Published in final edited form as:
J Am Stat Assoc. 2014 June 1; 109(506): 514–524. doi:10.1080/01621459.2014.881739.

Estimating Risk with Time-to-Event Data: An Application to the Women's Health Initiative

Dandan Liu[a], **Yingye Zheng**[b], **Ross L. Prentice**[b], and **Li Hsu**[b,†]

[a]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232

[b]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Abstract

Accurate and individualized risk prediction is critical for population control of chronic diseases such as cancer and cardiovascular disease. Large cohort studies provide valuable resources for building risk prediction models, as the risk factors are collected at the baseline and subjects are followed over time until disease occurrence or termination of the study. However, for rare diseases the baseline risk may not be estimated reliably based on cohort data only, due to sparse events. In this paper, we propose to make use of external information to improve efficiency for estimating time-dependent absolute risk. We derive the relationship between external disease incidence rates and the baseline risk, and incorporate the external disease incidence information into estimation of absolute risks, while allowing for potential difference of disease incidence rates between cohort and external sources. The asymptotic properties, namely, uniform consistency and weak convergence, of the proposed estimators are established. Simulation results show that the proposed estimator for absolute risk is more efficient than that based on the Breslow estimator, which does not utilize external disease incidence rates. A large cohort study, the Women's Health Initiative Observational Study, is used to illustrate the proposed method.

Keywords

absolute risk; attributable risk; cohort data; colorectal cancer; external disease incidence rate 1

1. INTRODUCTION

Accurate and individualized risk prediction is valuable for population control of chronic diseases such as cancer and cardiovascular diseases. To be clinically useful, prediction of risks needs to reflect an individual's true risk state accurately and to have small statistical variation. Developing such a risk model can be a major undertaking, because it often requires assembling large prospective cohorts with many individuals assessed for genetic and environmental risk factors at baseline and followed up over time until the occurrence of clinical events of interest or termination of the follow-up period.

†Corresponding author (lih@fhcrc.org) .

Our research is motivated by the need to provide accurate and efficient risk prediction for rare chronic diseases. The Women's Health Initiative (WHI), launched in 1991, is one of the largest U.S. disease prevention programs addressing heart disease, breast and colorectal cancer, osteoporotic fractures, and other clinical outcomes in postmenopausal women. Between 1993 and 1998, a total of 93,676 women aged 50 to 79 years were recruited into the WHI observational cohort study throughout the US (Women's Health Initiative Study Group, 1998). The study collects information on sociodemographic and epidemiologic factors using standardized questionnaires, and biological samples are collected at clinic visits. While the WHI constitutes one of the largest research cohorts in the US, the efficiency of individual-level risk estimates is still a concern due to the relatively low incidence of disease during cohort follow-up and the large number of risk factors under consideration.

The probability of developing disease by time t given the subject is disease free at t_0 and his/her risk factor profile at t_0 can be generally decomposed into two components: the time-dependent disease risk for a baseline risk factor profile, and the relative risk of developing disease for a particular risk factor profile compared to the baseline. Therefore, a natural approach for estimating the disease risk is to fit a regression model to obtain relative risk estimators and then estimate the time-dependent baseline risk nonparametrically by using for example, the Breslow estimator (Breslow, 1972; Breslow and Crowley, 1974), in the setting of the proportional hazards model (Cox, 1972). One potential weakness for such an approach is that the baseline risk may be estimated with large variation due to limited sample sizes and sparse events in the follow-up period.

Suitable external disease incidence rates from a national registry or other large cohort studies could be used to obtain estimates of the baseline risk (e.g., Gail et al., 1989; Chen et al., 2006). The key is to link the baseline risks with the external disease incidence rates. This can be achieved by using a time-dependent attributable risk function, which is defined as the proportion of disease incidence rate attributed to risk factors. If one knows the distribution of risk factors in the population, then the attributable risk function at time t can be obtained by one minus the inverse of the integrated relative risk function over the risk factor distribution of subjects who are at-risk at time t. However, the information about risk factor distribution is usually not readily available, particularly if the risk factor profile includes novel risk factors and biomarkers. Several population-based studies may need to be combined to estimate a joint distribution of risk factors. See for example, Chen et al. (2006), for an e ort of this type to predict breast cancer risk from mammographic density and other risk factors. Alternatively one can estimate the attributable risk function from more readily available disease case data only, assuming the attributable risk to be constant over time (Bruzzi et al., 1985). This approach does not require known distribution of risk factors in the at-risk population, and it is easy to implement. However, since the estimator employs information from cases only, it is not efficient if non-case data are available and it also requires a strong independent censoring assumption. Moreover, when the attributable risk function varies with time, it leads to a biased baseline risk estimator.

The goal of this paper is to make use of disease incidence data from external sources to obtain an efficient baseline risk estimator. One example of external sources for cancer

outcomes is the Surveillance Epidemiology and End Results Registry (SEER) data, a primary source for cancer statistics representing 26% of the US population and providing incidence rates for various cancer sites. We propose a novel approach to estimating the absolute disease risk, extending the idea of Gail et al. (1989) to allow the attributable risk function to vary flexibly with time, and the censoring to be conditionally independent of failure times given the risk factors. We expect the use of the external disease incidence information to improve the efficiency of the risk function compared to the estimator based on the Breslow baseline hazard estimator from the cohort data. However, there is also a question of the extent to which the baseline risk of cohort participants is same as the external sources. For example, subjects in the WHI may differ from the general population because of eligibility criteria that included a clinically assessed predicted survivorship of at least 3 years from enrollment, and characteristics related to their self-selection for the study. The key to our proposal is to allow for such difference when estimating the attributable risk function using all the cohort data. In addition, our proposed estimator does not require the joint distribution of risk factors to be known in the population, and therefore can be readily applied in practice.

The rest of the article is organized as follows. We describe estimation methods for risk prediction in Section 2 and the large sample properties in Section 3. The finite sample performance of the proposed estimator is assessed through a simulation study in Section 4. The proposed method is applied to risk prediction for colorectal cancer using the WHI observational study in Section 5. Finally, we conclude the article with some remarks in Section 6. Technical details are provided in the Supplementary Materials.

2. METHODS DEVELOPMENT

2.1 Notation and Model

Since for many chronic diseases the time to a clinical event can be censored by competing risk events such as death, it is necessary to take this aspect into consideration when projecting disease risk particularly for long intervals (Gail, 2011). We, therefore, adopt the competing risks framework and model cause-specific hazards for the cause of interest and competing risk events (Kalbfleisch and Prentice, 2002). Without loss of generality, we assume there are two types of events with $\epsilon = 1$ for the cause of interest and $\epsilon = 2$ for competing causes. Let T be the failure time from all causes. We specify the cause-specific hazard for (T, $\epsilon = 1$) given Z by the commonly used proportional hazards model (Cox, 1972):

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}\right), \quad (1)$$

where $\lambda(t|\boldsymbol{Z}) = \lim_{dt\to 0}\Pr(t<T<t+dt, \epsilon=1|T\geq t, \boldsymbol{Z})/dt$, $\lambda_0(t)$ is an unspecified baseline cause-specific hazard function, and β_0 is a p-vector of regression coefficients. Let $\lambda^{\dagger}(t\boldsymbol{Z})$ be the cause-specific hazard function for competing causes $\epsilon = 2$.

Let $R(t|t_0; \boldsymbol{Z})$ denote a subject's risk of developing disease by a future time t given he/she is disease-free at current age t_0 and has risk profile Z. Then $R(t|t_0; \boldsymbol{Z})$, also called the absolute risk by Gail (2011), can be defined as

$$\begin{aligned} R(t|t_0;\boldsymbol{Z}) &= \Pr(t_0 \le T < t, \epsilon = 1 | T \ge t_0, \boldsymbol{Z}) \\ &= \int_{t_0}^{t} \lambda(u|\boldsymbol{Z}) \exp\left[-\int_{t_0}^{u} \left\{\lambda(s|\boldsymbol{Z}) + \lambda^{\dagger}(s|\boldsymbol{Z})\right\} ds\right] du. \end{aligned} \quad (2)$$

The second equation in (2) follows integration of instantaneous probabilities of developing disease between time t_0 and time t, where the instantaneous probability at time u, $t_0 < u < t$, is the hazard of developing disease at u multiplied by the probability that the subject is free of disease and other competing risks (the exponential term) at that time. The presence of competing risks reduces the absolute risk for developing the disease, as the subject may suffer other competing causes before he/she has a chance to develop the disease. In the example of WHI, the outcome of interest is colorectal cancer (CRC) diagnosis and the competing risk event is death due to causes other than CRC. The absolute risk is defined as the probability of developing CRC in the next $(t - t_0)$ year interval given a subject is CRC-free and alive at age t_0 and her risk profile at age t_0.

