Published in final edited form as:
Biometrics. 2010 December ; 66(4): 1012–1023. doi:10.1111/j.1541-0420.2009.01372.x.

Modelling familial association of ages at onset of disease in the presence of competing risk

Joanna H. Shih[1,*] and **Paul S. Albert**[2,**]

[1]Biometric Research Branch, National Cancer Institute, 6130 Executive Boulevard, Rm 8132, Rockville MD, 20852

[2]Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Boulevard, Rockville MD, 20852, U.S.A

Abstract

In genetic family studies, ages at onset of diseases are routinely collected. Often one is interested in assessing the familial association of ages at onset of a certain disease type. However, when a competing risk is present and is related to the disease of interest, the usual measure of association by treating the competing event as an independent censoring event is biased. We propose a bivariate model that incorporates two types of association: one is between the first event time of paired members, and the other is between the failure types given the first event time. We consider flexible measures for both types of association, and estimate the corresponding association parameters by adopting the two-stage estimation of Shih and Louis (1995) and Nan et al. (2006). The proposed method is illustrated using the kinship data from the Washington Ashkenazi Study.

Keywords

cause-specific cross-ratio; competing risk; familial association; odds-ratio

1. Introduction

In family studies, often one is interested in assessing the familial association in diseases with ages at onset. When only one disease type is considered (e.g. cancer), there exists an extensive literature for the analysis of correlated failure time data. These methods can be used to model and estimate the association between ages at onset of family members (Hougaard, 2001). For example, Shih and Louis (1995) proposed a two-stage estimation procedure for estimating the association of paired failure times. Hougaard et al. (1992) proposed bivariate survival models which were used to measure the similarities between the lifetimes of adult Danish twins. Li et al. (1998) proposed a parametric likelihood approach to study the familial aggregation in lung cancer risk from case-control family studies. When the competing risk is present, all the above mentioned methods censor the failure of interest at the time of the competing event and proceed with the proposed modeling and analysis of correlated failure time data. Such censoring implicitly assumes that the competing events are independent and that the estimate of the marginal distribution for the event of interest (e.g. the Kaplan-Meier estimate) is consistent (Kalbfleish and Prentice, 2002). Bandeen-Roche and Liang (2002), Chatterjee et al. (2003), and Bandeen-Roche and Ning (2008), showed that censored observations due to a competing risk may affect the estimation of association

[*]jshih@mail.nih.gov. [**]albertp@mail.nih.gov.

of correlated failure times of the event of interest. It is well recognized that if the competing events are dependent, the marginal distribution of the failure time for the event of interest is not identifiable, because only the failure time of the first event is observable. Instead of the marginal distribution, the univariate cause-specific hazard (Prentice et al.,1978) and incidence function are observable quantities, and have been used in making cause-specific inference (Kalbfleish and Prentice, 2002). Bandeen-Roche and Liang (2002) extended the concept of univariate cause-specific hazard to a bivariate setting. They proposed a cause-specific cross-ratio which accommodates competing risks in measuring the association of paired failure times. Recently Bandeen-Roche and Ning (2008) proposed a nonparametric estimation method to estimate the constant cause-specific cross-ratio. Cheng and Fine (2008) gave an alternative representation of the cause-specific cross-ratio, and proposed a simple plug-in nonparametric estimate.

In this article, we are interested in assessing the familial association of disease in the presence of competing risk of the kinship data collected from the Washington Ashkenazi Study (Struewing et al., 1997). In this study, more than 5000 volunteer Ashkenazi Jews living in Washington D.C. provided blood samples for genotyping of BRCA1/BRCA2 mutations. They also gave family history information on cancers and mortality. One primary interest of the study was to estimate the breast cancer risk among the carriers and non-carriers of the gene mutations. For our application, we consider female first-degree relatives of the volunteer participants, and are interested in studying their familial association of ages at onset of cancer and non-cancer mortality. Here cancer and non-cancer mortality are competing risks. To this end, we propose a bivariate survival model for correlated failure times of relatives which incorporates competing risks. The proposed model is decomposed into two parts: one based on the time to first event, and the other is based on the event type. For the former, we generalize the bivariate survival model of Clayton (1978) to one which allows for piecewise constant cross ratios. For the latter, we generalize the bivariate logistic model to one with piecewise constant odds ratios. Different from the methods of Bandeen-Roche and Ning (2008) and Cheng and Fine (2008) which treat each bivariate failure type one at a time, we incorporate all possible bivariate failure types, namely both cancers, both non-cancer deaths, and one cancer and one non-cancer death, to model the association related to failure types. In addition, since we separately model the association of times to first event and failure types, we are able to assess the effects of these two types of associations on the cause-specific cross-ratio. One important feature of our model is that even though both cross-ratio and odds-ratio are piecewise constant, the cause-specific cross-ratio generally is non-piecewise constant.

The remainder of the paper is organized as follows. In Section 2 we propose bivariate survival models for the association of correlated failure times and in the presence of competing risks. In Section 3, we propose a quasi-likelihood estimation procedure for familial association. In Section 4 and 5, we use simulations and the WAS study data to illustrate the proposed methodology. A discussion follows in Section 6.

2. Model

2.1 Setup

Consider two competing events. Under the usual assumption of competing risks, for any given subject, only the first event is observable. For each individual, in the absence of censoring, the observable data is time to the first event and failure type. For a pair of individuals, $j = 1, 2$, let T_j define the time to the first event for individual j and Y_j the failure type, where $Y_j = 1$ if failure type one is observed, and 2 if failure type 2 is observed. In terms of our WAS study application, cancer is failure type one and non-cancer mortality is failure type two. We first model the distribution of (T_j, Y_j) for each pair member $j = 1, 2$. Let $S_j(t)$

and $f_j(t)$ denote the marginal survival function and probability density function of T_j. Let $f_j^k(t)=\lim_{\Delta t\to 0}\Delta t^{-1}Pr(t \le T_j<t+\Delta t, Y_j=k)$ denote the sub-density function for failure type k. The cause-specific hazard (Prentice et al., 1978), defined as

$$\lambda_j^k(t)=\lim_{\Delta t\to 0}\Delta t^{-1}Pr(t \le T_j<t+\Delta t, Y_j=k|T_j \ge t)$$

equals $f_j^k(t)/S_j(t)$. Then $f_j^k(t)$ equals $f_j(t)\lambda_j^k(t)/\lambda_j(t)$, where $\lambda_j(t)=\sum_k \lambda_j^k(t)$ is the hazard function of T_j. In the case of only one event type, $\lambda_j(t)=\lambda_j^1(t)$. For convenience, we write $p_j^k(t)$ for $\lambda_j^k(t)/\lambda_j(t)$ to denote the probability of the event being type k given the failure time at t for pair member j.

2.2 Association

The next stage of the model involves specifying a dependency structure between the members of the pair. We consider two types of association: one is between the first event times (T_1 and T_2), and the other is between the failure types (Y_1 and Y_2) given (T_1, T_2). For the measure of association between T_1 and T_2, we consider the cross ratio (Oakes, 1989) defined by

$$\begin{aligned}\theta(t_1,t_2) &= \frac{\lambda_1(t_1|T_2=t_2)}{\lambda_1(t_1|T_2>t_2)} \\ &= \frac{S(t_1,t_2)f(t_1,t_2)}{\frac{\partial}{\partial s_1}S(s_1,t_1)\,|_{s_1=t_1}\frac{\partial}{\partial s_2}S(t_1,s_2)\,|_{s_2=t_2}},\end{aligned} \quad (1)$$

where S is the joint survival function of T_1 and T_2, and f is the joint density function. Most research has imposed models or constraints on $\theta(t_1, t_2)$ such as $\theta(t_1, t_2)$ is constant or S belongs to the archimedean copulas (Oakes, 1989). Nan et al. (2006) proposed a piecewise constant cross-ratio model which allows $\theta(t_1, t_2)$ to be piecewise constant in one dimension of time. For our application, since each pair consists of first-degree relatives to each other and their event times are paired in an arbitrary order, it is more reasonable to allow the cross ratios to vary in both time dimensions. Therefore, in this work, we model the cross-ratio as a piecewise constant function of the event times of both members in the pair. Specifically the cross-ratio is modelled as $\theta(t_1,t_2)=\sum_{i=1}^{K}\sum_{j=1}^{K}\theta_{ij}I(w_{i-1} \le t_1<w_i)I(w_{j-1} \le t_2<w_j)$, where $0 = w_0 < w_1 < \cdots < w_K$ are a set of pre-specified knots in the appropriate age range of interest. Extending the work of Nan et al. (2006) which assumes θ changes only in one dimension of time, it follows that under the assumption that $\theta(t_1, t_2) = \theta_{ij}$ for $(t_1, t_2) \in A_{ij} = [w_{i-1},w_i) \times [w_{j-1},w_j)$, the bivariate survival function is given by

$$S_{A_{ij}}(t_1,t_2)=[S_{A_{ij}}(t_1,w_{j-1})^{-(\theta_{ij}-1)}+S_{A_{ij}}(w_{i-1},t_2)^{-(\theta_{ij}-1)}-S_{A_{ij}}(w_{i-1},w_{j-1})^{-(\theta_{ij}-1)}]^{-1/(\theta_{ij}-1)}. \quad (2)$$

When $\theta(t) \equiv \theta$ for all $t_1 > 0$, $t_2 > 0$, the above bivariate survival function is equivalent to the Clayton model (1978). When $\theta(t_1, t_2) \equiv 1$, T_1 and T_2 are independent, and $S(t_1, t_2) = S_1(t_1)S_2(t_2)$. In the appendix, we show the bivariate survival function S is determined by the marginal survival functions S_1, S_2 and θ_{ij}'s.

