STATISTICS IN MEDICINE

Statist. Med. 2000; **00**:1–26 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Assessing Cumulative Incidence Functions under the Semiparametric Additive Risk Model

Seunggeun Hyun[1], Yanqing Sun[2] and Rajeshwari Sundaram[3,*]

[1] *Division of Mathematics and Computer Science, University of South Carolina Upstate, Spartanburg, SC 29303,* [2] *Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, NC 28223,* [3] *Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research,* Eunice Kennedy Shriver *National Institute of Child Health and Human Development, National Institutes of Health, 6100 Executive Boulevard, Rockville, MD 20852*

SUMMARY

In analyzing competing risks data, a quantity of considerable interest is the cumulative incidence function. Often the effect of covariates on the cumulative incidence function is modeled via the proportional hazards model for the cause-specific hazard function. As the proportionality assumption may be too restrictive in practice, we consider an alternative more flexible semiparametric additive hazards model of [1] for the cause-specific hazard. This model specifies the effect of covariates on the cause-specific hazard to be additive as well as allows the effect of some covariates to be fixed and that of others to be time-varying. We present an approach for constructing confidence intervals as well as confidence bands for the cause-specific cumulative incidence function of subjects with given values

*Correspondence to: Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, 6100 Executive Boulevard, Rockville, MD 20852

Received

Revised

of the covariates. Furthermore, we also present an approach for constructing confidence intervals and confidence bands for comparing two cumulative incidence functions given values of the covariates. The finite sample property of the proposed estimators is investigated through simulations. We conclude our paper with an analysis of the well-known malignant melanoma data using our method.

1. INTRODUCTION

Competing risks data are typically encountered in medical studies, for example, in studies dealing with time to progression of spontaneous labor, where labor due to medical intervention (example: delivery by cesarean) or membrane rupture leading to labor are treated as other causes. Another example can be found in the well-known malignant melanoma data [2], where patients with malignant melanoma were either at risk to die from malignant melanoma or from other causes. Typically response to a treatment can be classified in terms of failure from disease of interest and/or non-disease related causes. So in competing risks framework, each individual is exposed to K distinct types of risks and the eventual failure can be attributed to precisely one of the risks. Suppose each subject has an underlying continuous failure time $\tilde{T}$ that may be subject to censoring. The cause of failure $J \in \{1, \ldots, K\}$ is observed along with covariate information. The cause-specific hazard function for a subject with a covariate vector z is defined by

$$\lambda_k(t|z) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\left(t \leq \tilde{T} < t + \Delta t, J = k | \tilde{T} \geq t, z\right)$$

for $k = 1, \ldots, K$. The cause-specific hazard function $\lambda_k(t|z)$ is the instantaneous rate of occurrence of the k-th failure cause at time t in the presence of all causes of failure given z. To predict the survival probability of the k-th type of failure time with a particular set of

covariates z, $S_k(t|z) = \exp\left(-\int_0^t \lambda_k(u|z)\,du\right)$ is not meaningful as it has no simple probability interpretation when the competing risks are dependent [3]. The more appropriate function is the cumulative incidence function, defined by

$$F_k(t|z) = P\left(\tilde{T} \leq t, J = k|z\right).$$

The cumulative incidence function $F_k(t|z)$ is the probability of a subject failing from cause k at time t in the presence of all the competing risks given z.

Note that $F_k(t|z)$ can be expressed in terms of $\lambda_l(t|z)$, where $l = 1, \ldots, K$. It is easy to show that

$$F_k(t|z) = \int_0^t S(u|z)\lambda_k(u|z)\,du$$

with the overall survival function $S(t|z) = \exp\left(-\int_0^t \sum_{l=1}^K \lambda_l(u|z)\,du\right)$. Thus, it is natural to estimate the cumulative incidence function through the cause-specific hazard function. Typically, one models the effect of covariates using the proportional hazards model ([4], [5]). Cheng *et al.* have studied the estimation of the cumulative incidence function based on Cox's regression model in a competing risks model [6]. However, the proportionality assumption under the proportional hazards model may be too restrictive in practice. For instance, the proportionality assumption for the malignant melanoma data fails based on the score test provided by [7] (see the numerical section for further details). Thus, it is of interest to investigate alternative approach and an important alternative is the additive risk model.

Shen and Cheng [8] presented an approach to estimating the cumulative incidence function under the additive risk model first proposed by [9]

$$\lambda_k(t|z) = \lambda_{0k}(t) + z^T\beta_k, \tag{1}$$

where $\lambda_{0k}(\cdot)$ is an unspecified baseline hazard function for the k-th failure type and β_k an

unknown parameter vector. Often, this model only provides a rough summary of the effect of covariates, because in their model the influence of all the covariates was restricted to be constant, i.e., time-varying effects of covariates cannot be captured by this model. Aalen *et al.* [10] have considered a fully nonparametric additive model for the cause-specific transition hazards, where the effects of all the covariates are assumed to be time-varying in dealing with Markov chain models.

In many instances in practice, we may know that some of the covariates will have time-constant effects (for example, demographic characteristics) and some may be suspected to have time-varying effect (for example, treatment). In particular, using the test provided by [11] for checking the time-varying effect or constant effect of a covariate based on the additive model , we found that age and sex have time constant effect, and thickness of the tumor has time-varying effect for the malignant melanoma data. Motivated by this, we study a flexible additive risk model which allows for some covariates to be modeled parametrically and others to be modeled nonparametrically. The suggested approach is to study the semiparametric additive model based on [1] in the competing risks setup which is more appropriate in some applications.

To be specific, under the semiparametric additive risk model the cause-specific hazard function for cause k, given covariates x and z takes the form

$$\lambda_k(t|x,z) = x^T\alpha_k(t) + z^T\beta_k,$$

where the covariates are partitioned into x, a p-dimensional covariate vector with time-varying effects and z, a q-dimensional covariate vector with time constant effects, $\alpha_k(t)$ is a p-dimensional locally integrable function, and β_k is a q-dimensional regression vector. The first component of x may be set to 1 to allow for a general baseline hazard. When some

covariates are known or anticipated to yield time-varying effects, they will be investigated nonparametrically. In this paper, we present a way to construct confidence interval for the cumulative incidence function for a specific cause given specific value of covariates. We also present an approach to constructing confidence intervals and bands for cumulative incidence functions under semi-parametric additive model. Furthermore, we also propose methods to compare two cumulative incidence functions by constructing confidence interval and band of their ratio or difference.

The paper is structured as follows. In Section 2 we present our methods for constructing confidence intervals and bands for the cumulative incidence function and for the difference and the ratio of two cumulative incidence functions. The finite sample property of the proposed estimators is investigated extensively through simulations in Section 3. In Section 4 we illustrate the proposed method with data from malignant melanoma study. The details of our asymptotic results are made explicit in the Appendix.

2. Estimation of the Cumulative Incidence Functions

Suppose that there are K distinct failure types. Let T_{ki} be the k-th latent failure time for individual i, where $k = 1, \ldots, K$ and $i = 1, \ldots, n$. τ is the end of follow-up. Then, one can only observe $(T_i, J_i, \delta_i, x_i, z_i)$, where $T_i = \min(\tilde{T}_i, C_i)$, $\tilde{T}_i = \min_k\{T_{ki}, k = 1, \ldots, K\}$ and (x_i, z_i) are covariates for the i-th subject. Here, J_i is the failure type indicator ($J_i = k$ if $\tilde{T}_i = T_{ki}$), δ_i is the censoring indicator ($\delta_i = 1$ if $\tilde{T}_i$ is observed and 0 otherwise), and C_i is the independent censoring variable. We assume that T_{ki} does not equal T_{li} for $k \neq l$, $k, l = 1, \ldots, K$. In this paper, we assume that the cause-specific hazard function for T_{ki}, given covariates x_i and z_i is

modeled by

$$\lambda_{ki}(t|x_i, z_i) = x_i^T \alpha_k(t) + z_i^T \beta_k, \tag{2}$$

for $k \in \{1, 2, \cdots, K\}$. Let $Y_i(t) = I(T_i \geq t)$ indicate the at-risk process of the i-th individual, where $I(\cdot)$ is the indicator function. We organize the covariates into design matrices

$$X(t) = \big(Y_1(t)x_1, \ldots, Y_n(t)x_n\big)^T \text{ and } Z(t) = \big(Y_1(t)z_1, \ldots, Y_n(t)z_n\big)^T.$$

