HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

Stat Methods Med Res. 2020 January ; 29(1): 243–257. doi:10.1177/0962280219831725.

Adjusted Time-varying Population Attributable Hazard in Case-Control Studies

Wei Zhao[1,†], **Jiayin Zheng**[2,†], **Ying Qing Chen**[2], **Li Hsu**[2]

[1]Department of Biostatistics, University of Washington, Seattle, WA, USA

[2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Abstract

Population attributable fraction (PAF) is a widely used measure for quantifying the disease burden associated with a modifiable exposure of interest at the population level. It has been extended to a time-varying measure, population attributable hazard (PAH) function, to provide additional information on when and how the exposure's impact varies over time. However, like the classic PAF, the PAH is generally biased if confounders are present. In this article, we provide a natural definition of adjusted PAH to take into account the effects of confounders, and its alternative that is identifiable from case-control studies under the rare disease assumption. We propose a novel estimator, which combines the odds ratio estimator from logistic regression model, and the conditional density function estimator of the exposure and confounding variables distribution given the failure times of cases or the current times of controls from a kernel smoother. We show that the proposed estimators are consistent and asymptotically normal with variance that can be estimated empirically from the data. Simulation studies demonstrate that the proposed estimators perform well in finite sample sizes. Finally, we illustrate the method by an analysis of a case-control study of colorectal cancer. Supplementary materials for this article are available online.

Keywords

Confounding; Cox Proportional hazards model; Kernel smoother; Population attributable fraction

1 Introduction

For evidence-based disease prevention, the population attributable fraction (PAF) has been widely used as an indispensable metric to assess the impact of modifiable exposures on disease burden in populations to help plan and prioritize public health strategies[1,2]. First introduced by Levin[3], the PAF is defined as the proportion of potential preventable disease cases, had the exposure been eliminated. Compared to association measures such as relative risk or rate ratio, the PAF is a more appropriate epidemiological measure for quantifying the

Corresponding author: Li Hsu, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98109. lih@fredhutch.org.

†Co-first author, contributed equally to this work.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

population impact because it integrates both the strength of association between the exposure and the disease and the prevalence of the exposure. The implication, estimation, and application of the PAF have been extensively studied for various epidemiological sampling designs, see, for example, Walter[4], Whittemore[5], Greenland[6], Benichou and Gail[7], Benichou[8,9], Kooperberg and Petitti[10].

PAF is a static measure, which evaluates the impact of exposure on binary disease outcome. In practice, mortality and morbidity incidences are often recorded as time-to-events. Therefore, rather than Levin's PAF, a time-varying PAF can help researchers and policy makers better understand how the impact of risk exposure on disease varies over time, and provide guidance on the timing of actions or interventions. Recently, extensions of the PAF to right-censored failure time have been proposed, including the time-dependent population attributable hazard function (PAH)[11-13] and time-dependent population attributable disease probability[14-18]. In survival analysis, the hazard function is a key concept that quantifies the impact of risk factors on the rate of developing the disease among subjects at risk at time t. Here we focus on the PAH because it directly assesses the instantaneous effect of eliminating risk factors on the hazard function.

Specifically, let T be the failure time and $\mathrm{Z} = (Z_1, \ldots, Z_p)^T$ be a p-vector of time-independent risk factors, then the PAH[11] is defined as

$$\phi(t) = \frac{\lambda(t) - \lambda(t \mid \mathrm{Z} = 0)}{\lambda(t)}, \tag{1}$$

where $\lambda(t)$ is the hazard function for the population of interest, which is the instantaneous rate of failing at time t, that is, $\lambda(t) = \lim_{\delta t \to 0} \Pr(t < T \leq t + \delta t)/\Pr(T > t)$; and $\lambda(t|\mathrm{Z} = 0)$ is the conditional hazard function given Z = 0. Estimation of $\lambda(t)$ requires estimates of hazard ratios, for which methods are well established for cohort data[19], and estimates of the distribution of exposure and failure time[11,15].

In the past several decades, the case-control study design is one of the most commonly used designs in epidemiological cancer research, because of its efficiency in logistic conduct, time and cost. In particular, Anderson[20] and Prentice and Pyke[21] provided theoretical basis that under the rare disease assumption the relative risk of exposure on disease occurrence can be obtained from the retrospective case-control data as if they were prospectively collected under the logistic regression model. For the time-to-event outcome, the hazard ratios under the Cox proportional hazards model can also be obtained from the (time-matched) case-control data[22]. As a result, many case-control studies have been conducted. For example, in our motivating real data example, the Genetics and Epidemiology Colorectal Cancer Consortium (GECCO) is consisted of dozens of well-characterized case-control studies of colorectal cancer where extensive risk factor information has been collected. It is therefore important to make use of these rich data when estimating the PAH. Bruzzi et al.[23] showed that if one has the relative risk estimates, the PAF can be obtained from the distribution of exposure among the cases only. In a similar fashion, for time-varying PAH, Zhao et al.[13] showed at any specific time t, PAH can be estimated consistently with the hazards ratio estimated from case-control data[22] and the conditional distribution of exposure estimated from the cases at time t.

Confounders are nevertheless ubiquitous in observational studies. If they were not properly accounted for, the association of exposure of interest and outcome and the PAF would be biased and could be misleading in some situations[4,24]. The unadjusted PAH proposed by Zhao et al.[13] has the same issue. For instance, obesity is a potential confounder for the association between diabetes and colorectal cancer, because obese patients are more likely to have diabetes[25] and also more likely to develop colorectal cancer[26]. Thus, without adjusting for obesity, the preventable effect of eliminating diabetes may be overestimated. Therefore, in order for the estimated PAH to have a meaningful interpretation, it is necessary to adjust for confounders. To this end, a natural generalization of the confounder-adjusted PAF for binary outcome[8] to time-to-event outcome is to substitute $\lambda(t|Z = 0)$ in (1) by $\widetilde{E}_{U \mid T \geq t}\{\lambda(t \mid Z = 0, U)\}$, where U are confounding variables and $\widetilde{E}_{U \mid T \geq t}$ is the expectation of U given the subject at risk at time t in a ideal population where Z is eliminated at baseline. However, this quantity is not identifiable from case-control data because cases are over-sampled and the distribution of (Z, U) in the sample does not reflect the distribution in the population.

The goal of the paper is to provide an alternative adjusted PAH that approximates well to $\widetilde{E}_{U \mid T \geq t}$ under the rare disease assumption, and a novel kernel-based estimator integrating the information from both cases and controls. The rest of this article is organized as follows: in Section 2, we lay out the basic formulation and the kernel-based estimator for the adjusted PAH for case-control data. We establish the large sample properties and provide asymptotic-based variance estimators. In Section 3, we present simulation studies and the performance of the proposed estimator in finite sample sizes. We show an application of the proposed estimator to a case-control study of colorectal cancer in Section 4. Finally, we provide some concluding remarks in Section 5. Some technical details are included in the appendices.

2 Formulation and Inference

2.1 Definition of Population Attributable Hazard Function Adjusting for Confounders

We first describe the data and notation. Consider a case-control study which consists of n_1 cases and n_2 controls, in total n ($n = n_1 + n_2$) subjects. Let Δ be the binary disease status (1: case and 0: control) and $X = \min(T, C)$ be the observed age, which is the age-at-onset *i.e.* failure time T if the individual is diseased (case) and age at examination *i.e.* censoring time C if the individual is not diseased (control). Further we denote Z a vector of time-independent covariates, which are of interest, and U a vector of time-independent confounders. We are interested in the proportion of the time-varying hazard attributed to Z while adjusting for potential confounders U.

Before we describe the PAH, we first review classic model-based adjusted PAF for binary disease status in the presence of confounders. Suppose the vector of time-independent confounders U form J levels and time-independent covariate of interest Z has $I + 1$ levels, then the adjusted PAF[5,23] is defined as

$$\begin{aligned} \mathrm{PAF} &= \frac{P(\Delta = 1) - \sum_{j=1}^{J} P(\Delta = 1 \mid \mathrm{Z} = 0, \mathrm{U} = \mathrm{u}_j) P(\mathrm{U} = \mathrm{u}_j)}{P(\Delta = 1)} \\ &= 1 - \sum_{j=1}^{J} \sum_{i=0}^{I} \frac{P(\mathrm{Z} = \mathrm{z}_i, \mathrm{U} = \mathrm{u}_j \mid \Delta = 1)}{RR_{i \mid j}} \end{aligned} \tag{2}$$

where $RR_{i|j} = P(\Delta = 1 \rightarrow \mathrm{Z} = \mathrm{z}_i, \mathrm{U} = \mathrm{u}_j)/P(\Delta = 1 | \mathrm{Z} = 0, \mathrm{U} = \mathrm{u}_j)$. The re-expression of the PAF in the second equation as a function of $f(\mathrm{Z}, \mathrm{U}|\Delta)$ and RR indicates that both quantities can be estimated from the retrospectively collected case-control data under the rare disease assumption. As a result the adjusted PAF can also be estimated from case-control data.

