Published in final edited form as:

J Am Stat Assoc. 2018 ; 113(522): 560–570. doi:10.1080/01621459.2017.1356315.

On Estimation of the Hazard Function from Population-based Case-Control Studies

Li Hsu,

Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center

Malka Gorfine, and

Department of Statistics and Operations Research, Tel Aviv University

David M. Zucker

Department of Statistics, Hebrew University

Abstract

The population-based case-control study design has been widely used for studying the etiology of chronic diseases. It is well established that the Cox proportional hazards model can be adapted to the case-control study and hazard ratios can be estimated by (conditional) logistic regression model with time as either a matched set or a covariate (Prentice and Breslow, 1978). However, the baseline hazard function, a critical component in absolute risk assessment, is unidentifiable, because the ratio of cases and controls is controlled by the investigators and does not reflect the true disease incidence rate in the population. In this paper we propose a simple and innovative approach, which makes use of routinely collected family history information, to estimate the baseline hazard function for any logistic regression model that is fit to the risk factor data collected on cases and controls. We establish that the proposed baseline hazard function estimator is consistent and asymptotically normal and show via simulation that it performs well in finite samples. We illustrate the proposed method by a population-based case-control study of prostate cancer where the association of various risk factors is assessed and the family history information is used to estimate the baseline hazard function.

Keywords

Family history; Multivariate survival analysis; Marginal hazard function; Copula model

1. Introduction

The population-based case-control study design has been widely used in epidemiologic studies of chronic diseases such as cancer and coronary heart disease because of its time and cost efficiencies. Under this design, cases (subjects who developed the disease) and controls (subjects who don't have the disease) are ascertained from the population, and their detailed risk factors and family history are recorded. Case-control data have been used to estimate the

---

SUPPLEMENTARY MATERIAL

Title: Asymptotics and Additional Simulation Results

effects of various genetic, lifestyle and environmental factors on diseases. When cases and controls are matched on time, the regression coefficients in the conditional logistic regression model, with matched cases and controls as sets, have the same interpretation as the hazard ratios in the Cox proportional hazards model (Prentice and Breslow, 1978). In common practice, a conventional logistic regression analysis with time as a covariate is used, and this approach usually provides a close approximation, with lower variance (cf. Prentice and Breslow, 1978, Section 5). However, because the ratio of cases and controls is controlled by the investigators, it does not reflect the true disease incidence rate in the population. As a result, the baseline hazard function is not identifiable from the data on cases and controls only. As the baseline hazard function is a critical component for estimating absolute risk, the impossibility of estimating it under the case-control design limits the utility of this design for broader applications.

We propose to utilize the family history information of cases and control to recover the baseline hazard function. Family history information (e.g., disease status and failure times of relatives) is routinely collected, but it is most often used as a risk factor for studying the disease risk. Since the case: control ratio among the relatives is not controlled by the investigators, in principle this information can be used to estimate the population-based hazard function, provided that the failure time distribution of relatives can be assumed to be same as that of the population from which cases and control are drawn. Multivariate survival analysis techniques can be used to jointly model the failure time distribution of the relatives, while taking into account the case-control sampling (Shih and Chatterjee, 2002; Hsu et al. 2004). However, all of the existing approaches require that the same extensive risk factors (e.g., dietary variables, smoking, screening) as collected on the cases and controls also be collected on the relatives. Unfortunately, in the vast majority of case-control studies, the risk factor data on the relatives are not typically collected due to both logistical and cost issues. For these existing approaches to accommodate missing risk factor information on the relatives, a joint distribution of risk factors between cases/controls and relatives is needed. However, such joint distribution is generally not available, except for genotypes where Mendel's law of inheritance can be applied to obtain the joint genotypic distribution of all family members (Chatterjee et al. 2006; Chen et al. 2009). In this paper, we will address this most common situation that no risk factors are collected on relatives, and develop a method for baseline hazard function estimation that use only the failure times and disease status of the relatives, and the risk factor information of cases and controls.

The work is motivated by a large population-based case-control study of prostate cancer conducted at the Fred Hutchinson Cancer Research Center (Stanford et al., 1999). Cases were identified from the Seattle-Puget Sound Surveillance, Epidemiology, and End Results (SEER) cancer registry. Controls were identified by use of random-digit dialing and they were frequency matched to case participants by age (within 5-year age groups). In this study, detailed risk factors were collected on cases and controls, including demographic characteristics, medical history, family history of prostate cancer, screening history, and dietary variables. The information collected on the relatives was the age at diagnosis if the relative had prostate cancer or age at the last observation if the relative did not have prostate cancer. No risk factors were collected on the relatives. The investigators fit a (conditional) logistic regression model to estimate the effect of various risk factors on prostate cancer risk

using case-control data. The goal of this paper is to obtain the baseline hazard function estimate for the model that the investigators fit to the case-control data with various risk factors by making use of the family history information.

Our novel estimation procedure includes two steps. In the first step, we propose to estimate a “composite” hazard function, the hazard function of a group of individuals with mixed risk profiles, based on the failure times and disease status of the relatives. In the second step, through a time-dependent attributable hazard function, we combine the composite hazard function estimate from the first step with the hazard ratio estimates obtained from the case-control data to construct a baseline hazard function estimator. The proposed method addresses the important issue of baseline hazard function estimation from case-control data by making use of routinely collected family history information (e.g., Malone et al. 1996; Stanford et al. 1999). Considering the many case-control studies that have been conducted over the last several decades, and the fact that many genetic and biomarker studies have been re-using these samples, our method offers new possibilities for using case-control data for absolute-risk estimation without resorting to other existing cohorts or establishing new ones. The proposed method is simple and easy to implement. As the inference is based on weighted bootstrap, it can be implemented with minimal modification to the estimation procedure. Our approach offers a practical and rigorous solution to estimating the baseline hazard function from case-control data. To our knowledge, this is the first paper on providing a baseline hazard function estimator based on such family history information that no risk factors are needed for the relatives.

The rest of the paper is organized as follows. The methods and the estimation procedure are described in Section 2, with the large-sample results given in Section 3. The finite-sample performance is evaluated by an extensive simulation study and the results are summarized in Section 4. Section 5 presents the application of the proposed method to the aforementioned prostate cancer study. Some final remarks are presented in Section 6.

2. Methods

2.1. Notation and model

Before describing the case-control data, we first introduce some basic notation and present the model that we consider. Let $(T_0, T_1, \dots, T_M)$ be multivariate failure times (or ages at disease onset) of $M+1$ related individuals, such as family members, on the support $[0, \tau]^{M+1}$, where M is a finite integer and $\tau < \infty$. Assume that $T_m, m=0, \dots, M$, have the same continuous failure-time density function $f(\cdot)$. The hazard function is given by $\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Pr(t \leq T_m < t + \Delta t | T_m \geq t) = f(t)/S(t)$, where $S(t) = \Pr(T_m > t) = \exp\{-\Lambda(t)\}$ with $\Lambda(t) = \int_0^t \lambda(s) ds$. We also assume that the joint survival distribution $S(t_0, t_1, \dots, t_M) = \Pr(T_0 > t_0, T_1 > t_1, \dots, T_M > t_M)$ follows a copula model

$$S(t_0, t_1, \dots, t_M) = h\{S(t_0), S(t_1), \dots, S(t_M); \theta_0\}, \quad (1)$$

for $t_m \in [0, \tau]$ and $m = 0, \dots, M$, where $h(\cdot)$ is a known function on $[0, 1]^{M+1}$, and θ_0 is a vector of unknown parameters that characterize the dependence between the failure times among family members.

Let Z_m be the vector of risk factors of the m th individual, which can include environmental risk factors and genetic variants. The covariate-dependent hazard function is defined as $\lambda(t|z) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Pr(t \leq T_m < t + \Delta t | T_m \geq t, Z_m = z)$, for $m = 0, \dots, M$. We call $\lambda(\cdot|\cdot)$ the “hazard function” and $\lambda(\cdot)$ the “composite hazard function”; the latter can be considered as the overall hazard function for subjects of all risk profiles. We assume that the effect of risk factors on the hazard function follows the Cox proportional hazards model (Cox, 1972)

$$\lambda(t|z) = \lambda_0(t) \exp(\beta_0' z), \quad (2)$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and β_0 is a vector of regression coefficients that quantify the effect of risk factors on the hazard function.

It is common in survival analysis that the failure times are subject to right censoring. Let $C_m, m = 1, \dots, M+1$, denote the corresponding censoring times (or age at last observation) of the $M+1$ related subjects. The observable random variables are then the disease status $\delta_m = I(T_m \leq C_m)$, where $I(\cdot)$ is an indicator function, and the observed time $X_m = T_m \wedge C_m$, the minimum between the failure and censoring times, $m = 0, 1, \dots, M$.

Now consider case-control data with n cases and controls, referred to as *probands* hereafter. For the i th ($i = 1, \dots, n$) proband, let δ_{i0} denote the disease status and X_{i0} the observed time, which is age at disease onset if $\delta_{i0} = 1$ (case), and age at last observation if $\delta_{i0} = 0$ (control). The risk factor vector of the i th proband is denoted by Z_{i0} . Often in epidemiologic studies, the probands are asked for disease outcome information about all of their relatives. We denote the failure times and disease status of the relatives of these probands by $\{(X_{im}, \delta_{im}), i = 1, \dots, n, m = 1, \dots, M\}$ with subscript im denoting the m th relative of the i th proband. Note that no risk factor data are collected on any of the relatives. For simplicity of presentation, we first focus on only one relative for each proband, i.e., $M = 1$, and provide an extension to accommodate general family structures at the end of Section 2.2.

In this paper, we are interested in not only the log hazard ratios, β , but also the baseline hazard function, $\lambda_0(\cdot)$, in (2). We briefly describe the idea and then present the estimation procedure in the following sections. It has been well established that β can be estimated consistently by maximizing the conditional logistic regression likelihood of matched case-control data (Prentice and Breslow, 1978). In contrast, $\lambda_0(\cdot)$ cannot be identified from case-control data only because the sampling is conditional on the disease status. However, the possibly censored failure times of the relatives are random, suggesting that if we account for the sampling scheme of the probands, we can make use of the relatives' failure time information to estimate the hazard function. Since there is no risk factor information on the relatives, we can obtain a consistent estimator only for the composite hazard function $\lambda(\cdot)$ with no covariates. But the baseline and composite hazard function are related through the equation $\lambda_0(t) = \lambda(t)\varnothing(t)$, where $\varnothing(\cdot)$ is one minus the attributable hazard function, which

can be consistently estimated from case-control data. In the following sections, we first describe the method for estimating the composite hazard function $\lambda(\cdot)$, and then the method for the attributable hazard function $\lambda_0(\cdot)$, and hence $\lambda_0(\cdot)$.