Let $\delta_{ki} = \delta_i I(J_i = k)$. Then $N_{ki}(t) = \delta_{ki} I(T_i \leq t)$ indicates whether an event of type k has been observed by time t for individual i. Let

$$N_k(t) = \big(N_{k1}(t), \ldots, N_{kn}(t)\big)^T \text{ and } \lambda_k(t) = \big(\lambda_{k1}(t), \ldots, \lambda_{kn}(t)\big)^T$$

be the n-dimensional counting process and its intensity. Denote by $\omega_k(t)$, an $n \times n$ matrix given by $\text{diag}\{1/\lambda_{k1}, \ldots, 1/\lambda_{kn}\}$ for $k \in \{1, \ldots, K\}$. Estimation of (2) for the cause-specific function can be carried out by using the estimation procedures proposed in [1]. Each β_k and the cumulative time-varying effect $A_k(t) = \int_0^t \alpha_k(u)\, du$, $k = 1, \ldots, K$ can be consistently estimated by $\widehat{\beta}_k$ and $\widehat{A}_k(t)$, while treating all the failure time T_i with $J_i \neq k$ as censored observation. To be specific,

$$\widehat{\beta}_k = \left[\int_0^\tau Z^T(t) H_k(t) Z(t)\, dt\right]^{-1} \int_0^\tau Z^T(t) H_k(t)\, dN_k(t)$$

and

$$\widehat{A}_k(t) = \int_0^t X_k^-(u) \left(dN_k(u) - Z(u)\widehat{\beta}_k\, du \right),$$

where $H_k(t) = \omega_k(t) - \omega_k(t)X(t)W_k(t)^{-1}X^T(t)\omega_k(t)$, $X_k^-(t) = W_k(t)^{-1}X^T(t)\omega_k(t)$, and $W_k(t) = X^T(t)\omega_k(t)X(t)$.

However, the above estimates involve the weight function $\omega_k(t)$. As suggested in [1], one can replace $\omega_k(t)$ by an identity matrix and calculate the estimates for β_k and $A_k(\cdot)$ and use them to estimate the weight matrix $\omega_k(\cdot)$ and calculate the consistent estimates $\hat{\beta}_k$ and $\widehat{A}_k(\cdot)$.

By combining these we can consistently estimate the cumulative incidence function for the cause 1, for example, with a particular set of covariates x_0 and z_0 by

$$\widehat{F}_1(t|x_0,z_0)=\int_0^t \widehat{S}(u|x_0,z_0)\,d\widehat{\Lambda}_1(u|x_0,z_0)$$

where $\widehat{S}(t|x_0,z_0) = \exp\left(-\sum_{k=1}^K \widehat{\Lambda}_k(t|x_0,z_0)\right)$ is an estimator of the overall survival probability $S(t|x_0,z_0)$ and $\widehat{\Lambda}_k$ is the estimator of the cumulative cause-specific hazard function $\widehat{\Lambda}_k(t|x_0,z_0)=\int_0^t x_0^T\,d\widehat{A}_k(u)+tz_0^T\widehat{\beta}_k$. We next present our approach for constructing confidence interval and confidence band for a specific cause-k cumulative incidence function conditional on specific values of the covariate.

2.1. Confidence interval (band) for a specific cumulative incidence function

In the Appendix, we show that $\sqrt{n}\left(\widehat{F}_1(t|x_0,z_0)-F_1(t|x_0,z_0)\right)$ is asymptotically equivalent to a sum of square integrable martingales $U_1(t|x_0,z_0)$ given by

$$\begin{aligned}
&U_1(t|x_0,z_0)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_{1i}(t|x_0,z_0)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big\{\int_0^t S(u|x_0,z_0)x_0^T\left(n^{-1}W_1(u)\right)^{-1}\omega_{1i}(u)x_i\,dM_{1i}(u)\\
&\qquad\qquad-\int_0^t S(u|x_0,z_0)x_0^T X_1^-(u)Z(u)\,duC_1^{-1}D_{1i}\\
&\qquad\qquad-\int_0^t S(u|x_0,z_0)\,duz_0^TC_1^{-1}D_{1i}\\
&\qquad-\sum_{k=1}^K\Big(\int_0^t F_1^C(t,u)x_0^T\left(n^{-1}W_k(u)\right)^{-1}\omega_{ki}(u)x_i\,dM_{ki}(u)\\
&\qquad\qquad\qquad-\int_0^t F_1^C(t,u)x_0^T\omega_k(u)X^-(u)Z(u)\,duC_k^{-1}D_{ki}\\
&\qquad\qquad\qquad-\int_0^t F_1^C(t,u)\,duz_0^TC_k^{-1}D_{ki}\Big)\Big\},
\end{aligned}\tag{3}$$

where

$$\begin{aligned} F_k^C(t,u) &= F_k(t|x_0,z_0) - F_k(u|x_0,z_0) \\ C_k &= \frac{1}{n}\int_0^\tau Z^T(t)H_k(t)Z(t)\,dt \\ D_{ki} &= \int_0^\tau \left\{ z_i\omega_{ki}(t) - Z^T(t)\omega_k(t)X(t)W_k(t)^{-1}\omega_{ki}(t)x_i \right\} dM_{ki}(t) \\ M_{ki}(t) &= N_{ki}(t) - \int_0^t Y_i(u)\{\, x_i^T\alpha_k(u) + z_i^T\beta_k\}\,du \end{aligned}$$

for $k \in \{1,2,\ldots,K\}$ and $i = 1,2,\ldots,n$. Furthermore using the martingale central limit theorem, it converges in distribution to a Gaussian process. The martingale representation of $U_1(t|x_0,z_0)$ can also be used to construct a consistent estimator of the asymptotic variance function. The variance function at time t can be consistently estimated by

$$\widehat{\sigma}_1^2(t|x_0,z_0) = \frac{1}{n}\sum_{i=1}^n \big(\widehat{\epsilon}_{1i}(t|x_0,z_0)\big)^2$$

obtained by replacing all the terms with their empirical version and the parameters β_k and A_k by their consistent estimates $\widehat{\beta}_k$ and $\widehat{A}_k$.

Next, combining the asymptotic normality of $\widehat{F}_1(t|x_0,z_0)$ and the consistent estimator $\widehat{\sigma}_1^2(t|x_0,z_0)$ for the asymptotic variance, we can construct an $(1-\alpha)\times 100\%$ confidence interval for $F_1(t;x_0,z_0)$ as follows:

$$g^{-1}\left(g(\widehat{F}_1(t|x_0,z_0)) \pm n^{-1/2} g'(\widehat{F}_1(t|x_0,z_0))\widehat{\sigma}_1(t|x_0,z_0)z_{\alpha/2} \right),$$

where g is a smooth function chosen such that it retains the range of the distribution F_1, g' its continuous derivative and g^{-1} denotes the inverse function of g. Typically, the transformation g is chosen for constructing confidence interval for distributions so as to retain their range as well as to improve the coverage probability. Observe that using the functional delta method, the asymptotic normality of $g(\widehat{F}_1(t|x_0,z_0))$ follows from that of $\widehat{F}_1(t|x_0,z_0)$, but with asymptotic variance given by $(g^{'}(F_1(t|x_0,z_0)))^2\sigma_1^2(t|x_0,z_0)$.

Another quantity of interest is the confidence band for the cumulative incidence function. However, in order to construct an $(1-\alpha)\times 100\%$ confidence band for $F_1(t|x_0, z_0)$ simultaneously for $t \in [0, \tau]$, we need to investigate the distribution of the supremum of the process $U_1(t|x_0, z_0)$, $0 \leq t \leq \tau$. This is analytically challenging as U_1 does not have an independent increment structure. Alternatively, we adapt the general procedure suggested by [7] to get an approximation of the distribution of the process U_1. We approximate the distribution of $U_1(t|x_0, z_0)$ by a zero-mean Gaussian process, denoted by $\widehat{U}_1(t|x_0, z_0)$, whose distribution can be easily generated through simulations. We replace $\{M_{ki}(t)\}$ in (3) with $\{N_{ki}(t)G_i\}$, and the other unknown quantities in (3) with their respective sample estimators to yield $\widehat{U}_1(t|x_0, z_0)$, where $\{G_i : i = 1, \ldots, n\}$ are independent standard normal variables. The asymptotic equivalence of $U_1(t|x_0, z_0)$ and $\widehat{U}_1(t|x_0, z_0)$ is established by the fact that

$$E\{M_{ki}(t)\} = 0, \quad Var\{M_{ki}(t)\} = E\{N_{ki}(t)\}.$$

To approximate the distribution of $U_1(t|x_0, z_0)$, we simply obtain a large number, say M, of realizations from $\widehat{U}_1(t|x_0, z_0)$ by repeatedly generating random samples $\{G_i\}$, while fixing the data $\{(T_i, \delta_{ki}, x_i, z_i) : k = 1, \ldots, K, i = 1, \ldots, n\}$ at their observed values.