For the measure of association for failure types Y_1 and Y_2 given the event times T_1, T_2, we consider the odds ratio defined by

$$\phi(t_1, t_2)=\frac{p^{22}(t_1, t_2)p^{11}(t_1, t_2)}{p^{12}(t_1, t_2)p^{21}(t_1, t_2)},$$

where $p^{k_1k_2}(t_1, t_2) = Pr(Y_1 = k_1, Y_2 = k_2|T_1 = t_1, T_2 = t_2) = f^{k_1k_2}(t_1, t_2)/f(t_1, t_2)$ with $f^{k_1k_2}(t_1, t_2) = f(t_1, t_2, Y_1 = k_1, Y_2 = k_2)$.

Let $p_j^k(t_1, t_2)$ denote the conditional probability $Pr(Y_j = k \mid T_1 = t_1, T_2 = t_2)$, $j = 1, 2$, $k = 1, 2$. The joint probability of the failure types, $p^{k_1k_2}(t_1, t_2)$, is determined by $p_j^{k_j}(t_1, t_2)$, $j=1, 2$ and ϕ, and given by

$$p^{k_1k_2}(t_1, t_2;\phi)=\begin{cases} [h^{k_1k_2}(t_1, t_2) - \{h^{k_1k_2}(t_1, t_2)^2 - 4\phi(\phi - 1)(p_1^{k_1}(t_1, t_2)p_2^{k_2}(t_1, t_2))\}^{0.5}]/(2(\phi - 1)) \\ \text{if } \phi \neq 1 \\ p_1^{k_1}(t_1, t_2)p_2^{k_2}(t_1, t_2) \text{ otherwise,} \end{cases} \quad (3)$$

where $h^{k_1k_2}(t_1, t_2)=1 - (1 - \phi)\{p_1^{k_1}(t_1, t_2)+p_2^{k_2}(t_1, t_2)\}$. In order to allow for flexibility, we will model the odds ratio as a piecewise constant function of the event times, expressed by

$$\phi(t_1, t_2)=\sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{ij}I(w_{i-1} \leq t_1<w_i)I(w_{j-1} \leq t_2<w_j),$$

The cause-specific bivariate distribution of event times can be characterized with the above marginal distributions and two types of association θ and ϕ. Specifically, we can combine these two types of association to derive cause-specific time-dependent measures of association of practical interest. It may be more meaningful to evaluate the association with the cause-specific cross-ratio (Bandeen-Roche and Liang, 2002), since the above cross-ratios θ_{ij}s are associated with times to first failure and provide little practical interest. For a given (t_1, t_2), the cause-specific cross-ratio is defined by

$$\theta^{k_1k_2}(t_1, t_2)=\frac{\lambda_1^{k_1}(t_1 \mid T_2=t_2, Y_2=k_2)}{\lambda_1^{k_1}(t_1 \mid T_2>t_2)},$$

where $\lambda_1^{k_1}(t_1 \mid T_2=t_2, Y_2=k_2)$ is the conditional hazard at time t_1 with failure type k_1 given the other pair member has an event at time t_2 with failure type k_2 and $\lambda_1^{k_1}(t_1 \mid T_2>t_2)$ is the conditional hazard at time t_1 with failure type k_1 given the other pair member has survived time t_2. It is straightforward to show that $\theta^{k_1k_2}(t_1, t_2)$ can be alternatively represented by

$$\theta^{k_1k_2}(t_1, t_2)=\theta(t_1, t_2)\frac{p^{k_1,k_2}(t_1, t_2)}{Pr(Y_1=k_1 \mid T_1=t_1, T_2> t_2)Pr(Y_2=k_2 \mid T_2=t_2, T_1> t_1)}. \quad (4)$$

Thus $\theta^{k_1k_2}(t)$ is decomposed into the product of two terms: the first term is the overall cross-ratio, and the second term provides additional contribution to the dependency of cause-specific event times arising from the dependency of failure types between correlated members. Note that according to (4), even if $\theta(t_1, t_2)$ and $\phi(t_1, t_2)$ are both piecewise constant, $\theta^{k_1k_2}(t_1, t_2)$ may not be.

Although the cause-specific cross-ratio is a meaningful association measure, it is expressed as an instantaneous odds-ratio which may be difficult to interpret for practitioners. We consider another function, namely the conditional cumulative incidence defined as $Pr(T_1 <= t_1, Y_1 = k_1 \mid a \le T_2 < b, Y_2 = k_2)$, which is simpler to interpret, since it simply measures the cumulative cause-specific incidence of one pair member given the other member's failure time and failure type. The impact of the other pair member's failure time and failure type on the cumulative incidence can be assessed by comparing the conditional cumulative incidence with its marginal cumulative incidence counterpart.

3. Estimation

Let T_{ij} and C_{ij} denote the failure and censoring times to the first event for the jth member of the ith family, $j = 1, \cdots, m_i$ $i = 1, \cdots, n$. Assume T_{ij} and C_{ij} are independent and let $X_{ij} = \min(T_{ij}, C_{ij})$ denote the observed failure time and Y_{ij} the failure type, where $Y_{ij} = k$, $k = 1, 2$, if failure type k is observed, and 0, if the failure time is censored. We consider a setting suitable for the application of the WAS study. We assume that individuals within a family have a common marginal distribution. Thus we suppress the subscript denoting pair membership to simplify the notation presented in the previous sections. Specifically let λ denote the overall hazard, f^k and λ^k the sub-density and cause-specific hazard for failure time of failure type, and $p^k(t)$ the probability of failure type given the failure time at t, $k = 1, 2$. We assume that the joint distribution of the failure times and failure types for any pair of individuals in the same family follows the model proposed in the previous section. Assumptions on third or higher order dependency structures are left unspecified. Specifically, we formulate the composite likelihood as in Chatterjee et al. (2006), where number of observations in each family were broken into doublets and each doublet was treated independent of the others, ignoring possible dependence between doublets of the same family.

3.1 Estimating θ_{ij}s

We first consider estimating the piecewise constant cross-ratios θ_{ij}s by adopting the two-stage estimation method of Shih and Louis (1995) and sequential two-stage method of Nan et al. (2006). For estimation of each θ_{lm}, we use the subset D_{lm} of paired data $\{X_{ip}, X_{iq}, Y_{ip}, Y_{iq}\}$, $p, q = 1, \cdots, m_i, j \neq k$, $i = 1, \cdots, n$ in the study cohort such that $X_{ip} \ge w_{l-1}$ and $X_{iq} \ge w_{m-1}$. For each observation in D_{lm}, we right censor the observed failure time at the upper bound of the region A_{lm} as $\tilde{X}_{ip} = \min(X_{ip}, w_l)$, $\tilde{\delta}_{ip} = I[w_{l-1} \le X_{ip} < w_l, Y_{ip} > 0]$ for the first pair member and $\tilde{X}_{iq} = \min(X_{iq}, w_m)$, $\tilde{\delta}_{iq} = I[w_{m-1} \le X_{iq} < w_m, Y_{iq} > 0]$ for the second pair member. In our application of the WAS data, since the bivariate survival function is symmetric, θ_{ij} is symmetric, i.e. $\theta_{ij} = \theta_{ij}$, and hence for $l \neq m$ we use both data sets D_{lm} and D_{ml} to estimate θ_{lm}. We estimate each θ_{ij} in the same order as laid out in the appendix for computing the bivariate survival function. That is, we begin with θ_{11} and proceed with the estimation for θ_{1j}, $j = 2, \cdots, K$ sequentially. After all the θ_{1i} 's are estimated, we proceed with the estimation for θ_{22} followed by θ_{2i}, $k = 3, \cdots, K$. We continue the process until the last cross-ratio parameter θ_{KK} is estimated.

