Published in final edited form as:
Stat Methods Med Res. 2011 June ; 20(3): 261–274. doi:10.1177/0962280209347046.

Estimating the personal cure rate of cancer patients using population-based grouped cancer survival data

Binbing Yu, **Ram C. Tiwari***, and **Eric J. Feuer**
Laboratory of Epidemiology, Demography and Biometry National Institute on Aging, Bethesda, MD 20892

Office of Biostatistics, Center for Drug Evaluation and Research Food and Drug Adminstration, Silver Spring, MD 20993

Statistical Research and Applications Branch National Cancer Institute, Bethesda, MD 20892

Abstract

Cancer patients are subject to multiple competing risks of death and may die from causes other than the cancer diagnosed. The probability of not dying from the cancer diagnosed, which is one of the patients' main concerns, is sometimes called the "personal cure" rate. Two approaches of modeling competing-risk survival data, namely the cause-specific hazards approach and the mixture model approach, have been used to model competing-risk survival data. In this article, we first show the connection and differences between crude cause-specific survival in the presence of other causes and net survival in the absence of other causes. The mixture survival model is extended to population-based grouped survival data to estimate the personal cure rate. Using the colorectal cancer survival data from the Surveillance, Epidemiology and End Results (SEER) Program, we estimate the probabilities of dying from colorectal cancer, heart disease, and other causes by age at diagnosis, race and American Joint Committee on Cancer (AJCC) stage.

Keywords

Competing risks; grouped survival data; mixture model; personal cure; SEER Program

1 Introduction

Net survival, i.e., survival in the absence of other causes, is a measure of excess mortality due to cancer. This is a hypothetical measure of survival if all causes of death other than cancer of interest were to be eliminated. Net survival is a desirable measure to evaluate the progress of cancer treatment and control efforts since the interpretation as excess mortality due to cancer is not a ected by changes in mortality due to other diseases.1 However, net survival does not represent the actual survival patterns observed in a cohort of cancer patients. Comorbidity for cancer patients may limit treatment options and increase the risk of death from other causes. Usually comorbidity from competing causes increases with advancing age and is greater for patients in poor health. Thus, net survival may not be an ideal measure for assessing the impact of a cancer diagnosis in the presence of multiple competing risks. From another perspective, crude cause-specific probabilities of death provide measures of cause-specific mortality in the presence of causes of death in addition to cancer and reflect mortality patterns actually observed among patients.2 Thus, they are

*The research of Dr. Tiwari was conducted at National Cancer Institute and the view in the paper is solely his own and does not reflect that of the Food and Drug Administration.

appropriate measures when the focus is on inference and comparison of cause-specific failures under a variety of conditions. Such probabilities can be used to weigh the risks and benefits of various treatment options, particularly for patients diagnosed at older ages when comorbidity is high.

There has been considerable progress against cancer due to improvements in treatment, and the dissemination of early diagnosis and screening. Thus, successfully treated cancer patients may die from a cause other than the diagnosed cancer, which is called “personal cure.” The corresponding proportion of dying from causes other than the diagnosed cancer is defined as the personal cure rate. Gordon3 originally applied the mixture model to estimate the personal cure rate using breast cancer data from a clinical trial. To assess mortality from cancer at the population level, we extend the mixture model for competing-risk survival to population-based cancer survival data in order to calculate crude probabilities of dying from cancer and other competing risks. The rest of the paper is organized as follows: In Section 2, we first describe survival data with competing risks and review the mixture model for continuous survival data. Next, we extend the mixture model to grouped survival data, describe the estimation method and discuss the connection between net survival and crude survival. In Section 3, we apply the mixture model to colorectal cancer survival data to calculate the probabilities of dying from three competing causes of death. We discuss the potential use and limitations of the mixture model in the Discussion section.

2 The mixture model for grouped survival data with competing risks

We consider a patient subject to K mutually exclusive competing risks of death and assume that the primary outcome is a random pair (D, T), where D takes a value from the set $\{1, 2, ..., K\}$ indicating cause of death and T is a non-negative random variable representing time to death. Let z denote a vector of covariates. The cause-specific hazard rate, defined as the probability of dying from cause k alone in $[t, t + dt)$ in the presence of all acting risks, given $T \geq t$,4 is given by

$$h_k(t/z) = \lim_{dt \to 0} \frac{P(t \leq T < t+dt, D=k | T \geq t, z)}{dt}, \quad k=1, \ldots, K.$$

Let $h(t|z) = \sum_{k=1}^{K} h_k(t|z)$ and $S_k(t|z) = \exp\left\{-\int_0^t h_k(u|z)\, du\right\}$. The survival function for T is

$$S_T(t|z) = P(T>t) = \exp\left\{-\int_0^t h(u|z)\, du\right\} = \prod_{k=1}^{K} S_k(t|z).$$

David and Moeschberger5 consider the observed death time T as the minimum of K latent death times, i.e., $T = \min(T_1, \cdots, T_K)$. When the competing risks $T_1, \ldots, T_K$, are independent, $S_k(t|z)$ can be interpreted as the net survival function from cause k in the absence of other causes and $h_k(t|z)$ is the net hazard.6 Net survival can be estimated by the Kaplan-Meier method or the actuarial method by treating the other causes of deaths as censored.