An $(1 - \alpha) \times 100\%$ confidence band for $F_1(t|x_0, z_0)$ on $[t_1, t_2]$ is given by

$$\widehat{F}_1(t|x_0, z_0) \pm n^{-1/2} c_\alpha \widehat{\sigma}_1(t|x_0, z_0),$$

where the cutoff value c_α is given by

$$P\left[\sup_{t_1 \leq t \leq t_2} |\widehat{U}_1(t|x_0, z_0)/\widehat{\sigma}_1(t|x_0, z_0)| > c_\alpha\right] = \alpha$$

which can be estimated through replicates of $\widehat{U}_1(t|x_0, z_0)$ given the observed data. Here t_1 and t_2 $(t_1 < t_2)$ can be any time points between $(0, t_0)$, where $t_0 = \inf\{t : EY_1(t) = 0\}$.

2.2. Comparison of two cumulative incidence functions

In many practical situations, one is interested in comparing cumulative incidence functions for a specific cause given two different values of covariates or comparing cumulative incidence functions for two different causes given a specific value of the covariate. For instance in malignant melanoma data, the difference in cumulative incidence functions for malignant melanoma between females and males measures the difference in the probability of dying from malignant melanoma between females and males. Similarly, the difference between cumulative incidence functions at different values of tumor thickness shows how the probability of dying from malignant melanoma is changed by tumor thickness. While the difference of two cumulative incidence functions measures additional probability of survival/dying from a cause, the ratio shows a relative rate of survival/dying from a cause at two different covariate values.

We begin by presenting a way to compare cumulative incidence functions for k-th cause, say, the first cause given two different values for covariates $(x, z) = (x_1, z_1)$ and $(x, z) = (x_2, z_2)$. We will first present the method for calculating the confidence interval and confidence band for their ratio. Using the consistency of $\widehat{F}_k$ and boundedness in probability of $\sqrt{n}\big(F_k(t|x,z) - F_k(t|x,z)\big)$ for $t \in [0, \tau]$ established in previous section, it follows that

$$\sqrt{n}\left(\frac{\widehat{F}_1(t|x_1,z_1)}{\widehat{F}_1(t|x_2,z_2)} - \frac{F_1(t|x_1,z_1)}{F_1(t|x_2,z_2)}\right) \approx \frac{1}{F_1(t|x_2,z_2)}\sqrt{n}(\widehat{F}_1(t|x_1,z_1) - F_1(t|x_1,z_1)) - \frac{F_1(t|x_1,z_1)}{F_1^2(t|x_2,z_2)}\sqrt{n}(\widehat{F}_1(t|x_2,z_2) - F_1(t|x_2,z_2)),$$

where $\approx$ denotes asymptotic equivalence. For notational simplicity, we denote by $a(t) = 1/F_1(t|x_2,z_2)$ and by $b(t) = F_1(t|x_1,z_1)/F_1^2(t|x_2,z_2)$. Using similar arguments as in (3), it follows that $\sqrt{n}\left(\widehat{F}_1(t|x_1,z_1)/\widehat{F}_1(t|x_2,z_2) - F_1(t|x_1,z_1)/F_1(t|x_2,z_2)\right)$ is asymptotically

equivalent to a sum of square integrable martingales given by

$$
\begin{aligned}
&\widetilde{U}(t|x_1, z_1, x_2, z_2) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\epsilon}_i(t|x_1, z_1, x_2, z_2) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \int_0^t (a(u)S(u|x_1, z_1)x_1 - b(u)S(u|x_2, z_2)x_2)^T \left(n^{-1}W_1(u)\right)^{-1} \omega_{1i}(u)x_i \, dM_{1i}(u) \\
&\qquad\quad - \int_0^t (a(u)S(u|x_1, z_1)x_1 - b(u)S(u|x_2, z_2)x_2)^T X_1^-(u)Z(u) \, du C_1^{-1} D_{1i} \\
&\qquad\quad - \int_0^t (a(u)S(u|x_1, z_1)z_1 - b(u)S(u|x_2, z_2)z_2)^T \, du C_1^{-1} D_{1i} \\
&\qquad - \sum_{k=1}^{K} \Big(\int_0^t (a(u)F_1^C(t, u|x_1, z_1)x_1 - b(u)F_1^C(t, u|x_2, z_2)x_2)^T \left(n^{-1}W_k(u)\right)^{-1} \omega_{ki}(u)x_i \, dM_{ki}(u) \\
&\qquad\qquad\quad - \int_0^t (a(u)F_1^C(t, u|x_1, z_1)x_1 - b(u)F_1^C(t, u|x_2, z_2)x_2)^T X_k^-(u)Z(u) \, du C_k^{-1} D_{ki} \\
&\qquad\qquad\quad - \int_0^t (a(u)F_1^C(t, u|x_1, z_1)z_1 - b(u)F_1^C(t, u|x_2, z_2)z_2)^T \, du C_k^{-1} D_{ki} \Big) \Big\},
\end{aligned}
\tag{4}
$$

where $F_1^C(t, u|x_1, z_1) = F_1(t|x_1, z_1) - F_1(u|x_1, z_1)$, $F_1^C(t, u|x_2, z_2) = F_1(t|x_2, z_2) - F_1(u|x_2, z_2)$, C_k, ω_k and D_{ki} are defined in Section 2.1. Using the martingale central limit theorem, the above process converges in distribution to a Gaussian process. Furthermore, using the i.i.d. representation of $\widetilde{U}(t|x_1, z_1, x_2, z_2)$ we can construct a consistent estimator of the asymptotic variance function. The variance function at time t can be consistently estimated by

$$
\widehat{\widetilde{\sigma}}^2(t|x_1, z_1, x_2, z_2) = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\widetilde{\epsilon}}_i(t|x_1, z_1, x_2, z_2) \right)^2 \tag{5}
$$

obtained by replacing all the terms in (4) with their empirical version and $F_k(t, x_i, z_i)$ by their consistent estimator $\widehat{F}_k(t, x_i, z_i)$ and the parameters β_k and A_k by their consistent estimates $\widehat{\beta}_k$ and $\widehat{A}_k$.

Next, to construct an $(1 - \alpha) \times 100\%$ confidence band for $F_1(t|x_1, z_1)/F_1(t|x_2, z_2)$ for

$t \in [t_1, t_2]$, we propose a Gaussian multiplier approach. Let

$$\widehat{\widetilde{U}}(t|x_1, z_1, x_2, z_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\widetilde{\epsilon}}_i(t|x_1, z_1, x_2, z_2) G_i,$$

where $\{G_i : i = 1, \ldots, n\}$ are independent standard normal random variable. Observe that $\widehat{\widetilde{U}}(t|x_1, z_1, x_2, z_2)$ is asymptotically equivalent to $\widetilde{U}(t|x_1, z_1, x_2, z_2)$. We can construct an $(1-\alpha) \times 100\%$ confidence band for the ratio $F_1(t|x_1, z_1)/F_1(t|x_2, z_2)$ on $[t_1, t_2]$ as follows:

$$\frac{\widehat{F}_1(t|x_1, z_1)}{\widehat{F}_1(t|x_2, z_2)} \pm n^{-1/2} c_\alpha \widehat{\widetilde{\sigma}}(t|x_1, z_1, x_2, z_2),$$

where the cut-off value c_α given by

$$P\left[\sup_{t_1 \le t \le t_2} |\widehat{\widetilde{U}}(t|x_1, z_1, x_2, z_2)/\widehat{\widetilde{\sigma}}(t|x_1, z_1, x_2, z_2)| > c_\alpha\right] = \alpha.$$

The cut-off value can be estimated by approximating the distribution of $\widetilde{U}(t|x_1, z_1, x_2, z_2)$ by obtaining a large number of realizations of $\widehat{\widetilde{U}}(t|x_1, z_1, x_2, z_2)$ given the observed data by repeatedly generating random samples of $\{G_i : i = 1, \ldots, n\}$.