In the following, we describe how to estimate θ_{11}. At the first stage we estimate the common marginal survival function by the Nelson estimate ignoring dependency between failure times in the same family. The Nelson estimate specifies the probabilities for the bottom boundaries of A_{1j}, $j = 1, \cdots, K$ and left boundaries of A_{k1}, $k = 1, \cdots, K$. At the second stage, we treat the estimated probabilities at the bottom and left boundaries of A_{11} as known, and use data set D_{11} to obtain the estimate of θ_{11} which maximizes the following composite-likelihood with $l = 1$, $m = 1$

$$L(\theta_{lm})= \prod_{i=1}^{n} \prod_{(ip,iq)\in \mathscr{D}_{lm}\cup \mathscr{D}_{ml}}^{M_i} C(u_{ip}^{lm}, v_{iq}^{lm};\theta_{lm})^{(1-\tilde{\delta}_{ip})(1-\tilde{\delta}_{iq})} \left\{ \frac{-\partial C(u_{ip}^{lm}, v_{iq}^{lm};\theta_{lm})}{\partial u_{ip}} \right\}^{\tilde{\delta}_{ip}(1-\tilde{\delta}_{iq})} \times \left\{ \frac{-\partial C(u_{ip}^{lm}, v_{iq}^{lm};\theta_{lm})}{\partial u_{iq}} \right\}^{\tilde{\delta}_{iq}(1-\tilde{\delta}_{ip})} \left\{ \frac{-\partial^2 C(u_{ip}^{lm}, v_{iq}^{lm};\theta_{lm})}{\partial u_{ip}\partial u_{iq}} \right\}^{\tilde{\delta}_{ip}\tilde{\delta}_{iq}}, \quad (5)$$

where $\mathscr{D}_{ip}$ is the collection of pairs in data set D_{ip}, M_i is the number of doublets in the *ith* family, $C(u, v; \theta) = \{u^{-(\theta-1)} + v^{-(\theta-1)} - 1\}^{-1/(\theta-1)}$, $u_{ip}^{lm}=\widehat{S}_{A_{lm}}(\widehat{X}_{ip}, w_{m-1})/\widehat{S}_{A_{lm}}(w_{l-1}, w_{m-1})$, $v_{ip}^{lm}=\widehat{S}_{A_{lm}}(w_{l-1}, \widehat{X}_{ip})/\widehat{S}_{A_{lm}}(w_{l-1}, w_{m-1})$.. The function $C(u^{lm}, v_{lm}; \theta^{lm})$ is the conditional bivariate survival function given the pair having survived time (w_{l-1}, w_{m-1}). Note that for $l = m = 1$, $u_{ip}^{lm}=\widehat{S}_{A_{11}}(X_{ip}, 0)/\widehat{S}_{A_{11}}(0, 0)=\widehat{S}_{A_{11}}(X_{ip}, 0)$ and $v_{ip}^{lm}=\widehat{S}_{A_{11}}(0, X_{ip})/\widehat{S}_{A_{11}}(0, 0)=\widehat{S}_{A_{11}}(0, X_{ip})$ are the non-parametric Nelson's estimate of the univariate survival function. After the estimate of θ_{11} is obtained, $\hat{S}_{A11}(t_1, t_2)$ is readily calculated, and so are the survival probabilities on the upper and right boundaries of A_{11}. The probabilities on these boundaries in turn specify the survival probability $\hat{S}_{A12}(t_1, w_1)$ on the bottom boundary of A_{12} and $\hat{S}_{A21}(w_1, t_2)$ for the left boundary of A_{21}. Plugging these estimates along with $\hat{S}_{A12}(0, t_2)$ and $\hat{S}_{A12}(t_1, 0)$ into $L(\theta_{12})$, and maximizing $L(\theta_{12})$ with respect to θ_{12}, we obtain the quasi-MLE of θ_{12}. We iterate the above estimation until the last cross-ratio θ_{KK} is estimated.

Assume that the marginal survival function is continuous and $Pr(w_{i-1} \le T_1 < w_i, w_{j-1} \le T_2 < w_j) > 0$, for all $i = 1, \cdots, K, j = 1, \cdots, K$. The non-parametric Nelson's estimate of the marginal survival function is consistent in the support $t \in (0, \nu)$, where ν denotes the maximal follow-up time. Following Shih and Louis (1995), $\hat{\theta}_{11}$ is consistent and normally distributed with mean θ_{11} as the sample size $n \to \infty$. Then one can establish the consistency of the estimates of the survival functions in the upper and lower boundary in each grid element A_{ij}. Hence similar to the sequential estimation approach of Nan et al. (2006), as the number of pairs used in estimating θ_{ij} goes to ∞ proportionally as the sample size $n \to \infty$, $\hat{\theta}_{ij}$ converges to the true value θ_{ij} and $\sqrt{n}(\widehat{\theta_{ij}} - \theta_{ij})$ converges weakly to a mean 0 and normally distributed random variable under regularity conditions and model (2).

3.2 Estimating φ_{ij}s

To estimate the association of failure types between paired members, only paired data with failure observed in both members are used, because pairs with censored observations contain no information about the correlation. Specifically, for the estimation of each ϕ_{lm}, $l = 1, \cdots, K, m = 1, \cdots, K$, we use the subset $\bar{D}_{lm}$ of paired data $\{X_{ip}, X_{iq}, Y_{ip}, Y_{iq}\}$, $p, q = 1, \cdots, m_i, j \neq k, i = 1, \cdots, n$ in the study cohort such that $w_{l-1} \le X_{ip} < w_l, Y_{ip} > 0$ and $w_{m-1} \le X_{iq} < w_m, Y_{iq} > 0$. That is, the contribution to the estimation comes from a subset of the paired data where failures are observed in both members and occur in the region where ϕ_{lm} is defined. Similar to the set up of θ_{ij}'s, ϕ_{ij} is symmetric and hence for $l \neq m$ we use both $\bar{D}_{lm}$ and $\bar{D}_{ml}$ to estimate ϕ_{lm}. The estimate of ϕ_{lm}, denoted by $\hat{\phi}_{lm}$, is obtained by maximizing the following composite-likelihood

$$L(\phi_{lm})= \prod_{i=1}^{n} \prod_{(ip,iq)\in \bar{\mathscr{D}}_{lm}\cup \bar{\mathscr{D}}_{ml}}^{M_i} \tilde{p}^{11}(X_{ip}, X_{iq};\phi_{lm})^{(2-y_{ip})(2-y_{iq})} \tilde{p}^{12}(X_{ip}, X_{iq};\phi_{lm})^{(2-y_{ip})(y_{iq}-1)} \times \tilde{p}^{21}(X_{ip}, X_{iq};\phi_{lm})^{(y_{ip}-1)(2-y_{iq})} \tilde{p}^{22}(X_{ip}, X_{iq};\phi_{lm})^{(y_{ip}-1)(y_{iq}-1)}, \quad (6)$$

where $\mathcal{I}_{ij}$ is the collection of pairs in data set D_{ij}, and ˜ over $p^{k_1k_2}$ indicates that the probability of failure type $p_j^k(t_1, t_2)=Pr(Y_j=k \mid T_1=t_1, T_2=t_2)$, $k = 1, 2$ in (3) is substituted by its estimate. The non-parametric estimation technique used to calculate the probability of failure type for the univariate data (Kalbfleisch and Prentice, 2002) can be generalized here to estimate $p_j^k(t_1, t_2)$. Specifically let $(\tilde{t}_{i1}, \tilde{t}_{i2})$, i = 1, ⋯ , m denote the m distinct paired failure times in the data. Let d_i^{jk} denote the number of pairs with failure type j in pair member 1 and failure type k in pair member 2 at time $(\tilde{t}_{1i}, \tilde{t}_{2i})$. Then the non-parametric estimate of $p_1^k(\tilde{t}_{i1}, \tilde{t}_{i2})$ equals $(d_i^{k1}+d_i^{k2})/\sum_j\sum_k d_i^{jk}$, and the non-parametric estimate of $p_2^k(\tilde{t}_{i1}, \tilde{t}_{i2})$ equals $(d_i^{1k}+d_i^{2k})/\sum_j\sum_k d_i^{jk}$. In the case of the paired data being exchangeable as in the WAS data, the above non-parametric estimate is modified as follows. Let d_i^{jk} denote the number of pairs with failure type j in pair member 1 at time $\tilde{t}_{i1}$ and failure type k in pair member 2 at time $\tilde{t}_{i2}$, and $\bar{d}_i^{jk}$ denote the number of pairs with failure type j in pair member 2 at time $\tilde{t}_{i1}$ and failure type k in pair member 1 at time $\tilde{t}_{i2}$. Then the non-parametric estimate of $p_1^k(\tilde{t}_{i1}, \tilde{t}_{i2})$ equals $(d_i^{k1}+d_i^{k2}+\bar{d}_i^{k1}+\bar{d}_i^{k2})/\sum_j\sum_k(d_i^{jk}+\bar{d}_i^{jk})$, and the non-parametric estimate of $p_2^k(\tilde{t}_{i1}, \tilde{t}_{i2})$ equals $(d_i^{1k}+d_i^{2k}+\bar{d}_i^{1k}+\bar{d}_i^{2k})/\sum_j\sum_k(d_i^{jk}+\bar{d}_i^{jk})$. However, this non-parametric estimate may be unstable, because unless the data set is very large, the number of observations at each observed paired failure times is likely few. One can have a more stable estimate by assuming $p_j^k(t_1, t_2)$ to be piecewise constant in the same fashion as θ and ϕ are specified. Then the estimate of $p_j^k(t_1, t_2)$ is calculated according to how many pairs with failures occurring in a specific sub-region defined by the grid points. Alternatively one may assume the event type probability depends on individual's own failure time only, i.e. $p(Y_j = k \mid T_j = t_j, T_{j'} = t_{j'}) = p(Y_j = k \mid T_j = t_j) = p^k(t_j)$. In that case, its non-parametric estimate is readily calculated as $\widehat{p}^k(\tilde{t}_j)=d_j^k/(d_j^1+d_j^2)$ where d_j^k is the number of individuals with failure type k at time $\tilde{t}_j$, and where $\tilde{t}_1, \cdots, t_q$ are q distinct failure times in the data (Kalbfleisch and Prentice, 2002).

