11-11-2009

Analyzing Bivariate Survival Data with Interval Sampling and Application to Cancer Epidemiology

Hong Zhu
Johns Hopkins Bloomberg School of Public Health, hongzhu@jhsph.edu

Mei-Cheng Wang
Johns Hopkins Bloomberg School of Public Health, mcwang@jhsph.edu

Biometrics 000, 1–25 DOI: 000
000 0000

Analyzing Bivariate Survival Data with Interval Sampling and Application to Cancer Epidemiology

Hong Zhu* and Mei-Cheng Wang**

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205, U.S.A.

**email:* hongzhu@jhsph.edu

***email:* mcwang@jhsph.edu

Summary: In medical follow-up studies, ordered bivariate survival data are frequently encountered when bivariate failure events are used as the outcomes to identify the progression of a disease. In cancer studies interest could be focused on bivariate failure times, for example, time from birth to cancer onset and time from cancer onset to death. This paper considers a sampling scheme where the first failure event (cancer onset) is identified within a calendar time interval, the time of the initiating event (birth) can be retrospectively confirmed, and the occurrence of the second event (death) is observed subject to right censoring. To analyze this type of bivariate failure time data, it is important to recognize the presence of bias arising due to interval sampling. In this paper, nonparametric and semiparametric methods are developed to analyze the bivariate survival data with interval sampling under stationary and semi-stationary conditions. Numerical studies demonstrate the proposed estimating approaches perform well with practical sample sizes in different simulated models. We apply the proposed methods to SEER ovarian cancer registry data for illustration of the methods and theory.

Key words: Bivariate survival distributions; Copula; Interval sampling; Nonparametric and semiparametric; Stationarity and semi-stationarity.

1. Introduction

Ordered bivariate survival data arise frequently in medical follow-up studies when each subject may experience bivariate failure events, which are considered as the major outcomes to identify the progression of a disease. In cancer studies, for example, it is of interest to understand the process from birth to cancer onset, and to death. It is common to collect data with incidence of disease occurring within a calendar time interval. This type of sampling is referred to as interval sampling and we consider an interval sampling scheme in this paper.

Consider a case population where case refers to the first failure event and two failure events occur in a chronological order following the occurrence of the initiating event. Denote the calendar time of the initiating event by T, the time from the the initiating event to the first failure event by Y, and the time from the first event to the second by Z. The variables Y and Z are expected to be correlated because they come from the same subject. Bivariate failure times (Y, Z) are the outcome variables of interest in this paper. In statistical literature, Visser (1996), Wang and Wells (1998), Lin, Sun and Ying (1999), and Schaubel and Cai (2004) proposed various estimation methods for bivariate or multivariate survival data subject to right censoring. In this paper, we consider the problem of interval sampling and develop estimation approaches for analyzing the bivariate survival data with interval sampling.

We assume that the sample includes those subjects who have experienced the first failure event within a given time period, the initiating event of each subject can be retrospectively identified, and the occurrence of the second failure event is prospectively identifed. Also because of loss to follow-up or end-of-study, the observation of the second failure event is subject to right censoring. For example, let T, Y, and Z respectively represent the calendar time of birth, the time from birth to cancer onset, and the time from cancer onset to death for a subject, and Y and Z are the variables of interest. The study cohort is made up of

subjects whose first failure events occur within a calendar time interval $[0, T_0]$. Moreover, the observation of the second failure event is terminated at the calendar censoring time C ($C \leqslant T_0$). Bias arises due to interval sampling scheme, where the triplet (T, Y, Z) is observed subject to the constrains $-T \leqslant Y \leqslant T_0 - T$ and $Y + Z \leqslant C - T$. Figure 1 provides a simple explanatory plot for bivariate survival data with interval sampling with constant $C = T_0$.

[Figure 1 about here.]

The research is motivated by the problem in SEER cancer registry data, which also serves as an example to illustrate how interval sampling design arises. The SEER (Surveillance, Epidemiology, and End Results) program is an epidemiologic surveillance system consisting of population-based cancer registries designed to track cancer incidence and survival in the United States. Collection of the SEER data began from January 1, 1973 (Ries et al., 2002). The registries routinely collect information on newly diagnosed cancer patients residing in geographically defined areas representing 26 percent of the US population. The SEER data are released as the Patient Entitlement and Diagnosis Summary File for cancer cases diagnosed from 1973 to 2002. Basic diagnostic information is available for up to 10 diagnosed cancer cases for each person, such as breast cancer, lung cancer, ovarian cancer, etc. It contains information on each person's month and year of birth, date of cancer diagnosis, date of death, type of cancer, sex, race, state of residence etc. Taking ovarian cancer as our illustrative example, the cancer cases diagnosed from 1973 to 2002 are the cohort of interest under interval sampling, the initiating time is the birth time, and the bivariate failure events are the diagnosis of ovaian cancer and death.

2. Stationarity, Semi-Stationarity, and Non-Stationarity

The case population considered in this paper is a cohort of subjects whose first failure event occurs within a calendar time interval and then prospectively followed. We now introduce

notation and assumptions to facilitate the development of the proposed work. Assume that the initiating events occur over the calendar time with the intensity (or rate) function $\phi(t)$ for $t \leqslant T_0$. Let $f(y,z)$, $f_y(y)$, $f_z(z)$, and $\phi(t)$ denote the population joint density function of (Y, Z), marginal density of Y, Z, and the intensity function of T. Let $F_y(\cdot)$ and $F_z(\cdot)$ denote the cumulative distribution functions of $f_y(\cdot)$ and $f_z(\cdot)$ respectively, $y_- = inf\{y : F_y(y) > 0\}$, $y^+ = sup\{y : F_y(y) < 1\}$, $z_- = inf\{z : F_z(z) > 0\}$, $z^+ = sup\{z : F_z(z) < 1\}$ and $t_- = inf\{t : \phi(t) > 0\}$. To reduce the mathematical complexity in the discussion, assume the failure time Y has finite support, where $y^+ < \infty$ so that ϕ can be normalized as a probability density function. Let $g(t)$ denote the population density function of T in the interval $[-y^+, T_0 - y_- - z_-]$, derived as normalized $\phi(t)$:

$$g(t) = \phi(t) I(-y^+ \leqslant t \leqslant T_0 - y_- - z_-) / \int_{-y^+}^{T_0 - y_- - z_-} \phi(u) du \tag{1}$$

Let $G(\cdot)$ denote the cumulative distribution function of $g(\cdot)$. Assume $(T_1, Y_1, Z_1), \ldots,$ (T_n, Y_n, Z_n) are independent and identically distributed. Consider the following two assumptions:

S1. The disease process is independent of when the initiating event occurs. Or, equivalently, assume that T is independence of (Y, Z).

S2. The occurrence of the initiating event started in the distant past and the rate of occurrence has been stabilized. Or, quantitatively, assume that t_- is small enough so that $t_- \leqslant -y^+$, and that $\phi(t)$ is constant for $-y^+ \leqslant t \leqslant T_0 - y_- - z_-$ and $G(\cdot)$ is $Uniform\,[-y^+, T_0 - y_- - z_-]$.

The two conditions serve as the fundamental assumptions for studying the probability structures of the primary outcomes in this paper. We say that the model is stationary if both $(S1)$ and $(S2)$ are satisfied, semi-stationary if only $(S1)$ is satisfied, and non-stationary if neither $(S1)$ nor $(S2)$ is assumed. The discussion here is focused on the stationary and semi-stationary conditions. However, $(S1)$ and $(S2)$ may not always be valid, for example, if

new treatment becomes available, the incident rate of disease may change over time, which may also affect the distribution of (Y, Z). The non-stationary condition is beyond the scope of this paper and will be explored in the future.

The rest of the paper is organized as follows. In section 3, bias due to the interval sampling scheme is discussed, and a nonparametric model is developed to estimate the joint survival function of bivariate survival data with interval sampling under stationary condition. Section 4 proposes a semiparametric copula model of bivariate survival data under stationary condition to study the dependency structure. A semi-stationary model is presented in section 5. Numerical studies in section 6 demonstrate that the proposed estimating methods perform well with practical sample sizes in different simulated models. In section 7, a cohort of ovarian cancer cases from SEER data is analyzed for illustration. Finally, concluding remarks and discussion are included in section 8.

3. Nonparametric Estimation of Joint Survival Function under Stationary Condition

In this section, under stationary condition when both (S1) and (S2) hold, we develop a nonparametric approach for estimating joint survival function on the basis of uncensored data. For simplicity of the discussion, we first consider the case that the observation of the second failure time ends at calendar time C, where particularly $C = T_0$, a constant. This simple censoring mechanism can be replaced by random censoring.

First consider the case that only $(S1)$ is assumed, the joint density of uncensored (t, y, z) can be derived as the density of (T, Y, Z) conditional on $-T \leqslant Y \leqslant T_0 - T$ and $Y + Z \leqslant T_0 - T$:

$$\begin{aligned} p(t,y,z) &= P(T=t, Y=y, Z=z | -Y \leqslant T \leqslant T_0 - Y - Z) \\ &= \frac{g(t)f(y,z)I(-y \leqslant t \leqslant T_0 - y - z)}{P(-Y \leqslant T \leqslant T_0 - Y - Z)} \\ &= \left[\frac{g(t)I(-y \leqslant t \leqslant T_0 - y - z)}{G(T_0 - y - z) - G(-y)}\right] \cdot \left[\frac{\{G(T_0 - y - z) - G(-y)\}f(y,z)}{\int\{G(T_0 - u - v) - G(-u)\}f(u,v)dudv}\right] \end{aligned}$$

$$= \quad p_c(t|y,z)p(y,z) \tag{2}$$

The first bracket term above, which is denoted by $p_c(t|y,z)$, specifies the conditional density of the observed t given the observed uncensored (y,z); the second bracket term, denoted by $p(y,z)$ is the joint density of uncensored (y,z).

