}{\pi_{+0}^2} + \frac{\pi_{10}}{\pi_{0+}^2} + \pi_{11}\left(\frac{1}{\pi_{0+}^2} + \frac{1}{\pi_{+0}^2} + \frac{2}{\pi_{0+}\pi_{+0}}\right) - \left(\frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}} - 1\right)^2 \\
&= \frac{1}{\pi_{00}} + \frac{\pi_{01}+\pi_{11}-1}{\pi_{+0}^2} + \frac{\pi_{10}+\pi_{11}-1}{\pi_{0+}^2} \\
&\quad + 2\,\frac{(\pi_{11}-1)}{\pi_{0+}\pi_{+0}} + 2\left(\frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}}\right) - 1 .
\end{aligned} \tag{19}$$

Hence, the asymptotic standard error of $\ln(\hat{\psi})$ is

$$\begin{aligned}
\sigma(\ln(\hat{\psi})) &= \left(\frac{1}{n\pi_{00}} + \frac{\pi_{01}+\pi_{11}-1}{n\pi_{+0}^2} + \frac{\pi_{10}+\pi_{11}-1}{n\pi_{0+}^2}\right. \\
&\quad \left. + \frac{2}{n}\left(\frac{\pi_{11}-1}{\pi_{0+}\pi_{+0}} + \frac{1}{\pi_{0+}} + \frac{1}{\pi_{+0}}\right) - \frac{1}{n}\right)^{1/2} .
\end{aligned} \tag{20}$$

Proof of Proposition 1

Preliminaries

Consider a shared frailty model with cumulative baseline hazard functions $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$. We make use of the following two results from Farrington et al. (2010).

Fact 1: The relative frailty variance $a^*(t_1, t_2)$ may be written $a^*(t_1, t_2) = a(\Lambda_{01}(t_1) + \Lambda_{02}(t_2))$ for some function a, the scaled relative frailty variance. Thus $\theta^*(t_1, t_2) = 1 + a(\Lambda_{01}(t_1) + \Lambda_{02}(t_2))$.

Fact 2: Let $A(s) = \int_0^s a(t)\mathrm{d}t$. Suppose without loss of generality that the frailty has unit mean. Then the marginal survivor functions $S_1(t)$ and $S_2(t)$ and the joint survivor function $S(t,t)$ evaluated at (t,t), may be represented as follows:

$$S_1(t) = \exp\Big(-\int_0^{\Lambda_{01}(t)} \frac{1}{1+A(s)}\mathrm{d}s\Big) , \tag{21}$$

$$S_2(t) = \exp\Big(-\int_0^{\Lambda_{02}(t)} \frac{1}{1+A(s)}\mathrm{d}s\Big) , \tag{22}$$

$$S(t,t) = \exp\Big(-\int_0^{\Lambda_{01}(t)+\Lambda_{02}(t)} \frac{1}{1+A(s)}\mathrm{d}s\Big) . \tag{23}$$

The following proof of Proposition 1 (a) and (b) is for $\theta^*(t,t)$ monotone decreasing; that for monotone increasing $\theta^*(t,t)$ is analogous.

Proof of Proposition 1 (a)

Define

$$f(x,u) = \exp\Big(\int_0^x \frac{a(u)}{1+A(s)}\mathrm{d}s\Big) - 1$$

and note that $f(0,u) = 0$. Note also that $\phi(t)$ is defined implicitly by the relation

$$f(\Lambda_1(t), e^{\phi(t)}-1) + f(\Lambda_2(t), e^{\phi(t)}-1) - f(\Lambda_1(t)+\Lambda_2(t), e^{\phi(t)}-1) = 0.$$