In addition to the assumptions required to establish the consistency for the estimate of θ_{ij}, assume that $Pr(Y_1 = k_1, Y_2 = k_2 \mid T_1 = t_1, T_2 = t_2) > 0$, $(t_1, t_2) \in A_{ij}$ for all $i = 1, \cdots, K$, $j = 1, \cdots, K$, and $k_1 = 1, 2$, $k_2 = 1, 2$. Since the non-parametric estimate $\widehat{p}_j^k(t_1, t_2)$ of the probability of failure type being k for pair member j given $T_1 = t_1$, $T_2 = t_2$ is a continuous function of the Nelson-type estimate for the bivariate cause-specific hazard function (Cheng, Fine and Kosorok, 2009) which is consistent for $t_i \in (0, \nu)$, $i = 1, 2$, $\widehat{p}_j^k(t_1, t_2)$ is consistent. Following regularity conditions of asymptotic theory, $\sqrt{n}(\widehat{\phi}_{ij} - \phi_{ij})$ converges weakly to a mean 0 and normally distributed random variable under the piecewise-constant odds-ratio model,

$$\phi(t_1, t_2)=\sum_{i=1}^K\sum_{j=1}^K \phi_{ij} I(w_{i-1} \le t_1 < w_i) I(w_{j-1} \le t_2 < w_j).$$

With the above complex sequential estimation procedure, further work is needed to derive the asymptotic properties of the estimates of θ_{ij}'s and ϕ_{ij}. In this paper, we use the bootstrap approach (Efron and Tibshirani, 1993) with family as the bootstrap sampling units to obtain the variance of these semi-parametric estimates.

After these parameters are estimated, we are able to plug them in (2) to obtain the estimate of the cause-specific cross-ratios. We can also calculate the estimate of the conditional

cumulative incidence function $Pr(T_1 \le t_1, Y_1 = k_1 \mid a \le T_2 < b, Y_2 = k_2)$, $k_j = 1, 2$, $j = 1, 2$ given by

$$\widehat{Pr}(T_1 \le t_1, Y_1=k_1 \mid a \le T_2<b, Y_2=k_2)=\frac{\int_0^{t_1}\int_a^b \tilde{p}^{k_1k_2}(u_1,u_2;\widehat{\phi}(u_1,u2))\widehat{S}_{A_{i_{u_1}i_{u_2}}}(du_1,du_2)}{\int_a^b \widehat{f}^{k_2}(u)du},$$

where $\hat{S}$ is the semi-parametric estimate, the subscript $i_t = l$, if $w_{l-1} \le t < w_l$, and $\hat{f}^k(u) = \hat{S}(u, 0)\hat{\lambda}^k(u)$. Recently Cheng, Fine and Kosorok (2007,2009) proposed a non-parametric estimator of the bivariate cumulative incidence function, which offers an alternative to our semi-parametric estimator. As with the approach of Bandeen-Roche and Ning (2008), they consider one bivariate failure type (e.g. cancer-caner) at a time. The merits and disadvantages of our approach vs. theirs would be an interest for future study.

4. Example

We applied the proposed model and estimation method to the WAS study data. We analyzed a subset of the data in which the first-degree relatives of the female probands are also first-degree relatives to each other. That is, (mother, sister), (daughter, daughter) and (sister, sister) pairs were included in the analysis. The subset was chosen because the pairwise association should be similar among these pairs of first-degree relatives. Cancer and non-cancer mortality are the two competing events considered in this example. Among these women, the majority of the cancer incidences were breast and ovarian tumors. The ages of the relatives at the time of the interview of the proband define the censoring times. The data of 12,255 subjects coming from 4,235 distinct families were used to obtain the Nelson-type nonparametric estimate of the marginal survival of time to the first failure, overall hazard and cause-specific hazard, and the probability of failure type given the individual's failure time. The estimates of the probability of failure type being cancer is displayed in Figure 1. It shows that cancer risk is higher than non-cancer death in mid-age but lower in young and old ages.

Table 1 shows the estimation results for the association parameters. The number of pairs with both members having cancer(d_{11}), number of pairs with one member having cancer and the other member dead of non-cancer (d_{12}), and number of pairs with both members dead of non-cancer (d_{22}) are listed in the second column. The bootstrap standard errors were obtained from 500 bootstrap samples. In estimating the association parameters, to assure there are sufficient number of paired events in each sub-region to calculate the piecewise cross-ratios and odds-ratios, and to choose each region which is biologically meaningful, the number of knots was set at $K = 3$ with $w_1 = 50$, $w_2 = 70$, and $w_3 = \infty$ (see Table 1). The cut-off values 50 and 70 divide the cohort into young (< 50 years old), mid-age (50−70) and old (> 70) subgroups. The data of 13,962 pairs from 4,152 families were used to estimate the cross-ratios. All the piecewise cross-ratios are close to 1, indicating the association of times to first failure between first-degree relatives is modest and almost time invariant.

Nine hundred and fifty pairs with failures observed in both pair members were used to estimate the odds-ratios. Of these 950 pairs, there were 767 distinct paired failure times. Because at each of these distinct paired failure times, most of the time there was only one observation, the non-parametric estimate of the probability of failure type given the paired failure time, $p_j^k(t_1, t_2)$ is mostly 0 or 1. Thus, ϕ_{ij}s cannot be estimated reliably under this non-parametric estimation. Therefore, we considered the two alternative approaches for the estimation of $p_j^k(t_1, t_2)$ described in the previous Section: assuming $p_j^k(t_1, t_2)$ to be piecewise

constant vs. the probability of failure type depends only on the individuals' own failure time. The estimated odds-ratios in the six sub-regions as ordered in Table 1 for the two approaches are (4.25, 1.95, 0.86, 1.35, 1.63, 1.65) and (5.22, 1.20, 0.88, 1.57, 2.07, 1.97), respectively. Both approaches yielded similar estimates of the odds-ratios. The estimates based on the latter approach are shown in Table 1. The distributions of the bootstrap estimates of θ_{ij}s were close to normal, but those of ϕ_{ij}s were skewed. Hence the estimates of ϕ_{ij}s were log-transformed. Compared to the estimates of the cross-ratios, the odds-ratios for the association of failure types between the first-degree relatives vary in magnitude over different ranges of ages at onset. Of particular note is the large odds ratio for the ages at onset younger than 50. It implies that the probability of having cancer for a woman who had an event before age 50 is more than 5 times higher when her first-degree relative had cancer before age 50 than if her first-degree relative died of non-cancer before age 50. For women older than 70 years, there is a trend that her chance of developing cancer is doubled if her first-degree relative had developed breast cancer after age 50 than if her first-degree relative died of non-cancer after age 50 ($\phi_{22} = 2.07$, $\phi_{23} = 1.97$).

The estimates of cause-specific cross-ratios are displayed in Figure 2. The cross-ratio for cancer is high (> 2) when the ages at onset in both members are young and slightly elevated (1.5 − 2) when both members are old. However these elevations, likely due to few cancer cases in both pair member in these age ranges, are not statistically significant.

To see the impact of the failure time and failure type of one family member on the cumulative incidence of the other family member, we plot the conditional incidence function along with the unconditional counter part. The four plots in the left panel of Figure 3 display the marginal and conditional cumulative risk of a woman developing cancer, and the four plots in the right panel display the marginal and conditional cumulative risk of a woman dead of non-cancer. These plots show that a woman's cumulative risk of cancer, compared to the marginal cumulative risk, is increased if her first-degree relative had cancer before age 70, and decreased if her first-degree relative had cancer after age 70. If the first-degree relative died of non-cancer before age 50, then the woman's cumulative risk of cancer is still increased (top two plots in the left panel), although the magnitude of the increase is lower than if her first-degree had cancer. In contrast, if the first-degree relative died of non-cancer after age 50, the woman's cumulative risk of cancer is decreased slightly, and her cumulative risk of non-cancer death is increased.