Define the weight function $w(y,z) = G(T_0 - y - z) - G(-y)$, which describes the selection bias for observing (y,z). The value of the weight function coincides with the probability for the initiating events to occur within the 'window' $[-y, T_0 - y - z)$. Taking the ovarian cancer example to illustrate such weight, provided that the population of interest is a closed population, the weight function can be interpreted as the proportion of the subjects born in the interval $[-y, T_0 - y - z)$ from the total population born in $[-y^+, T_0 - y_- - z_-]$. As a result, shorter time of y and z are observed with the weight as the proportion of births from later calendar windows, and longer time of y and z are observed with the weight as the proportion of births from earlier calendar windows.

The joint density function of uncensored (y,z) can be expressed as $p(y,z) = \frac{w(y,z)f(y,z)}{\int\int w(u,v)f(u,v)dudv}$, so it is generally biased from its population density $f(y,z)$, and the direction of bias is determined by the weight function $w(y,z)$.

Then, assuming both $(S1)$ and $(S2)$ hold, the joint density of uncensored (y,z) can be further simplified as

$$p(y,z) = \frac{(T_0 - z)f(y,z)}{\int\int (T_0 - v)f(u,v)dudv} \tag{3}$$

Therefore, the weight function reduces to $T_0 - z$, and the nonparametric estimator of joint survival function of (Y,Z) can be simply derived as

$$\hat{S}(y,z) = \frac{\sum_{i=1}^{n}(T_0 - Z_i)^{-1}I(Y_i > y, Z_i > z)}{\sum_{i=1}^{n}(T_0 - Z_i)^{-1}} \tag{4}$$

where (Y_i, Z_i)'s are the uncensored bivariate failure times. $S(y,z)$ is identifiable on the domain $\{(y,z) : y + z \leqslant T_0 - t_-\}$ and this constrain will be redundant if $T_0 - t_- \geqslant y^+ + z^+$.

The above estimator can be proved to be the nonparametric maximum likelihood estimator (NPMLE) of $S(y,z)$, a special case under Vardi's selection bias models (1982, 1985). The asymptotic property of $\hat{S}(y,z)$ can be stated as follow.

Property 1. As $n \to \infty$, the process $\sqrt{n}\{\hat{S}(y,z) - S(y,z)\}$ converges weakly to a bivariate zero-mean Gaussian process with covariance function

$$\begin{aligned} \sigma &= W_{-1}W\{\frac{\int_y^\infty \int_z^\infty (T_0-v)^{-1} f(u,v)dudv}{W_{-1}}[1 - S(y^{'}, z^{'})] \\ &+ S(y,z)[S(y^{'}, z^{'}) - \frac{\int_{y^{'}}^\infty \int_{z^{'}}^\infty (T_0-v)^{-1} f(u,v)dudv}{W_{-1}}]\} \end{aligned} \tag{5}$$

where $W = \int\int (T_0 - v) f(u,v)dudv$, and $W_{-1} = \int\int (T_0 - v)^{-1} f(u,v)dudv$.

4. Semiparametric Copula Model under Stationary Condition

The nonparametric model discussed in section 3 only uses uncensored data. In this section, we consider a semiparametric copula model, where we impose a slightly stronger assumption on the dependency structure of the bivariate survival time of interest. The copula model approach is widely used to model the dependence in survival data (Genest, Ghoudi and Rivest, 1995; Li, Tiwari and Guha, 2007). By the proposed method we will be able to to fully utilize the information from both uncensored and censored data. As will be studied in this section, under stationary condition when both $(S1)$ and $(S2)$ are satisfied, 'double truncation' from interval sampling does not result in bias on the first failure time, and the second failure time is independently censored and can be treated as standard survival data. We take advantage of these properties to semiparametrically estimate joint survival distribution with copula model. The approach is attractive because it allows us to model and estimate the margins and dependency seperately.

We will investigate the semiparametric copula model by a 'two-stage' estimation approach similar to that of Genest et al. (1995) and Shih and Louis (1995). At the first stage, we explore the probability structure for each failure time marginally and obtain the nonparametric

consistent estimations for marginal survival functions under stationary condition, ignoring the dependence. At the second stage, these estimators are substituted into a conditional likelihood for the association parameter, yielding a pseudo likelihood (Gong and Samaniego, 1981). The association parameter is then estimated by solving the estimating equation derived from pseudo conditional likelihood.

4.1 *Failure Time Distributions under Stationary Condition*

First of all, we consider the first failure time Y separately, which is sampled given $-T \leqslant Y \leqslant T_0 - T$. The joint density of observed (t, y) can be written as

$$\begin{aligned} p(t,y) &= P(T=t, Y=y | -Y \leqslant T \leqslant T_0 - Y) \\ &= \frac{g(t) f_y(y) I(-y \leqslant t \leqslant T_0 - y)}{P(-Y \leqslant T \leqslant T_0 - Y)} \\ &= \left[\frac{g(t) I(-y \leqslant t \leqslant T_0 - y)}{G(T_0 - y) - G(-y)}\right] \cdot \left[\frac{\{G(T_0 - y) - G(-y)\} f_y(y)}{\int \{G(T_0 - u) - G(-u)\} f_y(u) du}\right] \\ &= p_c(t|y) p_y(y) \end{aligned}$$

Under stationary condition when G is uniformly distributed, the marginal density of observed y, $p_y(y)$ becomes $f_y(y)$, which means the density of observed y coincides with its population density and the 'double truncation' from interval sampling does not result in bias on Y. Therefore, the nonparametric estimation of survival function $S_y(y)$ of Y is simply the empirical survival function $\hat{S}_y(y) = \sum_{i=1}^{n} I(\tilde{Y}_i > y)$ where $\tilde{Y}_i$'s are the observed first failure time. $\hat{S}_y(y)$ is the nonparametric maximum likelihood estimator (NPMLE) of $S_y(y)$.

Then, for the second failure time Z, we investigate the probability structure of it and censoring time. We will explore when it remains a representative sample from the target population. We first consider the case that the observation of second failure time ends at calendar time C, where particularly $C = T_0$, a constant. This simple censoring mechanism can be replaced by random censoring C ($C \leqslant T_0$) and we will discuss it later. Let $W = T + Y$ denote the calendar time when the first failure event occurs. Let $\{(min(Z, C - W), I(Z \leqslant$

$C-W)) : C \geqslant W\}$ denote the observed second failure time and the corresponding censoring indicator. A question of interest is whether it is appropriate to apply standard methods to this survival data. It is known that the fundamental requirement for the validity of the usual survival analysis is the independence between Z and $C - W$.

When $(S1)$ holds, the density of Z conditional on $W = w$ is

$$\begin{aligned} p_z(z|w) &= \frac{P(Z = z, Y = w - T)}{P(T + Y = w)} \\ &= \frac{\int_{w-y^+}^{w} f(w - t, z)g(t)dt}{\int\int_{w-y^+}^{w} f(w - t, v)g(t)dtdv} \\ &= f_z(z)\frac{\int_{w-y^+}^{w} f_{y|z}(w - t, z)g(t)dt}{\int\int_{w-y^+}^{w} f(w - t, v)g(t)dtdv} \end{aligned}$$

For each z, $\int_{w-y^+}^{w} f_{y|z}(w - t, z)dt = 1$ and $\int\int_{w-y^+}^{w} f(w - t, v)dtdv = 1$. When $(S2)$ also holds, which means $g(t)$ is a constant, the density $p_z(z|w)$ of observed data is independent of w and equals the population density $f_z(z)$. Given that C is a constant, the above independence of Z and W results in the independence of Z and censoring time $C - W$. This result also extends to random censoring.

Consider the case that the calendar censoring time C is random. Assume that C is independent of (W, Z), that is, the censoring is independent of when the first failure event occurs and the second failure time. From the preceding discussion, we know that under stationary condition, the second failure time is a random sample from the population given that $C \geqslant W$. Let $p_c(z|w)$ be the density of Z given $W = w$ and $C \geqslant W$; then clearly, $p_c(z|w) = p(z|w)$ because C is independent of (W, Z). Then the density of the observed second failure time, $p_c(z|w)$ equals $f_z(z)$ and is independent of w. Further, the failure time Z is independent of the censoring time $C - W$ because Z is independent of (W, C). Therefore, the survival data $\{(min(z_i, c_i - w_i), I(z_i \leqslant c_i - w_i)) : c_i \geqslant w_i\}$ can be treated as the usual right-censored data for inferences of Z and the nonparametric maximum likelihood estimator (NPMLE) for the marginal survival function $S_z(z)$ of Z is the Kaplan-Meier estimator.