2.2 Estimation of composite hazard function $\lambda(t)$

The key to estimating $\lambda(\cdot)$ is that both the dependence of failure times among relatives and probands and the case-control ascertainment need to be taken into account. Assuming that the joint survival distribution of relatives and proband follows the copula model (1), we can derive the hazard functions of the relatives conditional on the probands' disease status and observation times.

Specifically, define $h^{(10)}(\cdot, \cdot)$, $h^{(01)}(\cdot, \cdot)$ and $h^{(11)}(\cdot, \cdot)$ to be the respective partial derivative of $h(\cdot, \cdot)$ with respect to the first argument but not the second, the second argument but not the first, and both first and second arguments. The survival function of a relative of a proband diseased at X_0 is $S(t|T_0 = X_0) \equiv \Pr(T_1 > t|T_0 = X_0) = h^{(10)}\{S(X_0), S(t); \theta_0\}$, and the survival function of a relative of a proband disease-free at X_0 is $S(t|T_0 > X_0) \equiv \Pr(T_1 > t|T_0 > X_0) = h\{S(X_0), S(t); \theta_0\}/S(X_0)$. Thus, the conditional hazard functions of a relative of a case or a control proband can be written, respectively, as

$$\lambda(t|T_0 = X_0) = \frac{h^{(11)}\{S(X_0), S(t); \theta_0\}}{h^{(10)}\{S(X_0), S(t); \theta_0\}} S(t)\lambda(t),$$

and

$$\lambda(t|T_0 > X_0) = \frac{h^{(01)}\{S(X_0), S(t); \theta_0\}}{h\{S(X_0), S(t); \theta_0\}} S(t)\lambda(t).$$

Let $\mathcal{R}\{S(X_0), S(t), \delta_0; \theta_0\} = \left[h^{(11)}\{S(X_0), S(t); \theta_0\} / h^{(10)}\{S(X_0), S(t); \theta_0\} \right]^{\delta_0} \left[h^{(01)}\{S(X_0), S(t); \theta_0\} / h\{S(X_0), S(t); \theta_0\} \right]^{1-\delta_0}$. We can then unify the presentation of the hazard function for the relative of a case or a control proband by

$$\lambda(t|X_0, \delta_0) = \lambda(t)\mathcal{R}\{S(X_0), S(t), \delta_0; \theta_0\}S(t).$$

This hazard function has some resemblance to the Cox model in the sense that $\lambda(\cdot)$ may be considered as the “baseline” hazard function and the remaining term as a time-dependent “risk function” in t . There is a complication, however, with this risk function: $\mathcal{R}\{S(X_0), S(t), \delta_0; \theta_0\}$ is not predictable at time t because it involves $S(X_0)$, and X_0 can be greater than t . To circumvent this unpredictability problem, we propose to use a two stage estimator in the spirit of Gorfine et al. (2009). Specifically, define counting processes $N_{i1}(t) = I(X_{i1} \leq t, \delta_{i1} = 1)$ and $Y_{i1}(t) = I(X_{i1} > t)$, for $i = 1, \dots, n$. The first-stage estimator, for a given value of θ , is defined as a step function whose jumps are given by

$$\Delta\tilde{\Lambda}(t) = \frac{\sum_{i=1}^n I(X_{i0} < t) N_{i1}(dt)}{\sum_{i=1}^n Y_{i1}(t) I(X_{i0} < t) \mathcal{R}\{\tilde{S}(X_{i0}), \tilde{S}(t-), \delta_{i0}; \theta\} \tilde{S}(t-)} \quad (3)$$

with $\tilde{S}(t) = \exp\{-\tilde{\Lambda}(t)\}$. The difference between $\tilde{\Lambda}(\cdot)$ and the usual Breslow-type estimator is that we include only relatives whose probands' observation time is less than time t , and thereby avoid the problem of unpredictability of the risk function. The first-stage estimator can then be estimated by an non-iterative procedure. Suppose $0 = t_{(0)} < t_{(1)} < t_{(2)} < \dots < t_{(n')}$ are n' ordered distinct observed failure times in relatives. The estimation procedure is described as follows:

- Step 1: The survival function at time origin $\tilde{S}(0) = 1$. By equation (3), the jump size at the first observed failure time $t_{(1)}$ is

$$\Delta\tilde{\Lambda}(t_{(1)}) = \frac{\sum_{i=1}^n I(X_{i0} < t_{(1)}) N_{i1}(dt_{(1)})}{\sum_{i=1}^n Y_{i1}(t_{(1)}) I(X_{i0} < t_{(1)}) \mathcal{R}\{\tilde{S}(0), \tilde{S}(0), \delta_{i0}; \theta\}},$$

and the cumulative hazard function $\tilde{\Lambda}(t_{(1)}) = \Delta\tilde{\Lambda}(t_{(1)})$.

- Step 2: At the second observed failure time $t_{(2)}$, the jump size is

$$\Delta\tilde{\Lambda}(t_{(2)}) = \frac{\sum_{i=1}^n I(X_{i0} < t_{(2)}) N_{i1}(dt_{(2)})}{\sum_{i=1}^n Y_{i1}(t_{(2)}) I(X_{i0} < t_{(2)}) \mathcal{R}\{\tilde{S}(X_{i0}), \tilde{S}(t_{(1)}), \delta_{i0}; \theta\}},$$

where $\tilde{S}(t_{(1)}) = \exp\{-\tilde{\Lambda}(t_{(1)})\}$. Then the cumulative hazard function $\tilde{\Lambda}(t_{(2)}) = \tilde{\Lambda}(t_{(1)}) + \Delta\tilde{\Lambda}(t_{(2)})$.

- The non-iterative procedure continues for the rest of the observed failure times $t_{(3)}, \dots, T_{(n')}$. This yields the first stage estimator $\tilde{\Lambda}(\cdot)$.

However, while restricting $I(X_{i0} < t)$ allows us to estimate $\Lambda(\cdot)$ non-iteratively, this restriction causes a loss in efficiency (see Table S1 in Section D of the Supplementary Materials). We therefore follow up with a second-stage estimator to recover efficiency. The second-stage estimator for a given value of θ is defined as

$$\Delta\hat{\Lambda}(t) = \frac{\sum_{i=1}^n N_{i1}(dt)}{\sum_{i=1}^n Y_{i1}(t) \mathcal{R}\{\hat{S}^\dagger(X_{i0}), \hat{S}(t-), \delta_{i0}; \theta\} \hat{S}(t-)} \quad (4)$$

Where $\hat{S}(t) = \exp\{-\hat{\Lambda}(t)\}$ and $\hat{S}^\dagger(X_{i0}) = \exp\{-\hat{\Lambda}^\dagger(X_{i0})\}$ with $\hat{\Lambda}^\dagger(X_{i0}) = \tilde{\Lambda}(X_{i0})$ and $X_{i0} < t$ and $\hat{\Lambda}(X_{i0})$ if $X_{i0} < t$. In other words, $\hat{\Lambda}(\cdot)$ on the right-hand side of (4) depends on $\hat{\Lambda}(s)$ values

only for event times prior to time t , and when $X_{i0} = t$, the first-stage estimator $\tilde{\Lambda}(\cdot)$ is used. Therefore, similar to the first-stage estimator, $\hat{\Lambda}(t)$ can also be estimated without any iteration starting from $t_{(1)}$ successively until $t_{(n)'}^*$, where the jump sizes at these observed failure times are defined in (4).

The above $\hat{\Lambda}(t)$ involves the unknown dependence parameter θ . We estimate θ by the maximum likelihood estimator (MLE) based on the conditional distribution of the failure times of the relatives given those of the probands. Since the distribution of the data being conditioned upon does not involve θ , we can estimate θ by solving the score equations based on only the log-likelihood of the joint failure time distribution of relatives and probands without being concerned with the case-control sampling. The score function can be written as

$$\sum_{i=1}^n \frac{\partial}{\partial \theta} \left[\delta_{i0} \delta_{i1} \log h^{(11)}\{S(X_{i0}), S(X_{i1}); \theta\} + (1 - \delta_{i0}) \delta_{i1} \log h^{(01)}\{S(X_{i0}), S(X_{i1}); \theta\} + \delta_{i0} (1 - \delta_{i1}) \log h^{(10)}\{S(X_{i0}), S(X_{i1}); \theta\} + (1 - \delta_{i0}) (1 - \delta_{i1}) \log h\{S(X_{i0}), S(X_{i1}); \theta\} \right].$$

Since the estimation of $\Lambda(\cdot)$ involves θ and the estimation of θ involves $\Lambda(\cdot)$, the two estimators need to be iterated until both converge.