Now we turn to the PAH. Let E denote the expectation with respect to the target population from which cases and controls are sampled, and $\widetilde{\mathrm{E}}$ denote the expectation with respect to a ideal population in which exposure Z was eliminated at baseline. Let $\lambda(t)$ be the hazard function of failure time T for the target population and $\lambda\{t|\mathrm{Z} = 0, \mathrm{U}\}$ be the hazard function of T given Z = 0 and U. In the same spirit as the adjusted PAF, we define a natural PAH of Z in the presence of U as

$$\Phi_{adj}(t) = \frac{\lambda(t) - \widetilde{\mathrm{E}}_{\mathrm{U} \mid T \geq t}\{\lambda(t \mid \mathrm{Z} = 0, \mathrm{U})\}}{\lambda(t)}, \tag{3}$$

which is the proportion of reduction of the hazard at time t from the current population to the ideal population that Z is eliminated at baseline, accounting for confounders U. The conditional expectation $\widetilde{\mathrm{E}}_{\mathrm{U} \mid T \geq t}$ with respect to the ideal population is taken over U among subjects at risk at time t. Note that this can be obtained based on $\widetilde{\mathrm{E}}_{\mathrm{U} \mid T \geq t}[\lambda\{t \mid \mathrm{Z} = 0, \mathrm{U}\}] = \widetilde{\mathrm{E}}_{\mathrm{U}}[f(T = t \mid \mathrm{Z} = 0, \mathrm{U})] / \widetilde{\mathrm{E}}_{\mathrm{U}}[\Pr(T \geq t \mid \mathrm{Z} = 0, \mathrm{U})]$ if $\lambda\{t \mid \mathrm{Z}, \mathrm{U}\}$ and the distribution function $F_{\mathrm{U}}(\mathrm{u})$ at the baseline are known, as $\widetilde{\mathrm{E}}_{\mathrm{U}}$ is the expectation taken over U at baseline.

Comparing (3) with the adjusted PAF (2), we see that in $\Phi_{adj}(t)$, the probability of being diseased is replaced with the hazard function at time t, and the probability mass function of confounders is replaced with the density of confounders for subjects who are at risk at time t in the ideal population. Thus, $\Phi_{adj}(t)$ can be regarded as the instantaneous evaluation of the confounder-adjusted PAF at t, and measures the impact on the disease development at time t due to eliminating exposure at baseline after taking into account confounding.

However, in case-control studies where cases are over sampled, $\lambda\{t|\mathrm{Z}, \mathrm{U}\}$ and the baseline distribution function of (Z, U) are not estimable, and thus, $\Phi_{adj}(t)$ can not be identified. Here we provide an alternative that is identifiable from the case-control data. The definition is as follows:

$$\phi_{adj}(t) = \frac{\lambda(t) - \mathrm{E}_{\mathrm{U} \mid T \geq t}\{\lambda(t \mid \mathrm{Z} = 0, \mathrm{U})\}}{\lambda(t)}. \tag{4}$$

It is worth noting that while $\mathrm{E}_{\mathrm{U}|T \geq t}$ is taken with respect to the original (natural history) population conditional on $\{T \geq t\}$ in this original population, the expectation $\widetilde{\mathrm{E}}_{\mathrm{U} \mid T \geq t}$ is

conditional on $\{T \geq t\}$ with respect to an ideal population. The difference between the original and ideal populations is that at baseline subjects who have exposure in the original population are free of exposure in the ideal population. Now let $f(\mathrm{u} \mid T \geq t)$ and $\tilde{f}(\mathrm{u} \mid T \geq t)$ be the density functions of U conditional on $T \geq t$ in the original and ideal populations, respectively. Under the rare disease assumption, both density functions are approximately $f(u)$, which implies $\phi_{adj}(t) \approx \Phi_{adj}(t)$.

To check the sensitivity of this approximation to the rare disease assumption, we calculated the differences of $\phi_{adj}(t)$ and $\Phi_{adj}(t)$ numerically under a wide range of scenarios. When the disease probability $\Pr(T < 70) \leq 0.05$, $\phi_{adj}(t)$ approximates well $\Phi_{adj}(t)$ with absolute differences below 0.05 for the vast majority scenarios. When the disease probability increases, additional parameter restrictions (e.g., log-hazard ratio for the confounder < 2 when $0.05 < \Pr(T < 70) \geq 0.1$) need to be imposed for a good approximation. Appendix E shows more detailed results of the approximation of $\phi_{adj}(t)$ and $\Phi_{adj}(t)$. In real-data-based simulations in Section 3 and 4.2, the difference between the two is no more than several thousandths across a very wide range of t. Thus $\phi_{adj}(t)$ can be a promising alternative to $\Phi_{adj}(t)$ for accommodating the case-control data. We focus on $\phi_{adj}(t)$ in this article.

Assume the effects of Z and U on the hazard function of the failure time T follow the Cox proportional hazards model [27],

$$\lambda(t \mid \mathrm{Z}, \mathrm{U}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathrm{Z} + \boldsymbol{\gamma}^T \mathrm{U}), \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is a vector of regression parameters for Z, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)^T$ is a vector of regression parameters for U, and $\lambda_0(t)$ is the baseline hazard function. For the target population, let $F_{\mathrm{U} \mid T \geq t}(\mathrm{u})$ be the distribution function of confounders U of subjects who are disease-free at time t with the corresponding density function $f_{\mathrm{U} \mid T \geq t}(\mathrm{u})$ if U is continuous, and $\mathcal{U}$ be the space for U (*i.e.*, $\Pr\{\mathrm{U} \in \mathcal{U}\} = 1$). Similarly we denote $F_{\mathrm{Z} \mid T}(\mathrm{z} \mid t)$ and $F_{\mathrm{Z} \mid T}(\mathrm{z} \mid t)$ the distribution function and the density function of Z given $T = t$, respectively, $F_{\mathrm{Z,U} \mid T \geq t}(\mathrm{z}, \mathrm{u})$ and $f_{\mathrm{Z,U} \mid T \geq t}$ the joint distribution function and the joint density function of Z and U given $T \geq t$, respectively, and $\mathcal{Z}$ the space for Z.

By using the relationship $\lambda(t) = f_T(t)/S_T(t)$ and applying the total probability theorem on $S_T(t)$, we can express as $\phi_{adj}(t)$ as

$$\begin{aligned} \phi_{adj}(t) &= 1 - \frac{\int_{\mathcal{U}} \lambda(t \mid \mathrm{Z} = 0, \mathrm{u}) dF_{\mathrm{U} \mid T \geq t}(\mathrm{u})}{\lambda(t)} \\ &= 1 - \frac{\lambda_0(t)}{\lambda(t)} \int_{\mathcal{U}} \exp(\boldsymbol{\gamma}^T \mathrm{u}) f_{\mathrm{U} \mid T \geq t}(\mathrm{u}) d\mathrm{u}\,. \\ &= 1 - \left\{ \int_{\mathcal{U}} \int_{\mathcal{Z}} \exp(\boldsymbol{\beta}^T \mathrm{z} + \boldsymbol{\gamma}^T \mathrm{u}) f_{\mathrm{Z,U} \mid T \geq t}(\mathrm{z}, \mathrm{u}) d\mathrm{z} d\mathrm{u} \right\}^{-1} \int_{\mathcal{U}} \exp(\boldsymbol{\gamma}^T \mathrm{u}) f_{\mathrm{U} \mid T \geq t}(\mathrm{u}) d\mathrm{u}\,. \end{aligned} \tag{6}$$

The above equation explicitly show the impact of U. For example, when $\boldsymbol{\beta} = 0$, $\phi_{adj}(t) = 0$ for all t. When $\boldsymbol{\gamma} = 0$, we have $\phi_{adj}(t) = \phi(t)$ the unadjusted PAH. Furthermore, we have

$$\begin{aligned}\int_{\mathcal{U}} \exp(\boldsymbol{\gamma}^T \mathrm{u}) f_{\mathrm{U} \mid T \geq t}(\mathrm{u}) d\mathrm{u} &= \int_{\mathcal{U}} \exp(\boldsymbol{\gamma}^T \mathrm{u}) \int_{\mathcal{Z}} f_{\mathrm{Z}, \mathrm{U} \mid T \geq t}(\mathrm{z}, \mathrm{u}) d\mathrm{z} d\mathrm{u} \\ &= \int_{\mathcal{U}} \int_{\mathcal{Z}} \exp(\boldsymbol{\gamma}^T \mathrm{u} + \boldsymbol{\beta}^T \mathrm{z}) \exp(-\boldsymbol{\beta}^T \mathrm{z}) f_{\mathrm{Z}, \mathrm{U} \mid T \geq t}(\mathrm{z}, \mathrm{u}) d\mathrm{z} d\mathrm{u} \\ &= \int_{\mathcal{U}} \int_{\mathcal{Z}} \frac{f_{T \mid \mathrm{Z}, \mathrm{U}}(t \mid \mathrm{z}, \mathrm{u})}{S(t \mid \mathrm{z}, \mathrm{u}) \lambda_0(t)} \exp(-\boldsymbol{\beta}^T \mathrm{z}) f_{\mathrm{Z}, \mathrm{U} \mid T \geq t}(\mathrm{z}, \mathrm{u}) d\mathrm{z} d\mathrm{u} \\ &= \int_{\mathcal{U}} \int_{\mathcal{Z}} \frac{f_{T, \mathrm{Z}, \mathrm{U}}(t, \mathrm{z}, \mathrm{u})}{S(t) \lambda_0(t)} \exp(-\boldsymbol{\beta}^T \mathrm{z}) d\mathrm{z} d\mathrm{u} \\ &= \frac{f_T(t)}{S(t) \lambda_0(t)} \int_{\mathcal{U}} \int_{\mathcal{Z}} \exp(-\boldsymbol{\beta}^T \mathrm{z}) f_{\mathrm{Z}, \mathrm{U} \mid T}(\mathrm{z}, \mathrm{u} \mid t) d\mathrm{z} d\mathrm{u} \\ &= \frac{\lambda(t)}{\lambda_0(t)} \int_{\mathcal{Z}} \exp(-\boldsymbol{\beta}^T \mathrm{z}) f_{\mathrm{Z} \mid T}(\mathrm{z} \mid t) d\mathrm{z},\end{aligned}$$

and thus we can also express (6) as

$$\phi_{adj}(t) = 1 - \int_{\mathcal{Z}} \exp(-\boldsymbol{\beta}^T \mathrm{z}) f_{\mathrm{Z} \mid T}(\mathrm{z} \mid t) d\mathrm{z}. \tag{7}$$

Equation (7) involves hazard ratios and conditional distribution of Z given $T = t$, i.e. cases who develop disease at time t. Complementary to (7), equation (6) involves hazard ratios and conditional distributions of $(\mathrm{Z}^T, \mathrm{U}^T)^T$ and U given $T > t$, *i.e.*, controls who are disease-free at time t.