4.2 *Copula Model and Two-stage Semiparametric Estimation*

Suppose bivariate failure times (Y, Z) come from the C_α copula for some association parameter α, where C_α is a distribution function with density c_α on $[0,1]^2$, then the joint survival function and density function of (Y, Z) are given by

$$S(y,z) = C_\alpha(S_y(y), S_z(z)), \;\; y,\, z \geqslant 0$$

$$f(y,z) = c_\alpha(S_y(y), S_z(z)) f_y(y) f_z(z), \;\; y,\, z \geqslant 0$$

The 'two-stage' estimating strategy and the conditional likelihood method are used to estimate the association parameter α. Conditional likelihood approaches in statistical literature are sometimes used as a tool to eliminate nuisance parameters. The conditional likelihood preserves most, if not all, of the information for the focused parameters if the conditional statistics are ancillary. For the observed data (t, y, x, δ) where $x = \min(z, c-t-y)$ and $\delta = I(z \leqslant c-t-y)$, the conditional likelihood function of $\{(y, x, \delta)\}$ given $\{t\}$ is

$$L_c(\alpha) = \prod_i \frac{f(y_i, x_i)^{\delta_i} \frac{\partial S(y_i, x_i)^{1-\delta_i}}{\partial y_i}}{S_y(c_i - t_i) - S_y(-t_i)}$$

Clearly, the distribution of T is eliminated by the conditioning procedure. We estimate two margins $S_y(y)$ and $S_z(z)$ by the empirical function $\hat{S}_y(y)$ and the Kaplan-Meier estimator $\hat{S}_z(z)$, respectively. Denote $(S_y(y_i), S_z(x_i))$ by (u_i, v_i) for $i = 1, \ldots, n$. Then given $(u_i, v_i, S_y(c_i - t_i), S_y(-t_i), \delta_i)$, the conditional likelihood of α is

$$L_c(\alpha) \propto \prod_{i=1}^{n} f(y_i, x_i)^{\delta_i} \frac{\partial S(y_i, x_i)^{1-\delta_i}}{\partial y_i} = \prod_{i=1}^{n} c_\alpha(u_i, v_i)^{\delta_i} \frac{\partial C_\alpha(u_i, v_i)^{1-\delta_i}}{\partial u_i} \tag{6}$$

Let $L(\alpha, u_i, v_i)$ denote $c_\alpha(u_i, v_i)^{\delta_i} \frac{\partial C_\alpha(u_i, v_i)^{1-\delta_i}}{\partial u_i}$. The semiparametric estimator $\hat{\alpha}$ for α is the solution to the estimating equation derived from the pseudo conditional likelihood

$$\begin{aligned} U_\alpha(\alpha, \hat{u}, \hat{v}) &= \frac{\partial}{\partial \alpha} \sum_{i=1}^{n} log\, L(\alpha, \hat{u}_i, \hat{v}_i) \\ &= \frac{\partial}{\partial \alpha} \left[\sum_{i=1}^{n} \delta_i log\{c_\alpha(\hat{u}_i, \hat{v}_i)\} + (1-\delta_i) log\{\frac{\partial C_\alpha(\hat{u}_i, \hat{v}_i)}{\partial u_i}\} \right] \\ &= \frac{\partial}{\partial \alpha} \left[\sum_{i=1}^{n} \delta_i log\{c_\alpha(\hat{S}(y_i), \hat{S}(x_i))\} + (1-\delta_i) log\{\frac{\partial C_\alpha(\hat{S}(y_i), \hat{S}(x_i))}{\partial u_i}\} \right] = 0 \end{aligned}$$

Under the regularity conditions stated in the Appendix the estimator of the association parameter has the following asymptotic property.

Property 2. As $n \to \infty$, $\sqrt{n}(\hat{\alpha} - \alpha)$ converges to normal distribution with mean zero and variance $\rho^2 = (\rho_1^2 + \rho_2^2)/\rho_1^4$.

The precise definitions of ρ_1^2 and ρ_2^2, together with the details of the proof can be found in the Appendix.

5. Semi-stationary Model

In section 3 and 4, Both (S1) and (S2) are assumed for the development of the statistical methods. This stationary condition typically holds for stable disease. In this section, we consider the situation when (S2) is violated and only (S1) is valid, and focus on a semi-stationary model based on uncensored data. Specifically, we consider a parametric density function of T, $g(t;\theta)$, where $\theta \in \Theta$ and Θ is an open set in R^k. For example, in cancer studies, g describes the growth of birth cohort for cases. Particular interest is focused on the estimations of parameter θ in $g(t;\theta)$ and joint survival function of (Y, Z). For simplicity, it is assumed that the observation of Z is censored only by the end of the calendar sampling time T_0.

5.1 *Estimation of* θ

The estimation of θ in $g(t;\theta)$ is also complicated by the bias from interval sampling. We explore the sampling bias on the distribution of T here. For given (y, z), the calendar time of the initiating event, t, is observable subject to the constraint $-y \leqslant t \leqslant T_0 - y - z$. Similar to the discussion shown in the formula (2) of section 3, the sampling density of T is generally biased. The conditional likelihood approach is used to estimate the parameter θ in g. When $(S1)$ is assumed, the conditional likelihood function of the observed $\{t\}$ given the observed

$\{(y,z)\}$ is

$$L_c(\theta) = \prod_{i=1}^{n} P_c(t_i|y_i, z_i, \theta) = \prod_{i=1}^{n} \{ \frac{g(t_i;\theta)}{G(T_0 - y_i - z_i;\theta) - G(-y_i;\theta)} \}$$

As a nice feature of this approach, the target parameter θ is the only parameter involved in conditional likelihood and the nuisance parameter $f(\cdot,\cdot)$ is eliminated by the conditioning procedures. The conditional maximum likelihood estimate of θ, denoted by $\hat{\theta}$, can be derived by maximizing $L_c(\theta)$ for $\theta \in \Theta$. Large sample properties of $\hat{\theta}$ can be obtained using techniques similar to those of Andersen (1970) or using techniques for M-estimators (Serfling, 1980). Under regularity conditions and as $n \to \infty$, the estimator $\hat{\theta}$ converges in probability to θ, and $\sqrt{n}(\hat{\theta} - \theta)$ converges weakly to a mean zero multivariate normal distribution with variance-covariance matrix I_c^{-1}, where $I_c = E[\{\frac{\partial}{\partial\theta} log\, p_c(T_i|Y_i, Z_i)\}\{\frac{\partial}{\partial\theta} log\, p_c(T_i|Y_i, Z_i)\}^t]$ is the Fisher information matrix for the conditional likelihood function $L_c(\theta)$.

5.2 *Estimation of Joint Survival Function S(y, z)*

We then study how to estimate the joint survival function of (Y, Z) under semi-stationary condition. The maximum likelihood approach in many situations produces efficient estimators of the model parameters. However, in the current model, the full likelihood function L, $L(\theta, f(\cdot,\cdot)) = L_c(\theta) L_{y,z}(\theta, f(\cdot,\cdot))$, does not factorize into simple terms. Although the maximum likelihood estimator from L is likely to be efficient under regularity conditions, numerical computation and inferences of the estimation procedures could be difficult to derive. Here, a method based on the joint probability structure of T and (Y, Z) is developed for the estimation of $S(y, z)$.

First consider the case when θ is known. As the previous discussion in the formula (2) of section 3, the joint density function $p(y, z)$ of observed uncensored (y, z) can be written as

$$p(y,z) \quad = \quad \frac{\{G(T_0 - y - z;\theta) - G(-y;\theta)\} f(y,z)}{\int\int \{G(T_0 - u - v;\theta) - G(-u;\theta)\} f(u,v) du dv}$$

Thus an estimator of joint survival function of (Y, Z) is

$$\hat{S}(y,z,\theta) = \frac{\sum_{i=1}^{n}\{G(T_0 - Y_i - Z_i;\theta) - G(-Y_i;\theta)\}^{-1} I(Y_i > y, Z_i > z)}{\sum_{i=1}^{n}\{G(T_0 - Y_i - Z_i;\theta) - G(-Y_i;\theta)\}^{-1}}$$

Assume θ is known, just as the similar conclusion in Section 3, the process $\sqrt{n}\{\hat{S}(y,z,\theta) - S(y,z)\}$ converges weakly to a bivariate zero-mean Gaussian process with covariance function

$$\begin{aligned}\sigma^* &= W_{-1}^* W^* \{\frac{\int_y^\infty \int_z^\infty \{G(T_0 - u - v;\theta) - G(-u;\theta)\}^{-1} f(u,v) du dv}{W_{-1}^*}[1 - S(y^{'}, z^{'})] \\ &+ S(y,z)[S(y^{'}, z^{'}) - \frac{\int_{y'}^\infty \int_{z'}^\infty \{G(T_0 - u - v;\theta) - G(-u;\theta)\}^{-1} f(u,v) du dv}{W_{-1}^*}]\}\end{aligned}$$

where $W^* = \int\int\{G(T_0 - u - v;\theta) - G(-u;\theta)\} f(u,v) du dv$, and $W_{-1}^* = \int\int\{G(T_0 - u - v;\theta) - G(-u;\theta)\}^{-1} f(u,v) du dv$.

Now we consider the general case when θ is an unknown parameter. We replace θ in $\hat{S}(y,z,\theta)$ by the conditional maximum likelihood estimator $\hat{\theta}$ and derive an estimator of $S(y,z)$ as $\hat{S}(y,z,\hat{\theta})$. Note that the error of $\hat{S}(y,z,\hat{\theta})$ can be decomposed into two terms:

$$\hat{S}(y,z,\hat{\theta}) - S(y,z) = \{\hat{S}(y,z,\theta) - S(y,z)\} + \{\hat{S}(y,z,\hat{\theta}) - \hat{S}(y,z,\theta)\}$$

where the first error term has been determined by σ^*. The error in the second term is generated by the use of $\hat{\theta}$ for estimating θ. The corresponding distributions of the two terms can be proven to be asymptotically orthogonal to each other because θ in the second term is estimated by the conditional likelihood estimator.