The above approach can be extended to a general family structure. We follow the idea of generalized estimating equations of Liang and Zeger (1986), and decompose each family into multiple relative-proband pairs. Suppose there are M relatives in the i th family and the dependence between the m th relative and the proband is captured by θ_m , $m = 1, \dots, M$. Then, the jump size for the first-stage composite hazard function estimator for a given value of $\theta = (\theta_1, \dots, \theta_M)$ becomes

$$\Delta \tilde{\Lambda}(t) = \frac{\sum_{i=1}^n \sum_{m=1}^M I(X_{i0} < t) N_{im}(dt)}{\sum_{i=1}^n \sum_{m=1}^M Y_{im}(t) I(X_{i0} < t) \mathcal{R}\{\tilde{S}(X_{i0}), \tilde{S}(t-), \delta_{i0}; \theta_m\} \tilde{S}(t-)}.$$

The second-stage composite hazard function estimator for a given value of θ is then

$$\Delta \hat{\Lambda}(t) = \frac{\sum_{i=1}^n \sum_{m=1}^M N_{im}(dt)}{\sum_{i=1}^n \sum_{m=1}^M Y_{im}(t) \mathcal{R}\{\hat{S}^\dagger(X_{i0}), \hat{S}(t-), \delta_{i0}; \theta_m\} \hat{S}(t-)} \quad (5)$$

Different family sizes are allowed and can be easily accommodated by setting the censoring times of missing relatives to 0. The dependence parameters θ can be estimated by maximizing a composite likelihood function, which is the product of the likelihood functions

of relative-proband pairs, or solving the estimating equations $U(\theta, \Lambda) = 0$ based on the partial derivative $U(\theta, \Lambda)$ of the log-composite likelihood with respect to θ , given by

$$U(\theta, \Lambda) = \sum_{i=1}^n \sum_{m=1}^M \frac{\partial}{\partial \theta} \left[\delta_{i0} \delta_{im} \log h^{(11)}\{S(X_{i0}), S(X_{im}); \theta_m\} + (1 - \delta_{i0}) \delta_{im} \log h^{(01)}\{S(X_{i0}), S(X_{im}); \theta_m\} + \delta_{i0} (1 - \delta_{im}) \log h^{(10)}\{S(X_{i0}), S(X_{im}); \theta_m\} + (1 - \delta_{i0}) (1 - \delta_{im}) \log h\{S(X_{i0}), S(X_{im}); \theta_m\} \right]. \quad (6)$$

If several relatives (e.g. sisters) share the same relation to the proband, i.e., a subset of the θ s are equal, a joint distribution of all of these relatives may be considered to increase the efficiency of the dependence parameter estimators (Hsu and Gorfine, 2006). Estimation of θ and $\Lambda(\cdot)$ is iterated until both estimators converge.

2.3. Estimation of the baseline hazard function $\lambda_0(\cdot)$

To estimate $\lambda_0(\cdot)$, we note that

$$\lambda_0(t) = \phi(t) \lambda(t), \quad (7)$$

where $\phi(\cdot)$ is one minus the attributable hazard function (Chen et al., 2006). This relation was also noted by Gail et al. (1989). In the previous section, we presented a two-stage estimator for $\lambda(\cdot)$ using only the failure time data of the relatives. Hence, in order to estimate $\lambda_0(\cdot)$, it remains to estimate $\phi(\cdot)$, and this will be done using the case-control data.

Let $f(t|z)$ and $S(t|z)$ denote the conditional density and survival function of $T_0 = t$ given the covariates $Z_0 = z$, respectively. Then

$$S(t) = \int S(t|z) f(z) dz = \int f(z, t) / \lambda(t|z) dz = f(t) \int f(z|t) / \lambda(t|z) dz$$

and thus

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{1}{\int f(z|t) / \lambda(t|z) dz}.$$

By definition, $\phi(t) = \lambda_0(t) / \lambda(t)$. It is easy to see that

$$\phi(t) = \int \frac{\lambda_0(t)}{\lambda(t|z)} f(z|t) dz = \int \exp(-\beta' z) f(z|t) dz.$$

The second equality follows from the Cox model (2) for $\lambda(t|Z)$. Examining the representation in the last equation, it is clear that both β and $\lambda(t)$ can be estimated from case-control data. We now discuss the estimation procedure in detail.

Estimation of the coefficient vector β involves standard methodology for case-control studies. In line with common practice, we propose to estimate β by a conventional logistic regression analysis adjusting for age, i.e., we maximize

$$L_{\text{logistic}}(\beta, \gamma) = \prod_{i=1}^n \left[\exp(\beta' Z_{i0} + \gamma X_{i0})^{\delta_{i0}} \{1 + \exp(\beta' Z_{i0} + \gamma X_{i0})\}^{-1} \right] \quad (8)$$

as a function of β and γ , and then take $(\hat{\beta}, \hat{\gamma})$ to be the maximizer (cf. Prentice and Breslow, 1978, Section 5). Defining $\ell(\beta, \gamma) = n^{-1} \log L_{\text{logistic}}(\beta, \gamma)$, we can write

$$\ell(\beta, \gamma) = n^{-1} \sum_{i=1}^n \ell^o(\beta, \gamma, X_{i0}, Z_{i0}) \text{ with } \ell^o(\beta, \gamma, t, z) = \delta_{i0}(\beta' z + \gamma t) - \log\{1 + \exp(\beta' z + \gamma t)\}.$$

Since the model for age may be misspecified, this approach involves some bias; that is, the estimator $\hat{\beta}$ converges to a limit $\bar{\beta}$ that in general will differ from the true coefficient vector β_0 . In our experience, based on extensive simulations in a variety of settings, the bias is typically very small. The bias can be reduced by expanding the model for the effect of age. An alternative would be to carry out an age-matched analysis (Prentice and Breslow, 1978, Equation (2)), but this results in a substantial increase in variance.

Regarding estimation of $\lambda(\cdot)$, a straightforward approach is to estimate $\lambda(t)$ by the average of $\exp(-\hat{\beta}' Z_{i0})$ among cases who failed at time t . However, the number of cases who failed at each time t is typically small. As a result, such an empirical estimator is inevitably unstable and inefficient. Borrowing information at neighboring time points will help improve the stability and efficiency of the estimator.

Towards this end, we propose to use a weighted local linear estimator for $\varphi(\cdot)$ (Fan, 1993). Let $K(\cdot)$ be a symmetric kernel function and put $K_h(t) = K(t/h)/h$, where $h > 0$ is a bandwidth of order $O(n^{-\nu})$ with $\nu \in (\frac{1}{4}, \frac{1}{2})$ (but we will suppress this dependence of h on n from the notation). Define $(\hat{b}_0, \hat{b}_1)$ be the value of (b_0, b_1) that minimizes the objective function

$$\phi(t; b_0, b_1) = \sum_{i=1}^n \delta_{i0} K_h(t - X_{i0}) \left[\exp(-\hat{\beta}' Z_i) - \{b_0 + b_1(t - X_{i0})\} \right]^2.$$

The local linear estimator of $\varphi(t)$ is then defined by $\hat{\phi}(t) = \hat{b}_0$. With $w_h(t) = \delta_{i0} K_h(t - X_{i0})$, we can write

$$\hat{\phi}(t) = \sum_i c_i(t) \exp(-\hat{\beta}' Z_i), \quad (9)$$

where

$$c_i(t) = \frac{w_i(t)}{\sum_j w_j(t)} - \frac{\bar{U}_w(t) w_i(t) \{U_i(t) - \bar{U}_w(t)\}}{\sum_j w_j(t) \{U_j(t) - \bar{U}_w(t)\}^2}$$

with $U_i(t) = t - X_{i0}$ and $\bar{U}_w(t) = \sum_i w_i(t) U_i(t) / \sum_i w_i(t)$.

Now plugging $\hat{\phi}(t)$ and $\hat{\Lambda}(t)$ into (7), we obtain the following estimator of $\Lambda_0(\cdot)$:

$$\hat{\Lambda}_0(t) = \int_0^t \hat{\phi}(u) \hat{\Lambda}(du). \quad (10)$$

There is a rich literature about kernel function and bandwidth selection, see e.g., Wand and Jones (1995) and Fan and Gijbels (1996). In the simulation we used the biweight kernel function $K(u) = (15/16)(1 - u^2)^2 I(|u| \leq 1)$ and cross-validation to select the bandwidth, and the estimators performed very well under a wide range of scenarios.

The large sample properties are presented in Section 3. Due to the complicated form of the asymptotic variance, especially for the composite hazard function estimator, we propose to use the weighted bootstrap for inference, which is easy to implement and may be preferred for data with small to medium sample sizes. Furthermore, as parallel computing becomes standard, the computing intensiveness of bootstrap techniques is no longer an issue. A theorem providing asymptotic justification for the weighted bootstrap is presented in the next section. Here we describe the procedure. For $b = 1, \dots, B$ bootstrap samples, generate $v_1^{(b)}, \dots, v_n^{(b)}$ i.i.d. from a known distribution with both mean and variance equal to 1, e.g., $\exp(1)$. For any given function g , the empirical sum $\sum_{i=1}^n g_i$ is replaced with its corresponding weighted empirical sum $\sum_{i=1}^n v_i^{(b)} g_i$ to produce the estimator $\hat{\Lambda}_0^{(b)}$. The variance estimator is then the empirical variance estimator calculated from $\Lambda_0^{(1)}, \dots, \Lambda_0^{(B)}$.

Taking everything together, we summarize the estimation procedure as follows.

1. Estimate $\hat{\Lambda}(\cdot)$ at the distinct failure times of the relatives using the two-stage estimator (5), and the dependence parameter θ by solving the equation $U(\theta, \Lambda) = 0$ (with $U(\theta, \Lambda)$ as in (6)), iterating back and forth.
2. Maximize the logistic likelihood (8) to obtain $\hat{\beta}$.
3. Obtain the attributable hazard function estimator $\hat{\phi}(\cdot)$ from (9).
4. Obtain $\hat{\Lambda}_0(\cdot)$ from (10).
5. Calculate the weighted bootstrap variance estimator.

3. Large sample properties

In this section, we state the consistency and asymptotic normality results for $\hat{\Lambda}_0(\cdot)$ and briefly describe the proofs. Let $X_j = \{(X_{j0}, \delta_{j0}, Z_{j0}), (X_{ij}, \delta_{ij}), j = 1, \dots, M\}$. The families are allowed to have different sizes, which is achieved by setting the respective censoring time to 0 whenever a subject is missing. Let $\mathcal{J}_k$, $k = 0, 1$, denote the set of families with $\delta_{0j} = k$, and let n_k be the number of such families. We assume that the probands are randomly selected from the diseased (cases) and non-diseased (controls) subpopulations, and that n_1 is equal to the integer part of αn for a fixed constant $\alpha \in (0, 1)$. Thus, for $k = 0, 1$, the X_j 's of the families in $\mathcal{J}_k$ are i.i.d., and the data on the $\mathcal{J}_0$ families are independent of the data on $\mathcal{J}_1$ families. The censoring mechanism is assumed to satisfy the independent censorship condition, i.e., the censoring times are independent of the survival times. For practical purposes we assume Z is bounded. Additional technical assumptions are listed in the Appendix.