Since $\theta^*(t,t)$ is a decreasing function of t, it follows that $a(s)$ is a decreasing function of s. Now

$$\frac{\partial^2 f}{\partial x^2}(x,u) = \frac{a(u)\{f(x,u)+1\}}{\{1+A(x)\}^2}\{a(u)-a(x)\} ,$$

thus, $\frac{\partial^2 f}{\partial x^2}(x,0) > 0$ and so $f(x,0)$ is superadditive on $[0, \Lambda_1(t)+\Lambda_2(t)]$. It follows that

$$f(\Lambda_1(t),0) + f(\Lambda_2(t),0) \le f(\Lambda_1(t)+\Lambda_2(t),0).$$

Also, $\frac{\partial^2 f}{\partial x^2}(x, \Lambda_1(t)+\Lambda_2(t)) < 0$ for $x \in [0, \Lambda_1(t)+\Lambda_2(t)]$, and so $f(x, \Lambda_1(t)+\Lambda_2(t))$ is subadditive on this interval. Thus,

$$f(\Lambda_1(t), \Lambda_1(t)+\Lambda_2(t)) + f(\Lambda_2(t), \Lambda_1(t)+\Lambda_2(t)) \ge f(\Lambda_1(t)+\Lambda_2(t), \Lambda_1(t)+\Lambda_2(t)).$$

Since f is continuous, by the Intermediate Value Theorem there exists $u(t) \in [0, \Lambda_1(t) + \Lambda_2(t)]$ such that

$$f(\Lambda_1(t), u(t)) + f(\Lambda_2(t), u(t)) = f(\Lambda_1(t) + \Lambda_2(t), u(t)).$$

We now verify that $u(t)$ is unique. Define

$$h(t, u) = f(\Lambda_1(t), u) + f(\Lambda_2(t), u) - f(\Lambda_1(t) + \Lambda_2(t), u).$$

Thus, $u(t)$ satisfies $h(t, u(t)) = 0$. Choose any such value u_0. We then have

$$\frac{\partial h(t,u)}{\partial u}(t, u_0) = \frac{a'(u_0)}{a(u_0)}\{\alpha_1 \ln(\alpha_1) + \alpha_2 \ln(\alpha_2) - (\alpha_1 + \alpha_2 - 1)\ln(\alpha_1 + \alpha_2 - 1)\} \ ,$$

where the prime denotes differentation with respect to the argument and

$$\alpha_i = \exp\Big(\int_0^{\Lambda_i(t)} \frac{a(u_0)}{1 + A(s)}\mathrm{d}s\Big), \quad i = 1, 2.$$

It is easily shown that $\alpha_1 \ln(\alpha_1) + \alpha_2 \ln(\alpha_2) - (\alpha_1 + \alpha_2 - 1)\ln(\alpha_1 + \alpha_2 - 1) < 0$, since $\alpha_1, \alpha_2 > 1$. Thus, since $a(s)$ is decreasing, $\frac{\partial h(t,u)}{\partial u}(t, u_0) > 0$. Hence the slope (with respect to u) of $h(t, u)$ at any u satisfying $h(t, u) = 0$ must be positive. It follows that $u(t)$ is unique and that $\phi(t) = \ln\{1 + a(u(t))\}$ is well-defined. Since $u(t) \in [0, \Lambda_1(t) + \Lambda_2(t)]$, as $t \to 0$, $u(t) \to 0$ and hence $\phi(0) = \ln(1 + a(0)) = \ln(\theta^*(0, 0))$. This completes the proof of Proposition 1 (a).

Proof of Proposition 1 (b)

The identifiability condition required is that, for any $u, t_1, t_2 > 0$ and $a(s)$ strictly monotone, $h(t_1, u) = h(t_2, u) = 0$ implies $t_1 = t_2$.