The joint survival function can be estimated by

$$\hat{S}(y,z,\hat{\theta}) = \frac{\sum_{i=1}^{n}\{G(T_0 - Y_i - Z_i;\hat{\theta}) - G(-Y_i;\hat{\theta})\}^{-1} I(Y_i > y, Z_i > z)}{\sum_{i=1}^{n}\{G(T_0 - Y_i - Z_i;\hat{\theta}) - G(-Y_i;\hat{\theta})\}^{-1}} \tag{7}$$

where (Y_i, Z_i)'s are the uncensored bivariate failure times. Again, $S(y,z)$ is identifiable on the domain $\{(y,z) : y + z \leqslant T_0 - t_-\}$. The proposed estimator $\hat{S}(y,z,\hat{\theta})$ has the desired asymptotic property as follow.

Property 3. As $n \to \infty$, the process $\sqrt{n}\{\hat{S}(y,z,\hat{\theta}) - S(y,z)\}$ converges weakly to a bivariate zero-mean Gaussian process with covariance function

$$\Sigma = \nabla_\theta \hat{S}(y,z,\theta)^T I_c^{-1} \nabla_\theta \hat{S}(y,z,\theta) + \sigma^* \tag{8}$$

The detail of the proof of Property 3 can be found in the Appendix.

It is true that the marginal survival function $S_y(y)$ of Y can be estimated directly by $\hat{S}(y, 0, \hat{\theta})$; while, we could also apply the same technique to model (T, Y) and estimate $S_y(y)$ based on the observed (t, y). To be specific, $S_y(y)$ can be estimated by

$$\hat{S}_y(y, \hat{\theta}_*) = \frac{\sum_{i=1}^{n}\{G(T_0 - \tilde{Y}_i; \hat{\theta}_*) - G(-\tilde{Y}_i; \hat{\theta}_*)\}^{-1} I(\tilde{Y}_i > y)}{\sum_{i=1}^{n}\{G(T_0 - \tilde{Y}_i; \hat{\theta}_*) - G(-\tilde{Y}_i; \hat{\theta}_*)\}^{-1}} \tag{9}$$

where $\tilde{Y}_i$'s are the observed first failure time and $\hat{\theta}_*$ is obtained by maximizing the conditional likelihood function of the observed $\{t\}$ given the observed $\{y\}$. It is noticed that $\tilde{Y}_i$'s contain more data points than Y_i's, which are from the uncensored bivariate failure times (Y_i, Z_i), thus the estimate $\hat{S}_y(y, \hat{\theta}_*)$ is expected to be more efficient than $\hat{S}(y, 0, \hat{\theta})$.

The marginal survival function for Z is generally not easy to be estimated under semi-stationary condition due to the induced sampling bias and dependence censoring; however, it is possible to estimate the conditional probability function

$$P(Z > z | y_1 < Y \leqslant y_2) = \frac{S(y_1, z) - S(y_2, z)}{S_y(y_1) - S_y(y_2)}$$

as long as $y + z \leqslant T_0 - t_-$. An estimator of $P(Z > z | y_1 < Y \leqslant y_2)$ is given by $\frac{\hat{S}(y_1, z, \hat{\theta}) - \hat{S}(y_2, z, \hat{\theta})}{\hat{S}_y(y_1, \hat{\theta}_*) - \hat{S}_y(y_2, \hat{\theta}_*)}$. Estimation of such a conditional survival function can be used to detect possible correlation between Y and Z.

6. Simulation Studies

6.1 *Nonparametric Estimation under Stationary Condition*

Two sets of simulations are carried out to assess the finite-sample performance of the nonparametric estimator $\hat{S}(y, z)$ of the joint survival function as well as its variance estimator under stationary condition.

The data $\{(t_1, y_1, z_1), \ldots, (t_n, y_n, z_n)\}$ are generated by the interval sampling scheme in the simulation studies. Let W be a random variable with $Uniform(0, 1)$ distribution and define $T = -13W + 9$. The bivariate failure times (Y, Z) are generated from Clayton's bivariate

survival function $S(y,z) = (S_y(y)^{-\alpha}+S_z(z)^{-\alpha}-1)^{-1/\alpha}$, $\alpha > 0$ with unit exponential margins. The value of the association parameter α is set to 0 and 2 in the first and second sets of simulations, respectively, corresponding to independent and correlated bivariate failure times (Y,Z). An observation (t,y,z) is included (untruncated and uncensored) in the data set if and only if $0 \leqslant t+y \leqslant 10$ and censored if $t+y+z \geqslant 10$. The proportion of untruncated and uncensored observations is around 0.7. In each scenario, 1000 simulated samples are generated, each with 400 subjects.

The findings of the simulations are shown in Table 1. For the joint survival function, the results are given at 16 selected bivariate time points (y,z), where y and z take values $0.2231, 0.5108, 0.9163$ and 1.6094, corresponding to marginal survival probabilities of $0.8, 0.6, 0.4$ and 0.2. Both the point estimator $\hat{S}(y,z)$ and its standard error estimator appear to be unbiased.

[Table 1 about here.]

6.2 *Semiparametric Copula Model under Stationary Condition*

The performance of the two-stage estimator in the semiparametric copula model under stationary condition is examined by simulations. We use unit exponential margins, and choose three values of α in each of the three Archimedean copula models as follows.

Clayton's Family(1978) : $C_\alpha(u,v) = (u^{-\alpha}+v^{-\alpha}-1)^{-1/\alpha}$, $\alpha > 0$. The failure times (Y,Z) are positively associated when $\alpha > 0$ and independent for $\alpha \to 0$.

Positive stable (Hougaard, 1986) : $C_\alpha(u,v) = exp(-[\{-log(u)\}^\alpha + \{-log(v)\}^\alpha]^{1/\alpha})$, $\alpha \geqslant 1$. The failure times (Y,Z) are positively associated when $\alpha > 1$ and independent for $\alpha \to 1$.

Frank's family (1979) : $C_\alpha(u,v) = -\frac{1}{\alpha}log\{1+\frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}\}$, $\alpha \neq 0$. The failure times (Y,Z) are positively associated when $\alpha > 0$, negatively associated when $\alpha < 0$ and independent for $\alpha \to 0$.

Two sampling schemes are explored - the random sampling and the interval sampling. A set of data $\{(t_1, y_1, z_1), \ldots, (t_n, y_n, z_n)\}$ is generated with interval sampling: define $T = -13W + 9$, where W follows $Uniform(0, 1)$ distribution, and let the bivariate failure times (Y, Z) be generated from the three aforementioned copula models. An observation (t, y, z) is included in the data set if and only if $0 \leqslant t + y \leqslant 10$ and censored if $t + y + z \geqslant 10$. For each value of α we generate 1000 simulated samples with $n = 400$.

Table 2 presents simulation results with data generated by random sampling and interval sampling. The mean of the proposed estimates is seen to be quite close to the true value of α. Furthermore, the variances of the estimators are reasonably small for different values of α. For the three models, the proposed method performs quite well for both sampling plans.

[Table 2 about here.]

6.3 *Semi-stationary Model*

Data $\{(t_1, y_1, z_1), \ldots, (t_n, y_n, z_n)\}$ with interval sampling are generated for the semi-stationary model. Define $T = -3W + 10$, where W follows $Exp(\theta)$ distribution, and let the bivariate failure times (Y, Z) be generated from Clayton's bivariate survival function with unit exponential margins. The association parameter α is set to 2 in the simulation, corresponding to the correlated bivariate failure times (Y, Z). An observation (t, y, z) is included in the data set if and only if $0 \leqslant t + y \leqslant 10$ and censored if $t + y + z \geqslant 10$. The proportion of untruncated and uncensored observations is around 0.6. 1000 simulated samples are generated, each with 400 subjects.

For the joint survival function, the results are given at 8 selected bivariate time points (y, z) as shown in Table 3. The conditional likelihood estimate of θ, $\hat{\theta}$, for each generated data set is calculated. The semiparametric estimate of $S(y, z)$, $\hat{S}(y, z, \hat{\theta})$, is calculated using formula (7) in section 5.2. Table 3 gives the simulation results including the Monte Carlo means of $\hat{\theta}$

and $\hat{S}(y, z, \hat{\theta})$, and the standard errors of $\hat{\theta}$ and $\hat{S}(y, z, \hat{\theta})$ based on 1000 replications for each choice of parameter, $\theta = 0.5, 1, 2$.

[Table 3 about here.]

7. Data Analysis: An Application to SEER Cancer Registry Data

7.1 *Analysis under Semi-stationary Condition*

We present an analysis of the ovarian cancer cases in the SEER registry data by the proposed semi-stationary model. The SEER data set used here consists of information from 36728 ovarian cancer patients diagnosed between 1973 and 2002, and the observations are subject to interval sampling. 24236 out of 36728 patients died before Dec 31, 2002. In the analysis of SEER data, the residual lifetime after cancer onset was typically analyzed by standard survival analysis methods and the onset age distribution was empirically estimated. The bias due to interval sampling was commonly ignore in both data analysis and research findings.