We will first focus on estimation of $\lambda(t|\boldsymbol{Z})$ for the event of interest by following the usual practice in epidemiologic studies for absolute risk estimation, assuming the competing risks are independent of $\boldsymbol{Z}$ (see e.g. Gail et al., 1989; Gail, 2011). We will then describe the generalization of the proposed approach to allow for the effect of $\boldsymbol{Z}$ on competing risks.

2.2 Proposed Estimation Method

To estimate $\lambda(t|\boldsymbol{Z})$, we need to estimate β_0 and $\lambda_0(t)$, both of which can be estimated in a standard way by treating failures from other causes as censoring. This is because under the competing risks framework, the likelihood factors into a separate component for each cause-specific hazard function, where the component can be considered as regarding failures from other causes as censored at the corresponding failure times, see equation (8.8) in Kalbfleisch and Prentice (2002). Left truncation is also considered here since it is frequently encountered in cohort studies as is shown in the motivating WHI study. Consider n subjects in a cohort study. Let T_i, L_i, and C_i be the minimum of failure times of all causes, left truncation time (e.g. study entry) and censoring time, respectively, for $i = 1, \ldots, n$, let $\epsilon_i \in \{1, 2\}$ be the cause of failure and $\boldsymbol{Z}_i$ be a $p \times 1$ covariate vector. Define $X_i = \min(T_i, C_i)$ if $X_i \ge L_i$, and the censoring indicator $\delta_i = I(L_i < T_i \le C_i) I(\epsilon_i = 1)$, where $I(\cdot)$ is an indicator function. Therefore δ_i equals 1 if the failure time of the event of interest is observed and 0 otherwise. We assume that $\{(X_i, \delta_i, \boldsymbol{Z}_i), i = 1, \ldots, n\}$ are independently and identically distributed (i.i.d.), and both L_i and C_i are independent of T_i conditional on $\boldsymbol{Z}_i$. Define the counting process $N_i(t) = I(L_i < X_i \le t, \delta_i = 1)$ and the at-risk process $Y_i(t) = I(L_i < t \le X_i)$. In addition, define $\overline{N}(t) = n^{-1} \sum_{i=1}^{n} N_i(t)$, $H(t) = n^{-1} \sum_{i=1}^{n} Y_i(t)$, and $\boldsymbol{H}_r(t;\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} Y_i(t) \boldsymbol{Z}_i^{\otimes r} \exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_i\right)$, where $r = 0, 1, 2$, $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1}$. The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ may be obtained by solving the following score equations

$$\boldsymbol{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t;\boldsymbol{\beta}) \right\} dN_i(t),$$

where $\overline{\boldsymbol{Z}}(t;\boldsymbol{\beta}) = \boldsymbol{H}_1(t;\boldsymbol{\beta}) / H_0(t;\boldsymbol{\beta})$ and τ is the end of the follow-up period. The cause-specific cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)$ can be estimated by the usual Breslow estimator, denoted by $\overline{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right) = \int_0^t d\overline{N}(u) / H_0\left(u;\hat{\boldsymbol{\beta}}\right)$. These estimators have been widely used for estimating β and $\Lambda_0(t)$ in the Cox model for cohort studies.

It is clear from (2) that to estimate the absolute risk both regression coefficients and baseline hazard function estimators are needed. However, since the Breslow estimator takes jumps at observed failure times, it can be quite inefficient due to sparse events or small at-risk sizes in cohort studies. As a result, it can lead to an inefficient estimation of $R(t|t_0; \boldsymbol{Z})$. In the following we propose an estimator for $\lambda_0(t)$, leveraging the information from the external disease incidence data.

Let $\lambda^*(t)$ and $\lambda(t)$ denote the cause-specific incidence rates for the event of interest from the external source and the cohort, respectively. Note that $\lambda^*(t)$ and $\lambda(t)$ are sometimes termed as the"composite" incidence rates (e.g., Gail et al., 1989; Gail, 2011), as they describe the marginal cause-specific incidence rates from pooling subjects with different risk profiles. Let $f(t)$ and $\overline{\boldsymbol{F}}(t)$ be the marginal cause-specific density and survival functions of T for the event of interest. In addition let $f(t|\boldsymbol{Z})$ and $\overline{F}(t|\boldsymbol{Z})$ be their counterparts conditional on $\boldsymbol{Z}$. Note that $\overline{F}(t|\boldsymbol{Z}) = \exp\left\{-\Lambda_0(t)\ \exp\left(\boldsymbol{\beta}_0^T \boldsymbol{Z}\right) - \Lambda^{\dagger}(t)\right\}$. Therefore, under the Cox model (1) we can write the composite cause-specific incidence rate as $\lambda(t) = f(t) / \overline{F}(t) = \int \lambda(t|z)\, \overline{F}(t|z)\, k(z)\, dz / \int \overline{F}(t|z)\, k(z)\, dz$, which, in the presence of non-differential competing risk events, can be written as

$$\lambda(t) = \lambda_0(t) \frac{\int \exp\left(\boldsymbol{\beta}_0^T z\right) \exp\left\{-\Lambda_0(t) \exp\left(\boldsymbol{\beta}_0^T z\right) - \Lambda^{\dagger}(t)\right\} k(z)\, dz}{\int \exp\left\{-\Lambda_0(t) \exp\left(\boldsymbol{\beta}^T z\right) - \Lambda^{\dagger}(t)\right\} k(z)\, dz},$$

where $k(z)$ is the density distribution for $\boldsymbol{Z}$ in the pertinent population. It then follows that $\lambda_0(t) = \varphi(t)\lambda(t)$, where

$$\phi(t) = \frac{\int \exp\left\{-\Lambda_0(t) \exp\left(\boldsymbol{\beta}_0^T z\right)\right\} k(z)\, dz}{\int \exp\left(\boldsymbol{\beta}_0^T z\right) \exp\left\{-\Lambda_0(t) \exp\left(\boldsymbol{\beta}_0^T z\right)\right\} k(z)\, dz}. \quad (3)$$

Note that the terms related to competing risk $\exp\{-\Lambda^{\dagger}(t)\}$ are canceled out in both numerator and denominator.

When a suitable external cause-specific composite incidence rate $\lambda^*(t)$ is available, we can plug it in for $\lambda(t)$, and then estimate $\lambda_0(t)$ by $\varphi(t)\lambda^*(t)$. However, the incidence rate in the cohort may differ from the external incidence rate because of eligibility criteria and

participant characteristics such that cohort participants may not be entirely representative of the population that the external incidence rate comes from. It is therefore important to allow for such difference. Towards this end, we propose to estimate

$$\lambda_0(t) = \rho_0 \phi(t) \lambda^*(t),$$

where $\rho_0 > 0$ with $\rho_0 = 1$ indicating no difference of disease incidence rates between the cohort and the external source. The key is then to estimate $\varphi(t)$ and ρ_0. Expression (3) naturally leads to an estimator for $\varphi(t)$, which entails estimation of $\exp\left\{-\Lambda_0(t)\exp\left(\boldsymbol{\beta}_0^T \boldsymbol{Z}\right)\right\}$ by replacing β_0 with $\hat{\boldsymbol{\beta}}$ and $\Lambda_0(t)$ with the Breslow estimator $\overline{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$. Plugging in an empirical estimator for $k(z)$, we can obtain an estimator for $\varphi(t)$ given by $\hat{\phi}\left(t;\hat{\boldsymbol{\beta}}\right) = G\left(t;\hat{\boldsymbol{\beta}}\right)/G_0\left(t;\hat{\boldsymbol{\beta}}\right)$, where $G(t;\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \exp\left\{-\overline{\Lambda}_0(t;\boldsymbol{\beta})\exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_i\right)\right\}$ and $G_0(t;\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_i\right)\exp\left\{-\overline{\Lambda}_0(t;\boldsymbol{\beta})\exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_i\right)\right\}$. A natural estimator for ρ_0 is

$$\hat{\rho} = \frac{\int_o^\tau \overline{N}(du)}{\int_0^\tau \sum_{i=1}^{n} Y_i(u) \exp\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_i\right) \hat{\phi}\left(u;\hat{\boldsymbol{\beta}}\right) \lambda^*(u)\, du}. \quad (4)$$