The asymptotic theory involves quantities of the form

$$\frac{1}{n} \sum_{i=1}^n R(X_i, t) \quad (11)$$

where $E[Y(X_i, t)|\delta_{i0}] = 0$. Such quantities can be re-written as

$$\frac{n_0}{n} \left[\frac{1}{n_0} \sum_{i=1}^{n_0} (1 - \delta_{i0}) R(X_i, t) \right] + \frac{n_1}{n} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{i0} R(X_i, t) \right]$$

where each of the bracketed terms is an average of mean-zero i.i.d. random variables and the bracketed terms are independent. This representation allows us to apply known empirical process theory.

It follows from maximum likelihood theory (Prentice & Pyke, 1979; White, 1982; van der Vaart, 1998, Chapter 5) that $(\hat{\beta}, \hat{\gamma})$ converges almost surely to a limit $(\bar{\beta}, \bar{\gamma})$ and is asymptotically normal. The quantity $\bar{\beta}$ appears in the theorems stated below.

Theorem 1. Assume Conditions C1–C7 in the Appendix. For a general β , define

$$\phi(t, \beta) = E\{\exp(-\beta' Z_{i0}) | X_{i0} = t, \delta_{i0} = 1\} = \int \exp(-\beta z) f(z|t) dz \quad (12)$$

$$\Lambda_0(t, \beta) = \int_0^t \frac{\phi(t, \beta)}{\phi(t, \beta_0)} \lambda_0(t) dt \quad (13)$$

Then the baseline hazard function estimator $\widehat{\Lambda}_0(\cdot)$ in (10) converges uniformly in probability to $\Lambda_0(\cdot, \bar{\beta})$, and $n^{1/2}\{\widehat{\Lambda}_0(\cdot) - \Lambda_0(\cdot, \bar{\beta})\}$ converges weakly to a Gaussian random process.

The outline of the proof is as follows. Letting $\tilde{\phi}$ denote the version of $\hat{\phi}$ with $\hat{\beta}$ replaced by $\bar{\beta}$, we can write

$$\begin{aligned}\widehat{\Lambda}_0(t) - \Lambda_0(t) &= \int_0^t \phi(u, \bar{\beta}) \{\widehat{\Lambda}(du) - \Lambda(du)\} + \int_0^t \{\tilde{\phi}(u) - \phi(u, \bar{\beta})\} \Lambda(du) \quad (14) \\ &+ \int_0^t \{\hat{\phi}(u) - \tilde{\phi}(u)\} \Lambda(du) + \int_0^t \{\hat{\phi}(u) - \phi(u, \bar{\beta})\} \{\widehat{\Lambda}(du) - \Lambda(du)\}\end{aligned}$$

In the Appendix we show that $\widehat{\Lambda}(\cdot)$ converges uniformly in probability to $\Lambda(\cdot)$, using similar arguments in Gorfine et al. (2009) and Spiekerman and Lin (1998). We show further that the first term in (14) can be approximated by a quantity of the form (11), and that $n^{1/2}$ times this term converges weakly to a Gaussian random process. We show next that the second term in (14) can also be approximated by a quantity of the form (11), and that $n^{1/2}$ times this term converges weakly to a Gaussian random process. The third term is driven by $\hat{\beta} - \bar{\beta}$, which is asymptotically equivalent to a quantity of the form (11) and asymptotically normally distributed. Finally, using martingale arguments, we show that the last term of (14) is $o_p(n^{-1/2})$. All this leads to a representation of $\widehat{\Lambda}_0(t) - \Lambda_0(t, \bar{\beta})$ as asymptotically equivalent to a quantity of the form (11) which converges weakly to a Gaussian process, with weak convergence defined in terms of the uniform metric. We thus obtain the desired result, finding in the process that the class of functions $Y^* = \{Y(\cdot, t), t \in [0, \tau]\}$, for both the data on the $\mathcal{J}_0$ families and the data on the $\mathcal{J}_1$ families, is a Donsker class.

Theorem 2. Let $v_1, \dots, v_n$ be n i.i.d. realizations of positive random weights generated from a known distribution with $E(v) = 1$ and $\text{var}(v) = 1$. The weighted bootstrap estimator $\widehat{\Lambda}_0^*(t)$ is obtained by replacing all empirical sums by weighted sums with weights $(v_1, \dots, v_n)$. Assuming Conditions C1–C7 in the Appendix, the asymptotic conditional distribution of $n^{1/2}\{\widehat{\Lambda}_0^*(t) - \widehat{\Lambda}_0(t)\}$ given the observed data is the same as the asymptotic distribution $n^{1/2}\{\widehat{\Lambda}_0(t) - \Lambda_0(t, \bar{\beta})\}$.

Since we have just stated that Y^* is Donsker, this result for the weighted bootstrap follows directly from Kosorok (2008, Theorem 2.6).

4. SIMULATION

We conducted an extensive simulation study to evaluate the finite sample performance of the proposed approach. To mimic the case-control study design, we first generated a large population and then selected cases and age-matched controls. Specifically, for each subject, we generated two covariates G and Z with the intention that G represents a genotype and Z represents a continuous environmental covariate. Given G and Z , the failure time followed the Cox proportional hazards model

$$\lambda(t; G, Z) = \lambda_0(t) \exp(\beta_1 G + \beta_2 Z), \quad (15)$$

where β_1 and β_2 are the log hazard ratios of G and Z , respectively. For each subject, we then generated the data on the relatives, allowing for possible dependence of the failure times among family members even after conditioning on G and Z . This is to account for many unobserved shared genetic and environmental risk factors that may cause the failure times of family members to be correlated. For the joint distribution, we used the ClaytonOakes model (Clayton, 1978; Oakes, 1989), the most commonly used copula model, in which

$$\Pr(T_0 > t_0, \dots, T_M > t_M) = \left\{ \sum_{m=0}^M \exp\{\theta_0 \Lambda_0(t_m) \exp(\beta_1 G + \beta_2 Z)\} - M + 1 \right\}^{-\frac{1}{\theta_0}}. \quad (16)$$

Specifically, we generated the data according to the following algorithm:

1. Generate G for both parents from a multinomial distribution with probabilities $(1-p)^2$, $2p(1-p)$, and p^2 , where p is the minor allele frequency. Generate the genotypes of the offspring by the Mendelian law, i.e., each allele has an equal chance to be transmitted. Generate the environmental covariate Z from a multivariate normal distribution with zero mean and fixed pairwise correlation coefficient ρ .
2. Generate the failure times of family members according to the Clayton-Oakes model(16) by using the conditional hazard function approach (Glidden and Self, 1999). Specifically, we assumed the baseline hazard function followed the Weibull distribution with scale and shape parameters 0.01 and 4.6, respectively. The failure time T then can be written as

$$T = \left[\frac{1}{\theta_0} \log\{-\theta_0 \log(1 - U)/\omega + 1\} \exp(-\beta_1 G - \beta_2 Z) \right]^{1/4 \cdot 6 / 0.01},$$

where $U \sim \text{Uniform}[0,1]$ and $\omega \sim \text{Gamma}$ with mean 1 and variance θ_0 . Finally, generate the censoring time C from $N(65,15)$ winsorized to lie in $[1,90]$.

The above data generation scheme was designed to mimic cancer or other chronic disease outcomes, which tend to have a late disease onset. Under this simulation scenario, the average age at onset was about 65 years with interquartile range from 52 to 75. We generated a large number of subjects and randomly selected 1,000 cases and 1,000 controls who were matched with cases within a year of age. For each subject, we generated one sibling and used the (censored) age at onset to estimate the composite hazard function. We took the allele frequency of G to be $p = 0.2$ with $\beta_1 = \log(1.5)$ and assumed that the variant had an additive effect on the hazard function, i.e., log hazard ratio for homozygous genotype is twice that for heterozygous genotype. The log hazard ratio for Z was $\beta_2 = \log(2.0)$ and the correlation ρ of Z among relatives was 0 or 0.5, representing no or high correlation,

respectively. We also considered different degrees of dependence of ages at onset among relatives, $\theta_0 = 0, 0.86$ and 2 . The corresponding Kendall's τ values are $0, 0.3$ and 0.5 , representing independence, moderate and strong dependence, respectively. Under each scenario, a total of $1,000$ simulated data sets were generated.

4.1. Composite hazard function $\Lambda(\cdot)$

We estimated $\Lambda(\cdot)$ by the proposed two-stage procedure (5). For comparison, we also estimated $\Lambda(\cdot)$ by the Nelson-Aalen estimator using the failure times of siblings, not accounting for the case-control sampling. Figure 1 shows both the proposed estimator (5) (dotted line) and the Nelson-Aalen estimator (dashed line) under various values of τ and ρ . When $\tau = 0$ and $\rho = 0$, i.e., cases and controls and their siblings are completely independent whether or not the risk factors are accounted for, both the proposed and the Nelson-Aalen estimators fall on the true values. However, even when $\tau = 0$ (no residual dependence) but $\rho = 0.5$, the Nelson-Aalen estimator has an upward bias, whereas the proposed estimator has no bias. The explanation for this is as follows. The composite hazard function does not account for covariates. If the covariates are correlated among family members, the unaccounted-for covariates will induce dependence among relatives. The Nelson-Aalen estimator ignores the dependence among relatives and the case-control ascertainment. As a result, the Nelson-Aalen estimator is biased, and the bias becomes more substantial as τ and ρ increase. We also present, in Table 1, the mean and empirical standard deviation (SD) of the Nelson-Aalen and the proposed estimates at selected ages over $1,000$ simulated data sets. Across the board, the SDs of proposed estimator are comparable to those of the Nelson-Aalen estimator, suggesting that the proposed estimator is as efficient as the Nelson-Aalen estimator.