2.2 Estimation and Large Sample Properties

Now we describe estimation of $\phi_{adj}(t)$ on the basis of (6) and (7) from case-control data. It is well established that if cases and controls are time-matched, hazard ratio parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ can be consistently estimated by maximizing conditional likelihood function with logistic regression model[22]. For unmatched case-control studies, $\boldsymbol{\theta}$ can be estimated by the convention maximum likelihood of a logistic regression model with time adjusted as a covariate. The approximation is generally not consistent, but yields little bias in practice. We denote the hazard ratio estimates by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)$. For $\phi_{adj}(t)$, what remains to be estimated is the conditional distribution function $F_{\mathrm{V} \mid T}(\mathrm{v} \mid t)$ and $F_{\mathrm{V} \mid T \geq t}(\mathrm{v})$. Here V represents Z, U, or $(\mathrm{Z}^T, \mathrm{U}^T)^T$.

Under the random censoring assumption that the censoring time is independent of the failure time, exposures and confounders as derived in Xu and O'Quigley[28], we can show that $F_{\mathrm{V} \mid T}(\mathrm{v} \mid t)$ equals to $F(\mathrm{v} \mid T = t, T \leq C) = P\{V \leq \mathrm{v} \mid T = t, T \leq C\}$, which can be estimated from the cases, and $F_{\mathrm{V} \mid T \geq t}(\mathrm{v})$ equals to $F(\mathrm{v} \mid C = t, T \geq C) = P\{\mathrm{V} \leq \mathrm{v} \mid C = t, T \geq C\}$, a quantity that can be estimated from the controls. Therefore, either $F_{\mathrm{V} \mid \mathrm{T}}(\mathrm{v} \mid t)$ or $F_{\mathrm{V} \mid T \geq t}(\mathrm{v})$ can be estimated by its empirical estimator. However, these empirical estimators would have poor performance because of few subjects observed at time t. To improve the performance, we use a kernel smoother to estimate these distribution functions. Let

$$\varphi(t) = \int_{\mathcal{Z}} exp(-\boldsymbol{\beta}^T \mathrm{z})\, f_{\mathrm{Z} \mid T}(\mathrm{z} \mid t) d\mathrm{z},$$
$$\psi(t) = \int_{\mathcal{U}} \int_{\mathcal{Z}} exp(\boldsymbol{\beta}^T \mathrm{z} + \boldsymbol{\gamma}^T \mathrm{u}) f_{\mathrm{Z,U} \mid T \geq t}(\mathrm{z,u}) d\mathrm{z} d\mathrm{u},$$
$$\upsilon(t) = \int_{\mathcal{U}} \exp(\boldsymbol{\gamma}^T \mathrm{u}) f_{\mathrm{U} \mid T \geq t}(\mathrm{u}) d\mathrm{u}\,.$$

Then $\phi_{adj}(t)$ can be expressed as $\phi_{adj}(t) = 1 - \varphi(t)$ or $\phi_{adj}(t) = 1 - \psi^{-1}(t)\upsilon(t)$.

Consider n subjects in a case-control study. Let X_i, Δ_i, Z_i and U_i be the observed time, the censoring indicator, the exposure and confounders, respectively, for $i = 1, \ldots, n$. Assume that $\{(X_i, \Delta_i, \mathrm{Z}_i, \mathrm{U}_i), i = 1, \ldots, n\}$ are independently and identically distributed. We propose the following estimators for $\varphi(t)$, $\psi(t)$, and $\upsilon(t)$, respectively:

$$\hat{\varphi}(t; \hat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^{n} \exp(-\hat{\boldsymbol{\beta}}^T \mathrm{Z}_i) \Delta_i K_h(t - X_i)}{\sum_{i=1}^{n} \Delta_i K_h(t - X_i)},$$
$$\hat{\psi}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \frac{\sum_{i=1}^{n} \exp(\hat{\boldsymbol{\beta}}^T \mathrm{Z}_i + \hat{\boldsymbol{\gamma}}^T \mathrm{U}_i)(1 - \Delta_i) K_h(t - X_i)}{\sum_{i=1}^{n} (1 - \Delta_i) K_h(t - X_i)},$$
$$\hat{\upsilon}(t; \hat{\boldsymbol{\gamma}}) = \frac{\sum_{i=1}^{n} \exp(\hat{\boldsymbol{\gamma}}^T \mathrm{U}_i)(1 - \Delta_i) K_h(t - X_i)}{\sum_{i=1}^{n} (1 - \Delta_i) K_h(t - X_i)},$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a kernel function that is a weighting function that satisfies $\int K(x)dx = 1$, and h is the bandwidth that controls the spread of weighting window.

We propose two estimators for $\phi_{adj}(t)$, corresponding to equations (7) and (6), respectively:

$$\hat{\phi}_{adj+}(t; \hat{\boldsymbol{\beta}}) = 1 - \hat{\varphi}(t; \hat{\boldsymbol{\beta}}), \tag{8}$$

and

$$\hat{\phi}_{adj-}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = 1 - \hat{\psi}^{-1}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \hat{\upsilon}(t; \hat{\boldsymbol{\gamma}})\,. \tag{9}$$

Since the kernel estimators of $\hat{\phi}_{adj+}(t; \hat{\boldsymbol{\beta}})$ and $\hat{\phi}_{adj-}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ are based on cases and controls separately, combining the two estimators could potentially improve the efficiency. We therefore propose a weighted estimator as follows:

$$\hat{\phi}_{adjw}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = w(t)\hat{\phi}_{adj+}(t; \hat{\boldsymbol{\beta}}) + \{1 - w(t)\}\hat{\phi}_{adj-}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\,.$$

where $w(t)$ is a weighting function with value between 0 and 1. A natural choice of the weight is i $w(t) = \pi$q, which is proportion of cases in the sample. Then the weighted estimator is

$$\hat{\phi}_{adjw}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \pi_0 \hat{\phi}_{adj+}(t; \hat{\boldsymbol{\beta}}) + (1 - \pi_0)\hat{\phi}_{adj-}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\,. \tag{10}$$

Next we derive the asymptotic properties of $\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$. Assume the number of cases n_1 and the total number of subjects n satisfy $n_1/n \to \pi_0$ as $n \to \infty$, where $0 < \pi_0 < 1$. In addition, we assume the following regularity conditions:

A1. The time t is in a range of $[0, \tau]$ for a constant $\tau > 0$ such that the density of failure time $f_T(t)$, the density of censoring time $f_C(t)$ and their survival functions, $S_T(t)$ and $S_C(t)$ all take positive real values on $[0, \tau]$.

A2. The density of failure time $f_T(t)$ and the density of censoring time $f_C(t)$ are both continuous, uniformly bounded, and have second derivatives on $[0, \tau]$.

A3. Random censoring: the censoring time C is independent of the failure time T, exposure Z and confounder U for $t \in [0, \tau]$.

A4. The bandwidth satisfies $h = n^d h_0$ for constants $-1/2 < d < -1/5$ and $h_0 > 0$.

A5. The kernel function $K(\cdot)$ has bounded variation and satisfies the following conditions,

$$\int_{-\infty}^{\infty} K(u)d\mathrm{u} = 1, \quad \int_{-\infty}^{\infty} K^2(u)d\mathrm{u} < \infty,$$
$$\int_{-\infty}^{\infty} uK(u)d\mathrm{u} = 0, \quad \int_{-\infty}^{\infty} u^2K(u)d\mathrm{u} < \infty.$$

A6. Z and U are bounded almost surely and have uniformly bounded total variation on $[0, \tau]$.