The crude cumulative probability or cumulative incidence function (CIF) of dying from cause k in the presence of other causes is

$$F_k^*(t|z) = P(T \leq t, D=k|z) = \int_0^t h_k(u|z)\, S(u|z)\, du. \quad (1)$$

The functions $F_k^*(t|z)$ is also called sub-distribution function for cause k, $k = 1, ..., K$. Let $\pi_k(z) = P(D = k|z)$ be the probability of dying from cause k and $Q_k(t|z) = P(T > t|D = k, z)$ be the conditional crude survival function. Then $F_k^*(t) = \pi_k(z)\{1 - Q_k(t|z)\}$.

2.1 Review of mixture model for competing risk data

The mixture model7 specifies the death probabilities $\pi_k(z)$ and the crude survival $P(T > t|D = k, z)$, $k = 1, ..., K$. It does not require the independence assumption among competing risks. The cause of death D follows a multinomial distribution with probabilities $P_1(z), ..., P_K(z)$ with

$$\pi_k(z) = P(D=k|z) = \frac{\exp\left(\mu_k + \gamma_k^T z\right)}{\sum_{l=1}^{K} \exp\left(\mu_l + \gamma_l^T z\right)}, \tag{2}$$

where μ_k is a scalar constant and γ_k is a vector of regression coefficients. Because $\sum_k P_k(z) = 1$, we set $\mu_K = 0$ and $\gamma_K = 0$ for identifiability purpose. We also assume cause $D = 1$ is death due to cancer of interest. The personal cure rate is then calculated as $1 - P_1(z)$.

The functions $Q_k(t|z) = P(T > t|D = k, z)$, $k = 1, ..., K$ are called crude cause-specific survival functions.1 Larson and Dinse7 use a proportional hazards (PH) model for $Q_k(t|z)$:

$$Q_k(t|z) = \exp\left\{-\int_0^t q_k(u|z)\,du\right\}, \tag{3}$$

where $q_k(t|z) = q_k(t)\exp\left(\beta_k^T z\right)$, $q_k(t)$ is the baseline hazard function and β_k is a vector of regression coefficients. The overall survival function can also be expressed as

$$S_T(t|z) = \sum_{k=1}^{K} \pi_k(z)\, Q_k(t|z). \tag{4}$$

It can be shown that $S_k(t|z) \geq Q_k(t|z)$ (see Appendix I) for all t unless $P_j(z) = 0$ for $j \neq k$. The inequality implies that, under the independent competing-risks assumption, the net survival function in the absence of other causes is always greater than the crude cause-specific survival function in the presence of other causes. Hence, the Kaplan-Meier and actuarial estimates always overestimate the conditional crude cause-specific survival probability.

2.2 The mixture model for grouped survival data

Survival times from population-based cancer registries are usually grouped into annual or monthly intervals, $I_j = (t_{j-1}, t_j]$ for $j = 1, ..., J$, where $t_0 = 0$ and $t_J = \tau$ denote the beginning and end of follow-up, respectively. For the cohort with covariates z, let n_{jz} be the number of people alive at the beginning of interval I_j, d_{kjz} be the number of people who die from cause k, $k = 1, ..., K$ and let l_{jz} be the number of people lost to follow-up in the interval. For simplicity of notation, we omit the subscript z and denote the observed data as $\mathcal{D} = \left(n_j, l_j, d_{kj}, k = 1, \ldots, K; j = 1, \ldots, J\right)$. The total number of people who die during interval I_j is $d_j = \sum_{k=1}^{K} d_{kj}$.