One may be also interested in the conditional cumulative incidence given the failure type of the first-degree relative. In Figure 4, the top panel shows that a woman's conditional cumulative cancer incidence increased if her first-degree relative had cancer, and decreased if her first-degree relative had died of non-cancer. The lifetime (up to age 100) cumulative incidence of cancer increased from 40% to 46%, if a woman's first-degree relative had cancer. Such a pattern also holds for non-cancer mortality. The lower panel of Figure 4 shows that the conditional cumulative non-cancer increased if the first-degree relative had died of non-cancer, and decreased if the first-degree relative had cancer. The lifetime risk of non-cancer death increased from 58% to 61%, if a woman's first-degree relative had died of non-cancer.

5. Simulations

We used simulations to study the performance of the estimation procedure proposed in the previous section. Initially, we generated 5,000 pairs of (T_1, T_2) and (Y_1, Y_2) according to the proposed model described in Section 2. Specifically, we first generated (T_1, T_2) followed by (Y_1, Y_2) conditional on (T_1, T_2). We assumed the marginal distribution of T_1 and T_2 follows a Weibull model with the shape and scale parameters estimated from the WAS data. The

number of knots and cutoff values used to define the piecewise cross-rations and odds-ratios were set at the same values as in the WAS data analysis ($w_1 = 50$, $w_2 = 70$, $w_3 = \infty$). We took the constant value of 1.2 for all the cross-ratios, because it is close to the estimates seen in the WAS study. We generated the failure time t_1 for the first member in each pair from the Weibull model and the bivariate survival function values on the bivariate knots on each grid in the sample space in the order laid out in the Appendix. Then we generated the event time for the second pair member, t_2, by solving

$$u=P(T_2>t_2 \mid T_1=t_1)=S_{A_{i_1 i_2}}(t_1,t_2)^{\theta i_1 i_2}\prod_{k=2}^{i_2} S_{A_{i_1 k}}(t_1,w_{k-1})^{-\{\theta_{i_1 k}-\theta_{i_1(k-1)}\}} S(t_1)^{-\theta_{i_1 1}},$$

where u is a uniform deviate and i_1, i_2 are the knots such that $(t_1, t_2) \in A_{i1,i2}$. We fitted a fourth-degree polynomial model to the probability of failure type being cancer (Figure 1) and used the fitted model to generate bivariate failure types. The fitted model is given by

$$P(Y_j=1 \mid T_j=t)=0.11+1.37t^{*}+6.57t^{*2}+-17.48t^{*3}+9.43t^{*4},$$

where $t^* = t/100$. The failure type for the first member of each pair was generated from the bernoulli distribution with $p = P(Y_1 = 1 \mid T_1 = t_1)$. The failure type for the second member was generated from the bernoulli distribution with $p = P(Y_2 = 1 \mid T_1 = t_1, T_2 = t_2, Y_1 = k; \phi_{i1i2})$, where k is the failure type of the first member. Each individual is subject to independent censoring, where for the first member of each pair the censoring variable follows a normal distribution $N(82, 14)$ corresponding to 40% censoring and $N(60, 10)$ for the second member corresponding to 75% censoring. The censoring pattern was devised to mimic that of the (mother, sister) pairs in the WAS study. Randomly generated failure times were truncated to integer to represent ages at onset as recorded in the WAS study.

The simulations were repeated 1,000 times and the results were summarized in Table 2. It shows that most of the Monte-Carlo means are close to the true parameter values. One exception is ϕ_{13}, where $\hat{\phi}_{13}$ is slightly over-estimated. Consistent with the observations seen in the example, the standard errors of odds-ratios overall are larger than those of cross-ratios. This is due to the fact that for the cross-ratios all paired data, whether censored or not, contribute to the estimation, but for the odds-ratios, only pairs with both failures observed enter the likelihood for estimation. In addition, the estimate of the marginal probability of failure type given the individual's failure time is highly variable if the number of failures at that failure time is small. Therefore, in order to obtain a stable estimate of ϕ_{ij}, an adequate number of bivariate failures for each bivariate failure type in each grid region is essential. The coverage probability for most of the parameters are close to the 95% nominal level. Hence the parameter estimators are approximately normally distributed.

We performed additional simulations to compare the performance of our estimator of the cause-specific cross-ratio with that of Bandeen-Roche and Ning (2008). Their method was an extension of the non-parametric estimator of the Kendall's tau ignoring competing risks (Oakes, 1982). In their approach, only data with failure observed in both pair members and with failure type of interest were included. For example, if one is interested in cancer-cancer cross-ratio, their method would eliminate all censored pairs, pairs with cancer vs. non-cancer death, and pairs with non-cancer death in both members.

We first generated the bivariate competing risk data using the same scenario as presented above. Under this scenario, the probability of event type is time-dependent, and the odds-ratios for the bivariate failure types are piecewise constant. As a result, the cause-specific

cross-ratio according to (4) is time-varying. In our approach, the cause-specific cross-ratio in each grid element was averaged with the bivariate incidence function as its weight to represent the cause-specific cross-ration in that grid element, whereas the Bandeen-Roche and Ning approach used the probability of discordance as the weight. The simulation results are presented in Table 3. The cause-specific cross-ratios for bivariate failure type (1, 1) was underestimated by the Bandeen-Roche and Ning approach except in the last grid element $[70, \infty) \times [70, \infty)$ where the estimate was overestimated. However, as one referee pointed out, since the two approaches use different weights and estimate different quantities under the time-vary cause-specific cross-ratio model, the two estimates are not directly comparable. Nevertheless, the estimate from our approach had substantially lower standard deviation.

Second, we generated bivariate failure time data with constant cause-specific cross-ratio. Specifically, bivariate failure times to first event were generated according to the gamma frailty model with same parameter values as our original simulation study presented in the manuscript (i.e. Weibull marginal survival and constant cross-ratio $\theta = 1.2$). The bivariate failure types were generated from the beta-bernoulli distribution as described in Bandeen-Roche and Liang (2002) with scale parameter 1 and the probability of failure type being 1 equal to 0.2. Under this scenario, the cause-specific cross-ratio for bivariate failure types (1, 1) equals 3.6. In the case of no censoring, the sample size was set at $n = 1000$. With censoring, censoring time was generated from normal with mean 65 and standard deviation 10. Approximately 25% of the failure times were censored and the sample size was set at $n = 2000$. Since the probability of failure type being 1 does not change with time, the estimate of the probability of failure type given the failure time from the simulated data fluctuates around the true value. Hence in our estimation procedure for the odds ratio, the marginal probability of failure type was assumed constant. We mimicked the simulation study of Bandeen-Roche and Ning (2008) in setting the bins by bisecting each time dimension at its median in each simulation run. Under this scenario, the averaged cause-specific cross-ratio in each grid is constant regardless of the weight used in averaging and equals 3.6. The simulation results are summarized in Table 4. It shows that when there is no censoring, the estimate of the cross-ratio for bivariate failure types (1, 1) has little bias for both approaches. Consistent with the findings in the second simulation study above, the variance of the Bandeen-Roche and Ning' approach is considerably larger than that of our approach. Such an inflation in the variance is due to the fact that our approach uses all data in the estimation, while the Bandeen-Roche and Ning's appraoch only uses the pairs with bivariate failure types (1,1), which on average consists of only 12% of the total number of pairs. With bivariate failure types (2,2) which consists of about 75% of the data, the variability of the estimate of the Bandeen-Roche and Ning approach is still larger but closer to our approach (results not shown). In the presence of censoring, The Bandeen-Roche and Ning's appraoch is more biased and the variance of their estimate is more inflated than for the uncensored case.

In summary, these simulation studies indicated that, in general, our proposed approach has less bias and is more efficient than Bandeen-Roche and Ning's approach.

6. Discussion

In this paper, we have proposed a flexible model for bivariate competing risk data. In our approach, we decompose the bivariate competing risk data into bivariate times to first failure and bivariate failure types given the failure times. To each bivariate component, we apply a piecewise-constant association model to measure the dependency. We have shown that with a given marginal distribution of time to first failure and piecewise constant cross-ratios, the bivariate distribution is determined. With this approach, we are able to assess the

contribution of each component of the bivariate data to the cause-specific association. In practice, the knots and the number of knots used to specify the piecewise-constant cross-ratios and odds-ratios are primarily determined by the data such that sufficient number of paired failures in each grid elements are observed. Such an approach is often necessary in the analysis of rare events such as cancer.