Similarly, we can construct confidence interval and confidence band for the difference between cumulative incidence functions for a specific cause-k given different values of covariates, i.e., for $F_k(t|x_1, z_1) - F_k(t|x_2, z_2)$. Notice that

$$\sqrt{n}\left(\widehat{F}_1(t|x_1, z_1) - F_1(t|x_1, z_1)\right) - \sqrt{n}\left(\widehat{F}_1(t|x_2, z_2) - F_1(t|x_2, z_2)\right)$$

is asymptotically equivalent to $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\epsilon}_i(t|x_1, z_1, x_2, z_2)$ with $a(t) = b(t) = 1$ for all $t \in [0, t_0]$. Hence, an $(1-\alpha)\times 100\%$ confidence interval and confidence band in $t \in [t_1, t_2]$ for the difference can be constructed as

$$\left(\widehat{F}_1(t|x_1, z_1) - \widehat{F}_1(t|x_2, z_2)\right) \pm n^{-1/2} c_\alpha \widehat{\widetilde{\sigma}}(t|x_1, z_1, x_2, z_2),$$

where c_α and $\widehat{\widetilde{\sigma}}(\cdot)$ can be calculated as mentioned above with $a(t) = 1$ and $b(t) = 1$ for $t \in [0, t_0]$ with $t_0 = \inf\{t : EY_1(t) = 0\}$.

Next, we discuss constructing confidence interval and band for comparing cumulative incidence functions for two different causes given a specific value for the covariate. We begin by considering their ratio i.e., $F_1(t|x_0,z_0)/F_2(t|x_0,z_0)$. Using the consistency of $\widehat{F}_k$ and boundedness in probability of $\sqrt{n}\big(F_k(t|x,z)-F_k(t|x,z)\big)$ for $t\in[0,t_0]$ established in previous section, it follows that

$$\sqrt{n}\left(\frac{\widehat{F}_1(t|x_0,z_0)}{\widehat{F}_2(t|x_0,z_0)}-\frac{F_1(t|x_0,z_0)}{F_2(t|x_0,z_0)}\right)\approx\frac{1}{F_2(t|x_0,z_0)}\sqrt{n}(\widehat{F}_1(t|x_0,z_0)-F_1(t|x_0,z_0))$$
$$-\frac{F_1(t|x_0,z_0)}{F_2^2(t|x_0,z_0)}\sqrt{n}(\widehat{F}_2(t|x_0,z_0)-F_2(t|x_0,z_0)).$$

For notational simplicity, we denote by $a(t)=1/F_1(t|x_0,z_0)$ and by $b(t)=F_1(t|x_0,z_0)/F_2^2(t|x_0,z_0)$. Using similar arguments as in (4), we can show that $\sqrt{n}\left(\widehat{F}_1(t|x_0,z_0)/\widehat{F}_2(t|x_0,z_0)-F_1(t|x_0,z_0)/F_2(t|x_0,z_0)\right)$ is asymptotically equivalent to a sum of square integrable martingales given by

$$\begin{aligned}
&\widetilde{U}(t|x_0,z_0)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{\epsilon}_i(t|x_0,z_0)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\int_0^t S(u|x_0,z_0)x_0^T(a(u)\left(n^{-1}W_1(u)\right)^{-1}\omega_{1i}(u)x_i\,dM_{1i}(u)\\
&\qquad-b(u)\left(n^{-1}W_2(u)\right)^{-1}\omega_{2i}(u)x_i\,dM_{2i}(u))\\
&\qquad-\int_0^t S(u|x_0,z_0)x_0^T(a(u)X_1^-(u)Z(u)C_1^{-1}D_{1i}-b(u)X_2^-(u)Z(u)C_2^{-1}D_{2i})\,du\\
&\qquad-\int_0^t S(u|x_0,z_0)z_0^T(a(u)C_1^{-1}D_{1i}-b(u)C_2^{-1}D_{2i})\,du\\
&\quad-\sum_{k=1}^{K}\Big(\int_0^t(a(u)F_1^C(t,u)-b(u)F_2^C(t,u))x_0^T\left(n^{-1}W_k(u)\right)^{-1}\omega_{ki}(u)x_i\,dM_{ki}(u)\\
&\qquad-\int_0^t(a(u)F_1^C(t,u)-b(u)F_2^C(t,u))x_0^TX_k^-(u)Z(u)\,duC_k^{-1}D_{ki}\\
&\qquad-\int_0^t(a(u)F_1^C(t,u)-b(u)F_2^C(t,u))\,duz_0^TC_k^{-1}D_{ki}\Big)\Big\}.
\end{aligned}\tag{6}$$

Using the martingale central limit theorem, the above process converges in distribution to

a Gaussian process. Furthermore, using the martingale representation of $\widetilde{U}(t|x_0, z_0)$ we can construct a consistent estimator of the asymptotic variance function. The variance function at time t can be consistently estimated by

$$\widehat{\widetilde{\sigma}}^2(t|x_0, z_0) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\widetilde{\epsilon}}_i(t|x_0, z_0)\right)^2$$

obtained by replacing all the terms with their empirical version and $F_k(t|x_0, z_0)$ by their consistent estimator $\widehat{F}_k(t|x_0, z_0)$ and the parameters β_k and A_k by their consistent estimates $\widehat{\beta}_k$ and $\widehat{A}_k$. The rest is similar to constructing an $(1-\alpha)\times 100\%$ confidence interval and band for comparing cumulative incidence functions for a cause-k given two different values of covariates.

3. Simulation Study

We investigate the finite sample properties of the various proposed confidence intervals and confidence bands for the cumulative incidence function(s) and apply the methods in analyzing a real data. We first begin with simulation studies conducted to evaluate the performance of our proposed method. We generated a semiparametric additive risks model (2) with two causes. The covariate x associated with the time-varying parameter was chosen to be a 0/1 variable each with probability .5. We also included a covariate z, associated with the time constant effect generated from uniform distribution on $[0, 1]$. The $K = 2$ latent failure times were taken to be conditionally independent with cause-specific hazard functions given by

$$\lambda_k(t|x, z) = x\alpha_k(t) + z\beta_k$$

where for $k = 1$ (cause 1), $\alpha_1(t) = 2.0$ for $0 \le t \le 0.5$ and 1.0 for $t > 0.5$ and $\beta_1 = 1.0$; for $k = 2$ (cause 2), $\alpha_2(t) = 1.0$ for $t > 0$ and $\beta_2 = 1.5$. The censoring C was taken to

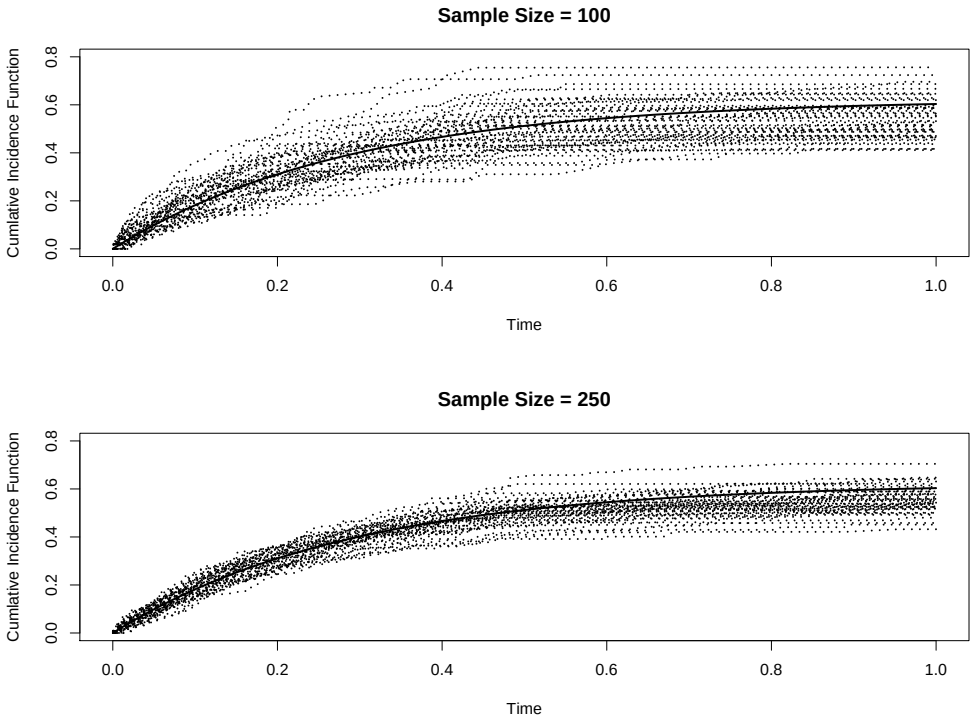

Figure 1. Cumulative incidence function of cause 1 (solid line) at $(x, z) = (1, .5)$ with 50 random estimates (dotted line) for sample sizes 100 and 250 with 30% censoring.