We define the following notation.

$$B_n(t) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i K_h(t - X_i),$$
$$\mathrm{D}_n(t;\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i K_h(t - X_i)\mathrm{Z}_i \exp(-\boldsymbol{\beta}^T \mathrm{Z}_i),$$
$$\bar{A}_n(t;\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} (1-\Delta_i) K_h(t - X_i) \exp(\boldsymbol{\beta}^T \mathrm{Z}_i + \boldsymbol{\gamma}^T \mathrm{U}_i),$$
$$\bar{B}_n(t) = \frac{1}{n}\sum_{i=1}^{n} (1-\Delta_i) K_h(t - X_i),$$
$$\bar{\mathrm{D}}_n(t;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} (1-\Delta_i) K_h(t - X_i)\mathrm{V}_i \exp(\boldsymbol{\theta}^T \mathrm{V}_i),$$
$$\bar{\mathrm{G}}_n(t;\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} (1-\Delta_i) K_h(t - X_i)\mathrm{U}_i \exp(\boldsymbol{\gamma}^T \mathrm{U}_i).$$

We use P_0 and E_0 to denote the probability and expectation with respect to the target population from which cases and controls are sampled. As shown in Appendix A and B, it is also useful to regard cases and controls as members of a second, hypothetical population of individuals whose disease probability is given by π_0[13]. We use P^* and E^* to denote the probability and expectation with respect to this hypothetical population. Let $p_0 = P_0(T \le C)$,

$\pi_0 = P^*(T \le C)$. Denote the limits of $B_n(t)$, $D_n(t;\boldsymbol{\beta})$, $\bar{A}_n(t;\boldsymbol{\beta},\boldsymbol{\gamma})$, $\bar{B}_n(t)$, $\bar{\mathrm{D}}_n(t;\boldsymbol{\theta})$ and $\bar{\mathrm{G}}_n(t;\boldsymbol{\gamma})$ by $B(t)$, $D(t;\boldsymbol{\beta})$, $\bar{A}(t)$, $\bar{B}(t)$, $\bar{\mathrm{D}}(t;\boldsymbol{\theta})$ and $\bar{\mathrm{G}}(t;\boldsymbol{\gamma})$ the detailed expressions of which can be found in Appendix A. The main results, namely consistency and asymptotic normality of the proposed estimators, are summarized in the following two theorems.

Theorem 1. Consistency. Suppose assumptions A1-A6 are satisfied. Then as $n \to \infty$, $\hat{\phi}_{adj+}(t;\hat{\boldsymbol{\beta}})$, $\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ and $\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ are uniformly consistent for $\phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)$ for $t \in [0,\tau]$, where $\phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)$ is the true value of $\phi_{adj}(t)$ defined in (4) under the Cox proportional hazards model (5).

Theorem 2. Asymptotic Normality. Suppose assumptions A1-A6 are satisfied. Then

$$\sqrt{nh}\{\hat{\phi}_{adj+}(t;\hat{\boldsymbol{\beta}}) - \phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\} \to_d N(0,\sigma_+^2(t)), \qquad \text{for } t \in [0,\tau],$$

where the limiting variance $\sigma_+^2(t) = \int K^2(u)du \int \{\exp(-\boldsymbol{\beta}_0^T \mathrm{z}) - \phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\}^2 dF_{\mathrm{Z}\mid T}(\mathrm{z}) \Big/ B(t)$. We also have

$$\sqrt{nh}\{\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - \phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\} \to_d N(0,\sigma_-^2(t)), \qquad \text{for } t \in [0,\tau],$$

where the limiting variance $\sigma_-^2(t) = \bar{B}(t)\bar{A}^{-2}(t)\int K^2(u)du \int\int \{\exp(\boldsymbol{\gamma}_0^T \mathrm{u}) - \psi^{-1}(t)\upsilon(t)\exp(\boldsymbol{\beta}_0^T \mathrm{z} + \boldsymbol{\gamma}_0^T \mathrm{u})\}^2 f_{\mathrm{Z,U}\mid T\ge t}(\mathrm{z},\mathrm{u})d\mathrm{z}d\mathrm{u}$. The asymptotic normality of the weighted estimator $\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ then follows, that is, $\sqrt{nh}\{\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - \phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\} \to_d N(0,\sigma_w^2(t))$ for $t \in [0,\tau]$, where $\sigma_w^2(t) = \pi_0^2\sigma_+^2(t) + (1-\pi_0)^2\sigma_-^2(t)$.

The proofs of Theorem 1 and 2 are provided in Appendix A and B.

2.3 Variance Estimation with Correction for Finite Sample

One natural estimator for $\sigma_+^2(t)$, $\sigma_-^2(t)$ or $\sigma_w^2(t)$, denoted as $\hat{\sigma}_+^2(t), \hat{\sigma}_-^2(t)$ and $\pi_0^2\hat{\sigma}_+^2(t) + (1-\pi_0)^2\hat{\sigma}_-^2(t)$, can be obtained by replacing the expectation components with the corresponding empirical estimators, and replacing parameters $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)$ with their estimators. However, variance estimates relying solely on the asymptotic results may perform poorly due to the slow vanishing rate of $\sqrt{h}$. In fact, in real practice with finite samples, $\sqrt{h}$ might not be close to zero even with large sample size. In this case some components in $\sqrt{nh}\{\hat{\phi}_{adj+}(t;\hat{\boldsymbol{\beta}}) - \phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\}$, $\sqrt{nh}\{\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - \phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\}$, and $\sqrt{nh}\{\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - \phi_{adj}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\}$ cannot be ignored even though their asymptotic contribution to the overall variance is negligible. Instead, we need to consider their contributions to the asymptotic results by taking the bandwidth h as a known non-zero constant. The variance estimators with finite sampling correction are

$$\hat{\sigma}_+^{*2}(t;h) = \hat{\sigma}_+^2(t) + h\mathrm{D}_n(t;\hat{\boldsymbol{\beta}})^T\hat{\mathrm{I}}^{-1}(\hat{\boldsymbol{\beta}})\mathrm{D}_n(t;\hat{\boldsymbol{\beta}}) \,/\, B_n^2(t)$$
$$+ \frac{2h}{n}\sum_{i=1}^{n}\Delta_i K_h(t-X_i)\{1-\hat{\phi}_{adj+}(t;\hat{\boldsymbol{\beta}}) - exp(-\hat{\boldsymbol{\beta}}^T\mathrm{Z}_i)\}\hat{\mathrm{I}}_{\boldsymbol{\beta}}(X_i)^T\mathrm{D}_n(t;\hat{\boldsymbol{\beta}}) \,/\, B_n^2(t),$$

$$\hat{\sigma}_-^{*2}(t;h) = \hat{\sigma}_-^2(t) + \frac{h}{n}\sum_{i=1}^{n}[\hat{\mathrm{I}}_{\boldsymbol{\gamma}}(X_i)^T\overline{\mathrm{G}}_n(t;\hat{\boldsymbol{\gamma}}) - \{1-\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}\hat{\mathrm{I}}_{\boldsymbol{\theta}}(X_i)^T\overline{\mathrm{D}}_n(t;\hat{\boldsymbol{\theta}})]^2 \,/\, \overline{A}_n^2(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$$
$$+ \frac{2h}{n}\sum_{i=1}^{n}(1-\Delta_i)K_h(t-X_i)\left[\exp(\hat{\boldsymbol{\gamma}}^T\mathrm{U}_i) - \{1-\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}\exp(\hat{\boldsymbol{\beta}}^T\mathrm{Z}_i + \hat{\boldsymbol{\gamma}}^T\mathrm{U}_i)\right]$$
$$\times[\hat{\mathrm{I}}_{\boldsymbol{\gamma}}(X_i)^T\overline{\mathrm{G}}_n(t;\hat{\boldsymbol{\gamma}}) - \{1-\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}\hat{\mathrm{I}}_{\boldsymbol{\theta}}(X_i)^T\overline{\mathrm{D}}_n(t;\hat{\boldsymbol{\theta}})] \,/\, \overline{A}_n^2(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}),$$

$$\hat{\sigma}_w^{*2}(t;h) = \pi_0^2\hat{\sigma}_+^{*2}(t;h) + (1-\pi_0)^2\hat{\sigma}_-^{*2}(t;h) + 2\pi_0(1-\pi_0)\hat{\sigma}_{+-}^*(t;h),$$

where $\hat{\mathrm{I}}^{-1}(\hat{\boldsymbol{\beta}})$ is the estimated information matrix for $\boldsymbol{\beta}$, $\hat{\mathrm{I}}_{\boldsymbol{\beta}}(X_i)$, $\hat{\mathrm{I}}_{\boldsymbol{\gamma}}(X_i)$ and $\hat{\mathrm{I}}_{\boldsymbol{\theta}}(X_i)$ are the estimated efficient influence functions for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, respectively, and

$$\hat{\sigma}_{+-}^*(t;h) = -\{\hat{\sigma}_{AF}(t;h)\overline{A}_n^{-1}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})B_n^{-1}(t) + \hat{\sigma}_{CE}(t;h)\overline{A}_n^{-1}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})B_n^{-1}(t)$$
$$+ \hat{\sigma}_{CF}(t;h)\overline{A}_n^{-1}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})B_n^{-1}(t)\}$$

$$\hat{\sigma}_{AF}(t;h) = \frac{h}{n}\sum_{i=1}^{n}\Delta_i K_h(t-X_i)\{\hat{\varphi}(t;\hat{\boldsymbol{\beta}}) - \exp(-\hat{\boldsymbol{\beta}}^T\mathrm{Z}_i)\}$$
$$\times[\hat{\mathrm{I}}_{\boldsymbol{\gamma}}(X_i)^T\overline{\mathrm{G}}_n(t;\hat{\boldsymbol{\gamma}}) - \{1-\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}\hat{\mathrm{I}}_{\boldsymbol{\theta}}(X_i)^T\overline{\mathrm{D}}_n(t;\hat{\boldsymbol{\theta}})],$$

$$\hat{\sigma}_{CE}(t;h) = \frac{h}{n}\sum_{i=1}^{n}(1-\Delta_i)K_h(t-X_i)\left[\exp(\hat{\boldsymbol{\gamma}}^T\mathrm{U}_i) - \{1-\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}\exp(\hat{\boldsymbol{\beta}}^T\mathrm{Z}_i + \hat{\boldsymbol{\gamma}}^T\mathrm{U}_i)\right]$$
$$\times\{\hat{\mathrm{I}}_{\boldsymbol{\beta}}(X_i)^T\mathrm{D}_n(t;\hat{\boldsymbol{\gamma}})\},$$

$$\hat{\sigma}_{CF}(t;h) = \frac{h}{n}\sum_{i=1}^{n}\{\hat{\mathrm{I}}_{\boldsymbol{\beta}}(X_i)^T\mathrm{D}_n(t;\hat{\boldsymbol{\gamma}})\}[\hat{\mathrm{I}}_{\boldsymbol{\gamma}}(X_i)^T\overline{\mathrm{G}}_n(t;\hat{\boldsymbol{\gamma}}) - \{1-\hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}\hat{\mathrm{I}}_{\boldsymbol{\theta}}(X_i)^T\overline{\mathrm{D}}_n(t;\hat{\boldsymbol{\theta}})].$$