In the analysis we assume that the joint distribution of the age of cancer onset and the residual lifetime is independent of the birth time of the study cohort. We apply the proposed method in the semi-stationary model to the SEER ovarian cancer data. In our analysis, the variable T represents the birth time of ovarian cancer patient, Y represents the patient's age of cancer onset, and Z represents residual lifetime. All the variables are analyzed by a continuous scale in years. Figure 2 shows the exploratory plots of ovarian cancer statistics, which includes the kernel density estimate for T, the empirical distribution function estimate for Y and the Kaplan-Meier estimate for Z based on the observed data. The density estimate for T is likely to be biased due to interval sampling, although the y-plot is expected to be close to the true curve when the stationary condition is approximately satisfied. To estimate the t-distribution, we use two polynomial density models for (1) in section 2: a linear model $\phi(t) = c_0 + \theta_1 t$, and a quadratic polynomial model $\phi(t) = c_0 + \theta_1 t + \theta_2 t^2$, where in both

models c_0 is a given positive-valued constant. The joint survival function $S(y, z)$ is estimated based on the semi-stationary model.

[Figure 2 about here.]

In the data the earliest birth is $t = -89.4$ and the latest is $t = 7.30$ with the time origin 0 corresponds to Jan 1, 1973. The model-based density plots of t are shown in Figure 3 (a), where the difference between the linear and quadratic models is considerably small. By comparing the model-based density plots with the kernel density plot, it demonstrates the huge bias in estimating the birth density by the empirical estimate. Interestingly, an increasing trend in birth cohort over the calendar time is found in both models. Such a trend could be explained by the effect of Post-World War II baby boom or the improvement of ovarian cancer screening techniques, or other unclear factors.

[Figure 3 about here.]

Given the small difference of the two models, the linear model is chosen as the birth density in the analysis. The proposed estimates for θ and $S(y, z)$ are calculated, with the corresponding standard error estimates by 500 bootstrap samples. The estimate of θ together with its standard error estimate is $\hat{\theta} = 3.914$ ($s.e. = 0.030$). Table 4 summarizes the proposed estimates for the joint survival function $S(y, z)$ at 9 selected bivariate time points where $y = 62.2, 69.8, 77.5$ years and $z = 0.25, 1.58, 4.58$ years, corresponding to the 1st, 2nd, and 3rd quartiles of the observed age of cancer onset and residual lifetime. The result shows that the joint survival functions are generally underestimated by the empirical estimate, and the magnitude of bias is non-ignorable when it is compared to the proposed estimate. For example, the estimated proportions of patient who was diagnosed later than 62 years old and survived longer than 4.6 years by the empirical estimate and proposed one are 16% vs 22%.

[Table 4 about here.]

It is of interest to study the marginal distribution of age of cancer onset. The estimated marginal distributions by the empirical and proposed method are plotted in Figure 3 (b). The estimated median age of ovarian cancer onset by the proposed method is 77.0 years, older than the observed 69.8 years. The impact of the age of cancer onset on the residual lifetime is explored and demonstrated in Figure 4, by comparing $P(Z > z|y_1 < Y \leqslant y_2)$, the conditional probability functions of residual lifetime giving different onset age subgroups: $(y_1, y_2) = (0, 60)$ for onset age less than or equal to 60 years, $(y_1, y_2) = (60, 70)$ for onset age between 60 and 70 years, and $(y_1, y_2) = (70, \infty)$ for onset age greater than 70 years. It is observed that, given older age of onset, the probability of survival after cancer onset is lower; therefore a negative association between age of onset and residual lifetime is presented. The result is sort of just as we expected because of the biological limitation of the overall lifetime.

[Figure 4 about here.]

Table 4 also provides estimated joint survival distributions by race subgroups. It is shown that the white are likely to be diagnosed at older age and survive longer than the non-white, which is a consistent result with the findings in the literature. Figure 4 also shows the negative associations between age of onset and residual lifetime for both race subgroups: white and non-white. The method provides an exploratory tool to compare the failure time performance for different risk subgroups.

7.2 *Example of Copula Model under Stationary Condition*

In this section we present an example to illustrate the proposed method in the semiparametric copula model considered in section 4, and study the dependency structure of the bivariate survival times (Y, Z).

The SEER ovarian cancer data file consists of 36728 patients diagnosed between 1973 and

2002. We assume constant birth rate of the studying cohort and analyze the data by the proposed semiparametric copula model. The analytical result in section 7.1 shows that there is a negative association between the age of ovarian cancer onset and the residual lifetime. The copula model allows us to quantitatively examine the association. The marginal survival functions of the age of onset and the residual lifetime are estimated by empirical survival function and Kaplan-Meier estimate respectively, and the dependency structure is fitted by copula model of Frank's family. For the overall ovarian cancer patients, the estimated association parameter $\hat{\alpha}$ is -4.747 with 95% bootstrap percentile confidence interval (-4.815, -4.656), and the corresponding estimated Kandall's tau, the rank correlation coefficient, $\hat{\tau}$ is -0.440 with 95% bootstrap percentile confidence interval (-0.445, -0.434), by the formula of $\tau(\alpha) = 1 + \frac{4}{\alpha}(\int_0^\alpha \frac{t}{\alpha(e^t-1)}dt - 1)$ for Frank's family. For those who are white, $\hat{\alpha}$ is -4.790 with 95% bootstrap percentile confidence intervals (-4.874, -4.698), and $\hat{\tau}$ is -0.443 with 95% bootstrap percentile confidence interval (-0.449, -0.437). For those who are non-white, $\hat{\alpha}$ is -4.504 with 95% bootstrap percentile confidence interval (-4.752, -4.248), and $\hat{\tau}$ is -0.424 with 95% bootstrap percentile confidence interval (-0.441, -0.406). The result also suggests a significant negative association between the age of cancer onset and the residual lifetime, for the overall, white and non-white ovarian cancer patients respectively. And the magnitude of association is slightly different between the white and the non-white. While it is noticed that because it often takes a long time for the ovarian cancer occur, the stationary assumption of constant birth rate over time may be inappropriate. Therefore, the future development of semiparametric copula model under semi-stationary condition will be important for SEER ovarian cancer data and other bivariate survival data problems.

8. Concluding Remarks

In collection of registry or surveillance data of a disease, it is common to identify incidence of disease within a calendar time interval and subsequently collect bivariate or multivariate

survival data as end points for progression of the disease. This paper considers statistical issues which arise due to the use of interval sampling, and develops nonparametric and semiparametric methods for bivariate survival data with interval sampling. The copula model approach is proposed to study the dependency structure of the bivariate survival data under stationary condition. However, we recognize that the assumption of stationarity does not always hold, and it will be very interesting to relax the assumption and extend method in the copula model to more general cases for future research.

Moreover, the assessment of risk factors or treatment is crucial in biomedical studies, so it would be worthwhile to develop efficient estimating methods for the regression model for bivariate survival data with interval sampling. The covariates involved in the regression model could be defined at baseline or time-dependent. In sometime applications, information about time-dependent variables would become available only after a certain time point. For example, the treatment information of SEER ovarian or breast cancer patients is provided by SEER-Medicare Link Data (Warren et al., 2002) which were collected from 1986 instead of 1973. Therefore a prevalent sample is involved and this further complicates the analysis. In such model settings, methods need to be developed to address the problems and bias arising from both interval and prevalent sampling. Furthermore, copula model approach could also be extended to accommodate covariates with regression model in the study of the association.

Acknowledgements

References

Andersen, E. B. (1970). Asymptotic properties of conditional likelihood estimators. *Journal of the Royal Statistical Society, Series B (Methodological)* **32,** 283–301.

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

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae* **19,** 194–226.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82,** 543–552.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Annals of Statistics* **9,** 861-869.

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

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* New York: Springer.

Li, Y., Tiwari, R., Guha, S. (2007). Mixture cure survival models dependent censoring. *Journal of the Royal Statistical Society, Series B (Methodological)* **69,** 285–306.

Lin, D.-Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of gap time distributions for serial events with censored data. *Biometrika* **86,** 59-70.

Ries, LAG, Eisner, M. P., Kosary, C. L., Hankey, B. F., Miller, B. A., Clegg, L., Mariotto, A., Feuer, E. J., Edwards, B. K. (eds). *SEER Cancer Statistics Review*, 1975-2002, National Cancer Institute. Bethesda, MD.

Schaubel, D. E. and Cai, J. (2004). Nonparametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine* **23,** 1885–1900.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* New York: Wiley.

Shih, J. H, and Louis, T. A. (1995). Inferences on the association parameters in copula models for bivariate survival data. *Biometrics* **51,** 1384–1399.

Vardi, Y. (1982). Nonparametric estimation in the presence of length-bias. *Annals of Statistics* **10,** 616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics* **13,**

178–203.

Visser, M. (1996). Nonparametric estimation on the bivariate survival function with application to vertically transmitted AIDS. *Biometrika* **83,** 507–518.

Wang, W.-J. and Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* **85,** 561–572.

Warren, J. L., Klabunde, C. N., Schrag, D., Bach, P. B. and Riley, G. F. (2002). Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical Care* **40,** 3–18.