For a given value t_0 of t, we have $h(0, u(t_0)) = h(t_0, u(t_0)) = 0$. Also, for $x \in [0, u(t_0)]$ the function $f(x, u(t_0))$ is subadditive, so for t in the neighbourhood of zero such that $\Lambda_1(t) + \Lambda_2(t) \leq u(t_0)$, $h(t, u(t_0)) > 0$. Thus, provided that $t = t_0$ is the only positive value such that $h(t, u(t_0)) = 0$, then

$$\frac{\partial h(t,u)}{\partial t}(t_0, u(t_0)) \leq 0 \ .$$

By the Implicit Function Theorem, it then follows that

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = \Big(-\frac{\partial h(t,u)}{\partial u}(t, u(t))\Big)^{-1} \frac{\partial h(t,u)}{\partial t}(t, u(t)) \geq 0$$

and hence $u(t)$ is non-decreasing with t. This implies that $\phi(t)$ is monotone.
Suppose now that $\Lambda_1(t)$ and $\Lambda_2(t)$ are unbounded, but that $u(t)$ is bounded, that is, $u(t) \to u_0$ as $t \to \infty$. We necessarily have $a(u_0) > 0$. Thus, for large t, $a(u_0) - a(u(t)) = o(1)$ in t and so

$$h(t, u_0) = h(t, u(t)) + o(1) = o(1).$$

Thus, as $t \to \infty$, the frailty tends to the gamma with $a(t) = a_0$. It follows that if $\ln(\theta^*(t,t))$ tends to ∞ or 0, then $u(t)$ cannot be bounded and hence $u(t) \to \infty$, and so $\phi(t)$ tends to the same limit.
Suppose now that $\ln(\theta^*(t,t)) \to c$ where $0 < c < \infty$. If $u(t) \to \infty$ as $t \to \infty$, then $\phi(t)$ also tends to c. If, on the other hand, $u(t) \to u_0 < \infty$, the values $\theta = a(u_0)$ and $\theta = c$ both asymptotically (as $t \to \infty$) satisfy the relation

$$\exp\Big(\int_0^{\Lambda_1(t)} \frac{\theta}{1+A(s)}\mathrm{d}s\Big) + \exp\Big(\int_0^{\Lambda_2(t)} \frac{\theta}{1+A(s)}\mathrm{d}s\Big) - 1 = \exp\Big(\int_0^{\Lambda_1(t)+\Lambda_2(t)} \frac{\theta}{1+A(s)}\mathrm{d}s\Big).$$

It follows that $a(u_0) = c$ so that $\phi(t)$ again tends to the same limit as $\ln(\theta^*(t,t))$. This finishes the proof of Proposition 1 (b).

Asymptotic standard error for $\hat{\phi}$

Suppose that x is a monitoring time. Define

$$f(\phi, \pi_{00}(x), S_1(x), S_2(x)) = \Big(S_1(x)^{1-e^\phi} + S_2(x)^{1-e^\phi} - 1\Big)^{1/(1-e^\phi)} - \pi_{00}(x) \ .$$

Thus $\phi(x)$ is defined implicitly by the equation

$$f(\phi(x), \pi_{00}(x), S_1(x), S_2(x)) = 0 \ .$$

Note that x plays no role in these relationships. In some of what follows, we suppress reference to x and regard f as a function of the four variables ϕ, S_1, S_2 and π_{00}.
A Taylor series representation of

$$f(\phi, \pi_{00}, S_1, S_2) = \Big(S_1^{1-e^\phi} + S_2^{1-e^\phi} - 1\Big)^{\frac{1}{1-e^\phi}} - \pi_{00} \tag{24}$$

about $(\phi^0, \pi_{00}^0, S_1^0, S_2^0)$ is

$$\begin{aligned} f(\phi, \pi_{00}, S_1, S_2) &= f(\phi^0, \pi_{00}^0, S_1^0, S_2^0) + (\phi - \phi^0)\frac{\partial f}{\partial \phi}\bigg|_0 + (\pi_{00} - \pi_{00}^0) + \frac{\partial f}{\partial \pi_{00}}\bigg|_0 \\ &\quad + (S_1 - S_1^0)\frac{\partial f}{\partial S_1}\bigg|_0 + (S_2 - S_2^0)\frac{\partial f}{\partial S_2}\bigg|_0 + \cdots \ . \end{aligned} \tag{25}$$