The probability of dying from cause k during the interval I_j is

$$P\left(t_{j-1}<T \le t_j | D=k, z\right)=Q_k\left(t_{j-1}|z\right)-Q_k\left(t_j|z\right). \tag{5}$$

Because some people are lost to follow-up during the interval, a widely used technique is to adjust the person-years at risk as $n_j^*=n_j-0.5*l_j$,[6,8,9] and the resulting estimate is called the actuarial estimate. Gail6 showed that the actuarial estimate is a good approximation of the maximum likelihood estimate (MLE) of $S(t)$ under the assumption that time when lost to follow-up and time of death from competing risks are independent. The actuarial estimate can also be justified by assuming that time when lost to follow-up is uniform in interval I_j. Then the number of people who are censored at time t_j, i.e., $T > t_j$, is

$$c_j=\begin{cases} n_j^*-d_j-n_{j+1}^* & \text{when} \quad j<J, \\ n_j^*-d_j & \text{when} \quad j=J. \end{cases}$$

Let $\theta_0 = (\mu_k, \gamma_k, k = 1, ..., K)$ be the parameters in the logistic model (2) for the cause of death and let θ_k be the parameters for the crude survival functions $Q_k(t)$, $k = 1, ..., K$. The likelihood function for observed competing-risk survival data $\mathcal{D}$ is

$$L\left(\theta|\mathcal{D}\right)=\prod_{j=1}^{J}\left[S\left(t_j|z\right)^{c_j}\prod_{k=1}^{K}\left\{\pi_k\left(z\right)\left(Q_k\left(t_{j-1}|z\right)-Q_k\left(t_j|z\right)\right)\right\}^{d_{kj}}\right],$$

where $\theta = (\theta_0, \theta_1, ..., \theta_K)$. As the Newton-Raphson method requires the calculation of a complex Hessian matrix, the Expectation-Maximization (EM) algorithm is used to find the MLEs of θ.

The complete data are $(n_j^*, d_{kj}, c_{kj}, k = 1, ..., K, j = 1, ..., J)$, where c_{kj} is the number of people censored at time t_j who would ultimately die from cause k. Using Equation (4), the loglikelihood for the complete data is $\log L\left(\theta\right)=\ell\left(\pi\right)+\sum_{k=1}^{K}\ell_k\left(Q\right)$, where

$$\begin{aligned} \ell\left(\pi\right) &= \sum_{k=1}^{K}\sum_{j=1}^{J}\left(d_{kj}+c_{kj}\right)\log\pi_k\left(z\right) \\ \ell_k\left(Q\right) &= \sum_{j=1}^{J}\left[d_{kj}\log\left\{Q_k\left(t_{j-1}|z\right)-Q_k\left(t_j|z\right)\right\}+c_{kj}\log Q_k\left(t_j|z\right)\right] \end{aligned}$$

The E-step assigns the censored observations, i.e., the people who are lost to follow-up or who are still alive at the end of the study, into one of the K causes of death according to their conditional probabilities $P(D = k|z, t > t_j)$. The expected number of deaths due to causes k in interval I_j for those censored people is

$$c_{kj}=c_jP\left(D=k|z,T>t_j\right)=c_j\frac{\pi_k\left(z\right)Q_k\left(t_j|z\right)}{S_T\left(t_j|z\right)}. \tag{6}$$

The M-step involves maximizing the loglikelihood functions $\ell\left(\pi\right)$ and $\ell_k\left(Q\right)$, $k = 1, ..., K$.

The MLEs of the parameters in $\pi_k(z)$ can be obtained by multinomial logistic regression, and the estimation of the parameters in $Q_k(t|z)$ depends on the model specifications for $Q_k(t|z)$. The popular models include the Weibull and Gompertz models and the semi-parametric proportional hazards model. For arbitrary interval-censored survival data, various methods are proposed by Finkelstein10, Pan and Chappell11 and Goetghebeur and Ryan12. For grouped survival data, we follow Prentice and Gloeckler13 and write the loglikelihood $\ell_k(Q)$ as

$$\ell(Q_k) = \sum_{j=1}^{J} \left[d_{kj} \log \left\{ 1 - p_{kj}(z) \right\} + \left(r_{kj} - d_{kj} \right) \log p_{kj}(z) \right]$$

where $p_{kj}(z) = Q_k(t_j|z)/Q_k(t_{j-1}|z)$ and $r_{kj} = \sum_{l=j}^{J} (d_{kl} + c_{kl})$. Let $\alpha_{kj} = \log \left\{ \int_{t_{j-1}}^{t_j} h_k(u)\, du \right\}$, $j = 1, \ldots, J$. Then,

$$p_{kj}(z) = \exp \left\{ -\exp \left(\alpha_{kj} + \beta_k z \right) \right\}. \tag{7}$$

The estimates of $\theta_k = (\alpha_{k1}, \ldots, \alpha_{kJ}, \beta_k)$ can be obtained by SAS PROC LOGISTIC.