We have proposed a semi-parametric two-stage procedure for estimation. For the estimation of cross-ratios θ_{ij}s, at the first stage, the marginal distribution of the first failure time is estimated non-parametrically by the Nelson estimate. In the second stage, these cross-ratios are estimated sequentially in each coordinate such that the estimation of each cross-ratio depends on the estimates of the cross-ratio and in turn the bivariate survival function at times preceding it. For the estimation of odds-ratios ϕ_{ij}s, at the first stage, the probability of event type given the bivariate failure times is estimated non-parametrically. In the WAS study, the integer age was recorded and thus the probability of failure type was estimated at each observed age at onset. In the case of continuous time, the failure time needs to be binned in order to reliably estimate the probability of failure type. At the second stage, the odds-ratio in each region is estimated assuming the event type probability is known and using the paired data with failures observed in that regions for both pair members. Since the number of observations used in estimating each odds-ratio is smaller than that used in estimating the cross-ratio, the variation of the odds-ratio estimates tends to be larger than that of the cross-ratio. This is seen in the analysis of the WAS data as well as in simulated data. Therefore, in order to be able to assess the effect of the event type of one member on that of the other family member, it is necessary to observe sufficient number of failures for each bivariate failure type. The number of bivariate failures required to have a reliable estimate for each odds-ratio depends on the probability of failure type in each sub-region. If the probability does not change with time, then the rule of thumb for the estimation of odds-ratio in a 2×2 table applies here. In this case, a minimal number of 30 pairs would usually be sufficient. However, if the failure type probability changes with time, as the case for the WAS study shown in Figure 1, a larger sample size is required. This is because the non-parametric estimate of the failure type probability is more variable and hence the estimate of the odds-ratio is more variable as well.

Because of the complex nested structure of the piecewise-constant bivariate survival model for the times to first failure and sequential dependency of the estimation of cross-ratios, derivation of the theoretical properties of the proposed estimates is complex and a topic for future research. We used the bootstrap to estimate the standard errors and to draw inference. For the WAS study data, it took about one minute of CPU to estimate the parameters. With 500 bootstrap samples, it took about 8 hours to compute the standard errors. We used simulations to study the properties of the proposed parameter estimates. The simulation result shows that there is little bias for the parameter estimates, and the estimators are close to be normally distributed.

There are several interesting findings from our analysis of the WAS data. In this cohort, the majority of the cancers among women are breast and ovarian cancers, and the probability of having cancer is higher than non-cancer death when the age at onset is between 30 to 70 years old. An individual's failure type also affects her first-degree relative's failure type especially at young ages. Based on $\hat{\phi}_{11} = 5.22$, the probability of having cancer for a woman who had an event before age 50 is more than 5 times higher when her first-degree relative had cancer before age 50 than if her first-degree relative died of non-cancer before age 50. However, such a large increase in the conditional probability of failure type being cancer before age 50 does not result in the same degree of increase in the cumulative risk of cancer. For example, a woman's cumulative risk of having cancer at age 30–50 increases 1.4–1.9 times if her first-degree relative had cancer between age 30 and 50 then if her first-degree

relative died of non-cancer between age 30 and 50. Overall, a woman's cumulative risk at all ages is increased if her failure type is the same as her first-degree relative's. This translates into 6% and 3% increase of lifetime risk of cancer and non-cancer mortality, respectively.

Acknowledgments

The authors wish to thank Dr. Nilanjan Chatterjee for kindly providing the WAS study data used in this paper. The authors thank the associate editor and referees for their thorough review and insightful comments.

REFERENCES

Bandeen-Roche K, Liang KY. Modelling multivariate failure time associations in the presence of a competing risk. Biometrika. 2002; 89:299–314.

Bandeen-Roche K, Ning J. Nonparametric estimation of bivariate failure time associations in the presence of a competing risk. Biometrika. 2008; 95:221–232. [PubMed: 20305739]

Chatterjee N, Hartge P, Wacholder S. Adjustment for competing risk in kin-cohort estimation. Genetic Epidemiology. 2003; 25:303–313. [PubMed: 14639700]

Chatterjee N, Kalaylioglu Z, Shih JH, Gail M. Casecontrol and case-only designs with genotype and family history data: Estimating relative risk, residual familial aggregation, and cumulative risk. Biometrics. 2006; 62:36–48. [PubMed: 16542227]

Cheng, Yu; Fine, Jason P.; Kosorok, Michael R. Nonparametric Association Analysis of Bivariate Competing-Risks Data Journal of the American Statistical Association. 2007; 102:1407–1415.

Cheng Y, Fine JP. Nonparametric estimation of cause-specific cross hazard ratio with bivariate competing risks data. Biometrika. 2008; 95:233–240.

Cheng Y, Fine JP, Kosorok MR. Nonparametric association analysis of exchangeable clustered competing risks data. Biometrics. 2009; Vol. 65:385–393. [PubMed: 18549422]

Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978; 65:141–151.

Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. New York: Chapman and Hall; 1993.

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

Hougaard, P. Analysis of multivariate survival data. Springer-Verlag Inc; 2001.

Li H, Yang P, Schwartz AG. Analysis of age of onset data from case-control family studies. Biometrics. 1998; 54:1030–1039. [PubMed: 9750249]

Kalbfleisch, JD.; Prentice, RL. The statistical analysis of failure time data. John Wiley & Sons; 2002.

Nan B, Lin X, Lisabeth LD, Harlow SD. Piecewise constant cross-ratio estimation for association of age at a marker event and age at menopause. Journal of the American Statistical Association. 2006; 101:65–77.

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

Oakes D. Bivariate survival models induced by frailties. Journal of the American Statistical Association. 1989; 84:487–493.

Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978; 34:541–554. [PubMed: 373811]

Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. Biometrics. 1995; 51:1384–1399. [PubMed: 8589230]

Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Lawrence BC, Tucker MA. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. The New England Journal of Medicine. 1997; 336:1401–1408. [PubMed: 9145676]

APPENDIX

Construction of the bivariate survival function of times to first event

We arrange the sample space of T_1 and T_2 into a grid with elements A_{ij}, $i = 1, \cdots, K$, $j = 1, \cdots, K$ where $A_{ij} = [w_{i-1}, w_i) \times [w_{j-1}, w_j)$ as shown in Figure 5.

We show that the bivariate survival function (2) is completely determined by the two marginal survival functions S_1, S_2 and the piecewise constant cross-ratios θ_{ij}'s. Notice that if the survival functions on the bottom and left boundaries in each region A_{ij} are known, the joint survival function in that region is determined according to (2). Hence it is sufficient to show that the survival functions on the bottom and left boundaries for each A_{ij} are determined by the marginal survival functions S_1, S_2 and θ_{ij}'s. Note that for each A_{ij}, $i > 1$, $j > 1$, the bottom boundary $[w_{i-1}, w_i) \times w_{j-1}$ is the upper boundary of its adjacent lower region $A_{i,j-1}$ on which the joint survival function is determined by $S(t_1, w_{j-2})$ on the bottom boundary $[w_{i-1}, w_i) \times w_{j-2}$, $S(w_{i-1}, t_2)$ on the left boundary $w_{i-1} \times [w_{j-2}, w_{j-1})$, and $\theta_{i,j-1}$ for $(t_1, t_2) \in A_{i,j-1}$. Similarly the left boundary $w_{i-1} \times [w_{j-1}, w_j)$ in A_{ij} is the right boundary of its adjacent left region $A_{i-1,j}$ on which the joint survival function is determined by $S(t_1, w_{j-1})$ on the bottom boundary $[w_{i-2}, w_{i-1}) \times w_{j-1}$, $S(w_{i-2}, t_2)$ on the left boundary $w_{i-2} \times [w_{j-1}, w_j)$, and $\theta_{i-1,j}$ for $(t_1, t_2) \in A_{i-1,j}$. These lower and left boundaries of $A_{i-1,j}$ and $A_{i,j-1}$ in turn are the right and upper boundaries of the adjacent lower and left regions. We repeat the iterations until we reach the bottom and left boundaries of A_{11} on which the survival function is equal to the marginal survival function $S_1(t_1)$ and $S_2(t_2)$ respectively for $(t_1, t_2) \in A_{11}$. This shows that the survival functions on the boundaries are determined by the marginal survival functions and θ_{ij}'s. For a $K \times K$ grid on the first quadrant for the sample space of (T_1, T_2), it contains a total of $K(K + 1)/2$ boundaries. Of these boundaries K lie on the first coordinate (i.e. $t_2 = 0$) and similarly K lie on the second coordinate (i.e. $t_1 = 0$), where their corresponding survival probabilities are equal to the respective marginal survival probabilities. The rest of the boundaries are either the upper or right boundary of a grid element. In terms of actual computation of the survival probabilities on these boundaries, we start from the upper and right boundaries of A_{11} because the survival probabilities on the bottom and left boundaries are known and equal to the marginal survival probabilities. We use (2) to calculate the survival probabilities on the right and upper boundaries of A_{11}. Equating these probabilities to the survival probabilities on the left and lower boundaries of A_{21} and A_{12}, using (2) again, the survival probabilities on the upper and right boundaries of A_{12} and A_{21} are readily computed. We iterate the above computation procedure by sequentially moving the grid element horizontally and vertically until all the boundary probabilities are computed. Figure A1 is used to illustrate the order of constructing the bivariate survival function on the boundaries. It begins with the upper and right boundaries of A_{11}. Then we compute the joint survival on upper and right boundaries for $A_{12}, \cdots, A_{1K}$ and $A_{21}, \cdots, A_{k1}$ sequentially. After the boundaries probabilities for the grid regions on the first row and first column are computed, we computes the boundary probabilities for A_{22} next, and repeats the procedure until the boundaries probabilities of A_{KK} are obtained.