Consequently, we can obtain the cause-specific baseline risk estimator by $\hat{\lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right) = \hat{\rho}\hat{\phi}\left(t;\hat{\boldsymbol{\beta}}\right)\lambda^*(t)$, and the corresponding estimator for the absolute risk $R(t|t_0; \boldsymbol{Z})$ as

$$\hat{R}\left(t|t_0;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right) = \int_{t_0}^{t} \hat{\lambda}_0\left(u;\hat{\boldsymbol{\beta}}\right) \exp\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}\right) \exp\left[-\int_{t_0}^{u}\left\{\hat{\lambda}_0\left(s;\hat{\boldsymbol{\beta}}\right)\exp\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}\right) + \lambda^\dagger(s)\right\} ds\right] du.$$

Our proposed absolute risk estimator differs from previous approaches that also use external incidence rates for obtaining baseline risk. Our approach does not require known risk factor distribution $f(Z|T \geq t)$ as in Chen et al. (2006), which overcomes the potential difficulty for obtaining such distribution in the population. Compared to the case-only estimator of $\varphi(t)$ (Gail et al., 1989), our estimator uses not only the cases but also subjects who have not developed diseases in the cohort when estimating $\varphi(t)$; and hence, can be expected to improve the efficiency for estimating $\varphi_0(t)$ compared to the case-only estimator. Furthermore, since the proposed approach makes use of known external disease incidence rates, it also improves efficiency compared to the Breslow estimator that uses cohort data only. Lastly, we propose a scale factor ρ to allow for potential difference between the cohort and the external incidence rate, which ensures a broad applicability of our proposed estimator.

2.3 Effect of Covariates on Competing Risks

In settings that competing risk events are associated with covariates in the risk model, we can model the cause-specific hazard for the competing risks (T, $\epsilon = 2$) given a $q\times 1$ covariate $\boldsymbol{Z}^{\dagger}$ by using the proportional hazards model:

$$\lambda^{\dagger}\left(t|\boldsymbol{Z}^{\dagger}\right)=\lambda_0^{\dagger}(t)\,\exp\left(\gamma_0^T\boldsymbol{Z}^{\dagger}\right),$$

where $\lambda^{\dagger}\left(t|\boldsymbol{Z}^{\dagger}\right)=\lim_{dt\to 0}\Pr\left(t<T<t+dt,\epsilon=2|T\geq t,\boldsymbol{Z}^{\dagger}\right)/dt$, $\lambda_0^{\dagger}(t)$ is an unspecified baseline cause-specific hazard function for competing risks, and γ_0 is a $q\times 1$ vector of regression coefficients. The function $\varphi(t)$ becomes

$$\phi(t)=\frac{\int\int\exp\left\{-\Lambda_0(t)\,\exp\left(\beta_0^T z\right)-\Lambda_0^{\dagger}(t)\,\exp\left(\gamma_0^T z^{\dagger}\right)\right\}k\left(z,z^{\dagger}\right)dzdz^{\dagger}}{\int\int\exp\left(\beta_0^T z\right)\exp\left\{-\Lambda_0(t)\,\exp\left(\beta_0^T z\right)-\Lambda_0^{\dagger}(t)\,\exp\left(\gamma_0^T z^{\dagger}\right)\right\}k\left(z,z^{\dagger}\right)dzdz^{\dagger}},\quad(5)$$

which could be estimated by additionally plugging in the Breslow estimator for $\lambda_0^{\dagger}(t)$ and the maximum partial-likelihood estimator for γ_0. The estimator for $\gamma_0(t)$ can then be obtained following the same approach as in Section 2.2 by multiplying the external incidence rate with the modified estimator for $\varphi(t)$ in (5), and $\hat{\rho}$ in (4).

To increase efficiency for estimating $\lambda_0^{\dagger}(t)$, we can take the same [hatwide] approach as for estimating $\lambda_0(t)$ by borrowing the incidence rate information from the external source. In this case, $\varphi^{\dagger}(t)$ is defined as in (5) except that in the denominator the first term $\exp\left(\boldsymbol{\beta}_0^T z\right)$ is replaced by $\left(\gamma_0^T z^{\dagger}\right)$ and $\rho_0^{\dagger}$ is the multiplicative factor allowing for difference in cause-specific incidence rates for competing risks between cohort and the external source. Then $\lambda_0^{\dagger}(t)$ can be estimated by multiplying $\rho_0^{\dagger}$, $\varphi^{\dagger}(t)$ and $\lambda^{\dagger}(t)$, where $\lambda^{\dagger}(t)$ is the external incidence rates for competing risks.

3. LARGE SAMPLE RESULTS

In this section we establish the large sample results of the proposed estimators and show that $\hat{\Lambda}_0(t)$ is consistent to $\Lambda_0(t)$ uniformly and asymptotically normal, and the corresponding risk estimator $\hat{R}\left(t|t_0;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right)$ is also consistent uniformly and asymptotically normal.

First we define the following frequently used notation in survival analysis. Let $M_i(t)=N_i(t)-Y_i(t)\,\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}_i\right)\Lambda_0(t)$ be the martingale process with respect to the filtration $\mathscr{F}_t=\sigma\left\{N_i(u),Y_i(u),\boldsymbol{Z}_i;i=1,\ldots,n,\quad 0\leq u\leq t\right\}$, where $t\in(0,\ \tau]$ with τ being the maximum follow-up time and $\Pr(X\geq\tau)>0$. Let $\boldsymbol{I}(\boldsymbol{\beta})=-\partial\mathscr{U}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ denote the observed information matrix and $\boldsymbol{A}=\mathrm{E}\{n^{-1}\boldsymbol{I}(\beta_0)\}$. In addition, we let $\overline{z}(t;\boldsymbol{\beta}_0)$ denote the limit value of $\overline{\boldsymbol{Z}}(t;\boldsymbol{\beta}_0)$ and $\boldsymbol{b}(t;\boldsymbol{\beta}_0)=\int_0^t\overline{z}(u)\,\boldsymbol{\beta}_0\,d\Lambda_0(u)$. In addition we define the following requisite notation:

$$\boldsymbol{G}_r(t;\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \boldsymbol{Z}_i^{\otimes r} e^{\boldsymbol{\beta}^T \boldsymbol{Z}_i - \overline{\Lambda}_0(t;\boldsymbol{\beta})\exp(\boldsymbol{\beta}^T \boldsymbol{Z}_i)},$$
$$\boldsymbol{Q}_r(t;\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \boldsymbol{Z}_i^{\otimes r} e^{2\boldsymbol{\beta}^T \boldsymbol{Z}_i - \overline{\Lambda}_0(t;\boldsymbol{\beta})\exp(\boldsymbol{\beta}^T \boldsymbol{Z}_i)}.$$

Furthermore we let $h(t;\beta)$, $\boldsymbol{h}_r(t;\beta)$, $\boldsymbol{g}(t;\beta)$, $\boldsymbol{g}_r(t;\beta)$ and $\boldsymbol{q}_r(t;\beta)$ be the limit of $H(t;\beta)$, $\boldsymbol{H}_r(t;\beta)$, $G(t;\beta)$, $\boldsymbol{G}_r(t;\beta)$ and $\boldsymbol{Q}_r(t;\beta)$, $r = 0, 1$, respectively.

It has been shown in Andersen and Gill (1982) that $\hat{\boldsymbol{\beta}}$ and $\overline{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$ are strongly consistent estimators for β_0 and $\Lambda_0(t)$ over$(0, \tau]$, respectively. In addition, they showed that

$$n^{\frac{1}{2}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \simeq n^{-\frac{1}{2}}\sum_{i=1}^{n} W_i, \qquad n^{\frac{1}{2}}\left\{\overline{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right) - \Lambda_0(t)\right\} \simeq n^{-\frac{1}{2}}\sum_{i=1}^{n} W_{\Lambda_0,i}^{B}(t),$$

where $W_i = A^{-1}\int_0^{\tau}\left\{\boldsymbol{Z}_i - \bar{z}(u;\boldsymbol{\beta}_0)\right\} dM_i(u)$ and
$W_{\Lambda_0,i}^{B}(t) = \int_0^t dM_i(u)/h_0(u;\boldsymbol{\beta}_0) - \boldsymbol{b}(t;\boldsymbol{\beta}_0)^T W_i$.