Several factors complicate variance estimation for the parameter estimates. First the dimension of parameters is large for the semiparametric model. Second variance estimates do not come from the EM algorithm as a byproduct. Several approaches to calculate the observed information matrix in an EM context have been proposed.14,15 But, these approaches involve tedious algebra and are analytically intractable. Another variance estimator, which is simple to compute and also turns out to have good small sample properties, is based on multiple imputation.12,16 First the expected numbers of deaths due to cause k in interval I_j, c_{kj}, $j = 1, \ldots, J$, are imputed M times using a multinomial distribution with conditional probabilities given in (6), after the final step of the EM algorithm is completed. Then for each imputed "complete" data set, a point estimate $\widehat{\theta}(m)$ and a variance estimate v_m of θ are calculated, $m = 1, \ldots, M$. Let $\bar{\theta} = \sum_{m=1}^{M} \widehat{\theta}_{(m)}/M$. The variance estimate for the MLE $\widehat{\theta}$ is given by

$$V\left(\widehat{\theta}\right) = (1 + 1/M) \frac{\sum_{m=1}^{M} \left(\widehat{\theta}_{(m)} - \bar{\theta} \right)^2}{M - 1} + \sum_{m=1}^{M} v_m/M. \tag{8}$$

This is a weighted sum of the within-imputation variance and the between-imputation variance.

2.3 Estimating net survival from the mixture model

The direct output from the mixture model consists of death probabilities from different causes and the conditional crude survival function. Under the assumption of independent competing risks, the net survival $S_k(t|z)$ can be derived from output from the mixture model. We assume that $h_k(t|z) = w_{kj}(z)h(t|z)$ for $t \in I_j$, where the weights $w_{kj}(z)$ are constants for each interval I_j and covariate z. From (1), we have

$$F_k^*\left(t_j\right) - F_k^*\left(t_{j-1}\right) = \int_{t_{j-1}}^{t_j} h_k\left(u|z\right) S_T\left(u|z\right) du = w_{kj}\left(z\right) \int_{t_{j-1}}^{t_j} h\left(u|z\right) S_T\left(u|z\right) du$$
$$= w_{kj}\left(z\right)\left\{S_T\left(t_{j-1}|z\right) - S_T\left(t_j|z\right)\right\},$$

hence

$$w_{kj}\left(z\right) = \frac{F_k^*\left(t_j|z\right) - F_k^*\left(t_{j-1}|z\right)}{S_T\left(t_{j-1}|z\right) - S_T\left(t_j|z\right)}. \tag{9}$$

Because

$$\frac{S_k\left(t_j|z\right)}{S_k\left(t_{j-1}|z\right)} = \exp\left\{-\int_{t_{j-1}}^{t_j} h_k\left(u|z\right) du\right\} = \left\{\frac{S_T\left(t_j|z\right)}{S_T\left(t_{j-1}|z\right)}\right\}^{w_{kj}(z)},$$

the net survival function $S_k(t_j|z)$ can be estimated by

$$S_k\left(t_j|z\right) = \prod_{l=1}^{j}\left\{\frac{S_T\left(t_l|z\right)}{S_T\left(t_{l-1}|z\right)}\right\}^{w_{kl}(z)}, \tag{10}$$

where w_{kl} is given by (9). The variance estimates for crude survival $Q_k(t|z)$ and net survival $S_k(t_j|z)$ are given in Appendix II. The relationship between $S_k(t|z)$ and $S_T(t|z)$ and $F_k^*(t|z)$ implies that the net survival functions can be calculated as a by-product of the mixture survival model under the independent competing risk assumption.

3 Application

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute is an authoritative source of information on cancer incidence and survival in the United States (http://www.seer.cancer.gov). SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 26 percent of the US population. SEER coverage includes 23 percent of African Americans, 40 percent of Hispanics, 42 percent of American Indians and Alaska Natives, 53 percent of Asians, and 70 percent of Hawaiian/Pacific Islanders. The SEER Program began collecting data on cancer cases in 1973 in the states of Connecticut, Iowa, New Mexico, Utah, and Hawaii and the metropolitan areas of Detroit and San Francisco-Oakland. In 1974-1975, the metropolitan area of Atlanta and the 13-county Seattle-Puget Sound area were added. These original 9 regions are referred to as the SEER 9 registries, covering 10% of the US population.

Colorectal cancer is the third most common cancer and the third leading cause of cancer-related mortality in the United States. Over the past decade, colorectal cancer incidence and mortality rates have modestly decreased or remained level. The most recent estimates from the American Cancer Society show that there are 106,100 new cases of colon cancer, 40,870 new cases of rectal cancer and 49,920 deaths from colorectal cancer for the year 2009. If diagnosed early and treated successfully, colorectal cancer can be cured. In fact, many

colorectal cancer patients live long enough to die ultimately from other causes, most commonly from heart disease.