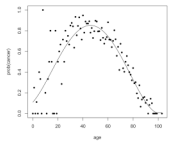

Figure 1.
The non-parametric estimate of the probability of failure type being cancer given the age at onset (solid circle) and a fitted line of a fourth-degree polynomial model to this non-parametric estimate. The fitted model is used in the simulation study to generate the failure type given the age at onset.

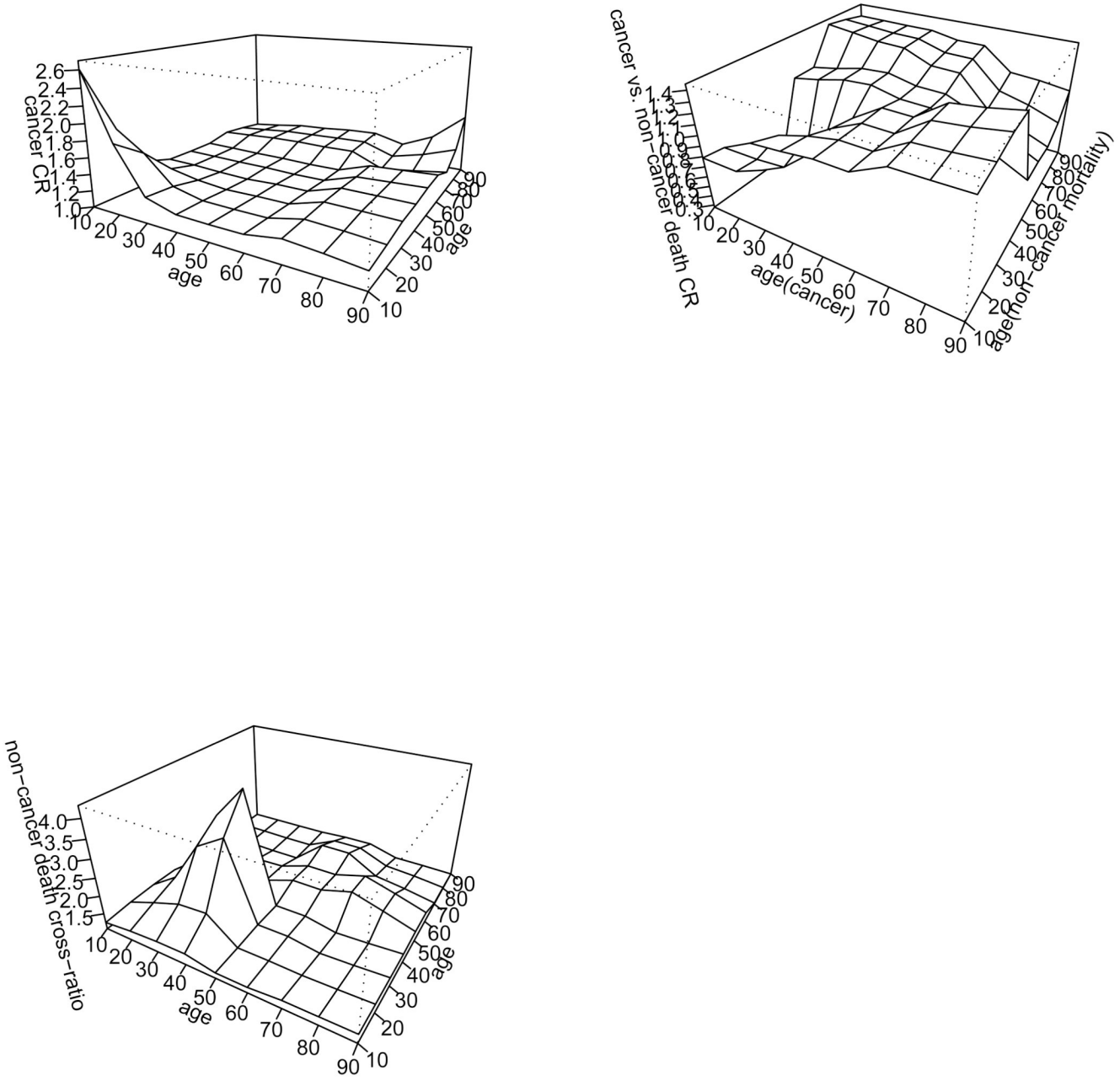

Figure 2.

Top left: cancer vs. cancer cross-ratio; top right: cancer vs. non-cancer death cross-ratio; bottom left: non-cancer death vs. non-cancer death cross-ratio. The averaged cancer-cancer cross-ratios in the six sub-regions ordered in Table 1 are 1.22, 1.19, 1.29, 1.20, 1.40 and 1.34 with corresponding standard error 0.17, 0.10, 0.11, 0.14, 0.12 and 0.25. The alternative Bandeen-Roche and Ning estimate described in the simulation study in Section 5 equals 2.36, 1.94, 1.80, 1.57, 1.45 and 1.94 with corresponding standard error 0.43, 0.24, 0.28, 0.31, 0.27, and 0.87.

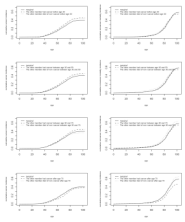

Figure 3.
Left panel: marginal and conditional cumulative cancer incidence; right panel: marginal and conditional cumulative non-cancer mortality incidence.

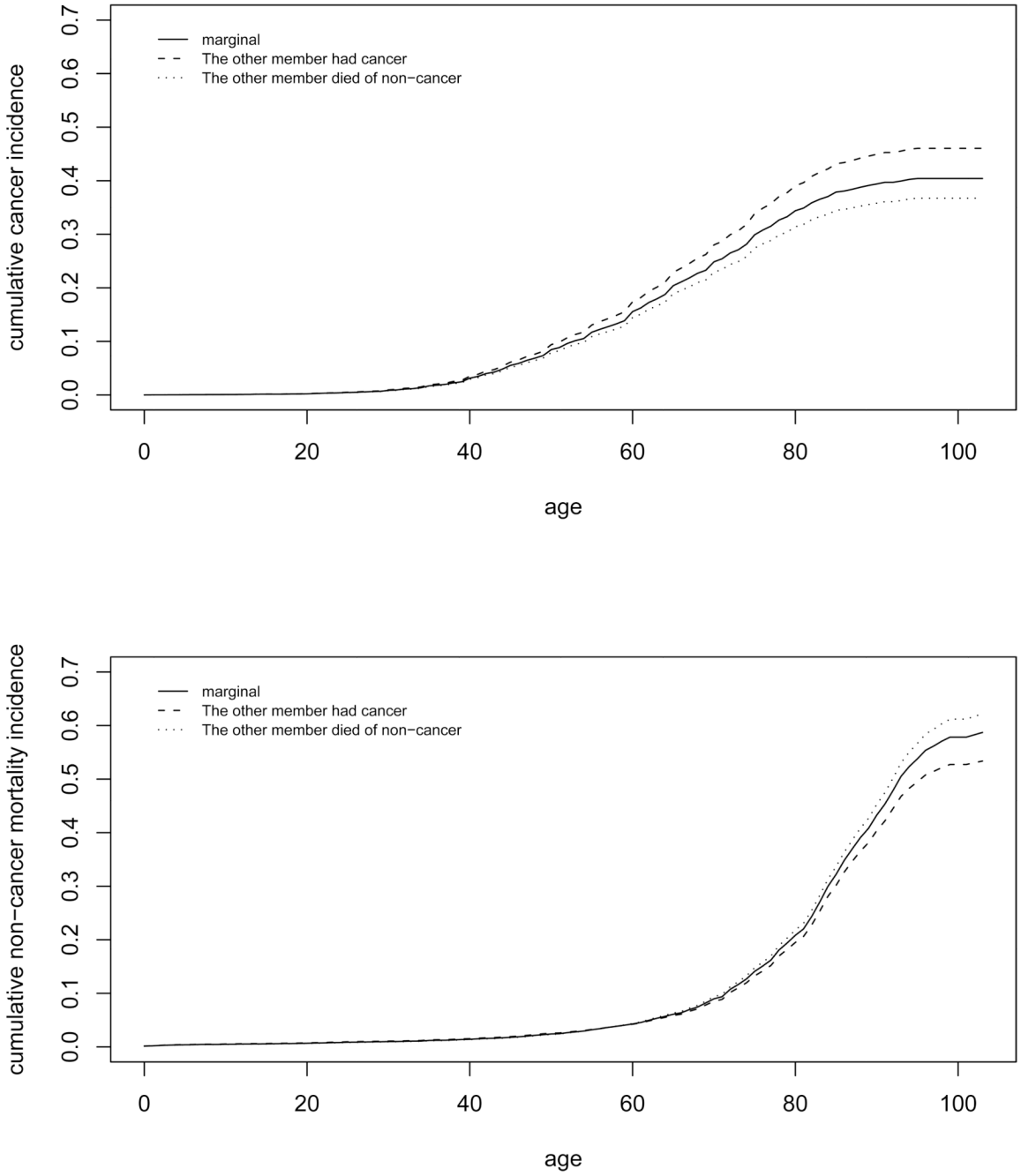

Figure 4.