In the following, we establish the asymptotic consistency and normality of the proposed cumulative baseline hazard function $\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$ (Theorem 1) and absolute risk estimators $\hat{R}\left(t|t_0;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right)$ (Theorem 2). Let $S(t|\boldsymbol{Z}) = \exp\left\{-\Lambda_0(t)\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}\right)\right\}$ and $S^{\dagger}(t) = \exp\{-\Lambda^{\dagger}(t)\}$.

Theorem 1

Under regularity conditions provided in Section A.1 of the Supplementary Materials, $\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$ *converges almost surely to* $\Lambda_0(t)$ *uniformly in* $t \in [0, \tau]$. *Moreover*, $n^{1/2}\left\{\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right) - \Lambda_0(t)\right\}$ converges in distribution to a zero mean Gaussian process with covariance function $\sum_{\Lambda_0}(u,t) = E\left\{W_{\Lambda_0,i}(u)\,W_{\Lambda_0,i}(t)\right\}$, *where*
$W_{\Lambda_0,t}(t) = \int_0^t \rho_0\lambda^*(u)\,W_{\phi,i}(u)\,du + W_{\rho,i}\int_0^t \lambda^*(u)\,\phi(u)\,du$, *with*
$W_{\phi,i}(t) = W_{\phi,i}^1(t) + W_{\phi,i}^2(t) + W_{\phi,i}^3(t) + W_{\phi,i}^4(t)$, $W_{\rho,i} = W_{\rho,i}^1 + W_{\rho,i}^2 + W_{\rho,i}^3$,

$$\begin{aligned}
W_{\phi,i}^1(t) &= \left\{\frac{1}{g_0(t;\boldsymbol{\beta})} - \frac{g(t;\boldsymbol{\beta}_0)}{g_0(t;\boldsymbol{\beta}_0)^2}\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}_i\right)\right\} e^{-\Lambda_0(t)\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}_i\right)},\\
W_{\phi,i}^2(t) &= \left\{\frac{g(t;\boldsymbol{\beta}_0)q_0(t;\boldsymbol{\beta}_0)}{g_0^2(t;\boldsymbol{\beta}_0)} - 1\right\}\int_0^t \frac{dM_i(u)}{h^{(0)}(u;\boldsymbol{\beta}_0)},\\
W_{\phi,i}^3(t) &= \left\{\boldsymbol{b}(t;\boldsymbol{\beta}_0) - \frac{g_1(t;\boldsymbol{\beta}_0)}{g_0(t;\boldsymbol{\beta}_0)}\Lambda_0(t)\right\}^T W_i,\\
W_{\phi,i}^4(t) &= -\boldsymbol{c}(t;\boldsymbol{\beta}_0)^T W_i,\\
W_{\rho,i}^1 &= E\{N_i(t)\}^{-1}\rho_0\int_0^{\tau} dM_i(t),\\
W_{\rho,i}^2 &= -E\{N_i(t)\}^{-1}\rho_0^2\int_0^{\tau} Y_i(t)\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}_i\right) W_{\phi,i}(t)\,\lambda^*(t)\,dt,\\
W_{\rho,i}^3 &= -E\{N_i(t)\}^{-1}\rho_0^2\int_0^{\tau} Y_i(t)\,\boldsymbol{Z}_i^T\phi(t)\,\lambda^*(t)\,dt W_i,
\end{aligned}$$

and $c(t;\beta) = g(t;\beta)/g_0(t;\beta)^2 \{\boldsymbol{g}_1(t;\beta) + q_0(t;\beta)h(t;\beta) - \boldsymbol{q}_1(t;\beta)\,\Lambda_0(t)\}$

To show the consistency of the proposed cumulative baseline hazard estimator $\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$, it suffices to prove the consistency of $\hat{\phi}_0\left(t;\hat{\boldsymbol{\beta}}\right)$ and $\hat{\rho}$. We decompose $\hat{\phi}_0\left(t;\hat{\boldsymbol{\beta}}\right) - \phi(t)$ into four components and show each of components converges to 0 almost surely. Similarly, we decompose $\hat{\rho} - \rho_0$ into two components, and show each component converges to 0 in probability. The asymptotic normality of $\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$ is proved by showing that the process $n^{\frac{1}{2}}\left\{\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right) - \Lambda_0(t)\right\}$ and $n^{\frac{1}{2}}(\hat{\rho} - \rho_0)$ are asymptotically equivalent to a summation of n i.i.d. quantities $n^{\frac{1}{2}}\sum_{i=1}^{n} W_{\Lambda_0,i}(t)$ and $n^{\frac{1}{2}}\sum_{i=1}^{n} W_{\rho,i}$, respectively, and appealing to the central limit theorem. The detailed proof is provided in Section A of the Supplementary Materials.

The following theorem establishes the large sample theory for the proposed absolute risk estimator based on $\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$.

Theorem 2

Under regularity conditions provided in Section A.1 of the Supplementary Materials, $\hat{R}\left(t|t_0;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right)$ *converges almost surely to* $R(t|t_0;Z)$ *uniformly in* $t \in [t_0, \tau]$. *Moreover*, $n^{1/2}\left\{\hat{R}\left(t|t_0;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right) - R\left(t|t_0;\boldsymbol{Z}\right)\right\}$ converges in distribution to a zero mean Gaussian process with covariance function $\sum_{R_{t_0}}(u,t) = E\left\{W_{R_{t_0},i}(u)\,W_{R_{t_0},i}(t)\right\}$ *where*

$$\begin{aligned}
W_{R_{t_0},i}(t) &= -\frac{R(t|t_0;\boldsymbol{Z})}{S(t_0|\boldsymbol{Z})}W_{S,i}(t_0) - \frac{\int_0^t S^{\dagger}(u)dW_{S,i}(u)}{S(t_0|\boldsymbol{Z})S^{\dagger}(t_0)},\\
W_{S,i}(t) &= -S(t|\boldsymbol{Z})\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}\right)\left\{W_{\Lambda_0,i}(t) + \Lambda_0(t)\,W_i^T\boldsymbol{Z}\right\},\\
dW_{S,i}(t) &= S(t|\boldsymbol{Z})\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}\right)\left[d\Lambda_0(t)\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}\right)\left\{W_{\Lambda_0,i}(t) + \Lambda_0(t)\,W_i^T\boldsymbol{Z}\right\}\right.\\
&\quad \left. - \left\{dW_{\Lambda_0,i}(t) + d\Lambda_0(t)\,W_i^T\boldsymbol{Z}\right\}\right].
\end{aligned}$$

The consistency of $\hat{R}\left(t|t_0;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right)$ follows directly from the convergence of $\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$ and $\hat{\boldsymbol{\beta}}$ to $\Lambda_0(t)$ and β_0, respectively. The asymptotic normality is shown by writing each component as summation of n i.i.d. quantities. A detailed proof is provided in Section A of Supplementary Materials.

4. SIMULATION STUDY

In the first simulation study, we evaluated the performance of estimators for $\varphi(t)$, ρ_0 and $\Lambda_0(t)$. Specifically, we generated failure times from a proportional hazards model $\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\beta_0\boldsymbol{Z})$, where $\lambda_0(t) = p\lambda(\lambda t)^{p-1}$, a Weibull distribution with $p = 2$ and $\lambda = 0.01$. We generated two exposure variables $Z = (Z_1, Z_2)$ with $Z_1 \sim$ *Bernoulli*(0.5) and $Z_2 \sim N(0, 1)$. The corresponding regression coefficients $\beta_0 = (\log 2, \log 2)$. We let $C = C^*I(1 \le C^* \le 100) + I(C^* < 1) + 100I(C^* > 100)$, where $C^* \sim N(\mu - 10Z_1, 15)$ and $\mu = 74.5$ and 51.5 were chosen

to yield the censoring rates of approximately 50% and 75%, representing moderate and high censoring, respectively. We generated 1000 replicates with cohort size $n = 1000$. We did not consider competing risks here since the focus for this set of simulation was to evaluate the performance of our proposed $\varphi(t)$, ρ and $\Lambda_0(t)$ estimators, for which only the regression coefficients and the Breslow estimators were needed where the competing risks event was treated as censoring. We considered $\rho_0 = 1, 1.5, 2$, i.e., the external incidence rate $\lambda^*(t) = \lambda(t)\rho_0$ and keep the simulated data the same regardless of the choice of ρ_0. Therefore, the result for $\hat{\phi}(t)$ does not change with ρ_0. Since $\lambda^*(t)$ is assumed known, the estimators for different ρ_0 only differ by a constant factor. To save space we only present the results for $\rho_0 = 1.5$.