Appendix

Proof of Property 2

Assuming that the joint distribution of (Y, Z) belongs to a copula model family, standard regularity conditions for maximum likelihood estimate hold and functions $W_\alpha(\alpha, S_y(y), S_z(z))$, $V_\alpha(\alpha, S_y(y), S_z(z))$, $V_{\alpha,1}(\alpha, S_y(y), S_z(z))$, and $V_{\alpha,2}(\alpha, S_y(y), S_z(z))$ are continuous and bounded for $(y, z) \in \mathcal{A} = [y_-, y_+] \times [z_-, z_+]$, where

$$W_\alpha(\alpha, S_y(y), S_z(z)) = \frac{\partial logL(\alpha, u, v)}{\partial \alpha}, \quad V_\alpha(\alpha, S_y(y), S_z(z)) = \frac{\partial^2 logL(\alpha, u, v)}{\partial \alpha^2}$$

$$V_{\alpha,1}(\alpha, S_y(y), S_z(z)) = \frac{\partial^2 logL(\alpha, u, v)}{\partial \alpha \partial u}, \quad V_{\alpha,2}(\alpha, S_y(y), S_z(z)) = \frac{\partial^2 logL(\alpha, u, v)}{\partial \alpha \partial v}$$

The above assumptions are used in the proof and the asymptotic normality of $\hat{\alpha}$ can be proved by the techniques outlined below.

Using Taylor expansion on the score function $U_\alpha(\alpha, \hat{S}_y, \hat{S}_z)$ around α_0, rearranging and evaluating it at $\alpha = \hat{\alpha}$, we get

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \cong \frac{-U_\alpha(\alpha_0, \hat{S}_y, \hat{S}_z)/\sqrt{n}}{\sum_{i=1}^n V_\alpha(\alpha_0, \hat{S}_y(Y_i), \hat{S}_z(X_i))/n}$$

Since $\hat{S}_y(\cdot)$ converges in probability to $S_y(\cdot)$ uniformly in $[y_-, y_+]$, $\hat{S}_z(\cdot)$ converges to $S_z(\cdot)$ uniformly in $[z_-, z_+]$, and $V_\alpha(\alpha, u, v)$ is a continous function of u and v, $|V_\alpha(\alpha_0, \hat{S}_y(y), \hat{S}_z(z)) - V_\alpha(\alpha_0, S_y(y), S_z(z))|$ converges in probability to zero for $(y, z) \in \mathcal{A} = [y_-, y_+] \times [z_-, z_+]$. Thus

$\sum_{i=1}^{n} V_\alpha(\alpha_0, \hat{S}_y(Y_i), \hat{S}_z(X_i))/n$ and $\sum_{i=1}^{n} V_\alpha(\alpha_0, S_y(Y_i), S_z(X_i))/n$ are asymptotically equivalent, which by the law of large numbers converges to ρ_1^2, specified as

$$\rho_1^2 = E[-V_\alpha(\alpha_0, S_y(Y_i), S_z(X_i))] = \int_{\mathcal{A}} -V_\alpha(\alpha_0, S_y(y), S_z(z)) dH_{\alpha_0}(y, z, \delta)$$

where H_{α_0} is the joint distribution of (Y, X, δ). Next, we have

$$\begin{aligned}
\sqrt{n}^{-1} U_\alpha(\alpha_0, \hat{S}_y, \hat{S}_z) &= \sqrt{n} \int_{\mathcal{A}} W_\alpha(\alpha_0, \hat{S}_y(y), \hat{S}_z(z)) dH_n(y, z, \delta) \\
&= \sqrt{n} \int_{\mathcal{A}} W_\alpha(\alpha_0, \hat{S}_y(y), \hat{S}_z(z)) dH_{\alpha_0}(y, z, \delta) \\
&+ \sqrt{n} \int_{\mathcal{A}} W_\alpha(\alpha_0, \hat{S}_y(y), \hat{S}_z(z)) (dH_n - dH_{\alpha_0})(y, z, \delta) \\
&= \pi_n(\alpha_0, \hat{S}_y, \hat{S}_z) + \eta_n(\alpha_0, \hat{S}_y, \hat{S}_z) \qquad \text{(A.1)}
\end{aligned}$$

where H_n is the empirical distribution of H_{α_0}. We further decompose η_n into two terms,

$$\begin{aligned}
\eta_n(\alpha_0, \hat{S}_y, \hat{S}_z) &= \sqrt{n} \int_{\mathcal{A}} [W_\alpha(\alpha_0, \hat{S}_y(y), \hat{S}_z(z)) - W_\alpha(\alpha_0, S_y(y), S_z(z))] (dH_n - dH_{\alpha_0})(y, z, \delta) \\
&+ \sqrt{n} \int_{\mathcal{A}} W_\alpha(\alpha_0, S_y(y), S_z(z)) (dH_n - dH_{\alpha_0})(y, z, \delta)
\end{aligned}$$

Because $\hat{S}_y \to S_y$, $\hat{S}_z \to S_z$, $\sqrt{n}(H_n - H) \to O_p(1)$, and W_α is continuous and bounded, by the dominated convergence theorem, the first term in η_n convergence to 0. The second term of η_n is a sum of n i.i.d. random variables of mean zero and variance ρ_1^2, so it converges to normal with mean zero and variance ρ_1^2 by the central limit theorem. Using Von Mises expansion on $\pi_n(\alpha_0, \hat{S}_y, \hat{S}_z)$ around S_y and S_z, we get

$$\begin{aligned}
\pi_n(\alpha_0, \hat{S}_y, \hat{S}_z) &\cong \pi_n(\alpha_0, S_y, S_z) + \sqrt{n} \int IC_y(y) d(\hat{S}_y - S_y)(y) + \sqrt{n} \int IC_z(z) d(\hat{S}_z - S_z)(z) \\
&= 0 + \sqrt{n} \int IC_y(y) d(\hat{S}_y - S_y)(y) + \sqrt{n} \int IC_z(z) d(\hat{S}_z - S_z)(z)
\end{aligned}$$

where IC_y and IC_z are obtained by differentiating $\pi(\alpha_0, (1-\varepsilon_1)S_y + \varepsilon_1 \hat{S}_y, (1-\varepsilon_2)S_z + \varepsilon_2 \hat{S}_z)$ with respect to ε_1 and ε_2 and evaluating at $\varepsilon_1 = \varepsilon_2 = 0$, and $IC_y(y) = -\int_0^y \int_0^{z_0} V_{\alpha,1}(\alpha_0, S_y(u), S_z(z)) h_{\alpha_0}(u, z, \delta) dz du$ and $IC_z(z) = -\int_0^z \int_0^{y_0} V_{\alpha,2}(\alpha_0, S_y(y), S_z(u)) h_{\alpha_0}(y, u, \delta) dy du$. By the counting process asymptotic techniques, $\sqrt{n}(\hat{S}_y(y) - S_y(y))$ is asymptotically equivalent to as a sum of n i.i.d. random variables $\sum_i I_1^0(Y_i)(y)/\sqrt{n}$. Similarly, $\sqrt{n}(\hat{S}_z(z) - S_z(z))$

is asymptotically equivalent to as a sum of n i.i.d. random variables $\sum_i I_2^0(X_i,\delta_i)(z)/\sqrt{n}$. I_1^0 and I_2^0 are martingales, defined as $I_1^0(Y_i)(y) = -S_y(y)[\int_0^y \frac{dN_{1i}(u)}{p(Y\geqslant u)} - \int_0^y \frac{I\{Y_i\geqslant u\}d\Lambda_1(u)}{p(Y\geqslant u)}]$ and $I_2^0(X_i,\delta_i)(z) = -S_z(z)[\int_0^z \frac{dN_{2i}(u)}{p(Z\geqslant u, C_2\geqslant u)} - \int_0^z \frac{I\{X_i\geqslant u\}d\Lambda_2(u)}{p(Z\geqslant u, C_2\geqslant u)}]$ where $C_2 = C - T - Y$, $N_{1i}(u) = I\{Y_i \leqslant u\}$, $N_{2i}(u) = I\{Z_i \leqslant u, \delta_i = 1\}$, and Λ_1 and Λ_2 are the cumulative hazard functions for Y and Z. Then we have

$$\begin{aligned}\pi_n(\alpha_0,\hat{S}_y,\hat{S}_z) &\cong \frac{1}{\sqrt{n}}[\sum_i \int_{\mathcal{A}} V_{\alpha,1}(\alpha_0,S_y(y),S_z(z))I_1^0(Y_i)(y)dH_{\alpha_0}(y,z,\delta)\\ &+ \int_{\mathcal{A}} V_{\alpha,2}(\alpha_0,S_y(y),S_z(z))I_2^0(X_i,\delta_i)(z)dH_{\alpha_0}(y,z,\delta)]\\ &= \frac{1}{\sqrt{n}}[\sum_i I_1(Y_i,\alpha_0) + I_2(X_i,\delta_i,\alpha_0)]\end{aligned}$$

which is a sum of n i.i.d. random variables. Since IC_y and IC_z are deterministic functions, the expectation of I_1 and I_2 are 0. By the central limit theorem, $\pi_n(\alpha_0,\hat{S}_y,\hat{S}_z)$ converges to normal with mean 0 and variance ρ_2^2, specified as

$$\rho_2^2 = E[\{I_1(Y,\alpha_0) + I_2(X,\delta,\alpha_0)\}^2] = \int_{\mathcal{A}}[I_1(y,\alpha_0) + I_2(z,\delta,\alpha_0)]^2 dH_{\alpha_0}(y,z,\delta)$$

Note that we have proved that $\pi_n(\alpha_0,\hat{S}_y,\hat{S}_z)$ is asymptotically equivalent to $\frac{1}{\sqrt{n}}[\sum_i I_1(Y_i,\alpha_0)+ I_2(X_i,\delta_i,\alpha_0)]$, and $\eta_n(\alpha_0,\hat{S}_y,\hat{S}_z)$ is asymptotically equivalent to $\frac{1}{\sqrt{n}}\sum_i W_\alpha(\alpha_0,S_y(Y_i),S_z(X_i))$. π_n and η_n are asymptotically independent as in the proof of Theorem 1 in Shih and Louis (1995). Hence, $\sqrt{n}(\hat{\alpha}-\alpha)$ converges to normal with mean zero and variance $\rho^2 = (\rho_1^2+\rho_2^2)/\rho_1^4$.