Top: marginal and conditional cumulative cancer incidence given the failure type of the other family member; bottom: marginal and conditional cumulative non-cancer mortality incidence given the failure type of the other family member.

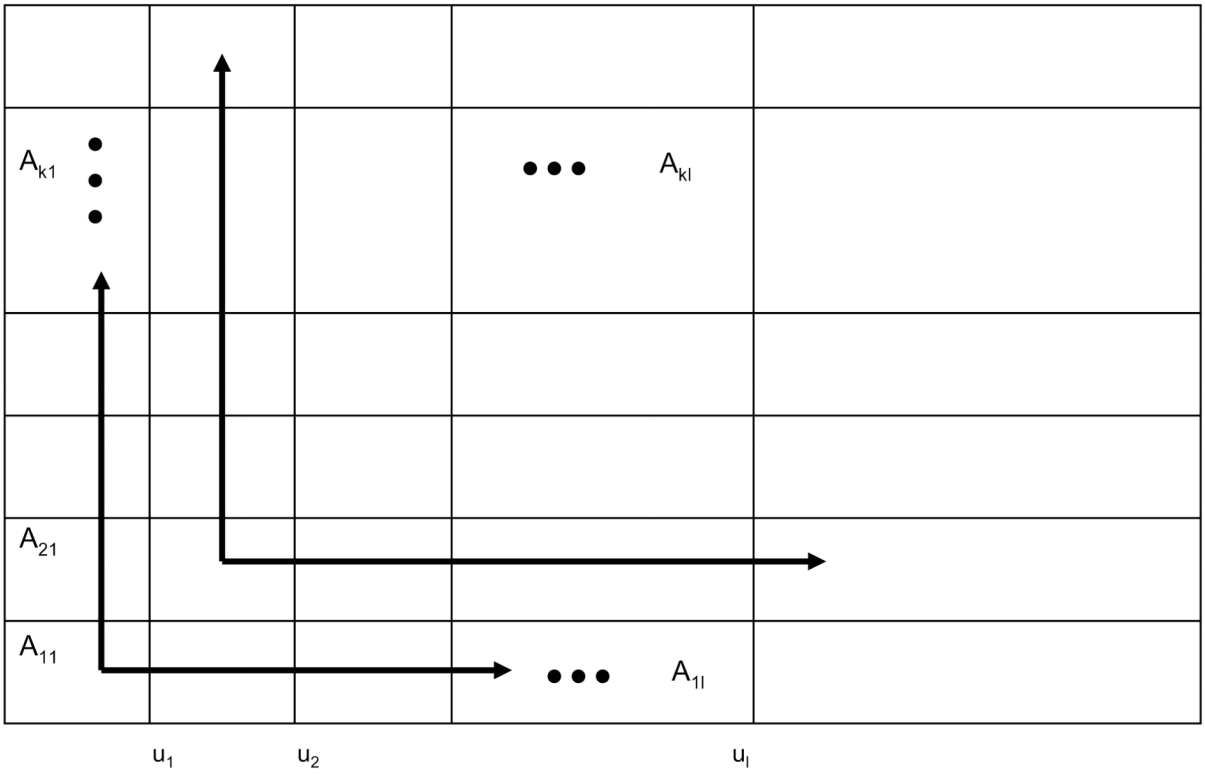

Figure 5.
A diagram used to illustrate the order of constructing the bivariate survival function on the boundaries.

Table 1

WAS data analysis

Ages at onset	# of paired events (d_{11}, d_{12}, d_{22})[1]	CR	Est.	SE	OR	Est.	SE	exp(Est.)
[0, 50) × [0, 50)	(39,21,12)	θ_{11}	1.15	0.13	log ϕ_{11}	1.65	0.83	5.22
[0, 50) × [50, 70)	(92,64,21)	θ_{12}	1.18	0.08	log ϕ_{12}	0.18	0.49	1.20
[0, 50) × [70, ∞)	(52,117,47)	θ_{13}	1.31	0.10	log ϕ_{13}	−0.13	0.39	0.88
[50, 70) × [50, 70)	(43,39,12)	θ_{22}	1.16	0.09	log ϕ_{22}	0.45	0.58	1.57
[50, 70) × [70, ∞)	(52,132,58)	θ_{23}	1.26	0.10	log ϕ_{23}	0.73	0.43	2.07
[70, ∞) × [70, ∞)	(17,56,76)	θ_{33}	1.05	0.32	log ϕ_{33}	0.68	0.55	1.97

Table 2

Simulation Study 1

Ages at onset	CR (θ)	Monte-Carlo mean	Monte-Carlo SE	Mean of SE	95% cov. prob.
[0, 50) × [0, 50)	1.20	1.16	0.13	0.13	91.9
[0, 50) × [50, 70)	1.20	1.19	0.08	0.08	96.2
[0, 50) × [70,∞)	1.20	1.20	0.11	0.11	94.6
[50, 70) × [50, 70)	1.20	1.14	0.09	0.09	89.4
[50, 70) × [70, ∞)	1.20	1.23	0.10	0.10	95.9
[70, ∞) × [70, ∞)	1.20	1.24	0.33	0.33	96.5

	OR (logϕ)	Median # of pairs	Monte-Carlo mean.	Monte-Carlo SE	Mean of SE	95% cov. prob.
[0, 50) × [0, 50)	1.65	80	1.69	0.78	0.82	96.3
[0, 50) × [50, 70)	0.18	238	0.18	0.39	0.41	96.9
[0, 50) × [70, ∞)	−0.13	137	−0.09	0.52	0.55	96.6
[50, 70) × [50, 70)	0.45	164	0.44	0.43	0.44	95.6
[50, 70) × [70, ∞)	0.73	157	0.78	0.48	0.49	96.2
[70, ∞) × [70, ∞)	0.68	17	0.64	1.23	1.22	96.1

Table 3

Simulation study 2.

(t_1, t_2)	$\theta^{11}(t_1, t_2)$	New approach			Bandeen-Roche and Ning		
		mean	SD	$\sqrt{\text{MSE}}$	mean	SD	$\sqrt{\text{MSE}}$
[0, 50) × [0, 50)	1.29	1.25	0.15	0.15	1.10	0.15	0.24
[0, 50) × [50, 70)	1.21	1.21	0.08	0.08	1.03	0.09	0.20
[0, 50) × [70,∞)	1.18	1.19	0.13	0.13	1.02	0.19	0.25
[50, 70) × [50, 70)	1.24	1.18	0.10	0.12	1.05	0.12	0.23
[50, 70) × [70, ∞)	1.34	1.38	0.15	0.15	1.11	0.21	0.31
[70, ∞) × [70, ∞)	1.53	1.55	0.74	0.74	1.67	2.10	2.10

Table 4

Simulation study 3.

$\theta^{11} = 3.6$	New approach			Bandeen-Roche and Ning		
	mean	SD	$\sqrt{\text{MSE}}$	mean	SD	$\sqrt{\text{MSE}}$
No censoring						
(t_1, t_2) < median	3.46	0.30	0.33	3.67	0.73	0.73
t_1 >= median, t_2 < median or t_2 >= median, t_1 < median	3.71	0.29	0.31	3.67	0.62	0.63
(t_1, t_2) >= median	3.52	0.33	0.34	3.76	0.91	0.93
With censoring						
(t_1, t_2) < median	3.50	0.34	0.36	3.16	0.67	0.80
t_1 >= median, t_2 < median or t_2 >= median, t_1 < median	3.75	0.37	0.40	3.16	0.83	0.94
(t_1, t_2) >= median	3.44	0.53	0.55	3.89	3.16	3.17