The consistency of $\hat{\sigma}_+^{*2}(t;h)$, $\hat{\sigma}_-^{*2}(t;h)$ and $\hat{\sigma}_w^{*2}(t;h)$ are summarized in the following theorem with the proof and derivation of the correction terms provided in Appendix C.

Theorem 3. Suppose that assumptions A1-A6 are satisfied. Then for $t \in [0, \tau]$, $\hat{\sigma}_+^{*2}(t;h)$, $\hat{\sigma}_-^{*2}(t;h)$ and $\hat{\sigma}_w^{*2}(t;h)$ are uniformly consistent for $\sigma_+^2(t)$, $\sigma_-^2(t)$ and $\sigma_w^2(t)$.

Based on the variance estimator, we can construct the point-wise $100(1-\alpha\%)$ confidence interval for the three estimators:

$$\begin{aligned}\hat{\phi}_{adj+}(t;\hat{\boldsymbol{\beta}}) &\mp z_{1-\alpha/2}(nh)^{-1/2}\hat{\sigma}_+^*(t;h),\\ \hat{\phi}_{adj-}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) &\mp z_{1-\alpha/2}(nh)^{-1/2}\hat{\sigma}_-^*(t;h),\\ \hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) &\mp z_{1-\alpha/2}(nh)^{-1/2}\hat{\sigma}_w^*(t;h),\end{aligned} \tag{11}$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard distribution.

In practice, it is also often of interest to construct simultaneous $100(1-\alpha)\%$ confidence bands. Here we use the bootstrap resampling approach to construct the confidence bands[29,30]. The original bootstrap method, random resampling with replacement, works well from the simulation studies. For simplicity, we only illustrate the construction of confidence

band for $\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$. Consider a process $\sqrt{nh}\{\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})-\phi_{adjw}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\}$. For each resampling, denote the resultant estimator as $\tilde{\phi}_{adjw}(t;\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}})$ and its corresponding variance estimator as $\tilde{\sigma}^{*2}_w(t;h)$, calculated using the resampled data. We use the process $\sqrt{nh}\{\tilde{\phi}_{adjw}(t;\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}})-\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}$ to approximate $\sqrt{nh}\{\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})-\phi_{adjw}(t;\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\}$ by simulating a large number of realizations through random resampling with replacement. Then the critical value for 100(1–α)th percentile simultaneous confidence band can be calculated based on these simulations:

$$\Pr\left\{\sup_{t\in[0,\tau]}\frac{|\sqrt{nh}\{\tilde{\phi}_{adjw}(t;\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}})-\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\}|}{\tilde{\sigma}^{*}_w(t;h)}\le\tilde{z}_{1-\alpha/2}\right\}-1-\alpha, \tag{12}$$

and the resulting confidence band for the time range of $[0,\ \tau]$ is

$$\hat{\phi}_{adjw}(t;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\ \mp\ \tilde{z}_{1-\alpha/2}(nh)^{-1/2}\hat{\sigma}^{*}_w(t;h)\,. \tag{13}$$

2.4 Kernel and Bandwidth Selection

In the kernel estimation, the performance of the estimator depends on choices of the kernel function and the bandwidth. While the kernel function has much less impact, it is known that bandwidth choice is critical to the adequacy of the estimator. Here we use the Epanechinikov kernel given its optimization property[31] and a "leave-one-out" least squares cross-validation approach which is a well-working approach for automatically optimizing the bandwidth[32,33]. The idea is to split the data into two parts, the larger part that contains all but one subject, and the smaller part that contains one subject, then use the larger part of the data for estimation and the smaller part for evaluation of accuracy. The optimal bandwidth would minimize the average prediction squared errors. In particular, let

$$\hat{\phi}^{(-j)}_{adj+}(X_j;h)=1-\frac{\sum_{i=1}^{n}I(i\neq j)\exp(-\hat{\boldsymbol{\beta}}^{T}Z_i)\Delta_i K_h(X_j-X_i)}{\sum_{i=1}^{n}I(i\neq j)\Delta_i K_h(X_j-X_i)},$$
$$\hat{\phi}^{(-j)}_{adj-}(X_j;h)=1-\frac{\sum_{i=1}^{n}I(i\neq j)\exp(\hat{\boldsymbol{\gamma}}^{T}U_i)(1-\Delta_i)K_h(X_j-X_i)}{\sum_{i=1}^{n}I(i\neq j)\exp(\hat{\boldsymbol{\beta}}^{T}Z_i+\hat{\boldsymbol{\gamma}}^{T}U_i)(1-\Delta_i)K_h(X_j-X_i)}.$$

which are the estimators of $\phi_{adj}(X_j)$ computed with all but the jth subjects. Then the cross-validation loss function is given by

$$CV(h)=\frac{1}{n}\sum_{j=1}^{n}\left\{1-\exp(-\hat{\boldsymbol{\beta}}^{T}Z_j)-\hat{\phi}^{(-j)}_{adjw}(X_j;h)\right\}^2,$$

where $\hat{\phi}^{(-j)}_{adjw}(X_j;h)$ is $\pi_0\hat{\phi}^{(-j)}_{adj+}(X_j;h)+(1-\pi_0)\hat{\phi}_{adj-}(X_j;h)$ if $\Delta_j=1$ and $\pi_0\hat{\phi}_{adj+}(X_j;h)+(1-\pi_0)\hat{\phi}^{(-j)}_{adj-}(X_j;h)$ if $\Delta_j=0$. A cross-validation bandwidth is obtained by minimizing $CV(h)$ with respect to h, $\hat{h}_{CV}=\arg\min_{h>0}CV(h)$. In practice, this approach generally works well.

3 Simulation Studies

We conducted simulation studies to evaluate the finite sample performance of the proposed estimators for case-control data. Specifically, we first generated a large population with 100000 subjects. For each subject, we generated an exposure and a confounder. We then generated a failure time T based on the Cox model (5) with a Weibull baseline hazard $\lambda_0(t) = (\nu/\eta)(t/\eta)^{\nu-1}$, where ν and η are shape and scale parameters, respectively. We also generated an independent censoring time C. The observed time X was the minimum of T and C, and the disease status $\Delta = 1$ if $T \leq C$ and $\Delta = 0$ if $T > C$. We obtained the case-control sample by randomly sampling 2000 cases ($\Delta = 1$), and 1000, 2000 or 4000 controls ($\Delta = 0$) with time matched to the cases within five-year intervals. Thus the observed data for analysis consist of X, Z, U and Δ. We estimated (β^T, γ^T) using conventional logistic regression adjusting for age. The proposed estimators were obtained using the Epanechnikov kernel (*i.e.*, $K(x) = 0.75(1 - x^2)I_{|x|<1}$) and the bandwidth was selected by the proposed automatic cross-validation approach. We also obtained estimators with a range of fixed bandwidths to evaluate the performance of the automatic bandwidth selector.

We set parameters in the baseline hazard function as $\eta = 180$ or 360, and $\nu = 2$. For generating (Z, U), we considered two scenarios. In scenario I, (Z, U) jointly follow Bernoulli distributions with the probabilities of (Z, U) being (1, 1), (1, 0), (0, 1), and (0, 0) as 0.35, 0.15, 0.15, and 0.35, respectively. In scenario II, we generated a continuous U which follows a normal distribution $N(0, 0.5)$ and then a binary Z was generated with the probability of $Z = 1$ as $\exp(U)/\{\exp(U) + 1\}$. The log-hazard ratios of Z and U were set to be log(3) and log(1.5), respectively. Censoring time was truncated normal between 1 and 100 with standard deviation of 30, and the mean was set to yield censoring probability of 70% or 80% for $\eta = 180$, 90% or 95% for $\eta = 360$. For each simulation scenario, a total of 2000 simulated datasets were generated. It can be seen from Table 1 that $\phi_{adj}(t)$ decreases monotonically, as subjects who had the exposure experienced diseases at an earlier age, leaving fewer subjects with exposure in the older age. The performance of the proposed estimators was assessed by following summary statistics: bias, empirical standard deviation (ESD), asymptotic-based standard error (ASE), and 95% coverage rate at selected age t. In addition, we calculated the coverage rate of the 95% simultaneous confidence band calculated by bootstrap with 200 repetitions, across the range of the observed ages. Specifically, the bias was calculated by taking the absolute difference between the mean of the point estimates and the true value of $\phi_{adj}(t)$. The ESD was the empirical standard deviation of the point estimates, and ASE was the average of the standard error estimators over the 2000 simulated datasets. The 95% pointwise coverage rate was the proportion of 95% estimated confidence intervals that covered the true value $\phi_{adj}(t)$ at time t.