be exponentially distributed with the rate parameter adjusted, so that 30% and 50% of the observations were censored. We simulated data of sample sizes $n = 100$ and $n = 250$, each replicated 1000 times. In order to illustrate the performance of the proposed estimator of the cumulative incidence functions, in Figure 1 we present a plot of the estimates of cumulative incidence function for cause 1 for covariate value $x = 1$ and $z = .5$ under the above model based on 50 replicates of the sample. The estimates are based on the sample sizes 100 and 250 with 30% censoring, respectively. Figure 1 shows that the estimates tend to get closer to the true cumulative incidence function as the sample size increases.

Next, in order to examine the performance of the estimates of cumulative incidence functions

and their confidence intervals at time points t for specific value of covariate, we examined their performance for t in $.1, .2, .3, .4, .5$ and $.6$, where $.60$ is close to the 70-th percentile of the underlying distribution for the time to first cause, i.e., $k = 1$. We also investigated the performance of confidence band for $t \in [.1, .6]$. The numbers reported in Table I are the bias, sampling standard deviation, estimated asymptotic standard deviation of the predicted cumulative incidence function for the cause 1 for covariates $(x, z) = (1, .5)$ at the time points mentioned above. Table I also presents the coverage probability of the 95% confidence interval for the cumulative incidence function for cause-1 at the above mentioned time points as well as the coverage probability for its 95% confidence band in the last column. These results are based on 1000 replications for each of the sample sizes ($n = 100, 250$) and censoring percentages (30% and 50%).

The numbers indicate overall small biases of the proposed cumulative incidence functions and that their sampling standard deviations decrease as the sample sizes increase for each time point and each censoring level. The performance remains same when the censoring percentage increases from 30% to 50%. The estimated asymptotic standard deviations are very close to the sampling standard deviations indicating the appropriateness of the finite sample approximations to calculate the estimate for the asymptotic standard deviation. Furthermore, the coverage probabilities for 95% confidence intervals for the cumulative incidence function is good for $n = 100$ and gets very close to the nominal level of 95% as the sample size increases to $n = 250$ for both censoring percentages 30% and 50%, respectively. We next studied the performance of proposed estimators for comparing cumulative incidence functions given different values of the covariates. In Table II, we present the bias, sampling standard deviation, estimated asymptotic standard deviation and the coverage probability of the 95%

Table I. Estimation of cumulative incidence function for cause 1 at $(x, z) = (1, .5)$.

		time						
Censoring = 30%		0.1	0.20	0.3	0.40	0.50	0.60	[.1, .6]
$n = 100$	Bias	-0.0045	-0.0101	-0.0150	-0.0189	-.0229	-.0384	
	SD	.0524	.0658	.0685	.0707	0.0712	.0710	
	ESD	.0541	.0666	.0721	.0746	.0757	.0763	
	CP	93.8	93.6	94.8	94.8	94.2	92.8	93.1
$n = 250$	Bias	-0.0018	-0.0036	-0.0066	-0.0072	-0.0078	-0.0229	
	SD	0.0348	0.0421	0.0439	0.0447	0.0446	0.0444	
	ESD	0.0348	0.0428	0.0462	0.0477	0.0483	0.0486	
	CP	94.1	95.1	96.0	94.9	95.4	93.8	95.4
		time						
Censoring = 50%		0.1	0.20	0.3	0.40	0.50	0.60	[.1, .6]
$n = 100$	Bias	-.0036	-.0084	-.0127	-.0194	-.0244	-.0422	
	SD	.0562	.0682	.0746	.0765	.0764	.0762	
	ESD	.0562	.0713	.0785	.0819	.0833	.0840	
	CP	92.5	94.3	95.2	94.3	94.8	93.7	92.2
$n = 250$	Bias	-0.0014	-0.0040	-0.0069	-0.0079	-0.0095	-0.0248	
	SD	0.0349	0.0446	0.0481	0.0489	0.0504	0.0507	
	ESD	0.0362	0.0456	0.0500	0.0522	0.0531	0.0537	
	CP	94.9	94.6	95.4	95.5	95.2	93.4	94.4

Table II. Estimation of the difference between cumulative incidence functions for cause 1 at $(x_1, z_1) = (1, .5)$ and $(x_2, z_2) = (0, .5)$.

		time						
Censoring = 30%		0.1	0.20	0.3	0.40	0.50	0.60	[.1, .6]
$n = 100$	Bias	0.0041	0.0084	0.0109	0.0124	0.0147	0.0296	
	SD	0.0534	0.0660	0.0705	0.0727	0.0738	0.0759	
	ESD	0.0533	0.0651	0.0700	0.0725	0.0737	0.0745	
	CP	93.0	93.1	93.5	93.6	94.4	92.4	92.8
$n = 250$	Bias	-0.0020	-0.0033	-0.0023	-0.0006	0.0025	0.0181	
	SD	0.0362	0.0448	0.0498	0.0517	0.0530	0.0534	
	ESD	0.0348	0.0423	0.0452	0.0466	0.0474	0.0479	
	CP	94.8	93.4	93.8	94.2	93.8	93.6	94.6
		time						
Censoring = 50%		0.1	0.20	0.3	0.40	0.50	0.60	[.1, .6]
$n = 100$	Bias	0.0027	0.0102	0.0136	0.0185	0.0232	0.0378	
	SD	0.0553	0.0689	0.0773	0.0818	0.0831	0.0862	
	ESD	0.0548	0.0675	0.0735	0.0767	0.0787	0.0799	
	CP	92.8	92.3	92.4	92.4	92.7	91.4	92.6
$n = 250$	Bias	0.0039	0.0050	0.0062	0.0074	0.0089	0.0230	
	SD	0.0355	0.0438	0.0498	0.0516	0.0553	0.0581	
	ESD	0.0350	0.0434	0.0473	0.0494	0.0508	0.0517	
	CP	93.7	95.3	93.9	93.5	94.1	93.5	94.4

confidence interval for the difference between cumulative incidence functions for cause-1 at $(x_1, z_1) = (1, .5)$ with that at $(x_2, z_2) = (0, .5)$, as well as its confidence band for $t \in [.1, .6]$. Overall, the performances are similar to that found in Table I indicating that the proposed estimation procedure for constructing confidence intervals and bands for comparing cumulative incidence functions seems to perform very well even with the weight function being chosen to be identity. Our experience with choice of weight function based on λ indicates that the performance improves from the point of view of efficiency and consequently the coverage probability.

4. A real data application: Malignant melanoma study

We now illustrate our methodology with an analysis of data on 205 malignant melanoma patients observed during the years 1962 to 1977 [2]. Among these 205 patients, 57 patients died from malignant melanoma, 14 patients died from causes other than malignant melanoma, and the remaining 134 patients were alive at the end of follow-up. The covariates considered here were tumor thickness (mean: 2.92 mm; standard deviation: 2.96 mm), ulceration status (90 present and 115 not present), age (mean: 52 years; standard deviation: 17 years) and gender (79 male and 126 female). In our analysis, the covariates of tumor thickness and age were standardized. The main interest in this melanoma study is to predict the patient-specific cumulative incidence probability for the melanoma death.