Truncating (25) after terms of first-order and setting $f \equiv 0$ gives the approximation:

$$(\phi - \phi^0)^2 \simeq \frac{1}{\left(\frac{\partial f}{\partial \phi}\big|_0\right)^2}\left((\pi_{00} - \pi_{00}^0)\frac{\partial f}{\partial \pi_{00}}\bigg|_0 + (S_1 - S_1^0)\frac{\partial f}{\partial S_1}\bigg|_0 + (S_2 - S_2^0)\frac{\partial f}{\partial S_2}\bigg|_0\right)^2 \ . \tag{26}$$

It follows from (26) that the asymptotic standard error of $\hat{\phi}$ is

$$\begin{aligned} \sigma(\hat{\phi}) &= \frac{1}{\frac{\partial f}{\partial \phi}\big|_0}\Bigg\{\mathrm{Var}(\pi_{00})\left[\frac{\partial f}{\partial \pi_{00}}\bigg|_0\right]^2 + \mathrm{Var}(S_1)\left[\frac{\partial f}{\partial S_1}\bigg|_0\right]^2 + \mathrm{Var}(S_2)\left[\frac{\partial f}{\partial S_2}\bigg|_0\right]^2 \\ &+ 2\mathrm{Cov}(\pi_{00}, S_1)\frac{\partial f}{\partial \pi_{00}}\bigg|_0\frac{\partial f}{\partial S_1}\bigg|_0 + 2\mathrm{Cov}(\pi_{00}, S_2)\frac{\partial f}{\partial \pi_{00}}\bigg|_0\frac{\partial f}{\partial S_2}\bigg|_0 \\ &+ 2\mathrm{Cov}(S_1, S_2)\frac{\partial f}{\partial S_1}\bigg|_0\frac{\partial f}{\partial S_2}\bigg|_0\Bigg\}^{1/2} \ . \end{aligned} \tag{27}$$

The variances, covariances and derivatives on the right hand side of (27) are

$$\mathrm{Var}(\pi_{00}) = \frac{1}{n}\pi_{00}(1 - \pi_{00}) \ , \tag{28}$$

$$\mathrm{Var}(S_1) = \frac{1}{n}\pi_{01}(1 - \pi_{01}) + \frac{1}{n}\pi_{00}(1 - \pi_{00}) - \frac{1}{n}\pi_{00}\pi_{01} \ , \tag{29}$$

$$\mathrm{Var}(S_2) = \frac{1}{n}\pi_{10}(1 - \pi_{10}) + \frac{1}{n}\pi_{00}(1 - \pi_{00}) - \frac{1}{n}\pi_{00}\pi_{10} \ , \tag{30}$$

$$\mathrm{Cov}(\pi_{00}, S_1) = \frac{1}{n}\pi_{00}(1 - S_1) \ , \tag{31}$$

$$\mathrm{Cov}(\pi_{00}, S_2) = \frac{1}{n}\pi_{00}(1 - S_2) \ , \tag{32}$$

$$\mathrm{Cov}(S_1, S_2) = \frac{1}{n}\pi_{00}(1 - \pi_{00}) - \frac{1}{n}\pi_{00}\pi_{10} - \frac{1}{n}\pi_{00}\pi_{01} - \frac{1}{n}\pi_{01}\pi_{10} \tag{33}$$

and

$$\frac{\partial f}{\partial \pi_{00}} = -1 \ , \tag{34}$$

$$\frac{\partial f}{\partial S_1} = B(\phi)^{\frac{e^\phi}{1-e^\phi}} S_1^{-e^\phi} \ , \tag{35}$$

$$\frac{\partial f}{\partial S_2} = B(\phi)^{\frac{e^\phi}{1-e^\phi}} S_2^{-e^\phi} \ , \tag{36}$$

$$\frac{\partial f}{\partial \phi} = B(\phi)^{C(\phi)} C'(\phi) \ln B(\phi) - C(\phi) B'(\phi) B(\phi)^{C(\phi)-1} \ , \tag{37}$$