Table 1 summarizes the results for the proposed estimators $\hat{\phi}\left(t, \hat{\boldsymbol{\beta}}\right)$, $\hat{\rho}$ and $\hat{\Lambda}_0(t)$, as well as the Breslow estimator $\overline{\Lambda}_0(t)$ at $t = 20, 40, 60$. All the estimators appear to be unbiased. The asymptotic standard errors (ASE) are close to their empirical standard deviations (ESD), and the resulting 95% coverage probabilities (CP) are close to the nominal level of 95%. The proposed estimator $\hat{\Lambda}_0(t)$ is more efficient than the Breslow estimator $\overline{\Lambda}_0(t)$. The estimated asymptotic relative efficiency (ARE) of $\hat{\Lambda}_0(t)$ to $\overline{\Lambda}_0(t)$ ranges from 1.21 to 2.38 when the censoring rate is 50%, and reduces to 1.16 to 1.94 when the censoring rate increases to 75%. The efficiency reduction for high censoring rate is because the proposed baseline hazard estimator $\hat{\Lambda}_0(t)$ also relies on $\hat{\rho}$ and $\hat{\phi}_0(t)$, both of which lose efficiency as the censoring rate increases. This is particularly the case for $\hat{\rho}$, as the estimator is proportional to the number of cases in the cohort, see equation (4). Under both censoring scenarios, the efficiency gain is the most at early time $t = 20$ with ARE=2.38 and 1.94 for censoring rate 50% and 75%, respectively, as there are few events in the early age interval.

In the second set of simulation, we considered a scenario that mimicked the WHI study on colorectal cancer in terms of incidence rates and sample size. The details of the study are provided in the following Application section. Specifically, we let $p = 2$ and $\lambda = 0.0026$ such that the crude incidence rate of colorectal cancer is $\rho_0 = 1.5$ times of the 2008 SEER incidence rate. We also simulated a competing risks event and let the time to the competing risks event follow a piecewise-constant exponential distribution with the hazard rate equivalent to the mortality rate from all causes (excluding colorectal cancer) obtained from the 2008 National Vital Statistics System (NVSS) (CDC and NCHS, 2012). The censoring time $C = C^* I(1 \leq C^* \leq 100) + I(C^* < 1) + 100 I(C^* > 100)$, where $C^* \sim N(\mu + 10\boldsymbol{Z}_1, 15)$ and $\mu = 70$ was chosen to yield the censoring rates of approximately 95%. We generated 1000 replicates with cohort size $n = 20,000$. The rest of the settings were kept the same as in the first simulation.

Table 2 summarizes the results of proposed $\hat{\phi}_0(t)$, $\hat{\rho}$ and $\hat{\Lambda}_0(t)$ as well as the Breslow estimator $\overline{\Lambda}_0(t)$. It also includes the results of both the proposed estimator $\hat{R}\left(a|t; \boldsymbol{Z}, \hat{\boldsymbol{\beta}}\right)$ and the Breslow-based estimator denoted by $\overline{R}\left(a|t; \boldsymbol{Z}, \hat{\boldsymbol{\beta}}\right)$ for baseline risks at $t_0 = 55, 60, 65$, $t = t_0 + 10$ and $\boldsymbol{Z} = (0, 0), (0, 1), (1, 1)$. The overall performance of the estimators for $\phi(t)$ and

$\Lambda_0(t)$ is similar to those in the first simulation, and the ARE of $\hat{\Lambda}_0(t)$ to $\overline{\Lambda}_0(t)$ ranges from 1.05 to 1.27, which is lower than that shown in Table 1. This is expected as the censoring rate in this simulation is 95%, which is considerably higher than the first simulation. Both absolute risk estimators are unbiased for all configurations. The ASEs and ESDs are close to each other and the CPs are close to the nominal level of 95%. The proposed absolute risk estimator is more efficient than the Breslow-based estimator with ARE of $\hat{R}\left(a|t;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right)$ to $\overline{R}\left(a|t;\boldsymbol{Z},\hat{\boldsymbol{\beta}}\right)$ ranging from 2.29 to 3.21.

5. APPLICATION

5.1 Study Population and Risk Factors

As described in the Introduction, WHI includes a large cohort study that follows postmenopausal women for various disease outcomes and mortality information. In this paper we focus on predicting absolute risks of CRC for white women aged 50 and above. A total of 76,733 subjects are included in the analysis. Among those, 1,073 (1.4%) developed CRC and 9,190 (12.0%) died during the follow-up of causes other than CRC. The disease outcome we considered is age (in integer years) at diagnosis of CRC, which is subject to left truncation (age at enrollment), right censoring (age at the end of follow-up or loss of follow-up) and competing risks (death).

The risk factors included in our model are selected based on the risk prediction model developed by Freedman et al. (2009). In their work, risk prediction models were obtained for colon (proximal and distal) and rectal cancers, respectively, using population-based case-control data. The risk factors included in their final models vary for different tumor sites. For the purpose of illustration, here we do not distinguish among tumor sites and include the risk factors for all cancer sites. These are history of endoscopy (SMC) and polyps in last 5 years (SMC and no polyps, SMC and polyps, no SMC, and unknown history); number of first-degree relatives with CRC (0, ≥1); current leisure-time vigorous activity (0, 0-2, > 2 hours per week); use of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs) (nonuser, regular user); vegetable consumption (< 2, ≥2 medium portion per day); BMI (< 30, ≥30kg/m^2); and estrogen status within the last 2 years (negative, positive). Risk factors are categorized analogous to Freedman et al. (2009) with some modifications made along with the WHI data collection regime. Different from the case-control study used in Freedman et al. (2009) which collected risk factors for cases at the time of CRC diagnosis, the information on risk factors for WHI women were collected prospectively at the time of enrollment.

The WHI study enrolled healthy post-menopausal women aged 50-79. Since no information is available for women aged less than 50, we will estimate absolute risk only starting at age 50. For women who enrolled after age 50, due to late entry, their failure times were treated as left truncated. For example, a women enrolled at age 60 and developed CRC at age 70 would only be in the at-risk set for developing disease between age 60 and 70. This problem can be appropriately handled by carefully defining the at-risk set for each subject as described in Section 2.2.

5.2 Results

We obtained the external composite incidence rate for age after 50 from the SEER. To check whether the SEER incidence rate is suitable for obtaining the baseline hazard function, we graphically examined the WHI incidence and cumulative incidence rates compared to the SEER rates (Figure 1). The WHI incidence rates are fairly close to the SEER incidence rates for age between 50 and 65 years old and higher when age is greater than 65 years old. Even though the 95% simultaneous confidence bands of the WHI incidence rates cover the SEER incidence rates for almost the entire range of the age, the confidence bands of the cumulative incidence rates fail to cover the SEER cumulative incidence rates for age greater than about 65 years old. We, therefore, assume that ρ_0 is piece-wise constant with age 65 as the cut-off. We extend constant ρ in equation (4) such that the integration from 0 to τ is replaced by pre-specified age intervals, in the WHI example, [50, 65] and 65+, to reflect different ρ values for before and after age 65 years old.

Since the Cox proportional hazards model was used to obtain the hazard ratio estimators, we also checked the proportional hazards assumption for the risk factors (Grambsch and Therneau, 1994). All p-values were greater than 0.05, suggesting the Cox model was appropriate. The competing risk incidence rates, i.e., mortality rates, were obtained from the 2008 National Vital Statistics System (CDC and NCHS, 2012) from all causes except CRC.