Because of the long history of the SEER 9 registries, we use the colorectal cancer survival data for illustration. We consider three competing risks ($K = 3$), namely, colorectal cancer death ($D = 1$), heart disease death ($D = 2$) and death due to other causes ($D = 3$). Here, we use the mixture model to estimate probabilities of dying from different causes given a patient's age at diagnosis, race and the American Joint Committee on Cancer (AJCC) stage. There are remarkable differences between racial and ethnic groups in both incidence and mortality. The mortality rate from colorectal cancer for African Americans is higher than that for whites (American Cancer Society, 2008). To confirm this conclusion, we also test the difference in probabilities of dying from multiple causes between whites and African Americans.

The SEER data we consider consist of 199,715 colorectal cancer cases diagnosed from 1975 to 2002. The end of followup is December, 2003 and the maximum followup time is 28 years. The survival data are then stratified by single age (50, 51,...,99, 100+), race (white, black) and AJCC stage (I, II, III, IV). The average age at diagnosis is 64, and 92% are whites and 8% are African Americans. The percentages of cancer stages I-IV are 20%, 33%, 25% and 22%, respectively.

In the first analysis, we fit separate mixture models for each combination of race and AJCC stage in order to estimate the probabilities of dying from different causes. Age at diagnosis is used as a covariate in equations (2) and (7). To account for the possible quadratic effect of age, the square of age is also included as a covariate. Figure 1 plots the observed and modeled probabilities of dying from colorectal cancer with respect to age at diagnosis. The observed probabilities are calculated as the proportion dying from colorectal cancer for each single age group $\tilde{\pi}_1 = \sum_{j=1}^{J} d_{1j} / \sum_{j=1}^{J} d_j$. The modeled probabilities are the estimates $\widehat{\pi}_1$ in Equation (2). For example, less than 40% of the patients diagnosed with Stage I colorectal cancer actually die from colorectal cancer. We see that the modeled probabilities fit the observed probabilities reasonably well. The probabilities of colorectal cancer death increase with AJCC stage. This makes sense as a diagnosis of more advanced colorectal cancer implies higher probability of death. Overall, the probability of colorectal cancer death decreases in older patients except for very old ages. Figure 1 also shows that the probabilities of dying from colorectal cancer are much lower for whites than blacks in the same cancer stage.

Figure 2 plots the nonparametric and modeled cumulative probabilities of death due to colorectal cancer within 5 years of cancer diagnosis by age at diagnosis. The nonparametric estimate17 is calculated as

$$\tilde{F}_1^*(t|z) = \sum_{t_j \le t} \frac{d_{1j}}{n_j} \tilde{S}\left(t_{j-1}|z\right),$$

where $\tilde{S}\left(t_{j-1}|z\right) = \prod_{l=1}^{j-1} \left(1 - d_l / n_l^*\right)$ is the actuarial survival. The modeled probabilities are the estimates of $\widehat{F}_1^*(t|z)$. We see that the 5-year probabilities of colorectal cancer death do not change much with age at diagnosis for stage I, and the corresponding probabilities increase slightly with age at diagnosis for Stage III and IV. This shows that patients with more severe diagnosis are more likely to die from cancer.

One may be interested in the actual survival pattern after diagnosis. As an example, we show the nonparametric and modeled conditional crude cause-specific survival estimates in the presence of competing risks for patients diagnosed at age 70 in Figure 3. We see that white patients have slightly higher crude survival rates than blacks. Figure 3 also shows that for stage IV cancer the conditional cause-specific survival rates are all less than 5% after 5 years of diagnosis. This implies that for patients diagnosed with stage IV cancer at age 70, 95% of cancer deaths would occur within 5 years after diagnosis.

The unconditional cumulative probability of death due to cancer versus other causes is useful to describe the experience of individual patients. For example, Figure 4 shows the cumulative probabilities of death due to three causes for white patients diagnosed with stage I cancer at age 70. The personal cure rate is about 90%. The cumulative colorectal cancer death probability levels off after about 10 years from diagnosis, but the probability of death due to other causes still increases. So if a patient does not die from colorectal cancer within 10 years, it is very likely that he will die from another cause.