where $B(\phi)$, $B'(\phi)$, $C(\phi)$ and $C'(\phi)$ are

$$B(\phi) = S_1^{1-e^\phi} + S_2^{1-e^\phi} - 1 \ , \tag{38}$$

$$B'(\phi) = -e^\phi\Big(\ln(S_1)S_1^{1-e^\phi} + \ln(S_2)S_2^{1-e^\phi}\Big) \ , \tag{39}$$

$$C(\phi) = (1-e^\phi)^{-1} \ , \tag{40}$$

$$C'(\phi) = e^\phi/(1-e^\phi)^2 \ . \tag{41}$$

In case $\phi = 0$, the following limits are used in (27) instead of (35)–(37):

$$\lim_{\phi\to 0}\frac{\partial f}{\partial S_1} = S_2 \ , \tag{42}$$

$$\lim_{\phi\to 0}\frac{\partial f}{\partial S_2} = S_1 \tag{43}$$

and

$$\lim_{\phi\to 0}\frac{\partial f}{\partial \phi} = S_1 S_2 \ln S_1 \ln S_2 \ , \tag{44}$$

by applying l'Hospital's rule.

Simulation study

We describe here a brief simulation study. First, 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 $\phi(x)$ 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 5000 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. Estimates of the association measure along with its standard errors are obtained for the 5000 replications using the procedure described in Subsection 3.2.

Figure 8 shows $\phi(x)$ along with the arithmetic mean of the $\hat{\phi}(x)$ for sample size $n_x = 50$ and $n_x = 400$. For $n_x = 50$, there is a slight upward bias for the gamma frailty (panel (i)), virtually no bias for the inverse gaussian (panel (ii)), and a slight downward bias for the compound Poisson (panel (iii)). For $x = 20$, Table 2 shows the bias of $\hat{\phi}$, its mean

standard error (s.e.), the empirical s.e. obtained as the standard deviation of the 5000 simulated values, and the coverage probability of the 95% confidence intervals (P95), that is, the proportion of the 5000 confidence intervals containing $\phi(20)$. The relative bias is less than 5% even for $n_x = 50$. The empirical standard error matches the mean of the asymptotic values. The coverage probabilities are close to the nominal 0.95, except for the compound Poisson model with $n_x = 50$. As one would expect, as the sample size increases the bias and variance of $\hat{\phi}(20)$ decreases, the only exception being the compound Poisson model. The results for the latter model are heavily influenced by the remedy of adding 0.5 to all counts if one of the values in the 4-tuple of observations is zero (cf. Section 6).

Table 2: Bias and variance of $\hat{\phi}$ for three shared frailty models and four different sample sizes evaluated at $x = 20$.

Frailty model	$\phi(20)$		$n_x = 50$	$n_x = 100$	$n_x = 200$	$n_x = 400$
$Z \sim \Gamma(0.1, 10)$	2.3979	bias	0.1041	0.0725	0.0391	0.0192
		mean s.e.	0.5010	0.3698	0.2547	0.1772
		empirical s.e.	0.4965	0.3858	0.2605	0.1775
		P95	0.9100	0.9556	0.9546	0.9552
$Z \sim InvG(1, 0.1)$	0.9637	bias	0.0486	0.0140	0.0097	0.0054
		mean s.e.	0.4083	0.2798	0.1953	0.1372
		empirical s.e.	0.4311	0.2870	0.1978	0.1383
		P95	0.9530	0.9508	0.9498	0.9520
$Z \sim CP(1, 10, 1.5)$	3.9826	bias	-0.1466	0.0849	0.1185	0.0681
		mean s.e.	0.4037	0.4620	0.4028	0.2848
		empirical s.e.	0.4222	0.4471	0.4228	0.3040
		P95	0.7718	0.9398	0.9236	0.9572