Table 1 shows the summary results of the simulations under scenarios I and II for 80% censoring and an equal number of cases and controls. Results of other combinations of censoring probability and case/control ratio have similar patterns and are presented in Appendix D. We see $\phi_{adj}(t)$ approximates $\Phi_{adj}(t)$ very well. The bias for the proposed estimators is small under both scenarios across a wide range of ages. The ASE is close to the ESD. The estimated pointwise coverage rates are generally close to 95%, so are the simultaneous confidence bands. This indicates that the proposed estimators and the

asymptotic-based variance estimators perform well. We also evaluated the proposed methods across a wide range of bandwidths and present the results in Appendix D. As expected from the bias-variance trade-off, when a small bandwidth is used, the bias tends to be smaller while the variance becomes larger, and when the bandwidth is larger, the bias tends to be greater especially at late ages, and the variance is smaller. The cross-validation selected bandwidths balance the bias and variance, yielding satisfactory estimates. As expected, $\hat{\phi}_{adjw}(t;\hat{\beta},\hat{\gamma})$ is more robust than either of $\hat{\phi}_{adj+}(t;\hat{\beta})$ and $\hat{\phi}_{adj-}(t;\hat{\beta},\hat{\gamma})$ in terms of bias, efficiency and coverage rates.

4 An Application to a Case-Control Study of Colorectal Cancer

4.1 Real Data Analysis

We apply our proposed methods to three case-control studies from the Genetics and Epidemiology Colorectal Cancer Consortium (GECCO). GECCO consists of dozens of well-characterized (nested) case-control studies of colorectal cancer (CRC)[34]. The broad objective of the consortium is to evaluate both lifestyle, environmental and genetic risk factors in relation to colorectal cancer risk. Key clinical and environmental data have been harmonized across all studies. Although CRC is one of the common cancers, it is still rare in the general population. As a result, the most common study design for studying CRC is the case-control study design. For illustration, the subset of the data we used includes 5498 subjects (2742 cases and 2756 controls) from three cohort-based nested case-control studies, for which controls are frequency matched on both age and gender and the risk factor information was collected at the study entry. All models are adjusted for study, age and gender. The aim of this data analysis is to estimate the adjusted PAHs of various risk factors to colorectal cancer.

In the analysis, we focus on history of diabetes (yes/no) and obesity (body mass index, BMI, >30kg/m^2). Both are risk factors for colorectal cancer and obesity is a risk factor for diabetes. It is therefore of interest to examine the adjusted PAH of obesity accounting for diabetes history, and vice versa. For ever-smokers, we also examine the adjusted PAH of years-since-quit-smoking (≤ 10 years vs > 10 years) while adjusting for pack-years (≥ 22.5 vs < 22.5), and vice versa.

The descriptive statistics of risk factors as well as matching variables age and gender by case-control status are summarized in Table 2. It also includes estimates of odds ratios, classic Levin's PAFs[7], and the model-based adjusted PAFs for each of the risk factors[35], treating the outcome colorectal cancer as binary (yes/no). By study design, age and gender are balanced between cases and controls. History of diabetes, obesity, and ever smoking are more common in cases than controls. For ever-smokers, cases also have a higher average of pack-year and shorter years-since-quit-smoking. History of diabetes has the highest OR estimate of 1.57, however, the prevalence of exposure is low with 7.5% in cases and 4.7% in controls, resulting in an estimated unadjusted PAF of only 0.029 (95% confidence interval (CI): 0.016–0.043). In contrast, the odds ratio estimate for obesity is 1.17 (95% CI: 1.04–1.32) and the prevalence is 29.5% in cases and 25.7% in controls. As a result, the unadjusted PAF is 0.051 (95% CI: 0.020-0.083), which is greater than obesity. The adjusted PAF

estimates are roughly the same, suggesting that both risk factors contribute to colorectal cancer risk. For ever-smokers, pack-year has an estimated unadjusted PAF of 0.153 (95% CI: 0.089-0.218) and years-since-quit-smoking has an estimate of 0.073 (95% CI: 0.024-0.121). After adjusting for pack-year, the adjusted PAF for years-since-quit-smoking is reduced roughly to half.

We estimated the adjusted PAH for each of the risk factors using the weighted estimator $\hat{\phi}_{adjw}(t;\hat{\beta},\hat{\gamma})$, because it has the most robust performance as shown in the simulation studies. The Epanechnikov kernel was used and the bandwidth was obtained by cross-validation. We fit a logistic regression model for obesity and diabetes history, and likewise for pack-year and years-since-quit-smoking in ever-smokers. As comparison, we also calculated the unadjusted PAH[13]. To test whether the constant hazard ratio hold for the underlying Cox model, an interaction term between time and the risk factor was added to the logistic model and the interaction term was not significantly different with 0 for each of the four risk factors at the 0.05 level.

It is worth noting that while obesity is a well-known confounder for the relationship between diabetes and CRC, it may not be true that diabetes is a confounder for obesity and CRC. Therefore, the PAH of obesity adjusting for diabetes need to be interpreted with caution. As diabetes may likely be on the pathway from obesity to CRC risk, such an adjusted PAH may be interpreted as the proportion of reduction of the hazard due to the direct effect of obesity with the indirect effect of obesity to CRC through diabetes unchanged. If one were to measure the impact of eliminating the total effect of obesity, adjusting for diabetes in the PAH of obesity might not be needed.

Figure 1 shows both unadjusted and adjusted PAHs for each of the four risk factors on CRC. The adjusted PAH for diabetes after adjusting for obesity is nearly unchanged compared to the unadjusted PAH, while the adjusted PAH for obesity after adjusting for diabetes is slightly lower than the unadjusted PAH and the mean difference across the age is about 0.02. The 95% confidence intervals for both adjusted PAHs of obesity and diabetes exclude 0, suggesting that despite somewhat reduced the PAH for obesity after adjusting for diabetes, both risk factors contribute to colorectal cancer risk. For ever-smokers, the PAH for pack-year after adjusting for years-since-quit-smoking is essentially the same as the unadjusted PAH. In contrast, the PAH for the years-since-quit-smoking is greatly reduced after adjusting for pack-year, and the 95% confidence intervals include 0 over time, suggesting that the PAH of years-since-quit-smoking is largely explained by pack-year.

The PAH for diabetes, year-since-quit-smoking and pack-years are approximately flat. The PAH for obesity decreases over time, from 8.8% at age 50 years old to 4.4% at age 80 years old, suggesting an early intervention may possibly reduce risk for early-onset colorectal cancer. Estimates of the adjusted PAHs and 95% confidence intervals at selected ages, and figures of the simultaneous confidence bands can be found in Appendix D.

4.2 Real Data Based Simulation

To assess the performance of the proposed weighted estimator for the real data, we conducted a simulation study mimicking the scenarios for diabetes and obesity as shown in

the real data example. Specifically, we generated two binary covariates (diabetes and obesity) from a multinomial distribution that the probabilities of having diabetes only, obesity only and both are 4.7%, 25.7% and 2.3%, respectively, same as what we observed in the controls of the GECCO data. The failure time T was generated based on the Cox proportional hazards model with log-hazard ratios as the estimated coefficients from logistic regression for diabetes (0.45) and obesity (0.15). The baseline hazard function was chosen such that the age-specific colorectal cancer incidence rates follow the Surveillance Epidemiology and End Results Registry (SEER) registry (https://seer.cancer.gov/data/). The probability of developing colorectal cancer is 2%, which can be considered as a rare disease scenario. We generated independent right censoring time as the minimum of current age and age at death obtained from the US lifetable. We generated a large population with 300000 subjects. The case-control sample was obtained by randomly sampling 2500 cases, and 1250, 2500 or 5000 controls with age matched to the cases within five-year intervals. The estimation and inference were identical as in Section 3. A total of 2000 simulated datasets were generated.

Two exposures were considered: (1) diabetes as the exposure and obesity as the confounder; (2) obesity as the exposure and diabetes as the confounder. The same summary statistics as in Section 3 were calculated to assess the performance of the proposed estimators. The results for equal number of cases and controls are presented in Table 3, and the results for other case/control ratios are presented in Appendix D. The bias for the proposed estimators is small across a wide range of ages. The ASE approximates the ESD well. The pointwise coverage rates are close to 95%, so are the simultaneous confidence bands. Note that mild under-coverage for the scenario of diabetes as the exposure at 40 and 80 years is mainly due to relatively small sample sizes. The average sample sizes within the kernel range at 40, 60 and 80 years are 714, 1746 and 753 for the method using only cases, 795, 1887 and 828 for the method using only controls, 1523, 3657 and 1597 for the method using both cases and controls, respectively. In this situation, the kernel bandwidth selection balances between bias and variance, which may lead to moderate bias and/or somewhat imprecise estimate of standard error. For age 40, the standard error approximates the empirical standard deviation well, but the bias is not very close to 0, resulting in under-coverage; for age 80, the bias is close to 0, but the estimated standard error is a bit lower than the empirical standard deviation, also resulting in under-coverage.