4.2. Baseline hazard function $\Lambda_0(\cdot)$

Table 2 shows the summary results for the proposed estimator of the cumulative baseline hazard function $\Lambda_0(\cdot)$. A total of 50 weighted bootstrap samples were generated for each data set to estimate the standard error (SE). For estimating $\hat{\phi}(\cdot)$ the biweight function was used as the kernel and the bandwidth was selected by cross-validation. It can be seen that the baseline hazard function estimator obtained through $\hat{\phi}(\cdot)$ was largely unbiased for all scenarios considered: various degrees of dependence of failure times among family members, and independent or correlated environmental covariates among family members. It is worth noting that the ages at onset of the relatives were generated under the Clayton-Oakes model (16) given covariates G and Z ; however, the estimation of the composite hazard function was under the unconditional model (1) with h being the Clayton-Oakes copula. Therefore, the composite hazard function estimator, and consequently the baseline hazard function estimator, were estimated under a misspecified joint distribution. However, as can be seen from Table 2, there is very little or no bias in $\hat{\Lambda}_0(t)$ over a wide range of ages except at 80 years where there is a small bias. Furthermore, the estimate of the standard deviation of $\hat{\Lambda}_0(t)$ obtained using the weighted bootstrap procedure is very close to the empirical standard deviation of the estimates under all the various scenarios studied, suggesting the weighted bootstrap variance estimator performs very well in finite samples.

4.3. Misspecification effect of the copula function h on $\hat{\Lambda}_0(\cdot)$

We further assessed the misspecification effect of the copula function h on the baseline hazard function estimator $\hat{\Lambda}_0(\cdot)$. Specifically, we first generated $M+1$ random variables from a multivariate normal distribution with zero mean and correlation, ρ and then transformed them to a set of random variables $U_0, \dots, U_{M+1}$ with each component having a $U(0,1)$ marginal distribution and the whole set having the same copula as the original multivariate normal variables.

The age at onset was then generated for each family member by $\{-\log(1-U_j) \exp(-\beta_1 G - \beta_2 Z)\}^{1/4.6} / 0.01, j = 0, \dots, M$. However, we continued using the Clayton-Oakes copula for estimating the composite hazard function. It can be seen from Table 3 that under zero to moderate dependence ($\tau = 0.3$), the bias is small. Under strong dependence ($\tau = 0.5$), $\hat{\Lambda}_0(\cdot)$ still has relatively low bias when age < 70 ; for later ages the bias becomes more noticeable, but is still within 10 to 15% of the true values.

We also considered other copulas, the inverse Gaussian and positive stable, for the true underlying multivariate survival distribution, and used the Clayton-Oakes copula model (i.e. Gamma frailty distribution) in the composite hazard function estimation (results are in Section E of Supplementary Materials). As before, we observed little bias at all ages under moderate dependence and at ages < 70 under strong dependence. For ages ≥ 70 under strong dependence, the bias is more noticeable, but still within 10–15% of the true values.

Overall, our proposed estimator of $\Lambda_0(\cdot)$ performs well under various scenarios with moderate sample size. The estimator is also robust under copula function misspecification.

5. An application to a case-control study of prostate cancer

We illustrate the method with a case-control study of prostate cancer conducted at the Fred Hutchinson Cancer Research Center over the period 1993–1996. Prostate cancer cases were identified from the Seattle-Puget Sound Surveillance, Epidemiology, and End Results (SEER) cancer registry. Controls were identified by use of random-digit dialing and they were frequency matched to case participants by age (within 5-year age groups). Information collected on cases and controls included demographic characteristics, medical history, screening history and dietary variables. Family history information collected on fathers, brothers and sons included the age at diagnosis of prostate cancer if the relative had prostate cancer, or age at the last observation if the relative did not have prostate cancer. See Stanford et al. (1999) and Cohen et al. (2000) for a detailed description of the study. In total there are 707 cases and 685 controls. Since prostate cancer is a late onset disease, we excluded sons as they were generally too young to have prostate cancer and in this data set only one son had prostate cancer. A total of 101 cases' fathers and 55 controls' fathers had prostate cancer. There were 2,300 brothers of these cases and controls, among which 42 case brothers and 17 control brothers had prostate cancer. Cases are twice as likely as controls to have a father or brother with prostate cancer.

Included in the analysis are variables that have been shown to be associated with prostate cancer risk. These are education (high school or lower, some college, college, graduate/professional); positive family history in the first degree relatives (yes/no); positive prostate-specific antigen (PSA) result in the last 5 years (yes/no); dietary variables such as calories (kcal) and cruciferous vegetables (# servings). Calories and cruciferous vegetables are continuous variables and centered by the mean of these variables in controls. We used the estimation procedure in Section 2 to estimate β and $\Lambda_0(\cdot)$. A total of 500 weighted bootstrap samples were generated to estimate the standard error of the estimates and the weight is randomly generated from an exponential distribution with mean 1. Since each family, not each family member, is an independent unit, we assigned the same weight to all family members in the same family to account for the intra-family correlation. Table 4 shows the estimates and the corresponding 95% confidence intervals (CI) for the odds ratio of the risk factors on prostate cancer risk. Having a high education and eating cruciferous vegetables reduce the risk, while having a positive family history and consuming high calories increase the risk for developing prostate cancer. Subjects who have a positive PSA testing result in the last 5 years have an increased risk with odds ratio about 5.04 (95% CI: 3.87–6.57).

We included both fathers and brothers in the estimation of composite hazard function. However, since they are in the different birth cohorts, we expanded the model to allow for different hazards for fathers and brothers by including a birth-cohort indicator such that fathers are 1 and brothers and probands are 0. We fit a Cox proportional hazards model including the indicator as a covariate using the estimation procedure described in Section 2.2, where the log-composite likelihood function (6) is modified to include the log hazard ratio for the birth cohort indicator. We further allowed for different dependence parameters for father-proband and brother-proband pairs. The dependence estimates are 1.09 (SE = 0.40) and 1.60 (SE = 0.93), respectively. The p-value for testing common dependence between father-proband and brother-proband pairs is 0.63, suggesting there is no substantial evidence for different dependence parameters. We subsequently assumed common dependence among all first-degree relatives while allowing for a birth cohort effect. The hazard ratio estimate for the birth cohort indicator is 0.52 (95% CI: 0.37–0.73), which shows that an earlier birth cohort (fathers) has lower prostate cancer incidence rates than the later cohort (probands and their brothers). This is consistent with an increasing trend of prostate cancer incidence rates around the time when this study was conducted, due to an improved cancer detection by PSA that was just then approved by the FDA to test men without symptoms for prostate cancer.

Figure 2 shows the plot of the estimated composite hazard function using both our proposed estimator and the naive Breslow estimator, which ignores familial dependency and case-control sampling and is obtained from the Cox model adjusting for birth cohort. Consistent with the observation in the simulation study, the naive baseline hazard estimate is greater than the proposed estimate, suggesting it may overestimate the composite hazard function. We also included the age-specific prostate cancer incidence rates from 1993 to 1996 (dotted line) for which the study was conducted from the SEER, a nation-wide cancer registry. It shows that our proposed estimator is closer to the SEER incidence rates, whereas the naive estimator considerably overestimates the incidence rates particularly at older ages.

We then obtained $\hat{\Lambda}_0(\cdot)$ from both the Breslow estimator, termed as “Naive”, and the proposed composite hazard function estimator, termed as “Proposed”, for the proband cohort. The results are presented in Table 5. It is clear that the naive baseline hazard function estimator has considerably larger values compared to the proposed estimator across all ages. This is again consistent to the observation of a potential upward bias for the naive estimator in the simulation study. Based on the proposed estimator, for a man in the same birth cohort as the probands with high school or lower education, no positive family history, no positive PSA results and average calories and cruciferous vegetable intake, the probability of developing prostate cancer by age 80 years old is $1 - \exp(-0.0825) = 7.9\%$. If the man has positive family history and also positive PSA results while the rest of factors are the same, the disease probability is $1 - \exp(-0.0825 \times 5.04 \times 1.77) = 52.1\%$. In contrast, the naive estimates of these disease probabilities are $1 - \exp(-0.1264) = 11.9\%$ and $1 - \exp(0.1264 \times 5.04 \times 1.77) = 67.6\%$, respectively, both of which are considerably higher than the estimates obtained from the proposed method.

To evaluate the stability of the standard error estimates, we increased the number of the weighted bootstrap samples to 10,000. Figure S1 (Supplementary Materials) shows the histogram of bootstrap estimates of hazard function at selected age 50, 60, 70, and 80 years old, as well as bootstrap estimates of birth cohort and the dependence parameter. The distributions of these estimates are roughly normal with slight skewness to right for $\Lambda(50)$, showing that bootstrap-based standard deviation estimator is likely to be a good approximation to true SE. We also evaluated the stability of bootstrap SE estimates over 500, 1000, 2000, and 4000 non-overlapping bootstrap samples (Table S4, Supplementary Materials). It can be seen that there is very little variation across the number of bootstrap samples.

6. Discussion

We proposed a novel method for estimating the baseline hazard function in case-control studies by using the (possibly censored) failure times of the relatives. The method uses multivariate analysis techniques to model the joint survival distribution of the failure times of a family, while accounting for the case-control ascertainment.

The theoretical consistency of the composite hazard function estimator requires correct specification of the copula model. There is an extensive literature on behaviour of various copula models (Hougaard, 2012, and references therein). A copula model may be chosen based on the dependence structure of the correlated failure times. For example, a relatively constant cross ratio (dependence) of paired failure times may suggest the Clayton-Oakes model and a rapidly decreasing dependence may suggest the positive stable model. The dependence structure can be empirically examined by a piecewise constant cross ratio estimator (Hsu et al., 1999). While it may seem important to choose a correct copula model, in our extensive simulations we found that the estimator is robust against misspecification of the copula model. This result is consistent with previous observations on robustness of hazard function estimators under misspecified joint distribution (Chatterjee et al., 2006; Hsu et al., 2007; Gorfine et al., 2012).

Care also needs to be taken regarding the accuracy of family history information, particularly if the information is reported by the probands. In a report based on the NHLBI Family Heart Study, using the relatives' self-report as the standard, sensitivity of the proband's report on their spouse, parent, and sibling was 87%, 85%, and 81% for coronary heart disease, 83%, 87%, and 72% for diabetes, 77%, 76%, and 56% for hypertension, and 66%, 53%, and 39% for asthma, respectively (Bensen et al., 1999). Most specificity values were above 90%. These results show that the accuracy may vary by the relative type and the disease, but by and large the family history information is accurate. Future research on how to improve the accuracy of family history information will be useful.