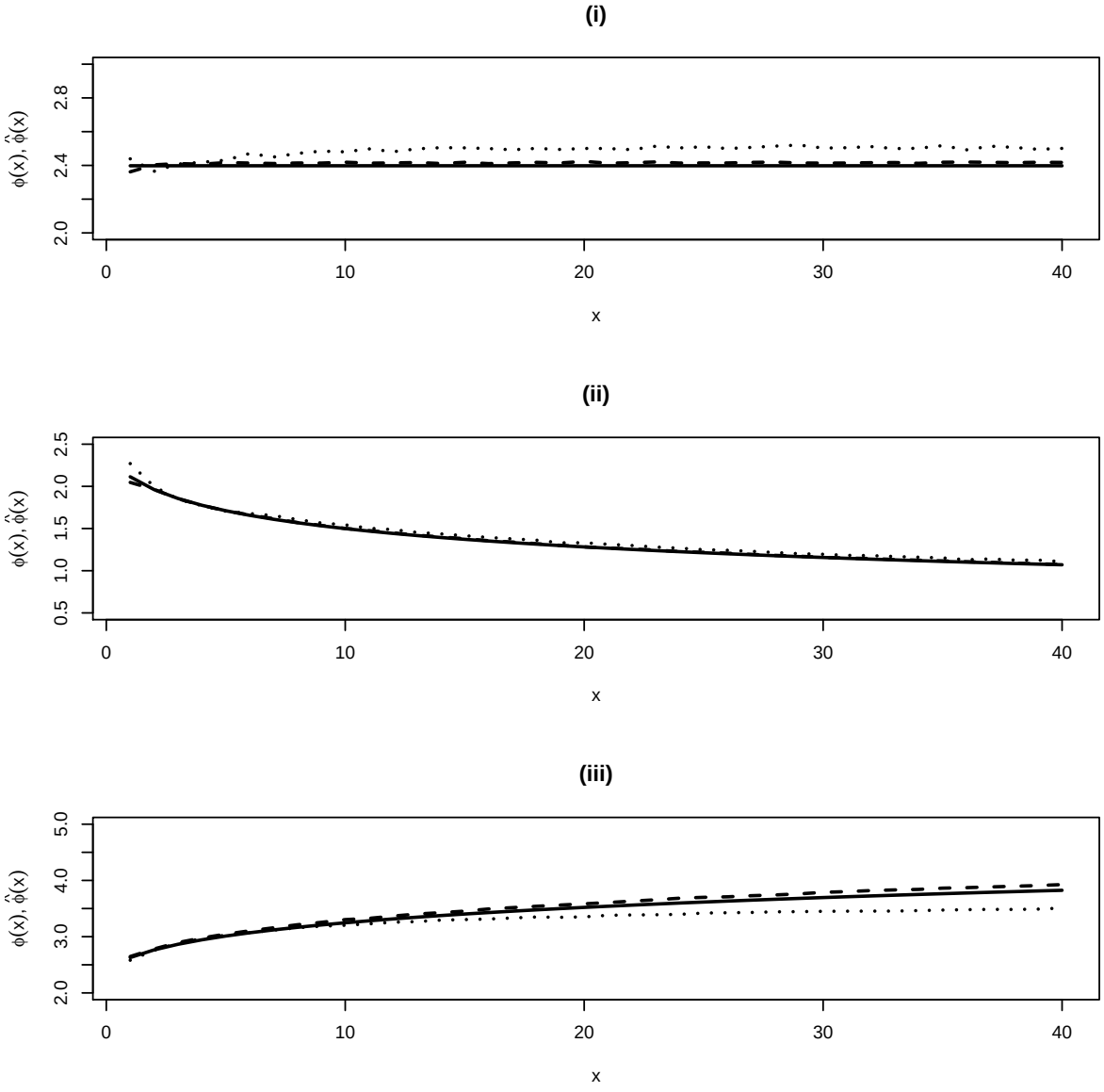

Figure 8: $\phi(x)$ (solid line) and (mean of) $\hat{\phi}(x)$ for $n_x = 50$ (dotted line) and $n_x = 400$ (dashed line) for the Gamma (i), inverse Gaussian (ii) and compound Poisson (iii) frailty model.