5 Discussion

In this article, we define an adjusted population attributable hazard function and propose a kernel-based estimator for the data from case-control studies. We establish the consistency and asymptotic normality of the proposed estimator, and show through extensive simulation the proposed estimator and the analytical variance estimator perform well in finite sample sizes. Our simulation also shows that the proposed estimator is robust with the proposed cross-validation bandwidth selection.

Our adjusted PAH for time-to-event outcome has connection to the adjusted PAF for binary outcome under the case-control study design. Estimation of the adjusted PAF from case-control data could be based on equation (2) where the odds ratio can be estimated by the

logistic regression model to approximate relative risk (by assuming rare disease), and the density of exposure and confounders in disease can be estimated by an empirical estimator[8,23]. In comparison, our estimator of the adjusted PAH is obtained by plugging in the odds ratio estimates from logistic regression model for log-hazard ratio in the Cox model and kernel estimators of the conditional density functions of exposure and confounders given $T = t$ or $T \geq t$. If a very large bandwidth is used to cover the entire time interval, the adjusted PAH becomes a flat line and approaches to the adjusted PAF proposed by Bruzzi et al.[23].

In observational studies, important assumptions underlie the unadjusted and adjusted PAFs, as well as their extensions to time-to-event. The first assumption is that removing the exposure does not change the distribution of other risk factors. It may not be true in real life, however. For example, quitting smoking may simultaneously decrease alcohol consumption due to improved health behavior, which makes interpretation of the smoking PAF for coronary deaths difficult[1]. The second assumption, which is untestable, is no unmeasured confounding, *i.e.*, measured covariates should be sufficient for confounding control[36]. While untestable, its plausibility may be determined on a case-by-case basis using subject matter knowledge. If it is violated, the PAF and extensions need to be interpreted with caution. The third assumption is that the exposure can be eradicated perfectly by an intervention. However, complete removal of an exposure is often unrealistic. To relax this assumption, one may consider the generalized impact fraction, which is the fractional reduction of cases that would result from changing the current level of exposure in the population to some modified (partially removed) level[37].

Our proposed estimator was derived under the Cox proportional hazards model. In practice, the Cox model has been shown fairly robust, as long as the proportionality of the hazards functions between exposed and non-exposed is not seriously violated or could be adequately accounted for by for example, interaction between time and covariates. However, such robustness is not necessarily guaranteed. A further extension to allow for non-proportionality would be of interest. Another potential extension is that since in practice it is not uncommon that many exposures and/or confounders change with time, accommodating time-dependent exposures and/or confounders is also of importance. However, the additional methodological development for time-dependent exposures and/or confounders requires special techniques such as a marginal structural model[38,39]. This is beyond the scope of the current manuscript and will be communicated in the future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank the GECCO Coordinating Center for their generosity of providing the data that is used for illustrating the methods.

Funding

This work was supported by the NIH [R01 CA172451, R01 CA189532, R01 CA 195789, P01 CA53996, U01 CA137088, R01 CA059045].

References

1. Mansournia MA and Altman DG. Population attributable fraction. BMJ 2018; 360: k757. [PubMed: 29472187]
2. Bray F and Soerjomataram I. Population attributable fractions continue to unmask the power of prevention. British Journal of Cancer 2018; 118: 1031–1032. [PubMed: 29567981]
3. Levin ML. The occurrence of lung cancer in man. Acta-Unio Internationalis Contra Cancrum 1953; 9(3): 531–541. [PubMed: 13124110]
4. Walter SD. The estimation and interpretation of attributable risk in health research. Biometrics 1976; : 829–849. [PubMed: 1009228]
5. Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. Statistics in Medicine 1982; 1(3): 229–243. [PubMed: 7187096]
6. Greenland S Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. Statistics in Medicine 1987; 6(6): 701–708. [PubMed: 2825320]
7. Benichou J and Gail MH. Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. Biometrics 1990; : 991–1003. [PubMed: 2085643]
8. Benichou J A review of adjusted estimators of attributable risk. Statistical Methods in Medical Research 2001; 10(3): 195–216. [PubMed: 11446148]
9. Biostatistics Benichou J. and epidemiology: measuring the risk attributable to an environmental or genetic factor. Comptes Rendus Biologies 2007; 330(4): 281–298. [PubMed: 17502285]
10. Kooperberg C and Petitti DB. Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study. Epidemiology 1991; : 363–366. [PubMed: 1742386]
11. Chen YQ, Hu C and Wang Y. Attributable risk function in the proportional hazards model for censored time-to-event. Biostatistics 2006; 7(4): 515–529. [PubMed: 16478758]
12. Samuelsen SO and Eide GE. Attributable fractions with survival data. Statistics in Medicine 2008; 27(9): 1447–1467. [PubMed: 17694507]
13. Zhao W, Chen YQ and Hsu L. On estimation of time-dependent attributable fraction from population-based case-control studies. Biometrics 2017; 73(3): 866–875. [PubMed: 28099992]
14. Cox C, Chu H and Muñoz A. Survival attributable to an exposure. Statistics in Medicine 2009; 28(26): 3276–3293. [PubMed: 19697303]
15. Chen L, Lin DY and Zeng D. Attributable fraction functions for censored event times. Biometrika 2010; 97(3): 713–726. [PubMed: 23956459]
16. Laaksonen MA, Knekt P, Härkänen T et al. Estimation of the population attributable fraction for mortality in a cohort study using a piecewise constant hazards model. American Journal of Epidemiology 2010; 171(7): 837–847. [PubMed: 20197386]
17. Sjölander A and Vansteelandt S. Doubly robust estimation of attributable fractions in survival analysis. Statistical Methods in Medical Research 2017; 26(2): 948–969. [PubMed: 25519888]
18. Gassama M, Bénichou J, Dartois L et al. Comparison of methods for estimating the attributable risk in the context of survival analysis. BMC Medical Research Methodology 2017; 17(1): 10. [PubMed: 28114895]
19. Andersen PK, Borgan O, Gill RD et al. Statistical models based on counting processes. Springer Science & Business Media, 2012.
20. Anderson JA. Separate sample logistic discrimination. Biometrika 1972; 59(1): 19–35.
21. Prentice RL and Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979; 66(3): 403–411.
22. Prentice RL and Breslow NE. Retrospective studies and failure time models. Biometrika 1978; 65(1): 153–158.

23. Bruzzi P, Green SB, Byar DP et al. Estimating the population attributable risk for multiple risk factors using case-control data. American Journal of Epidemiology 1985; 122(5): 904–914. [PubMed: 4050778]
24. Walter SD. Prevention for multifactorial diseases. American Journal of Epidemiology 1980; 112(3): 409–416. [PubMed: 7424889]
25. Lazar MA. How obesity causes diabetes: not a tall tale. Science 2005; 307(5708): 373–375. [PubMed: 15662001]
26. Larsson SC, Orsini N and Wolk A. Diabetes mellitus and risk of colorectal cancer: a meta-analysis. Journal of the National Cancer Institute 2005; 97(22): 1679–1687. [PubMed: 16288121]
27. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Statistical Methodological) 1972; 34(2): 87–22.
28. Xu R and O'Quigley J. Proportional hazards estimate of the conditional survival function. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2000; 62(4): 667–680.
29. Efron B and Tibshirani RJ. An introduction to the bootstrap. CRC press, 1994.
30. Efron B and Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method. Journal of the American Statistical Association 1997; 92(438): 548–560.
31. Nolan D and Marron JS. Uniform consistency of automatic and location-adaptive delta-sequence estimators. Probability Theory and Related Fields 1989; 80(4): 619–632.
32. Bierens HJ. Uniform consistency of kernel estimators of a regression function under generalized conditions. Journal of the American Statistical Association 1983; 78(383): 699–707.
33. Bierens HJ. Kernel estimators of regression functions. In Advances in Econometrics: Fifth World Congress, volume 1 pp. 99–144.
34. Peters U, Jiao S, Schumacher FR et al. Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. Gastroenterology 2013; 144(4): 799–807. [PubMed: 23266556]
35. Greenland S and Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. Biometrics 1993; : 865–872. [PubMed: 8241375]
36. Concepts Greenland S. and pitfalls in measuring and interpreting attributable fractions, prevented fractions, and causation probabilities. Annals of Epidemiology 2015; 25(3): 155–161. [PubMed: 25498918]
37. Morgenstern H and Bursic ES. A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. Journal of Community Health 1982; 7(4): 292–309. [PubMed: 7130448]
38. Hernán MÁ, Brumback B and Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. Epidemiology 2000; 11(5): 561–570. [PubMed: 10955409]
39. Bekaert M, Vansteelandt S and Mertens K. Adjusting for time-varying confounding in the subdistribution analysis of a competing risk. Lifetime Data Analysis 2010; 16(1): 45. [PubMed: 19821028]