The analysis above can be used to describe the survival patterns experienced by cancer patients. In contrast, a statistical model presents a great advantage when some form of inference is required. As we see from Figures 1 and 2, differences exist in survival patterns between racial and ethnic groups. In the second analysis, we perform a formal test to examine the effect of race on probabilities of dying from different causes for each cancer stage. The covariate z of interest is the indicator of being black. In logistic model (2), the parameters $\exp(\gamma_1)$ and $\exp(\gamma_2)$ represent the odds ratios of being blacks on the probabilities of dying from cancer and heart disease, respectively. For example, $\exp(\gamma_1) > 1$ means that blacks have higher probability of dying from colorectal cancer than whites. The estimates of odds ratio and the 95% confidence intervals (CI) are shown in Table 1. For stages I and II, blacks have a significantly higher risk of dying from colorectal cancer than whites. For Stage III, whites and blacks have similar cause-specific probabilities. For Stage IV, blacks have a lower probability of dying from cancer, but a higher probability of dying from heart disease.

By using a logistic model for probabilities of dying from different causes, the mixture model implicitly assumes that hazards from different causes are constant after the end of the follow-up. Usually, hazards due to causes other than cancer will increase remarkably as people get older, while the hazard due to cancer death may remain similar with respect to age. Misspecification of the logistic model might yield biased estimates. It is also necessary to have a sufficiently long follow-up time to observe most deaths and their corresponding causes, so that death probabilities can be modeled reliably.

As shown in Figures 1 and 2, the mixture model provides a reasonably good fit to the observed data. However, one can argue that most of the patterns and probabilities can be easily obtained by smoothing the raw data. For example, a multinomial logistic model with splines can be used to estimate π_k, the probabilities of dying from different causes. Here, we are trying to model the probabilities π_k and crude survival functions $Q_k(t)$ simultaneously. This provides a complete picture of survival patterns after cancer diagnosis.

4 Discussion

In this article, we apply the mixture model to grouped survival data with competing risks from population-based cancer registries. This model can be used to estimate probabilities of death due to different competing causes. The personal cure rate is helpful to describe the survival experience after cancer diagnosis. This model can also be applied to data from clinical trials. For example, one can compare the probabilities of different types of failures

to evaluate the risk and benefit of two treatment options. Physicians can determine the appropriate treatments for cancer patients based on their comorbidities and prognosis.

An alternative approach to competing-risk data is to model the net survival functions[5,13]. To ensure identifiability, the competing risks are assumed to be independent. The cancer patients, especially in their old ages, have higher comorbidity problems than the general cancer-free population. The mixture model assumes that the process of loss to follow-up is independent of the competing risks of death, but it does not require independence among the K competing risks. Another advantage of using the mixture model is that the net survival function can be derived as a by-product.

Acknowledgments

The research of Dr. Yu was carried out in part at the Information Management Services, Inc. and was supported in part by the contract with the National Cancer Institute and by the Intramural Research Program of the National Institute on Aging.

Appendix I. Proof of Sk(t|z)≥Qk(t|z)

Based on equation (3), we have $q_k(t|z) = -\frac{dQ_k(t|z)}{dt}/Q_k(t|z)$. Because $Q_k(t|z) = 1 - F_k^*(t|z)/\pi_k(z)$ and from (1),

$$q_k(t|z) = \frac{\frac{dF_k^*(t|z)}{dt}}{\pi_k(z)\,Q_k(t|z)} = \frac{h_k(t|z)\,S_T(t|z)}{\pi_k(z)\,Q_k(t|z)} \geq h_k(t|z).$$

Thus

$$S_k(t|z) = \exp\left(-\int_0^t h_k(r|z)\,dr\right) \geq \exp\left(-\int_0^t q_k(r|z)\,dr\right) = Q_k(t|z).$$

Appendix II. Variances for Qk(t|z) and Sk(t|z)

The variances of $Q_k(t|z)$ and $S_k(t|z)$ can be derived from the covariance matrix of θ in (8) using the delta method. Let $\theta_k = (\alpha_{k1}, ..., \alpha_{kJ}, \beta_k)$ be the parameters in $\theta_k(t|z)$ and $V(\theta_k)$ be the covariance matrix of θ_k. Because

$$\log Q_k\left(t_j|z\right) = \sum_{l=1}^{j} \log p_{kl}(z) = -\sum_{l=1}^{j} \exp\left(\alpha_{kj} + \beta_k z\right),$$

the variance of $Q_k(t_j|z)$ can be calculated as

$$\widehat{Var}\left[Q_k\left(t_j|z\right)\right] = \frac{1}{Q_k\left(t_j|z\right)^2}\,\Gamma_{kj}\,V(\theta_k)\,\Gamma_{kj}'\Big|_{\theta=\hat{\theta}},$$

where $\gamma_{kj} = (\log p_{k1}(z), ..., \log p_{kj}(z), z \log Q_k(t_j|z))$.