Proof of Property 3

It is crucial to study the asymptotic property of $\hat{S}(y,z,\hat{\theta})$. Observe that

$$\sqrt{n}\{\hat{S}(y,z,\hat{\theta}) - S(y,z)\} = \sqrt{n}\{\hat{S}(y,z,\theta) - S(y,z)\} + \sqrt{n}\{\hat{S}(y,z,\hat{\theta}) - \hat{S}(y,z,\theta)\} \quad (A.2)$$

Note that if θ is known, the property of $\hat{S}(y,z,\hat{\theta})$ follows from Vardi (1985) with a θ-involved weight function. As identified in Section 5.2, the process $\sqrt{n}\{\hat{S}(y,z,\theta) - S(y,z)\}$ converges weakly to a bivariate zero-mean Gaussian process with covariance function σ^*. By the counting processes methodology, the first term in $(A.2)$ can be approximated by

$$\sqrt{n}\{\hat{S}(y,z,\theta) - S(y,z)\} = n^{-1/2}\sum_{i=1}^{n}\phi(\theta, Y_i, Z_i, y, z) + o_p(1) \quad (A.3)$$

where $E[\phi(\theta, Y_i, Z_i, y, z)] = 0$ for each θ.

To develop the asymptotic result of the second term in $(A.2)$, the additional variation created by estimating θ by the use of $\hat{\theta}$ needs to be handled. Empirical process and semiparametric inference techniques are employed for the asymptotic properties of the second term in $(A.2)$. $\hat{S}(y,z,\theta)$ can be re-expressed as the empirical process $\hat{S}(y,z,\theta) = n^{-1}\sum_{i=1}^{n} I(Y_i > y, Z_i > z)r(Y_i, Z_i, \theta)$, where $r(Y_i, Z_i, \theta) = \frac{\{G(T_0-Y_i-Z_i;\theta)-G(-Y_i;\theta)\}^{-1}}{\sum_{i=1}^{n}\{G(T_0-Y_i-Z_i;\theta)-G(-Y_i;\theta)\}^{-1}}$. In Section (5.1), it has been shown that $\sqrt{n}(\hat{\theta}-\theta)$ converges in distribution to a mean zero multivariate normal distribution with variance-covariance matrix I_c^{-1}, where $\hat{\theta}$ is the MLE from the conditional likelihood function L_c. Therefore by the functional delta method for the empirical process (Kosorok, 2008), we get $\sqrt{n}\{\hat{S}(y,z,\hat{\theta}) - \hat{S}(y,z,\theta)\} \xrightarrow{D} N(0, \nabla_\theta \hat{S}(y,z,\theta)^T I_c^{-1} \nabla_\theta \hat{S}(y,z,\theta))$. Thus, the second term in $(A.2)$ can be approximated by

$$\begin{aligned}\sqrt{n}\{\hat{S}(y,z,\hat{\theta}) - \hat{S}(y,z,\theta)\} &= n^{-1/2}\nabla_\theta \hat{S}(y,z,\theta)^T I_c^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} logp_c(T_i|Y_i,Z_i) + o_p(1) \\ &= n^{-1/2}\nabla_\theta \hat{S}(y,z,\theta)^T I_c^{-1}\sum_{i=1}^{n}\varphi(T_i,Y_i,Z_i) + o_p(1) \quad (A.4)\end{aligned}$$

where $E[\varphi(T_i,Y_i,Z_i)] = E[\frac{\partial}{\partial\theta} logp_c(T_i|Y_i,Z_i)] = 0$. Combining the preceding results of $(A.3)$ and $(A.4)$, we get

$$\sqrt{n}\{\hat{S}(y,z,\hat{\theta}) - S(y,z)\} \cong n^{-1/2}\sum_{i=1}^{n}\phi(\theta,Y_i,Z_i,y,z) + n^{-1/2}\nabla_\theta \hat{S}(y,z,\theta)^T I_c^{-1}\sum_{i=1}^{n}\varphi(T_i,Y_i,Z_i) \quad (A.5)$$

Also the corresponding distributions of those two terms are asymptotically orthogonal to each other, since

$$E\{\phi(\theta,Y_i,Z_i,y,z)\varphi(T_i,Y_i,Z_i)\} = E\{\phi(\theta,Y_i,Z_i,y,z)E[\frac{\partial}{\partial\theta} logp_c(T_i|Y_i,Z_i)|Y_i,Z_i]\} = 0 \quad (A.6)$$

$(A.5)$ and $(A.6)$ imply that $\sqrt{n}\{\hat{S}(y,z,\hat{\theta}) - S(y,z)\}$ converges weakly to a bivariate zero-mean Gaussian process with covariance function, specified as $\nabla_\theta \hat{S}(y,z,\theta)^T I_c^{-1} \nabla_\theta \hat{S}(y,z,\theta) + \sigma^*$.

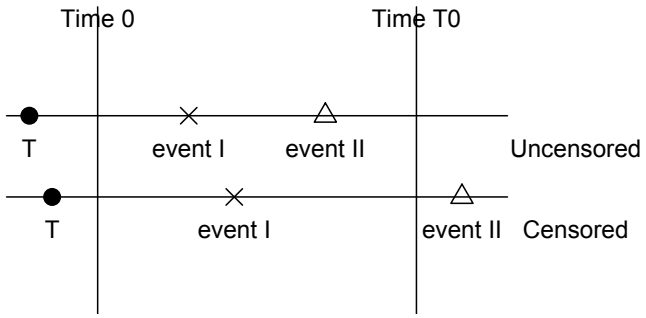

Figure 1. The interval sampling cohort

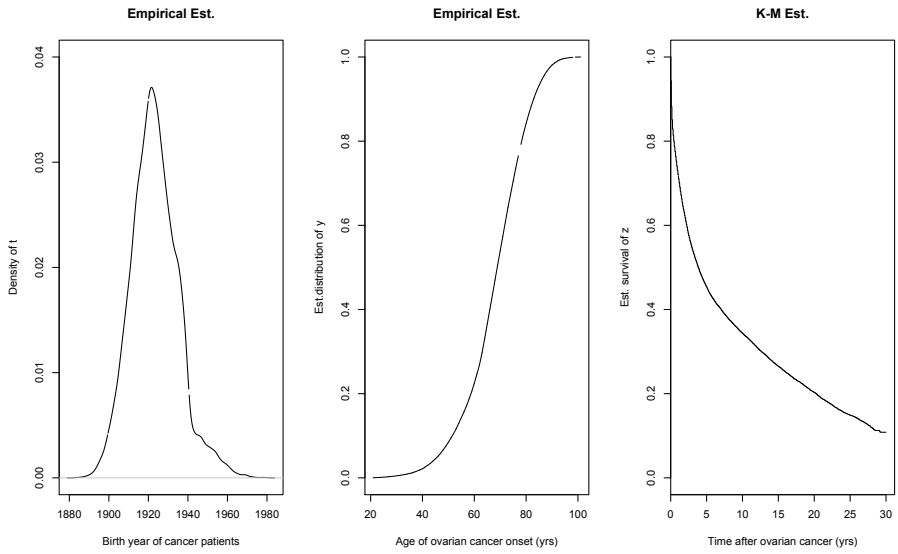

Figure 2. Exploratory plots of ovarian cancer statistics (Biased Estimates).

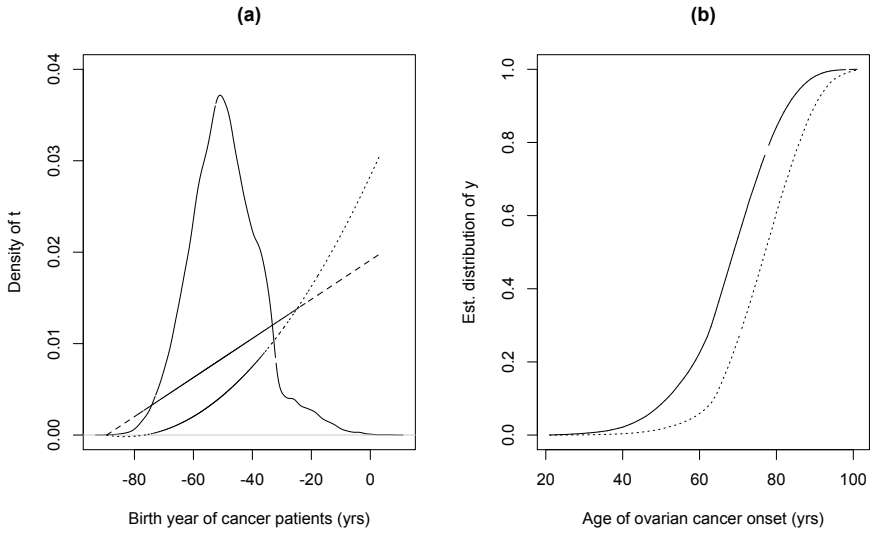

Figure 3. (a) Model-based birth density plots: solid line represents the biased empirical estimate, dashed line represents the estimate from linear model fit, and dotted line represents the estimate from quadratic model fit. (b) Estimated marginal distributions of age of cancer onset: solid line represents the empirical estimate, dotted line represents the proposed bias adjusted estimate.