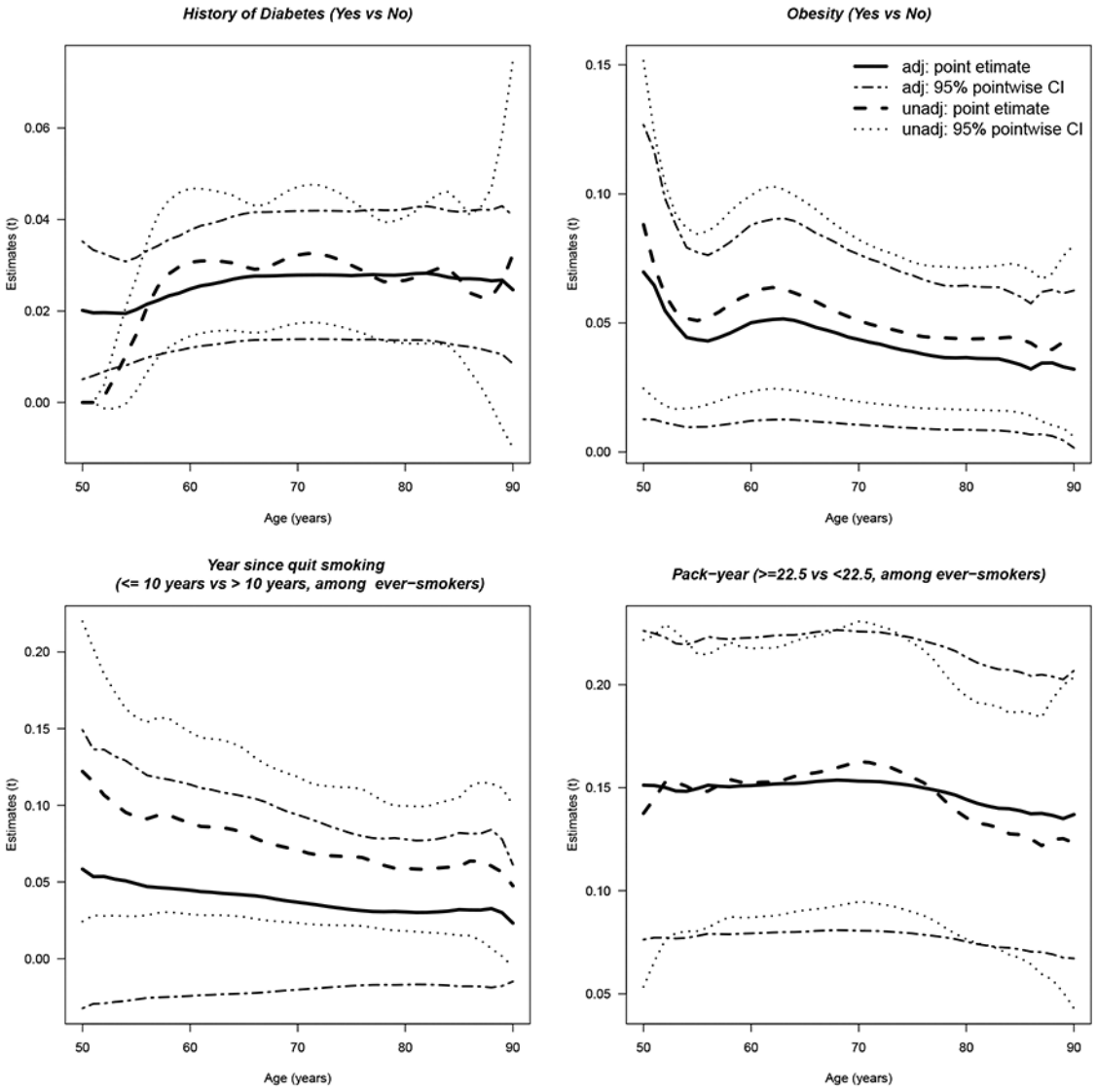

Figure 1.

Estimates of the unadjusted and adjusted PAHs with 95% pointwise confidence intervals versus time t (age in years) for various risk factors.

Table 1.

Summary statistics of the $\phi_{adj}(t)$ estimators under scenario I and II for 80% censoring and an equal number of cases and controls. Bias: absolute difference between the true value of $\phi_{adj}(t)$ and the mean of the point estimator. ESD: empirical standard deviation. ASE: mean of asymptotic-based standard error estimates. CR pointwise: coverage rate of 95% pointwise confidence intervals. CR: coverage rate of 95% simultaneous confidence bands.

Scenario I: binary *U*					**Scenario II: continuous *U***				
	Age	$\phi_{adj}(t)$	$\Phi_{adj}(t)$			Age	$\phi_{adj}(t)$	$\Phi_{adj}(t)$	
	30	0.509	0.507			30	0.512	0.510	
	50	0.491	0.486			50	0.497	0.492	
	70	0.462	0.450			70	0.473	0.463	
	Age	$\hat{\phi}_{adj+}$	$\hat{\phi}_{adj-}$	$\hat{\phi}_{adjw}$		Age	$\hat{\phi}_{adj+}$	$\hat{\phi}_{adj-}$	$\hat{\phi}_{adjw}$
Bias	30	0.000	−0.001	0.000	Bias	30	−0.001	−0.001	0.000
	50	0.000	−0.001	0.000		50	0.000	−0.001	0.000
	70	0.001	−0.001	0.000		70	0.001	−0.002	0.000
ESD	30	0.025	0.030	0.025	ESD	30	0.027	0.032	0.026
	50	0.027	0.031	0.025		50	0.027	0.031	0.026
	70	0.030	0.043	0.028		70	0.029	0.038	0.028
ASE	30	0.027	0.029	0.026	ASE	30	0.027	0.030	0.026
	50	0.027	0.030	0.026		50	0.027	0.030	0.026
	70	0.030	0.038	0.028		70	0.029	0.036	0.027
CR(%) pointwise	30	95.7	95.5	95.4	CR(%) pointwise	30	94.8	94.0	94.3
	50	95.1	94.8	95.1		50	94.6	95.3	94.9
	70	95.4	93.5	95.4		70	94.8	93.9	94.7
CR(%)	20:70	95.2	94.1	94.8	CR(%)	20:70	93.9	93.4	94.0

Table 2.

Summary statistics of risk factors by case-control status in GECCO data. Age: age at onset for cases and age at selection for controls. Years since quit smoking: <=10 years compared to >10 years since quit smoking, for ever-smokers only. Pack-year (>=22.5): for ever-smokers only.

Variables	cases	controls	OR(95% CI)	$\widehat{PAF}$ (95 % CI)	$\widehat{PAF}_{adj}$ (95 % CI)
Age (years) (Mean, range)	70.8 (50-91)	70.9 (50-91)	-	-	-
Gender (female)	73.9%	74.2%	-	-	-
History of Diabetes	7.5%	4.7%	1.57 (1.24, 1.98)	0.029 (0.016, 0.043)	0.027 (−0.002, 0.056)
Obesity	29.5%	25.7%	1.17 (1.04, 1.32)	0.051 (0.020, 0.083)	0.043 (0.002, 0.083)
Ever smoking	55.2%	50.8%	-	-	-
Years since quit smoking	35.0%	29.9%	1.12 (0.94, 1.33)	0.073 (0.024, 0.121)	0.037 (−0.029, 0.104)
Pack-year (>=22.5)	53.9%	46.2%	1.38 (1.16, 1.63)	0.153 (0.089, 0.218)	0.150 (0.074, 0.226)

Table 3.

Summary statistics of the estimators from simulated datasets based on real data for equal number of cases and controls. Bias: absolute difference between the true value of $\phi_{adj}(t)$ and the mean of the point estimator. ESD: sampling standard deviation. ASE: mean of asymptotic-based standard error estimates. CR pointwise: coverage rate of 95% pointwise confidence intervals. CR: coverage rate of 95% simultaneous confidence bands.

	Diabetes as the exposure					**Obesity as the exposure**			
	Age	$\phi_{adj}(t)$	$\Phi_{adj}(t)$			Age	$\phi_{adj}(t)$	$\Phi_{adj}(t)$	
	40	.0269	.0269			40	.0408	.0408	
	60	.0266	.0266			60	.0407	.0407	
	80	.0253	.0252			80	.0403	.0401	
	Age	$\hat{\phi}_{adj+}$	$\hat{\phi}_{adj-}$	$\hat{\phi}_{adjw}$		Age	$\hat{\phi}_{adj+}$	$\hat{\phi}_{adj-}$	$\hat{\phi}_{adjw}$
Bias	40	−.0004	−.0007	−.0007	Bias	40	.0003	.0003	.0002
	60	−.0006	−.0006	−.0005		60	.0003	.0003	.0003
	80	−.0001	−.0001	−.0002		80	.0005	.0005	.0004
ESD	40	.0084	.0085	.0081	ESD	40	.0172	.0172	.0171
	60	.0074	.0074	.0074		60	.0170	.0170	.0170
	80	.0083	.0083	.0079		80	.0172	.0171	.0170
ASE	40	.0083	.0084	.0080	ASE	40	.0173	.0173	.0172
	60	.0073	.0074	.0073		60	.0171	.0171	.0171
	80	.0079	.0081	.0077		80	.0171	.0171	.0170
CR(%) pointwise	40	93.6	92.8	93.9	CR(%) pointwise	40	95.7	95.5	95.7
	60	95.0	94.9	95.4		60	95.4	95.5	95.4
	80	93.3	93.8	93.7		80	95.4	95.7	95.3
CR(%)	40:80	92.5	92.0	92.7	CR(%)	40:80	94.6	94.5	94.9