The proposed method offers a practical approach for estimating the baseline hazard function. The approach utilizes routinely collected family history information rather than relying on external disease incidence rates. The method is particularly useful when the disease under study does not have a population disease registry. The R code that implements the proposed method will be available on <https://www.fredhutch.org/en/labs/profiles/hsu-li.html>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors gratefully acknowledge *National Institutes of Health grants P01CA53996, R01CA189532 and R01CA195789.*

Appendix: Technical Arguments

Proof of Consistency and Asymptotic Normality of $\hat{\Lambda}_0(\cdot)$.

Conditions.

Before presenting the proof of the asymptotic properties of $\hat{\Lambda}_0(\cdot)$, we first describe the conditions under which these large sample results have been proven.

C1. (Finite Interval) There exists a maximum follow-up time $\tau > 0$ and a constant c such that $\Pr(X_{ij} = \tau) < c < 1$.

C2. (Boundedness) The baseline hazard function $\lambda_0(\cdot)$ is twice differentiable with second derivatives over $[0, \tau]$ bounded by some fixed constant. Moreover,

$$\int_0^t \lambda_0(s)/\pi(s)ds < \infty \quad \forall t \in [0, \tau], \quad (\text{A.1})$$

Where $\pi(s) = E[I(X_{i0} < s) \sum_{j=1}^M Y_{ij}(s)]$.

C3. (Identifiability) There is a positive probability that two or more members in a family fail in the interval $[0, \tau]$.

C4. (Differentiability) The copula function $h(t_0, t_1, \dots, T_M, \theta)$ is continuously thrice differentiable with respect to each time component on $[c', 1]^{M+1}$, where $c' = \exp(-c)$ and c is defined in Condition C1. In addition, h and its second partial derivatives are twice differentiable with respect to θ , and the derivatives are also continuously differentiable with respect to each time component on $[c', 1]^{M+1}$.

C5. (Nondegeneracy) The function

$$\ell^*(\beta, \gamma) = (1 - \alpha)E\{\ell^o(\beta, \gamma, X_{i0}, Z_{i0})|\delta_{i0} = 0\} + \alpha E\{\ell^o(\beta, \gamma, X_{i0}, Z_{i0})|\delta_{i0} = 1\}$$

has a unique maximizer $(\bar{\beta}, \bar{\gamma})$. In addition, the limiting value of $U(\theta, \Lambda)/\theta$ is positive definite at θ_0 .

C6. (Kernel regularity condition) The kernel $K(\cdot)$ is symmetric, has support on $[-1, 1]$ and is differentiable on $(-\infty, \infty)$ such that $K(\cdot)$ decreases smoothly to 0 in the neighborhood of ± 1 . The bandwidth h satisfies $h \sim n^{-\nu}$ with $\nu \in (\frac{1}{4}, \frac{1}{2})$.

C7. (Density regularity condition) The conditional density $g(\cdot)$ of X_{i0} given $\delta_{i0} = 1$ is bounded below on $t \in [0, \tau]$ by some positive number g_{min} and $g(t)$ is thrice differentiable w.r.t. t over $[0, \tau]$ with bounded third derivative.

Condition C1 is a standard condition survival analysis. The condition (A.1) in Condition C2 is needed for the first-stage estimator, to ensure that the denominator in (3) does not tend too quickly to zero as $t \downarrow 0$. Condition C3 is necessary for the dependence parameter to be identifiable. The same condition has also been assumed in Nielsen et al. (1992) and Murphy (1994, 1995). Condition C4 is usually satisfied for commonly used copula functions such as the Clayton-Oakes model (Oakes, 1989) or the normal transformation model (Li and Lin, 2006). Regarding Condition C5, $\ell^*(\beta, \gamma)$ has a maximizer because it is concave, and the condition that the maximizer is unique simply rules out a degenerate case. Likewise, the matrix $U(\theta, \Lambda)/\theta$ is automatically nonnegative definite, and the assumption that it be positive definite at θ_0 simply rules out a degenerate case. Regarding the assumptions on the kernel function in Condition C6, it is not difficult to find kernels meeting these assumptions; one example is the biweight kernel $K(u) = (15/16)(1 - u^2)^2K(|u| - 1)$.

For a function $\varphi : [0, \tau] \rightarrow \mathcal{R}$, let $\|\varphi\| = \sup_{t \in [0, \tau]} |\varphi(t)|$.

Asymptotic properties of $\hat{\beta}$.

Under Condition C5, it follows from standard theory (White, 1982; van der Vaart, 1998, Chapter 5), that $(\hat{\beta}, \hat{\gamma})$ converges almost surely to $(\bar{\beta}, \bar{\gamma})$ and that $n^{1/2}\{(\hat{\beta}, \hat{\gamma}) - (\bar{\beta}, \bar{\gamma})\}$ is asymptotically normal. For use in proving the asymptotics of $\hat{\Lambda}_0(\cdot)$, we note that

$$\hat{\beta} - \bar{\beta} = \frac{1}{n} \sum_{i=1}^n e_1(\bar{\beta}, \bar{\gamma}) [\Psi(\bar{\beta}, \bar{\gamma}, X_{i0}, Z_{i0}) - E\{\Psi(\bar{\beta}, \bar{\gamma}, X_{i0}, Z_{i0})|\delta_{i0}\}] + o_p(n^{-1/2}) \quad (\text{A.2})$$

where $\Psi(\beta, \gamma, t, z)$ is the gradient vector of $\ell(\beta, \gamma, t, x)$ with respect to β and γ , and $e_1(\beta, \gamma)$ is -1 times the inverse of the Hessian matrix of $\ell(\beta, \gamma)$ with the last row deleted (the uniqueness of $(\bar{\beta}, \bar{\gamma})$ as a maximizer of ℓ^* implies that this Hessian matrix evaluated at $(\bar{\beta}, \bar{\gamma})$ is positive definite). This gives a representation of $\hat{\beta} - \bar{\beta}$ as asymptotically equivalent to a quantity of the form (11).

Asymptotic properties of $\hat{\Lambda}(\cdot, \hat{\theta})$.

We will show the consistency and asymptotic normality of $\hat{\Lambda}(\cdot, \hat{\theta})$ using some of the ideas presented in Gorfine et al. (2009) and Spiekerman and Lin (1998). Essentially, we can show that for a given θ ,

$$n^{1/2} \left\{ \hat{\Lambda}(t, \theta) - \Lambda(t, \theta) \right\} = \mathcal{M}_1(t) + \mathcal{M}_2(t) + o_p(1) \quad (\text{A.3})$$

with

$$\begin{aligned} \mathcal{M}_1(t) &= n^{-\frac{1}{2}} p_1(t) \int_0^t \frac{1}{p_1(s)h(s, \Lambda)} \sum_{i=1}^n \sum_{j=1}^M M_{ij}(ds) \\ \mathcal{M}_2(t) &= n^{-\frac{1}{2}} p_1(t) \int_0^\tau \frac{\mathcal{A}_1(t, u)}{p(u)q(u, \Lambda)} \sum_{i=1}^n \sum_{j=1}^M I(X_{i0} < u) M_{ij}(du), \end{aligned}$$

where the integrands in $\mathcal{M}_1(t)$ and $\mathcal{M}_2(t)$ are predictable processes. (The notation and detailed derivation are provided in Section A, the Supplementary Materials.) Hence by standard martingale theory, $\|\hat{\Lambda}(\cdot, \theta) - \Lambda(\cdot, \theta)\| \rightarrow_p 0$ and $n^{1/2} \left\{ \hat{\Lambda}(t, \theta) - \Lambda(t, \theta) \right\}$ converges weakly to a zero-mean Gaussian process, with weak convergence defined with respect to the uniform metric on $D[0, \tau]$ (Pollard, 1984, Section VIII.2).

To establish the consistency of $\hat{\theta}$, we note that under condition C4, $dh/d\theta$, $dh^{(01)}/d\theta$, $dh^{(10)}/d\theta$ and $dh^{(11)}/d\theta$ are also uniformly Lipschitz continuous with respect to $\Lambda(\cdot)$. Hence, from the uniform convergence of $\hat{\Lambda}(\cdot, \theta)$ to $\Lambda(\cdot, \theta)$ and the Lipschitz continuity, we have that $U\{\theta, \hat{\Lambda}(t, \theta)\}$ converges uniformly to $U\{\theta, \Lambda(t, \theta)\}$ in θ as $n \rightarrow \infty$. By the strong law of large numbers, $U\{\theta, \Lambda(\cdot, \theta)\}$ converges to a limit $u\{\theta, \Lambda(\cdot, \theta)\}$, which equals 0 at θ_0 . Finally, under condition C5 and by the Foutz theorem (1977), we conclude that there exists a unique root $\hat{\theta}$ to $U\{\hat{\theta}, \hat{\Lambda}(\cdot, \hat{\theta})\} = 0$ and that $\hat{\theta} \rightarrow \theta_0$ in probability, as $n \rightarrow \infty$. To show the asymptotic normality of $\hat{\theta}$, we expand

$$U\{\hat{\theta}, \hat{\Lambda}(\cdot, \hat{\theta})\} = U\{\theta_0, \Lambda(\cdot)\} + [U\{\theta_0, \hat{\Lambda}(\cdot, \theta_0)\} - U\{\theta_0, \Lambda(\cdot)\}] + [U\{\hat{\theta}, \hat{\Lambda}(\cdot, \hat{\theta})\} - U\{\theta_0, \hat{\Lambda}(\cdot, \theta_0)\}] = 0.$$

By a Taylor expansion, the second term is asymptotically equivalent to $\partial U\{\theta_0, \Lambda(\cdot)\} / \partial \Lambda(t) \{ \tilde{\Lambda}(\cdot; \theta_0) - \Lambda(\cdot) \}$ and the third term is asymptotically equivalent to $\partial U\{\theta, \Lambda(\cdot, \theta)\} / \partial \theta|_{\theta=\theta_0} (\hat{\theta} - \theta_0)$. By the martingale representation of $\hat{\Lambda}(\cdot, \theta_0) - \Lambda(\cdot)$ and

condition C5, we get that $n^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically equivalent to $n^{1/2}$ times a quantity of the form (11). By the central limit theorem, it is asymptotically normal with mean 0 and a covariance matrix that can be consistently estimated by a sandwich-type estimator. Similarly, we expand $\hat{\Lambda}(\cdot, \hat{\theta}) - \Lambda(\cdot) = \{\hat{\Lambda}(\cdot, \hat{\theta}) - \hat{\Lambda}(\cdot, \theta_0)\} + \{\hat{\Lambda}(\cdot, \theta_0) - \Lambda(\cdot)\}$, and approximate the first term by $\partial \hat{\Lambda}(\cdot, \theta) / \partial \theta|_{\theta = \theta_0}(\hat{\theta} - \theta_0)$. By the asymptotic results of $\hat{\theta}$, the martingale representation of $\hat{\Lambda}(\cdot, \theta)$ and condition C4, we also get that $\|\hat{\Lambda}(\cdot, \hat{\theta}) - \Lambda(\cdot)\| \rightarrow_p 0$ and $n^{1/2}\{\hat{\Lambda}(\cdot, \hat{\theta}) - \Lambda(\cdot)\}$ converges to a zero-mean Gaussian process.