We first began with fitting the [6] model for the cumulative incidence function by using the usual Cox's proportional hazards regression model to relate the cause-specific hazard function to the covariates for the malignant melanoma data. We used the score process test proposed by [7] to test for proportionality of the underlying covariates of interest. We found

that the proportionality assumption for age and sex were reasonable with p-values of 0.542 and 0.322, respectively. However, the p-values for the proportionality test for tumor thickness and ulceration status resulted in p-values of 0.022 and 0.004, respectively, indicating lack of fit. This indicates that the Cox's model may be inappropriate here.

Alternatively, the referee suggested using the approach laid out on pages 96-97 of [3]. Here, they propose the derived covariates approach to identify the time-varying effects. This approach requires knowing the shape of the time-varying effect, for instance, linearly changing with time $a + bt$ and also knowing the change point (if any) $a + bt$ for $t \leq t_0$ and $a + ct$ for $t > t_0$, where a, b, c are the regression coefficients. These may not be the appropriate shapes and one has to test for the shape being linear or knowing the change-point t_0 (i.e., when the shape changes). Alternatively, one can use the nonparametric approach, i.e., model the time-varying effect $\alpha_k(t)$ corresponding to a covariate x non-parametrically.

Next, we considered the popular alternative to the Cox model, the additive model as proposed by [9] and [8] to estimating the cumulative incidence function under the additive model. However, this model assume that covariates have time-constant effect only. We conducted goodness-of-fit tests of the additive model with the melanoma data using the score test procedures described below. We investigated the time-varying effect or constant effect of p-th covariate by conducting the following hypothesis test:

$$H_0 : \beta_p(t) = b,\ 0 \leq t \leq \tau$$

based on the semi-parametric additive model

$$\lambda(t) = x^T \beta(t) + z^T \gamma.$$

Defining the cumulative regression coefficient as $B_p(t) = \int_0^t \beta(u)du$, the above mentioned test

can be equivalently reformulated as

$$H_0 : B_p(t) = bt,\ 0 \leq t \leq \tau.$$

The following two tests were computed:

$$\text{Test 1} : \sup_{0 \leq t \leq \tau} |\hat{B}_p(t) - \hat{B}_p(\tau)\frac{t}{\tau}|$$

and

$$\text{Test 2} : \int_0^\tau \left(\hat{B}_p(t) - \hat{B}_p(\tau)\frac{t}{\tau}\right)^2 dt.$$

See [11] for details. The tests showed that tumor thickness does not have time-constant effect over time.

Test for time invariant effects: $H_0 : B(t) = bt$ based on Test 1 and Test 2

	Test 1		Test 2	
	Test Statistic	p-Value	Test Statistic	p-Value
(Intercept)	0.15100	0.473	2.90e-02	0.658
age	0.00398	0.331	5.18e-05	0.206
sex	0.12100	0.381	2.81e-02	0.364
thick	0.04600	0.004	5.62e-03	0.002
ulcer	0.13100	0.519	2.69e-02	0.612

Based on the above findings, we looked at more flexible model (2) which allows some covariates to have time-constant effect and others to have time-varying effect. Thus motivated by this, we have fitted the semiparametric additive hazard model proposed by [1] to the melanoma data, with the effects of age and sex as fixed and that of tumor thickness and ulceration to be time-varying. We also used the weight function introduced in Section 2 throughout this analysis as we found an improvement in efficiency with the usage of the weights in comparison to the estimates

based on weight matrix being identity. Based on this final model, the cumulative effects of the baseline estimate, tumor thickness, and ulceration, with 95% pointwise confidence intervals and bands, are given in Figure 2. The cumulative regression function for tumor thickness shows a positive effect in the first 4 years, which then seems to flatten out. Ulceration shows a positive effect within the first 4 years and a slower effect after that. The 95% confidence interval of the constant coefficients for gender and age are $(-.1021, .1450)$ and $(-.0535, .0691)$, respectively.

Next, we estimated the cumulative incidence function of malignant melanoma for a 52-year-old female patient with ulceration and tumor thickness of 6.76 mm (90-th percentile of tumor thickness), and display its 95% confidence intervals and bands under the semiparametric additive model in Figure 3. We used a transformation of $g(\cdot) = \log(-\log(\cdot))$.

Finally, in Figure 4 we estimate the difference between the cumulative incidence functions for malignant melanoma for different values of covariates. In Figure 4(a), in order to ascertain the effect of tumor thickness on the cumulative incidence function, we estimated the difference between cumulative incidence functions for a female of 52 years of age with ulceration and tumor thickness of one standard deviation above the observed mean tumor thickness with that for a female of same age with ulceration and tumor thickness at the mean value. In Figure 4(b), in order to ascertain the effect of ulceration on cumulative incidence function, we estimate the difference between cumulative incidence functions for a 52 year old female with mean value of tumor thickness and with ulceration with that for a 52 year old female with mean value of tumor thickness but no ulceration. In Figure 4(c), in order to ascertain the gender effect on cumulative incidence function, we estimated the difference between a 52 year old male with ulceration and mean tumor thickness with that of a 52 year old female with ulceration and mean tumor thickness. Observe that for a 52 years old female with ulceration,

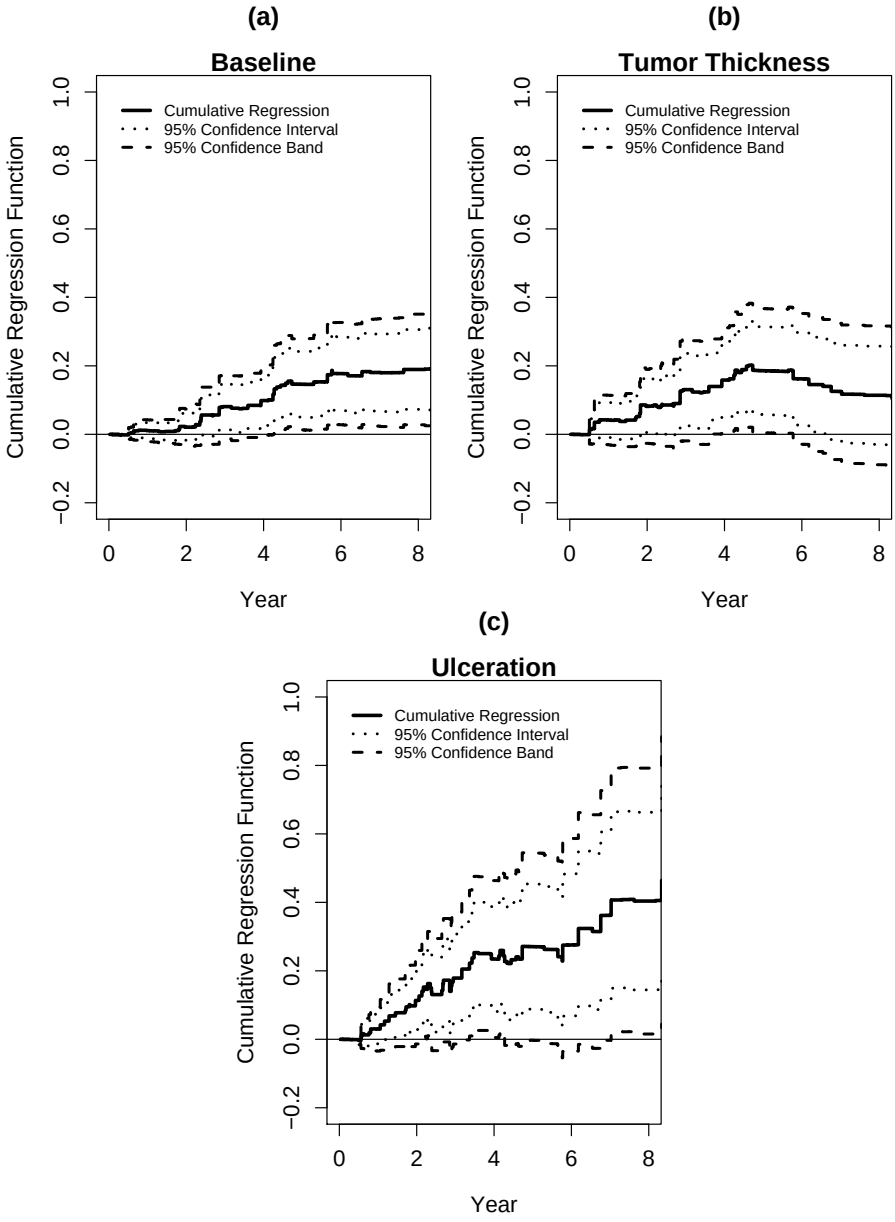

Figure 2. Cumulative time-varying effects for malignant melanoma death based on the semiparametric model with baseline, tumor thickness and ulceration as nonparametric part.