The hazard ratio estimates and the corresponding 95% confidence intervals (CI) are presented in Table S1 of the Supplementary Materials. Compared to women who had SMC and no colon or rectal polyps in the last 5 years, women who were found having polyps at screening were at higher risk for developing CRC (HR=1.16, 95% CI, 1.04 to 1.28), so did women not having SMC (HR=1.30; 95% CI, 1.22 to 1.39). Women with one or more first-degree relatives with CRC were more likely to be diagnosed with CRC than those without a positive family history (HR=1.23, 95% CI, 1.05 to 1.45). Those taking aspirin or NSAIDs at least once a week during the past 30 days (regular users) had lower risk of being diagnosed with CRC (RR=0.76, 95% CI, 0.70 to 0.83). Women having sufficient vegetable consumption (more than 2 medium portion/day) were also at lower risk of developing CRC (HR=0.94, 95% CI, 0.88 to 1.00). Obese women (BMI $\geq$30) had an increased risk of CRC (HR=1.38, 95% CI, 1.29 to 1.48). Women who used hormone- replacement therapy (HRT) during the past 2 years were less likely to develop CRC (HR=0.87, 95% CI, 0.81 to 0.93).

We then used the proposed method to estimate the baseline hazard function and the absolute risk. The proposed estimators $\hat{\rho}$ are 1.18 (95% CI: 1.01, 1.36) and 1.63 (95% CI: 1.52, 1.73) for age below 65 and age above 65, respectively. The cumulative baseline hazard function was slightly larger than the Breslow estimator with 95% CIs of the estimators overlapping with each other (Figure S1 in Supplementary Materials).

We selected 3 risk factor profiles representing lowest, medium and highest risk for each of the age groups 50, 55, 60 and 65, and calculated the 10- and 20-year absolute risk for developing CRC (Table 3). For example, the first profile describes a 55-year old white women categorized as lowest risk. She had SMC in the last 5 years but no polyps were found. She does not have any positive family history, exercises strenuously each week, takes aspirin daily, eats more than 2 medium portions of vegetable every day, took HRT in the

past 2 years and has a BMI of 28 kg/m^2. With this low-risk profile, her 10-year risk of developing CRC is only 0.38% (95%CI: 0.29% to 0.47%) and her 20-year absolute risk is 1.14% (95%CI: 0.87% to 1.40%). In contrast, the "averaged" 10-year and 20-year risk for women aged 50 is 0.52% and 1.86%, respectively, which grossly overestimates the risk for this low risk woman. The third profile represents a high risk profile for a woman who is also 55 years old and white. Compared with low-risk profile women in the aforementioned example, she did not have SMC nor HRT. She does not exercise or take aspirin/NSAIDs and eats only 1 medium portion of vegetable every day. Moreover she has positive family history of CRC and is obese with BMI of 32 kg/m^2. Her 10-year absolute risk of CRC is 1.59% (95%CI: 1.21% to 1.98%), and her 20-year absolute risk is 4.74% (95%CI: 3.67% to 5.80%); both her 10- and 20-year absolute risk estimates are more than 4 times higher than those of the low-risk women. If we were to use the averaged risk for these women, their risks would be greatly underestimated.

For comparison we also calculated the absolute risk based on the Breslow estimator, which does not make use of external incidence rates (Table 3). For the same two profiles shown above, the 10- and 20-year absolute risk estimates for a low-risk 50-year old woman are 0.29% (95% CI: 0.15 to 0.42) and 1.05% (95% CI: 0.77% to 1.33%), respectively; for a high-risk 50 year old woman, her 10- and 20- year absolute risk estimates are 1.22% (95% CI: 0.65% to 1.79%) and 4.38% (95% CI: 3.26% to 5.50%), respectively. The risk estimates are similar for both the proposed and the Breslow-based estimator with the proposed estimator having shorter confidence intervals, suggesting that our proposed estimates are more efficient than the Breslow-based estimators, although the efficiency gain of the proposed estimator diminishes at the older age.

To further demonstrate the patterns of absolute risks for subjects with different risk profiles, we plotted 5-20 years absolute risks in Figure 2 for white women aged 50, 55, 60 and 65 respectively with combined 4 levels of BMI (< 30 vs. ≥30) and estrogen status (negative vs. positive) while holding the rest of the risk factors constant: having SMC and no polyps, negative family history of CRC, less than 2 hours vigorous exercise every week, not taking aspirin or NSAIDs daily and eating less than 2 medium portion of vegetable every day. The absolute risks for the 4 combined levels of BMI and estrogen status present consistent patterns across the 4 age levels, with BMI less than 30 kg/m^2 and positive estrogen status being the lowest risk profile and BMI equal to or greater than 30 kg/m^2 and negative estrogen status being the highest risk profile. Clearly such information could help physicians and patients make more informed medical choices.

We further allowed for covariate effect on competing risks and assessed its impact by incorporating such effects in the modified absolute risk estimates proposed in Section 2.3. Specifically, we modeled the cause-specific hazards for competing risks using the Cox model with WHI data. Several risk factors for CRC have significant effects on the competing risks (death) including history of sigmoidoscopy and/or colonoscopy (SMC) and history of polyps in last 5 year (p-value=0.0003), NSAIDs (p-value < .0001), vegetable consumption (p-value< .0001) and BMI (p-value< .0001). We used the proposed method to estimate the cause-specific baseline hazard rate for competing risks. The external incidence rate for mortality due to causes other than CRC was extracted from 2008 NVSS. We plotted

the WHI mortality rate for non-CRC death in comparison with the 2008 NVSS in Figure S2 of the Supplementary Materials. The mortality rate for non-CRC from NVSS is generally higher than the WHI mortality rate, and is closed to the upper 95% simultaneous confidence band. The estimated multiplicative factor $\rho^{\dagger}$ is 0.96 (95% CI: 0.95 to 0.98). We re-calculated the absolute risks for the 12 profiles above using the covariate adjusted competing risk estimates. The results are shown in the Supplementary Materials Table S2. The absolute risk estimates are very closed to the estimates that didn't adjust for covariate effects on competing risks, suggesting that adjusting for covariates in the hazard function for competing risks does not have much influence on the absolute risk estimates in this application.

6. DISCUSSION

Large-scale prospective cohort studies are valuable resources for building risk prediction models. Risk factors are collected at the baseline and subjects are followed over time until the occurrence of diseases or termination of the study. However, for rare diseases, the number of events occurring during cohort follow up may be limited; as a result, the baseline hazard function may not be estimated reliably if using only the cohort data. Our novel approach utilizes external incidence rates to estimate the baseline hazard function, while allowing for differences between cohort and external incidence rates. We prove that the proposed estimator is uniformly consistent and converges weakly to a Gaussian process. We also show that in finite sample sizes the proposed estimator is susbstantially more efficient than the Breslow estimator that does not utilize the external incidence rates. Our proposed approach is aimed to improve the efficiency of baseline hazard estimation with the goal to improve the efficiency of absolute risk estimation. It does not affect the bias and efficiency of regression coefficients estimators which are obtained by maximizing the partial likelihood function of cohort data in a standard fashion.

We assumed the external disease incidence rate known from a population-based registry, for example, cancer incidence rates from SEER in our data example. Since the population-based registry typically involves a large number of subjects, the variability of the incidence rates are generally negligible compared to the variability of parameter estimates from the cohort. In some cases, we can also obtain the external disease incidence rate from another study. In these situations, the variability of the external disease incidence rate may not be negligible. This can be seen from the asymptotic distribution of $n^{1/2}\left\{\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)-\Lambda_0(t)\right\}$, which can be equivalently represented by

$$n^{1/2}\int_0^t\hat{\phi}\left(u;\hat{\boldsymbol{\beta}}\right)\left\{\hat{\Lambda}^*(du)-\Lambda^*(du)\right\}+n^{1/2}\int\lambda^*(u)\left\{\hat{\phi}\left(u;\hat{\boldsymbol{\beta}}\right)-\phi\left(u;\hat{\boldsymbol{\beta}}\right)\right\}du.$$

Since $\hat{\phi}\left(t;\hat{\boldsymbol{\beta}}\right)$ is uniformly consistent to $\varphi(t)$, the first term is asymptotically equivalent to $n^{1/2}\int_0^t\phi(u)\left\{\hat{\Lambda}^*(du)-\Lambda^*(du)\right\}$. Due to the fact that the external incidence rates are estimated from another study, the two terms are not correlated asymptotically. Therefore, the

asymptotic variance of $\hat{\Lambda}_0\left(t;\hat{\boldsymbol{\beta}}\right)$ is the sum of two variances: variance due to estimation of $\hat{\Lambda}^*(t)$ from external data and variance due to estimation of $\hat{\phi}\left(t;\hat{\boldsymbol{\beta}}\right)$ from the cohort.