Let $S_k^*(t|z) = \pi_k(z)\,Q_k(t|z)$. Then $S_T(t|z) = \sum_{k=1}^{K} S_k^*(t|z)$. Assuming the parameters $(\theta_k, \mu_k, \gamma_k)$, $k = 1, ..., K$, are functionally independent, we have

$$\frac{\partial S_T(t|z)}{\partial \theta_k} = \frac{\partial S_k^*(t|z)}{\partial \theta_k}, k=1,\ldots,K, \tag{11}$$

where the partial derivatives of $S_k^*(t_j|z)$ are:

$$\begin{aligned}
\frac{\partial S_k^*(t_j|z)}{\partial \mu_k} &= \frac{Q_k(t_j|z)}{\pi_k(z)\{1-\pi_k(z)\}},\\
\frac{\partial S_k^*(t_j|z)}{\partial \gamma_k} &= \frac{zQ_k(t_j|z)}{\pi_k(z)\{1-\pi_k(z)\}},\\
\frac{\partial S_k^*(t_j|z)}{\partial \alpha_{kl}} &= \pi_k(z)\,Q_k(t_j|z)\,\log p_{kl}(z), l=1,\ldots,J\\
\frac{\partial S_k^*(t_j|z)}{\partial \beta_k} &= z\pi_k(z)\,Q_k(t_j|z)\,\log Q_k(t_j|z).
\end{aligned}$$

The variance of $S_T(t_j|z)$ can be calculated as:

$$\widehat{Var}\left[\widehat{S}_T(t_j|z)\right] = \left(\frac{\partial S_T(t_j|z)}{\partial \theta}\right) V(\theta) \left(\frac{\partial S_T(t_j|z)}{\partial \theta}\right)'\Big|_{\theta=\widehat{\theta}},$$

where $V(\theta)$ is the variance of $\widehat{\theta}$.

The variance of $S_k(t_j|z)$ is given by $\widehat{Var}\left[\widehat{S}_k(t_j|z)\right] = \widehat{Var}\left[\log S_k(t_j|z)\right] / S_k(t_j|z)^2$, where

$$\widehat{Var}\left[\log \widehat{S}_k(t_j|z)\right] = \left(\frac{\partial \log S_k(t_j|z)}{\partial \theta}\right) V(\theta) \left(\frac{\partial \log S_k(t_j|z)}{\partial \theta}\right)'\Big|_{\theta=\widehat{\theta}}.$$

Based on Equation (10), $\log S_k(t_j|z) = \sum_{l=1}^{j} w_{kl}(z)\left[\log S_T(t_l|z) - \log S_T(t_{l-1}|z)\right]$. We have

$$\begin{aligned}
\frac{\partial \log S_k(t_j|z)}{\partial \theta} = \sum_{l=1}^{j} \Big\{ &\frac{\partial w_{kl}(z)}{\partial \theta}\left[\log S_T(t_l|z) - \log S_T(t_{l-1}|z)\right]\\
&+ w_{kl}(z)\left[\frac{1}{S_T(t_l|z)}\frac{\partial S_T(t_l|z)}{\partial \theta} - \frac{1}{S_T(t_{l-1}|z)}\frac{\partial S_T(t_{l-1}|z)}{\partial \theta}\right]\Big\},
\end{aligned}$$

where $\frac{\partial S_T(t_j|z)}{\partial \theta}$ is given in (11) and

$$\frac{\partial w_{kj}(z)}{\partial \theta} = \frac{\frac{\partial S_k^*(t_{j-1}|z)}{\partial \theta} - \frac{\partial S_k^*(t_j|z)}{\partial \theta}}{S_T(t_{j-1}|z) - S_T(t_j|z)} - \frac{S_k^*(t_{j-1}|z) - S_k^*(t_j|z)}{\left[S_T(t_{j-1}|z) - S_T(t_j|z)\right]^2}\left[\frac{\partial S_T(t_{j-1}|z)}{\partial \theta} - \frac{\partial S_T(t_j|z)}{\partial \theta}\right].$$

Note that $\frac{\partial S_k^*(t_{j-1}|z)}{\partial \theta_i} = 0$ for $i \neq k$.