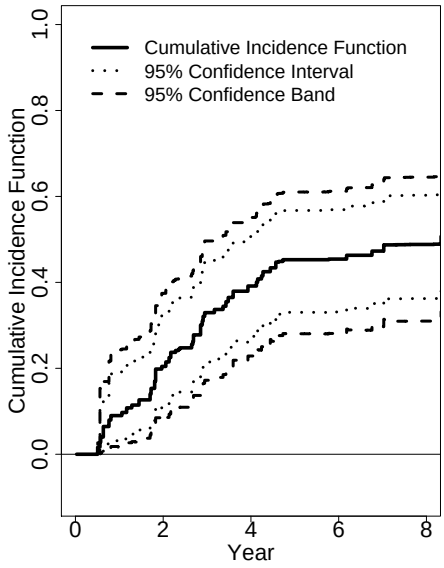

Figure 3. Predicted cumulative incidence function of malignant melanoma death for a 52-year-old female with ulceration and tumor thickness of 6.76 mm with 95% confidence intervals and bands.

Figure 4(a) estimates the additional probability of dying from malignant melanoma over time for those with tumor thickness of 5.88 mm (one standard deviation above the mean) compared to those with tumor thickness at 2.92 mm (the mean). For a 52 years old female with mean value of tumor thickness, Figure 4(b) shows the estimated increase in probability of dying due to malignant melanoma for those with ulceration compared to those without. Figure 4(c) indicates that men have higher probability of dying due to malignant melanoma than women with ulceration and with same mean level of tumor thickness.

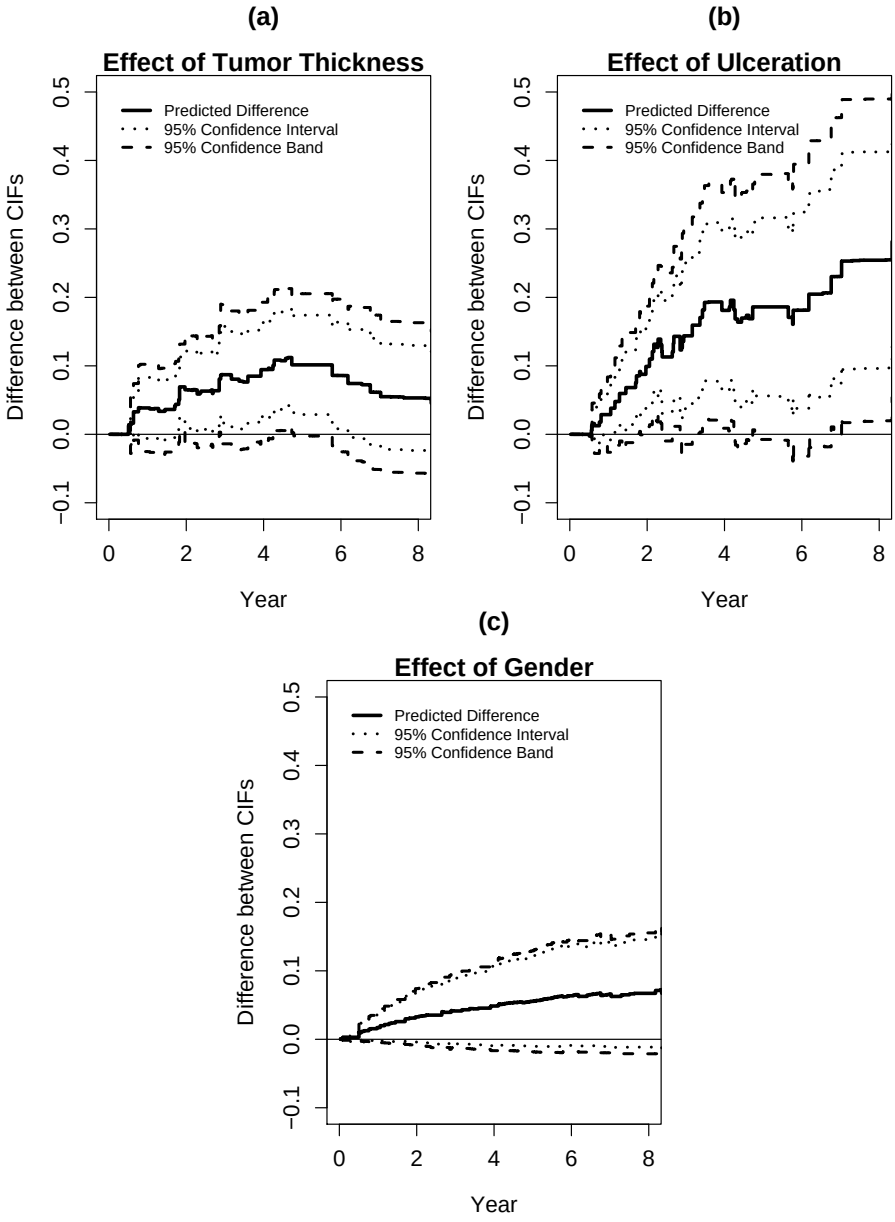

Figure 4. Predicted difference between two cumulative incidence functions of malignant melanoma death at three sets of covariate values specified in the last paragraph of Section 4.

5. Discussion

In this paper, we have proposed to use the semiparametric additive model for the prediction and comparison of cause-specific cumulative incidence functions for given values of the covariate. This model adds to the existing approaches for analyzing competing risks data which include multiplicative model of [6], Aalen's additive model (with only fixed effect of covariates) of [8] and Aalen's nonparametric additive model [10] among others. This gives a rich choice of modeling approaches which can all be applied when analyzing competing risks data, and one can then choose the best fit model. Most of the existing literatures study the cumulative incidence functions via modelling the cause-specific hazard functions. Direct modeling of cumulative incidence functions has recently been studied by [12]. This collection of models gives a rich variety from which a user can choose an appropriate model for analyzing the data. Under semiparametric additive hazard model, we developed statistical procedures to construct confidence intervals and bands for the difference and ratio of two cumulative incidence functions given the covariates. The proposed methods of this paper can be used to estimate the differential (or relative) probability of failure from one cause under two sets of covariate values over time. Unlike direct modeling of cumulative incidence functions, our approach is more flexible in that we do not need to assume simple forms for covariate effects on the cumulative incidence function. Thus, our approach can capture covariate effects that are nonlinear on the cumulative incidence function. Our R-package for the methods proposed is available upon request from the corresponding author and will be posted on a website.

ACKNOWLEDGEMENTS

The research of the first author was in part supported by the intramural research program of National

Institutes of Health and in part by NSF grants DMS-0304922 and DMS-0604576. The research of the second author was partially supported by NSF grant DMS-0604576 and NIH grant 2 RO1 AI054165-04. The research of third author was supported by the intramural research program of *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. The authorship is listed in alphabetical order.

REFERENCES

1. McKeague IW, Sasieni PD. A partly parametric additive risk model. *Biometrika*, 1994; **81**:501–514.
2. Andersen PK, Borgan Ø., Gill RD, Keiding N. *Statistical models based on counting processes*. Springer-verlag: New York, 1993.
3. Kalbfleisch JD, Prentice, RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
4. Cox DR. Regression models and life tables (with discussion). *Journal of royal statistical society, series –B*, 1972; **34**:187–220.
5. Andersen PK, Gill RD. Cox's regression model for counting processess: A large sample study. *Annals of Statistics*, 1982; **10**:1100–1120.
6. Cheng SC, Fine JP, Wei LJ. Prediction of cumulative incidence funciton under the proportional hazards model. *Biometrics*, 1998; **54**:219–228.
7. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 1993; **80**:557–572.
8. Shen Y, Cheng SC. Confidence Bands for Cumulative Incidence Curves under the Additive Risk Model. *Biometrics*, 1999; **55**:1093–1100.
9. Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika* 1994; **81**:61–71.
10. Aalen OO, Borgan O, Fekjaer H. Covariate adjustment of event histories estimated from Markov chains: The additive approach. *Biometrics*, 2001; **57**:993–1001.
11. Martinussen T, Scheike TH. *Dynamic Regression Models for Survival Data*. Springer: New York, 2006.
12. Scheike TH, Zhang MJ, Gerds T. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 2008;**95**: 1–16, DOI: 10.1093/biomet/asm096.