Asymptotic results of $\hat{\Lambda}_0(\cdot)$.

We now prove that $n^{1/2}\{\hat{\Lambda}_0(\cdot) - \Lambda_0(\cdot, \bar{\beta})\}$ converges weakly to a Gaussian random process (which implies uniform convergence in probability of $\hat{\Lambda}_0(t)$ to $\Lambda_0(t, \bar{\beta})$). Note that $\hat{\Lambda}_0(\cdot)$ is a function of θ , since it is a function of the composite hazard function $\hat{\Lambda}(\cdot, \theta)$, which depends on θ . For clarity of presentation, we now indicate θ explicitly in the notation, writing $\hat{\Lambda}_0(\cdot)$ as $\Lambda_0(\cdot, \theta)$. We have

$$\begin{aligned} \hat{\Lambda}_0(t) - \Lambda_0(t, \bar{\beta}) &= \int_0^t \phi(u, \bar{\beta}) \{\hat{\Lambda}(du, \hat{\theta}) - \Lambda(du)\} + \int_0^t \{\tilde{\phi}(u) - \phi(u, \bar{\beta})\} \Lambda(du) \\ &+ \int_0^t \{\hat{\phi}(u) - \tilde{\phi}(u)\} \Lambda(du) + \int_0^t \{\hat{\phi}(u) - \phi(u, \bar{\beta})\} \{\hat{\Lambda}(du, \hat{\theta}) - \Lambda(du)\} \equiv \text{(I)} + \text{(II)} + \text{(III)} + \\ &\text{(IV)}. \end{aligned}$$

(A.4)

Term (IV) is $O_p(n_1^{-1/2})$, and thus asymptotically negligible (For details, see Section C, Supplementary Materials). Regarding (I), by integration by parts, we can write

$$\text{(I)} = \phi(t, \bar{\beta}) \{\hat{\Lambda}(t, \hat{\theta}) - \Lambda(t)\} - \int_0^t \phi'(u, \bar{\beta}) \{\hat{\Lambda}(u, \hat{\theta}) - \Lambda(u)\} du = \text{(Ia)} + \text{(Ib)} + o_p(n^{-1/2})$$

with

$$\text{(Ia)} = \left[\phi(t, \bar{\beta}) \hat{\Lambda}'(t, \theta_0) - \int_0^t \phi'(u, \bar{\beta}) \hat{\Lambda}'(u, \theta_0) du \right] (\hat{\theta} - \theta_0)$$

and

$$\text{(Ib)} = \phi(t, \bar{\beta}) \{\hat{\Lambda}(t, \theta_0) - \Lambda(t)\} - \int_0^t \phi'(u, \bar{\beta}) \{\hat{\Lambda}(u, \theta_0) - \Lambda(u)\} du,$$

where $\widehat{\Lambda}'(t, \theta_0)$ is the derivative of $\widehat{\Lambda}(t, \theta)$ with respect to θ and evaluated at θ_0 , which can be shown to converge to $\Lambda'(t, \theta_0)$ uniformly in t using similar techniques as in the proof for $\widehat{\Lambda}(\cdot, \theta_0)$. The asymptotic results shown above for $\widehat{\theta}$ and $\widehat{\Lambda}(t, \theta)$, along with the continuous mapping theorem, imply that (I) is asymptotically equivalent to a quantity of the form (11) and that $n^{1/2}$ times (I) is weakly convergent.

We next show that (II) is asymptotically equivalent to the average of i.i.d. mean-zero terms involving the n_1 families only and that $n^{1/2}$ times (II) converges weakly to a Gaussian process. Here we sketch the argument; full details are provided in Section B, the Supplementary Materials.

Denoting $Y_i = \exp(-\bar{\beta}'Z_i)$, and $\Delta_i = Y_i - \phi(X_{i0}, \bar{\beta})$, it can be shown that

$$\int_0^t \{\tilde{\phi}(u) - \phi(u, \bar{\beta})\} \Lambda(du) = \frac{1}{n_1} \sum_{i=1}^n \delta_{i0} \Omega^\circ(t, X_{i0}) \Delta_i + o_p(n_1^{-1/2}), \quad (\text{A.5})$$

where $\Omega^\circ(t, x) = \frac{\lambda(x)}{g(x)} I(x \leq t)$. It is easy to see that the main term, which is an average of mean 0 i.i.d. processes in t , converges weakly in $\mathcal{C}^\infty[0, \tau]$ to $\mathbb{B}(V(t))$, where $\mathbb{B}$ is a Brownian motion process and the variance

$$V(t) = \int_0^t \left\{ \frac{\lambda(u)}{g(u)} \right\}^2 \sigma^2(u) g(u) du$$

with $\sigma^2(u) = \text{var}(Y_i | X_{i0} = u, \delta_{i0} = 1)$.

For term (III), a Taylor expansion gives (III) = $\mathcal{Q}(t, \bar{\beta})'(\hat{\beta} - \bar{\beta}) + o_p(n^{-1/2})$ where

$$\mathcal{Q}(t, \bar{\beta}) = \int_0^t \left[\sum_{i=1}^n \left\{ c_i(u) Z_i \exp(-\bar{\beta}'Z_i) \right\} \right] \Lambda(du)$$

Using Claim 11 in Section B, the Supplementary Materials, we find that $\mathcal{Q}(t)$ converges to a limit $\widehat{\mathcal{Q}}(t, \bar{\beta})$ uniformly in probability, and $\hat{\beta} - \bar{\beta}$ has already been shown above to be asymptotically equivalent to a quantity of the form (11).

Taking the results for (I)-(IV) together, and noting that the sum of tight processes is tight, we find that

$$\widehat{\Lambda}_0(t) - \Lambda_0(t, \bar{\beta}) = \frac{1}{n} \sum_{i=1}^n r(X_i, t) + o_p(n^{-1/2})$$

uniformly in t , where $E[Y(X_i, t)|\delta_{i0}] = 0$, and that $n^{1/2}$ times the main term on the right side converges weakly to a Gaussian process, with weak convergence defined in terms of the uniform metric. Theorem 1 is thus proved.

References

- Bensen JT, Liese AD, Rushing JT, Province M, Folsom AR, Rich SS and Higgins M (1999). Accuracy of proband reported family history: the NHLBI Family Heart Study (FHS). *Genetic Epidemiology* 17, 141–50. [PubMed: 10414557]
- Chatterjee N, Zeynep K, Shih JH and Gail M (2006). Case-control study with family history data: a combined approach of kin-cohort and case-control analysis. *Biometrics* 62, 36–48. [PubMed: 16542227]
- Chen L, Hsu L and Malone K (2009). A frailty-model based approach to estimating the age-dependent function of candidate genes using population-based casecontrol study designs: An application to data on BRCA1 gene. *Biometrics* 65, 1105–1114. [PubMed: 19210733]
- Chen YQ, Hu C and Wang Y (2006). Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics* 7, 515–529. [PubMed: 16478758]
- Clayton DG (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.
- Cohen JH, Kristal AR and Stanford JL (2000). Fruit and vegetable intakes and prostate cancer risk. *Journal of the National Cancer Institute* 92, 61–68. [PubMed: 10620635]
- Cox DR (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 34, 187–220.
- Fan J (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics* 21, 196–216.
- Fan J and Gijbels I (1996). Local polynomial modelling and its applications. Chapman & Hall.
- Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C and Mulvihill JJ (1989). Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *Journal of the National Cancer Institute* 81, 1879–1886. [PubMed: 2593165]
- Glidden DV and Self SG (1999). Semiparametric likelihood estimation in the Clayton-Oakes model. *Scandinavian Journal of Statistics* 26, 363–372.
- Gorfine M, De-Picciotto R and Hsu L (2012). Conditional and marginal estimates in case-control family data — Extensions and sensitivity analyses. *Journal of Computational Statistics and Simulation* 82(10), 1449–1470.
- Gorfine M, Zucker DM and Hsu L (2009). Case-control survival analysis with a general semiparametric shared frailty model—A pseudo full likelihood approach. *Annals of Statistics* 37, 1489–1517. [PubMed: 19684881]
- Hsu L, Prentice RL, Zhao LP and Fan JJ (1999). On dependence estimation using correlated failure time data from case-control family studies. *Biometrika* 86(4), 743–753.
- Hsu L, Chen L, Gorfine M and Malone K (2004). Semiparametric estimation of marginal hazard function from the case-control family studies. *Biometrics* 60, 936–944. [PubMed: 15606414]
- Hsu L and Gorfine M (2006). Multivariate survival analysis for case-control family data. *Biostatistics* 7, 387–398. [PubMed: 16368774]
- Hsu L, Gorfine M and Malone KE (2007). Effect of Frailty Distribution Misspecification on Marginal Regression Estimates and Hazard Functions in Multivariate Survival Analysis. *Statistics in Medicine* 26, 4657–4678. [PubMed: 17348081]
- Hougaard P (2012). *Analysis of Multivariate Survival Data*. Springer Science and Business Media.
- Kosorok MR (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.
- Liang K-Y and Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

- Malone KE, Daling JR, Weiss NS, McKnight B, White E, and Voigt LF (1996). Family history and survival of young women with invasive breast carcinoma. *Cancer* 78, 1417–1425. [PubMed: 8839546]
- Oakes D (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84, 487–493.
- Prentice RL and Breslow NE (1978). Retrospective studies and failure time models. *Biometrika* 65, 153–158.
- Prentice RL and Pyke R (1979). Logistic disease incidence models and case-control studies. *Biometrika* 66, 403–411.
- Shih JH and Chatterjee N (2002). Survival analysis of family data from case control studies. *Biometrics* 58, 502–509. [PubMed: 12229984]
- Spiekerman CF and Lin DY (1998). Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association* 93, 1164–1175.
- Stanford JL, Wicklund KG, McKnight B, Daling JR and Brawer MK (1999). Vasectomy and risk of prostate cancer. *Cancer Epidemiology, Biomarkers & Prevention* 8, 881–886.
- van der Vaart AW (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- Wand MP and Jones MC (1995). *Kernel Smoothing*. Chapman & Hall, London.
- White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25

Figure 1:

The composite hazard function $\Lambda(\cdot)$ for the true value (solid line), the NelsonAalen estimator (dashed line) and the proposed estimator (dotted line) averaged over 1,000 simulated data sets.