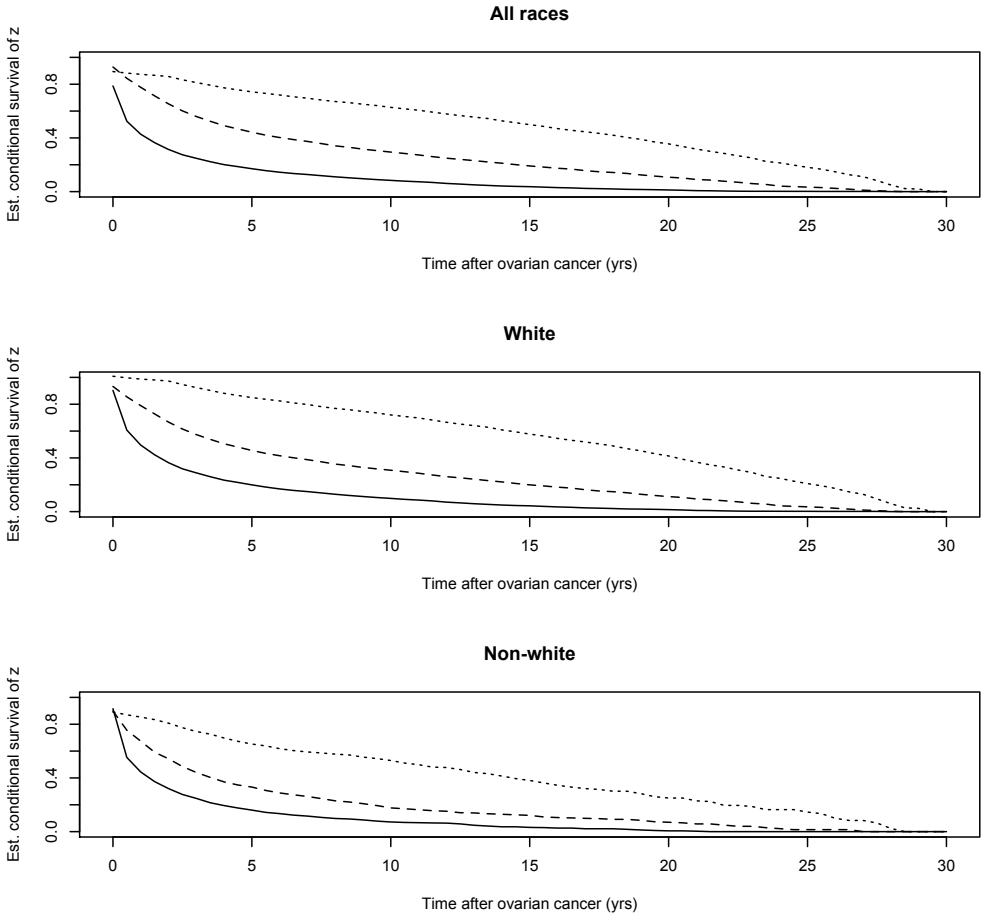

Figure 4. Estimated conditional survival functions of residual lifetime giving different onset age subgroups: solid line represents the subgroup of onset age >70 years, dashed line represents the subgroup of onset age 60-70 years, and dotted line represents the subgroup of onset age $\leqslant$60 years.

Table 1

Simulation summary statistics for $\hat{S}$ under stationary assumption: (a) true joint survival probabilities, (b) empirical means of estimated joint survival probabilities, (c) empirical standard errors of estimated joint survival probabilities, and (d) empirical means of standard error estimates.

	Independent				Correlated			
	z				z			
y	0.2231	0.5018	0.9163	1.6094	0.2231	0.5018	0.9163	1.6094
0.2231	(a)0.640	0.480	0.320	0.160	0.686	0.547	0.383	0.198
	(b)0.638	0.476	0.315	0.156	0.694	0.555	0.389	0.199
	(c)0.030	0.031	0.028	0.022	0.028	0.030	0.030	0.026
	(d)0.030	0.030	0.028	0.023	0.027	0.030	0.030	0.026
0.5108	(a)0.480	0.360	0.240	0.120	0.547	0.469	0.353	0.193
	(b)0.476	0.354	0.233	0.115	0.556	0.476	0.357	0.193
	(c)0.030	0.028	0.025	0.018	0.031	0.030	0.030	0.025
	(d)0.030	0.029	0.025	0.019	0.030	0.031	0.030	0.026
0.9163	(a)0.320	0.240	0.160	0.080	0.383	0.353	0.295	0.182
	(b)0.315	0.234	0.154	0.076	0.386	0.355	0.296	0.180
	(c)0.028	0.025	0.020	0.015	0.031	0.030	0.030	0.025
	(d)0.028	0.025	0.021	0.015	0.030	0.030	0.029	0.025
1.6094	(a)0.160	0.120	0.080	0.040	0.198	0.193	0.182	0.143
	(b)0.156	0.116	0.076	0.039	0.199	0.195	0.183	0.143
	(c)0.022	0.019	0.015	0.011	0.026	0.026	0.026	0.024
	(d)0.022	0.019	0.015	0.011	0.026	0.026	0.025	0.024

Table 2
Simulation summary statistics for $\hat{\alpha}$ under varying sampling schemes, from 1000 samples from Clayton's family , Positive stable frailties and Frank's family.

Model	α	Sampling	$Mean(\hat{\alpha})$	$SD(\hat{\alpha})$
Clayton's family	0.500	Random	0.488	0.093
		Interval	0.480	0.108
	1.333	Random	1.316	0.149
		Interval	1.283	0.179
	3.000	Random	2.946	0.261
		Interval	2.898	0.304
Positive stable frailties	1.250	Random	1.249	0.026
		Interval	1.256	0.058
	1.667	Random	1.667	0.067
		Interval	1.663	0.095
	2.500	Random	2.487	0.103
		Interval	2.478	0.155
Frank's family	2.000	Random	2.014	0.328
		Interval	1.986	0.357
	-1.000	Random	-0.994	0.280
		Interval	-1.010	0.347
	-2.000	Random	-1.977	0.320
		Interval	-1.984	0.372

Table 3
Simulation summary statistics for $\hat{S}$ under semi-stationary assumption: (a) true joint survival probabilities S, (b) means of estimated joint survival probabilities $\hat{S}$, and (c) standard errors of $\hat{S}$.

θ	$Mean(\hat{\theta})$	$SD(\hat{\theta})$	y	z			
				0.2231	0.5018	0.9163	1.6094
0.500	0.506	0.079					
			0.2231	(a)0.686	0.547	0.383	0.198
				(b)0.685	0.544	0.381	0.198
				(c)0.032	0.037	0.041	0.039
			0.9163	(a)0.383	0.353	0.295	0.182
				(b)0.381	0.351	0.294	0.182
				(c)0.041	0.041	0.041	0.038
1.000	1.008	0.090					
			0.2231	(a)0.686	0.547	0.383	0.198
				(b)0.685	0.546	0.381	0.197
				(c)0.037	0.045	0.052	0.054
			0.9163	(a)0.383	0.353	0.295	0.182
				(b)0.381	0.351	0.292	0.180
				(c)0.052	0.054	0.055	0.055
2.000	2.010	0.167					
			0.2231	(a)0.686	0.547	0.383	0.198
				(b)0.679	0.538	0.369	0.181
				(c)0.056	0.076	0.095	0.107
			0.9163	(a)0.383	0.353	0.295	0.182
				(b)0.373	0.342	0.281	0.165
				(c)0.094	0.098	0.103	0.108

Table 4
Analytical result from the SEER ovarian cancer data: (a) Empirical joint survival functions, (b) Proposed estimates of joint survival function $\hat{S}(y, z, \hat{\theta})$, and (c) Standard error estimates of $\hat{S}(y, z, \hat{\theta})$.

	All races			White			Non-white		
	z			z			z		
y	0.25	1.58	4.58	0.25	1.58	4.58	0.25	1.58	4.58
62.2	(a)0.622	0.367	0.160	0.631	0.375	0.163	0.569	0.307	0.121
	(b)0.623	0.402	0.222	0.629	0.407	0.227	0.573	0.344	0.175
	(c)0.049	0.030	0.017	0.053	0.035	0.020	0.003	0.002	0.001
69.8	(a)0.412	0.220	0.093	0.420	0.227	0.093	0.348	0.173	0.064
	(b)0.451	0.268	0.139	0.457	0.273	0.143	0.391	0.219	0.105
	(c)0.034	0.020	0.011	0.039	0.023	0.012	0.001	0.001	0.001
77.5	(a)0.189	0.086	0.034	0.192	0.089	0.034	0.144	0.064	0.022
	(b)0.243	0.128	0.062	0.248	0.130	0.063	0.196	0.100	0.045
	(c)0.018	0.010	0.005	0.021	0.011	0.005	0.002	0.001	0.001