An important issue for utilizing external disease incidence rates is that they may not be same as the disease incidence rate in the cohort. In this paper, we proposed to capture this potential difference by a scale factor ρ_0. Clearly, ρ_0 may vary with time, as shown in the example of the WHI study. In the extreme scenario of ρ_0 being completely nonparametric, the proposed estimator becomes the Breslow estimator, and there will be no gain in efficiency in estimating the baseline hazard function. In order to balance between bias and efficiency, we suggest to use a parametric function for ρ_0 such as the piece-wise constant function as used in the Application. The incidence rates of the cohort study and the external source should be carefully examined to determine the appropriate form of the parametric function for ρ_0.

Another important issue is whether the risk estimates obtained from the cohort study can be projected to the population that the external incidence rates come from, which is often of great interest in risk projection. We can test $H_0 : \rho_0 = 1$, and if there is not adequate evidence to reject the hypothesis, it may be reasonable to assume that there is no significant difference between the cohort and the external population, and apply the absolute risk estimates obtained from the cohort to this external population. If the hypothesis is rejected, we must be cautious about projecting the risk estimates to the population by simply using $\varphi(t)\lambda^*(t)$, because there may be several reasons that could give rise to discrepancy in the incidence rates between the external population and the cohort. For example, the discrepancy may be due to different risk factor distributions or even different hazard ratios. A much thorough investigation is needed to understand the underlying reasons for such discrepancy as each scenario requires different approach to handle misspecification. Methods for dealing this complex issue requires additional development, which is beyond the scope of this paper and will be communicated separately.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research was supported by the following grants from the National Institutes of Health: The work is in part supported by NIH grants (P01CA53996, U01-CA86368, R01AG14358 and R01GM085047). We are grateful to the generosity of WHI investigators for allowing us to use the WHI data to illustrate the method proposed in this paper. The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. The list of investigators is provided in Section B of the Supplementary Materials.

REFERENCES

Breslow N. Discussion following ”regression models and life tables by d. r. cox. Journal of the Royal Statistical Society B. 1972; 34:216–217.

Breslow N, Crowley J. A large sample study of the life table and product limit estimates under random censorship. The Annals of Statistics. 1974; 2(3):437–453.

Bruzzi P, Green S, Byar D, Brinton L, Schaiper C. Estimating the population attributable risk for multiple risk factors using case-control data. American journal of epidemiology. 1985; 122(5):904. [PubMed: 4050778]

CDC and NCHS. Multiple cause of death 1999-2010 on cdc wonder online database, released 2012. 2012

Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, Benichou J, Gail M. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. Journal of the National Cancer Institute. 2006; 98(17):1215–1226. [PubMed: 16954474]

Cox D. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological). 1972; 34(2):187–220.

Freedman A, Slattery M, Ballard-Barbash R, Willis G, Cann B, Pee D, Gail M, Pfeiffer R. Colorectal cancer risk prediction tool for white men and women without known susceptibility. Journal of Clinical Oncology. 2009; 27(5):686–693. [PubMed: 19114701]

Gail M. Personalized estimates of breast cancer risk in clinical practice and public health. Statistics in Medicine. 2011

Gail M, Brinton L, Byar D, Corle D, Green S, Schairer C, Mulvihill J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute. 1989; 81(24):1879. [PubMed: 2593165]

Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994; 81(3):515–526.

Kalbfleisch, JD.; Prentice, RL. The statistical analysis of failure time data. Second Edition. Vol. ume 360. John Wiley & Sons; 2002.

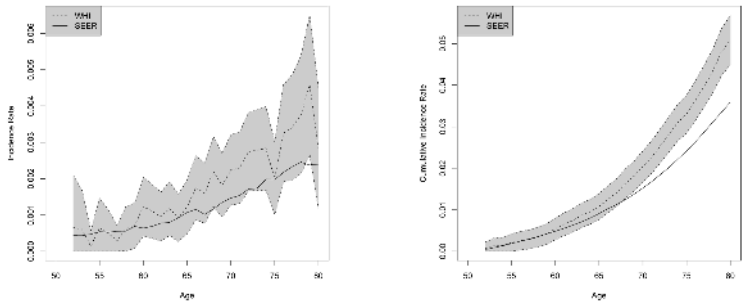

Figure 1.

Incidence and cumulative incidence rates from the Women's Health Initiative (WHI) and the SEER data. Shaded area: 95% simultaneous confidence band for the WHI estimates. (a) Incidence rate; (b) Cumulative incidence rate.

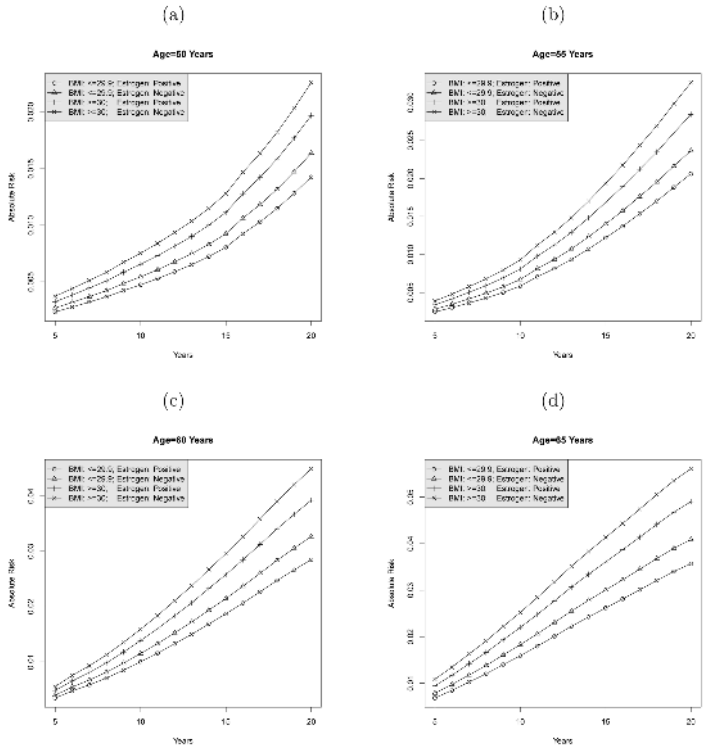

Figure 2.

Absolute risk estimates of colorectal cancer (CRC) from 5 to 20 years based on the WHI study for postmenopausal white women at 4 different ages: (a) 50 years old, (b) 55 years old, (c) 60 years old, and (d) 65 years old.

Table 1

Summary statistics of proposed estimators $\hat{\phi}(t)$, $\hat{\rho}$ and $\hat{\Lambda}_0(t)$ as well as the Breslow estimator for $\overline{\Lambda}_0(t)$ based on 1000 replicates with cohort size n = 1000. The bias (Bias), the empirical standard deviation of the 1000 estimators (ESD), the mean of the asymptotic standard errors (ASE), the coverage probability in percent (CP) of 95% nominal confidence intervals are presented for each estimator as well as the estimated asymptotic relative efficiency (ARE) of the proposed estimator $\hat{\Lambda}_0(t)$ compared with the Breslow estimator $\overline{\Lambda}_0(t)$.

(a) Proposed estimators $\hat{\phi}(t)$ and $\hat{\rho}$.

		$\hat{\phi}(t)$					$\hat{\rho}$				
CR	***t***	**True**	**Bias**	**ESD**	**ASE**	**CP**	**True**	**Bias**	**ESD**	**ASE**	**CP**
50%	20	0.555	0.000	0.039	0.039	95.9	1.5	0.000	0.005	0.006	95.7
	40	0.636	0.001	0.037	0.037	95.9					
	60	0.751	0.001	0.037	0.037	94.9					
70%	20	0.555	0.000	0.049	0.049	95.1	1.5	0.021	0.086	0.088	96.1
	40	0.636	0.001	0.046	0.046	95.1					
	60	0.751	0.004	0.045	0.045	95.8					

(b) Breslow estimator $\overline{\Lambda}_0(t)$ and proposed estimator $\hat{\Lambda}_0(t)$.