Figure 2:

The composite hazard function $\Lambda(\cdot)$ for the naive Breslow estimator (dashed line) and the proposed estimator (solid line) for the prostate cancer study. As reference, the age-specific SEER incidence rates (dotted line) is also included.

Table 1:

Summary statistics of the Nelson-Aalen estimator and the proposed estimator for the cumulative composite hazard function $\Lambda(\cdot)$. Mean and SD are the average and standard deviation of the estimates over 1,000 simulated data sets.

τ	Age (yrs)	$\rho = 0$						$\rho = 0.5$					
		Nelson-Aalen			Proposed			Nelson-Aalen			Proposed		
		True	Mean	SD	Mean	SD	Mean	Mean	SD	Mean	Mean	SD	SD
0	30	0.006	0.007	0.002	0.006	0.002	0.007	0.007	0.002	0.007	0.007	0.002	0.002
	40	0.024	0.024	0.004	0.024	0.004	0.026	0.026	0.004	0.024	0.024	0.004	0.004
	50	0.064	0.065	0.006	0.064	0.007	0.070	0.070	0.006	0.065	0.065	0.007	0.007
	60	0.143	0.143	0.011	0.141	0.011	0.155	0.155	0.011	0.143	0.143	0.012	0.012
	70	0.273	0.275	0.019	0.270	0.020	0.296	0.296	0.020	0.273	0.273	0.022	0.022
0.3	80	0.466	0.470	0.038	0.461	0.039	0.502	0.502	0.038	0.463	0.463	0.040	0.040
	30	0.006	0.008	0.002	0.007	0.002	0.009	0.009	0.002	0.007	0.007	0.002	0.002
	40	0.024	0.029	0.004	0.024	0.004	0.032	0.032	0.004	0.024	0.024	0.003	0.003
	50	0.064	0.079	0.007	0.065	0.007	0.086	0.086	0.007	0.064	0.064	0.007	0.007
	60	0.143	0.174	0.012	0.141	0.012	0.188	0.188	0.012	0.142	0.142	0.012	0.012
0.5	70	0.273	0.329	0.022	0.269	0.023	0.355	0.355	0.022	0.268	0.268	0.023	0.023
	80	0.466	0.555	0.042	0.455	0.043	0.597	0.597	0.044	0.453	0.453	0.043	0.043
	30	0.006	0.009	0.002	0.007	0.002	0.010	0.010	0.002	0.006	0.006	0.002	0.002
	40	0.024	0.035	0.004	0.024	0.004	0.037	0.037	0.005	0.024	0.024	0.003	0.003
	50	0.064	0.093	0.007	0.065	0.007	0.101	0.101	0.008	0.064	0.064	0.007	0.007
	60	0.143	0.202	0.013	0.141	0.013	0.219	0.219	0.013	0.139	0.139	0.013	0.013
	70	0.273	0.378	0.022	0.265	0.024	0.408	0.408	0.022	0.262	0.262	0.024	0.024
	80	0.466	0.625	0.045	0.442	0.045	0.669	0.669	0.045	0.434	0.434	0.044	0.044

Summary statistics of the proposed estimator for $\Lambda_0(\cdot)$. Mean and SD are the average and standard deviation of the estimates over 1,000 simulated data sets. SE is the average of the weighted bootstrap-based SE and 95% CP is the empirical coverage rate of 95% Wald-type confidence intervals. The ages at onset of relatives were generated following the Clayton-Oakes (Gamma frailty) distribution.

Table 2:

τ	Age(yrs)	True	$\rho = 0$				$\rho = 0.5$			
			Mean	SD(x10)	SE(x10)	95%CP(%)	Mean	SD(x10)	SE(x10)	95%CP(%)
0	30	0.004	0.004	0.013	0.012	92.1	0.004	0.013	0.012	94.4
	40	0.015	0.016	0.027	0.026	93.2	0.016	0.027	0.026	93.1
	50	0.041	0.042	0.049	0.048	93.9	0.043	0.051	0.049	93.1
	60	0.095	0.096	0.089	0.088	94.1	0.098	0.092	0.091	94.2
	70	0.194	0.192	0.169	0.166	93.7	0.194	0.177	0.172	93.2
0.3	80	0.358	0.349	0.351	0.337	91.0	0.351	0.346	0.349	93.1
	30	0.004	0.004	0.012	0.012	94.1	0.004	0.012	0.012	93.5
	40	0.015	0.016	0.026	0.025	93.8	0.016	0.026	0.025	93.7
	50	0.041	0.043	0.051	0.050	91.7	0.043	0.050	0.049	94.4
	60	0.095	0.097	0.098	0.094	93.6	0.097	0.097	0.095	94.2
0.5	70	0.194	0.192	0.191	0.179	93.3	0.191	0.185	0.180	92.5
	80	0.358	0.345	0.371	0.356	90.8	0.343	0.372	0.357	88.6
	30	0.004	0.004	0.011	0.011	95.1	0.004	0.011	0.011	94.8
	40	0.015	0.016	0.026	0.024	92.4	0.015	0.024	0.024	94.9
	50	0.041	0.043	0.051	0.049	93.8	0.042	0.049	0.050	94.6
	60	0.095	0.096	0.101	0.096	92.9	0.095	0.096	0.098	94.6
	70	0.194	0.189	0.193	0.186	92.2	0.186	0.190	0.190	90.5
	80	0.358	0.335	0.385	0.365	85.0	0.328	0.362	0.367	82.9

Table 3:

Summary statistics of the proposed estimator for $V_0(\cdot)$. Mean and SD are the average and standard deviation of the estimates over 1,000 simulated data sets. SE is the average of the weighted bootstrap-based SE and 95% CP is the empirical coverage rate of 95% Wald-type confidence intervals. The ages at onset of relatives were generated following the multivariate normal distribution.

τ	Age (yrs)	True	$\rho = 0$				$\rho = 0.5$			
			Mean	SD(x10)	SE(x10)	95%CP(%)	Mean	SD(x10)	SE(x10)	95%CP(%)
0	30	0.004	0.004	0.013	0.013	92.9	0.004	0.014	0.013	94.0
	40	0.015	0.016	0.027	0.026	93.2	0.016	0.028	0.026	94.0
	50	0.041	0.042	0.051	0.049	93.3	0.043	0.053	0.050	92.4
	60	0.095	0.097	0.089	0.086	93.9	0.098	0.093	0.089	93.0
	70	0.194	0.194	0.155	0.152	94.0	0.196	0.164	0.157	92.9
0.3	80	0.358	0.355	0.275	0.270	93.7	0.357	0.285	0.281	94.2
	30	0.004	0.005	0.013	0.013	92.3	0.005	0.013	0.012	92.0
	40	0.015	0.017	0.028	0.027	88.0	0.017	0.026	0.026	89.7
	50	0.041	0.045	0.051	0.051	91.0	0.045	0.051	0.051	90.8
	60	0.095	0.098	0.093	0.093	95.4	0.097	0.097	0.094	93.8
0.5	70	0.194	0.189	0.167	0.166	91.5	0.186	0.171	0.169	90.6
	80	0.358	0.331	0.294	0.295	82.2	0.325	0.295	0.300	77.2
	30	0.004	0.005	0.013	0.012	92.1	0.005	0.012	0.012	92.3
	40	0.015	0.017	0.027	0.027	88.3	0.017	0.027	0.026	89.1
	50	0.041	0.044	0.053	0.052	93.8	0.044	0.053	0.051	93.2
	60	0.095	0.092	0.102	0.098	91.1	0.094	0.096	0.093	92.1
	70	0.194	0.172	0.190	0.176	70.7	0.174	0.156	0.159	71.7
	80	0.358	0.291	0.323	0.307	41.2	0.293	0.259	0.262	32.3

Table 4:

The odds ratio estimates and the 95% confidence intervals (CI) of risk factors for prostate cancer risk

Variable	Odds ratio (95% CI)
Education	
high school or lower	1.00 (referent)
some college	0.77 (0.52–1.31)
college	0.89 (0.62–1.27)
graduate/professional	0.67 (0.47–0.96)
Family History	
no	1.00 (referent)
yes	1.77 (1.22–2.57)
PSA	
no	1.00 (referent)
yes	5.04 (3.87–6.57)
Calories (kcal/1000)	1.21 (1.02–1.43)
Cruciferous Veg	0.93 (0.87–0.99)

Table 5:

The cumulative baseline hazard estimates and the 95% confidence intervals (CI) at age 40, 50, 60, 70 and 80 years old

$\Lambda_0(t)$	Naive	Proposed
	Estimate (95% CI)	Estimate (95% CI)
$\Lambda_0(40)(\times 10^{-4})$	2.57 (0–6.99)	1.65 (0–8.24)
$\Lambda_0(50)(\times 10^{-4})$	14.96 (7.52–22.41)	9.57 (0–20.58)
$\Lambda_0(60)(\times 10^{-4})$	167.65 (120.58–214.71)	106.35 (54.52–158.18)
$\Lambda_0(70)(\times 10^{-4})$	525.51 (376.24–674.78)	337.88 (179.73–496.03)
$\Lambda_0(80)(\times 10^{-4})$	1263.59 (879.98–1647.2)	824.98 (403.84–1246.12)