APPENDIX: Proof of Asymptotic Results

The following results state that $\sqrt{n}\left(\widehat{F}_1(t|x_0,z_0) - F_1(t|x_0,z_0)\right)$ is asymptotically equivalent to a sum of martingale $U_1(t|x_0,z_0)$, and it converges in distribution to a Gaussian process. Using arguments similar to those given by [1] it can be shown that

$$\sqrt{n}\left(\widehat{\beta}_k - \beta_k\right) = C_k^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{ki} + o_p(1),$$

and

$$\sqrt{n}\left(\widehat{A}_k(t) - A_k(t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Q_{ki}(t) + o_p(1),$$

where

$$Q_{ki}(t) = \int_0^t \left(n^{-1}W_k(u)\right)^{-1}\omega_{ki}(u)x_i\, dM_{ki}(u) - \int_0^t X_k^-(u)Z(u)\, du C_k^{-1}D_{ki}.$$

Note that M_{ki} is a martingale with respect to the counting process N_{ki}.

Define $W_k(t|x_0,z_0) = \sqrt{n}\left(\widehat{\Lambda}_k(t|x_0,z_0) - \Lambda_k(t|x_0,z_0)\right)$. Then the process $W_k(t|x_0,z_0)$ is asymptotically equivalent to $\widetilde{W}_k(t|x_0,z_0)$, since

$$\begin{aligned} W_k(t|x_0,z_0) &= \sqrt{n}\left(x_0^T\widehat{A}_k(t) + tz_0^T\widehat{\beta}_k - x_0^T A_k(t) - tz_0^T\beta_k\right) \\ &= x_0^T\left[\sqrt{n}\left(\widehat{A}_k(t) - A_k(t)\right)\right] + tz_0^T\left[\sqrt{n}\left(\widehat{\beta}_k - \beta_k\right)\right] \\ &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_0^T Q_{ki}(t) + tz_0^T C_k^{-1}D_{ki}\right) \equiv \widetilde{W}_k(t|x_0,z_0). \end{aligned}$$

Note that

$$\begin{aligned} \widehat{F}_1(t|x_0,z_0) - F_1(t|x_0,z_0) = &\int_0^t S(u|x_0,z_0)\, d\{\widehat{\Lambda}_1(u|x_0,z_0) - \Lambda_1(u|x_0,z_0)\} \\ &+ \int_0^t \{\widehat{S}(u|x_0,z_0) - S(u|x_0,z_0)\}\, d\widehat{\Lambda}_1(u|x_0,z_0). \end{aligned}$$

Using the Taylor expansion to the second component in the preceding expression,

$$\begin{aligned} &\int_0^t \{\widehat{S}(u|x_0,z_0) - S(u|x_0,z_0)\}\, d\widehat{\Lambda}_1(u|x_0,z_0) \\ &= -\int_0^t S(u|x_0,z_0)\left(\sum_{k=1}^{K}\widehat{\Lambda}_k(u|x_0,z_0) - \sum_{k=1}^{K}\Lambda_k(u|x_0,z_0)\right) d\widehat{\Lambda}_1(u|x_0,z_0). \end{aligned}$$

Then it follows that

$$\begin{aligned} &\sqrt{n}\left(\widehat{F}_1(t|x_0,z_0) - F_1(t|x_0,z_0)\right) \\ &\approx \int_0^t S(u|x_0,z_0)\, d\widetilde{W}_1(u|x_0,z_0) - \sum_{k=1}^{K}\int_0^t S(u|x_0,z_0)\widetilde{W}_k(u|x_0,z_0)\, d\Lambda_1(u|x_0,z_0). \end{aligned} \tag{7}$$

From integration by parts,

$$\begin{aligned}
&\int_0^t S(u|x_0,z_0)\widetilde{W}_k(u|x_0,z_0)\,d\Lambda_1(u|x_0,z_0)\\
&\quad=\int_0^t \widetilde{W}_k(u|x_0,z_0)S(u|x_0,z_0)\,d\Lambda_1(u|x_0,z_0)\\
&\quad=\int_0^t \widetilde{W}_k(u|x_0,z_0)dF_1(u|x_0,z_0)\\
&\quad=F_1(t|x_0,z_0)\widetilde{W}_k(t|x_0,z_0)-\int_0^t F_1(u|x_0,z_0)d\widetilde{W}_k(u|x_0,z_0)\\
&\quad=\int_0^t \{F_1(t|x_0,z_0)-F_1(u|x_0,z_0)\}\,d\widetilde{W}_k(u|x_0,z_0)\\
&\quad=\int_0^t F_1^C(t,u)\,d\widetilde{W}_k(u|x_0,z_0).
\end{aligned}$$

Since

$$\widetilde{W}_k(t|x_0,z_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_0^T Q_{ki}(t)+tz_0^T C_k^{-1}D_{ki}\right),$$

it follows that

$$\begin{aligned}
&\int_0^t S(u|x_0,z_0)\widetilde{W}_k(u|x_0,z_0)\,d\Lambda_1(u|x_0,z_0)\\
&\quad=\int_0^t F_1^C(t,u)\,d\widetilde{W}_k(u|x_0,z_0)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(\int_0^t F_1^C(t,u)x_0^T\left(n^{-1}W_k(u)\right)^{-1}\omega_{ki}(u)x_i\,dM_{ki}(u)\\
&\qquad\qquad-\int_0^t F_1^C(t,u)x_0^T X_k^-(u)^T Z(u)\,du C_k^{-1}D_{ki}\\
&\qquad\qquad+\int_0^t F_1^C(t,u)\,duz_0^T C_k^{-1}D_{ki}\Bigg).
\end{aligned}\tag{8}$$

Then from (7) and (8) $\sqrt{n}\left(\widehat{F}_1(t|x_0,z_0)-F_1(t|x_0,z_0)\right)$ can be approximated by martingale processes

$$U_1(t|x_0,z_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_{1i}(t|x_0,z_0),\tag{9}$$

where

$$\begin{aligned}
&\epsilon_{1i}(t|x_0,z_0)\\
&\quad=\int_0^t S(u|x_0,z_0)x_0^T\left(n^{-1}W_1(u)\right)^{-1}\omega_{1i}(u)x_i\,dM_{1i}(u)\\
&\qquad\quad-\int_0^t S(u|x_0,z_0)x_0^T X_1^-(u)Z(u)\,duC_1^{-1}D_{1i}\\
&\qquad\quad-\int_0^t S(u|x_0,z_0)\,duz_0^T C_1^{-1}D_{1i}\\
&\qquad-\sum_{k=1}^K\left(\int_0^t F_1^C(t,u)x_0^T\left(n^{-1}W_k(u)\right)^{-1}\omega_{ki}(u)x_i\,dM_{ki}(u)\right.\\
&\qquad\qquad\quad-\int_0^t F_1^C(t,u)x_0^T X_k^-(u)Z(u)\,duC_k^{-1}D_{ki}\\
&\qquad\qquad\quad\left.-\int_0^t F_1^C(t,u)\,duz_0^T C_k^{-1}D_{ki}\right).
\end{aligned}\tag{10}$$

Under our assumption that T_{ki} does not equal T_{li} for $k\neq l$, $k,l=1,\ldots,K$, N_{ki} and N_{li} cannot jump at the same time, and $\{M_{ki}(t):k=1,\ldots,K\}$ are orthogonal martingales. An application of martingale central limit theorem [2] implies that the processes $U_1(t|x_0,z_0)$ converges weakly to a zero-mean Gaussian process with variance that can be consistently estimated with $\widehat{\sigma}_1^2(t|x_0,z_0)$, where

$$\widehat{\sigma}_1^2(t|x_0,z_0)=\frac{1}{n}\sum_{i=1}^n\left(\widehat{\epsilon}_{1i}(t|x_0,z_0)\right)^2,\tag{11}$$

and $\widehat{\epsilon}_{1i}(t|x_0,z_0)$ by plugging in estimates of the unknown quantities.