			$\overline{\Lambda}_0(t)$				$\hat{\Lambda}_0(t)$				
CR	*t*	$\Lambda_0(t)$	Bias	ESD	ASE	CP	Bias	ESD	ASE	CP	ARE
50%	20	0.04	0.000	0.005	0.006	95.7	0.001	0.003	0.004	95.4	2.38
	40	0.16	0.000	0.015	0.015	95.6	0.002	0.013	0.013	96.5	1.32
	60	0.36	0.001	0.029	0.028	95.1	0.004	0.026	0.028	97.1	1.21
70%	20	0.04	0.000	0.006	0.006	94.6	0.001	0.004	0.004	96.4	1.94
	40	0.16	0.001	0.017	0.017	96.1	0.004	0.016	0.017	96.8	1.16
	60	0.36	0.007	0.038	0.038	95.3	0.013	0.033	0.036	96.2	1.29

Table 2

Summary statistics of the proposed estimators for $\varphi(t)$, ρ_0 and $\Lambda_0(t)$ as well as 10-year absolute risk in percent $R(t|t_0]\ \mathbf{Z})$ based on 1000 replicates with cohort size $n = 20000$. This simulation mimics the WHI study on colorectal cancer. The bias (Bias), the empirical standard deviation of the 1000 estimators (ESD), the mean of the asymptotic standard errors (ASE), the coverage probability in percent (CP) of 95% confidence intervals are presented for each estimator.

(a) Proposed $\hat{\varphi}(t)$, $\hat{\rho}$ and $\hat{\Lambda}_0(t)$ and Breslow estimator $\bar{\Lambda}_0(t)$.

	$\hat{\varphi}(t)$					$\hat{\rho}$				
t	True	Bias	ESD	ASE	CP	True	Bias	ESD	ASE	CP
40	0.533	–0.001	0.027	0.027	95.4	1.5	0.007	0.045	0.047	95.5
60	0.543	–0.001	0.027	0.027	95.3					
75	0.553	–0.001	0.027	0.026	95.0					

		$\bar{\Lambda}_0(t)$				$\hat{\Lambda}_0(t)$				
t	$\Lambda_0(t)$	Bias	ESD	ASE	CP	Bias	ESD	ASE	CP	ARE
40	0.011	0.000	0.001	0.001	94.5	0.000	0.001	0.001	94.6	1.27
60	0.024	0.000	0.002	0.002	95.4	0.000	0.002	0.001	94.5	1.05
75	0.038	0.000	0.002	0.002	94.8	0.000	0.002	0.002	94.5	1.11

(b) Breslow-based absolute risk estimator $R(t \mid t_0; Z)$ and proposed estimator $\hat{R}(t \mid t_0; Z)$.

				$R(t \mid t_0; Z)$				$\hat{R}(t \mid t_0; Z)$				
t_0	Z_1	Z_2	$R(a\|t;Z)$	Bias	ESD	ASE	CP	Bias	ESD	ASE	CP	ARE
55	0	0	0.784	0.003	0.073	0.075	95.3	0.004	0.048	0.046	94.0	2.29
55	1	0	1.562	0.011	0.124	0.129	95.8	0.013	0.065	0.070	96.3	3.65
55	1	1	3.099	0.016	0.236	0.248	96.0	0.020	0.121	0.132	96.8	3.83
60	0	0	0.835	0.004	0.085	0.087	95.5	0.004	0.051	0.049	94.0	2.76
60	1	0	1.662	0.012	0.145	0.150	95.3	0.013	0.069	0.074	96.9	4.40
60	1	1	3.295	0.018	0.284	0.290	95.1	0.020	0.131	0.142	96.8	4.71
65	0	0	0.874	0.000	0.097	0.103	95.9	0.004	0.054	0.051	94.2	3.21
65	1	0	1.741	0.005	0.166	0.183	97.1	0.014	0.074	0.078	96.0	5.11
65	1	1	3.450	0.004	0.328	0.355	96.7	0.022	0.141	0.150	96.6	5.40

Table 3

The 10- and 20-year absolute risk estimates of colorectal cancer and 95% confidence intervals (CI) for the proposed and Breslow estimator for white women with representative risk profiles at different ages based on the Women's Health Initiative study.

Profile	Age (Years)	Sigmoidoscopy and / or Colonoscopy in the Last 5 Years	Polyp in the Last 5 Years	No. of First-Degree Relatives With CRC	Current Leisure-Time Activity (h/wk)	Regular User of Aspirin/ NSAIDs	Vegetable Intake (med/d)	Body Mass Index (kg/m^2)	Estrogen Status Within the Last 2 Years	The Proposed Estimates			
										10-Year		20-Year	
										%	95%CI	%	95%CI
1	50	Yes	No	0	3	Yes	2.5	28	Positive	0.38	(0.29, 0.47)	1.14	(0.87, 1.40)
2	50	Yes	Yes	1	1	Yes	2.5	29	Negative	0.62	(0.44, 0.79)	1.85	(1.33, 2.37)
3	50	No		2	0	No	1.3	32	Negative	1.59	(1.21, 1.98)	4.74	(3.67, 5.80)
4	55	Yes	No	0	3	Yes	2.1	28	Positive	0.47	(0.34, 0.59)	1.64	(1.26, 2.02)
5	55	Yes	Yes	1	1	No	1.3	31	Positive	1.72	(1.14, 2.31)	5.98	(4.18, 7.78)
6	55	No		2	0	No	1	32	Negative	1.96	(1.44, 2.48)	6.79	(5.29, 8.28)
7	60	Yes	No	0	3	Yes	2.5	28	Positive	0.80	(0.61, 0.99)	2.27	(1.75, 2.80)
8	60	Yes	Yes	1	1	No	1.3	32	Positive	2.94	(2.04, 3.85)	8.19	(5.76, 10.61)
9	60	No		2	0	No	1.3	32	Negative	3.35	(2.59, 4.11)	9.27	(7.27, 11.28)
10	65	Yes	No	0	3	Yes	2.5	28	Positive	1.27	(0.97, 1.57)	2.85	(2.19, 3.51)
11	65	Yes	Yes	1	1	Yes	2.5	28	Negative	2.06	(1.46, 2.66)	4.61	(3.29, 5.93)
12	65	No		2	0	No	1.3	32	Negative	5.28	(4.10, 6.45)	11.48	(9.02, 13.93)
										The Breslow Estimator Based Estimates			
1	50	Yes	No	0	3	Yes	2.5	28	Positive	0.29	(0.15, 0.42)	1.05	(0.77, 1.33)
2	50	Yes	Yes	1	1	Yes	2.5	29	Negative	0.47	(0.22, 0.72)	1.71	(1.16, 2.26)
3	50	No		2	0	No	1.3	32	Negative	1.22	(0.65, 1.79)	4.38	(3.26, 5.50)
4	55	Yes	No	0	3	Yes	2.1	28	Positive	0.47	(0.35, 0.60)	1.56	(1.19, 1.92)
5	55	Yes	Yes	1	1	No	1.3	31	Positive	1.74	(1.15, 2.33)	5.67	(3.95, 7.38)
6	55	No		2	0	No	1	32	Negative	1.98	(1.46, 2.51)	6.43	(4.99, 7.88)
7	60	Yes	No	0	3	Yes	2.5	28	Positive	0.80	(0.60, 1.00)	2.15	(1.65, 2.65)
8	60	Yes	Yes	1	1	No	1.3	32	Positive	2.95	(2.04, 3.86)	7.76	(5.51, 10.01)
9	60	No		2	0	No	1.3	32	Negative	3.36	(2.57, 4.15)	8.80	(6.91, 10.69)
10	65	Yes	No	0	3	Yes	2.5	28	Positive	1.17	(0.89, 1.45)	3.01	(2.30, 3.73)

Profile	Age (Years)	Sigmoidoscopy and / or Colonoscopy in the Last 5 Years	Polyp in the Last 5 Years	No. of First-Degree Relatives With CRC	Current Leisure-Time Activity (h/wk)	Regular User of Aspirin/ NSAIDs	Vegetable Intake (med/d)	Body Mass Index (kg/m^2)	Estrogen Status Within the Last 2 Years	The Proposed Estimates			
										10-Year		20-Year	
										%	95%CI	%	95%CI
11	65	Yes	Yes	1	1	Yes	2.5	28	Negative	1.90	(1.35, 2.45)	4.87	(3.52, 6.22)
12	65	No		2	0	No	1.3	32	Negative	4.87	(3.78, 5.95)	12.09	(9.55, 14.62)