References

1. Cronin A, Feuer EJ. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. Statistics in Medicine. 2000; 19:1729–1740. [PubMed: 10861774]
2. Shairer C, Mink PJ, Carroll L, Devesa SS. Probabilities of death from breast cancer and other causes among female breast cancer patients. Journal of the National Cancer Institute. 2004; 96:1311–1321. [PubMed: 15339969]
3. Gordon NH. Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer patients. Statistics in Medicine. 1990; 9:397–407. [PubMed: 2194263]
4. Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2nd ed.. Wiley; New York: 2002.
5. David, HA.; Moeschberger, ML. The Theory of Competing Risks. Griffin; London: 1978.
6. Gail M. A review and critique of some models used in competing risk analysis. Biometrics. 1975; 31:209–222. [PubMed: 1164533]
7. Larson MG, Dinse GE. A mixture model for the regression analysis of competing risks data. Applied Statistics. 1985; 34:201–211.
8. Cutler SJ, Ederer F. Maximum utilization of the life table method in analyzing survival. Journal of Chronical Disease. 1958; 8:699–712.
9. Chiang CL. Competing risks in mortality analysis. Annual Review of Public Health. 1991; 12:281–307.
10. Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics. 1986; 42:845–854. [PubMed: 3814726]
11. Pan W, Chappell R. Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. Biometrics. 2002; 58:64–70. [PubMed: 11890328]
12. Goetghebeur E, Ryan L. Semiparametric regression analysis of interval-censored data. Biometrics. 2000; 56:1139–1144. [PubMed: 11129472]
13. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. Biometrics. 1978; 34:57–67. [PubMed: 630037]
14. Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B. 1982; 44:226–233.
15. Meng XL, Rubin DB. Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. Journal of the American Statistical Association. 1991; 86:899–909.
16. Rubin DB, Schenker N. Multiple imputation in health-care data bases: An overview and some applications. Statistics in Medicine. 1991; 10:585–598. [PubMed: 2057657]
17. Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR, Straus DJ, Clarkson BD, Brennan MF. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. Journal of the American Statistical Association. 1993; 88:400–409.

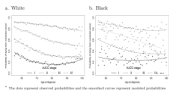

Figure 1.
Observed and modeled probabilities of dying from colorectal cancer by age at diagnosis *

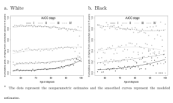

Figure 2.
Nonparametric and modeled cumulative probabilities of death due to colorectal cancer within 5 years of diagnosis *

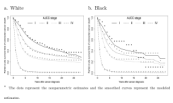

Figure 3.
Conditional crude survival estimates $Q_1(t)$ of colorectal cancer death in the presence of competing risks for patients diagnosed at age 70 *

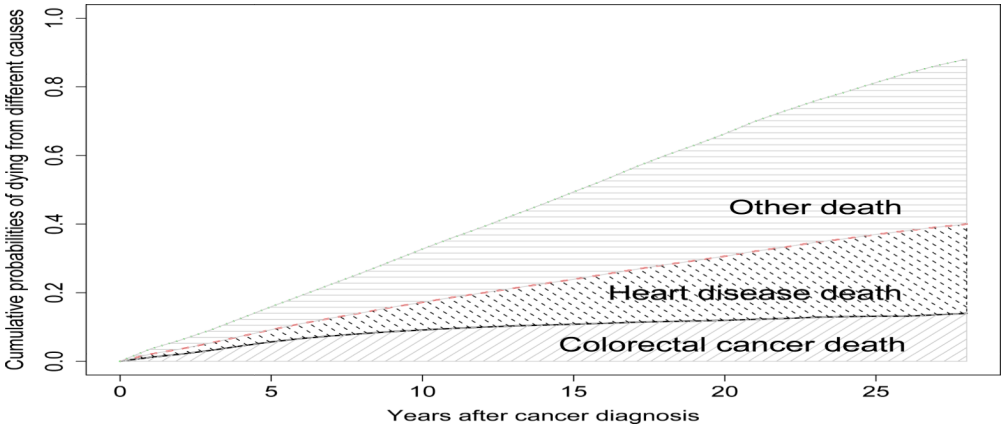

Figure 4.
Cumulative cause-specific probabilities of death for white patients diagnosed at age 70 with Stage I colorectal cancer

Table 1

Odds ratios of race on the probabilities of dying from different causes

Stage	Parameter	Cause of death	Estimate	95% CI	*p*-value
I	$\exp(\gamma_1)$	Cancer	1.322	(1.187, 1.472)	<0.001
	$\exp(\gamma_2)$	Heart Disease	1.202	(1.047, 1.381)	0.009
II	$\exp(\gamma_1)$	Cancer	1.229	(1.138, 1.328)	<0.001
	$\exp(\gamma_2)$	Heart Disease	1.046	(0.954, 1.147)	0.335
III	$\exp(\gamma_1)$	Cancer	1.060	(0.968, 1.161)	0.212
	$\exp(\gamma_2)$	Heart Disease	1.017	(0.896, 1.154)	0.797
IV	$\exp(\gamma_1)$	Cancer	0.807	(0.729, 0.894)	<0.001
	$\exp(\gamma_2)$	Heart Disease	1.223	(1.037, 1.441)	0.017