Published in final edited form as:

Lifetime Data Anal. 2018 July ; 24(3): 385–406. doi:10.1007/s10985-017-9407-2.

A Joint Model of Cancer Incidence, Metastasis, and Mortality

Qui Tran,

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.

Kelley M. Kidwell,

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.

Alex Tsodikov

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.

Abstract

Many diseases, especially cancer, are not static, but can be summarized as a series of events or stages (e.g. diagnosis, remission, recurrence, metastasis, death). Most available methods to analyze multi-stage type of data ignore intermediate events and focus on the terminal one or consider (time to) multiple events as independent. Competing-risk or semi-competing-risk models often fall short of adequate description of the complex relationship between disease progression events driven by the shared progression process. A multistage model only looks at two stages at a time and fails to capture the effect of one stage on the time spent between other stages. Moreover, most models do not account for latent stages. Our semi-parametric joint model of diagnosis and latent metastasis events leading to cancer death uses nonparametric maximum likelihood to estimate covariate effects on the risks of intermediate events and death and the dependence between them. We illustrate with Monte Carlo simulation and analysis of prostate cancer from SEER database.

Keywords

Disease natural history; Marked endpoints; Competing risks; Semiparametric regression; Survival analysis

1 Introduction

In a natural cancer progression, patients may be diagnosed with local disease and achieve remission for some time after treatment, potentially progress, develop metastasis and succumb to the cancer. Others may be diagnosed at a later stage and show metastasis at

---

quitran@umich.edu.

⁶Software

R code and sample input data is available on request from tsodikov@umich.edu.

⁷Supplementary Materials

Supplementary Materials are available online at <http://biostatistics.oxfordjournals.org>.

initial diagnosis. For the purpose of this paper, we focus on cancer diagnosis, metastasis, and death as a sequence of cancer events (Figure 1). Our motivating example is prostate cancer data from the Surveillance, Epidemiology, and End Results (SEER) registry. The cases file in SEER follows prostate cancer patients from diagnosis to death or end of study. SEER provides metastasis status at time of diagnosis (i.e. whether cancer had metastasized prior to patients being diagnosed), but not the exact time-of-metastasis-onset. Because SEER cases file only include subject diagnosed with cancer, we make use of a second SEER data that provide the population counts, matched by race and age, from which the cases file originated. Although it is more straightforward to only focus on the cases file, we want to make use of the available population data to (1) account for people who were not diagnosed of cancer and (2) gain information regarding the latent metastasis process which occurs before diagnosis.

As noted, the onset of metastasis is unobserved (latent) before the disease reaches a measureable size; therefore if metastasis is presented at the time of cancer diagnosis, the onset must have occurred at some prior time point. For prostate cancer, most deaths occur after cancer spread to other organs, thus our study tailor to setting where death-due-to-cancer can only occur as a result of metastasis (Figure 1). We assume that if a patient was first diagnosed without metastasis, metastasis occurs at some subsequent point prior to death. This tailoring does not limit the general applicability of our method since it can easily be removed by adding an additional log-likelihood term for the possibility of death without metastasis (i.e. connecting “diagnosis without metastasis” and “death” in Figure 1). All deaths not due to prostate cancer are censoring events.

As a result of the multi-stage nature of cancer, there are many issues that make conventional survival analysis with a single time-to-event outcome not appropriate. First, the occurrence of certain stages may prevent others from being observed (e.g. metastasis may occur prior to cancer diagnosis, preventing observance of time-to-diagnosis-without-metastasis). This censoring due to other events should be accounted for when analyzing each time-to-event out-come. Second, part of the natural disease progression may be latent and thus cannot be observed (e.g. metastasis onset). Third, the dependence between time-to-observable-events (e.g. diagnosis and death) and the latent disease leading up to them (e.g. metastasis) make it difficult, and potentially misleading, to separately model the outcomes. The treatment effect has multiple causal paths, one of them being the effect on the risk of death through the latent metastatic process.

In cancers where death can occur without metastasis (from local disease in a vital organ), death and metastasis are semi-competing risks (Fine et al 2001) and this area has received considerable attention. Because of the non-identifiability issue from competing risks, a full nonparametric solution is not available (Peng and Fine 2007). Many have combined the semi-parametric approach with a copula model (Nelsen 1999) to account for dependent competing risks. Clayton copula (Clayton 1978) and bivariate survival models (Oakes 1989) are popular choices for this approach. Recently, Chen (2012) proposed using a semi-parametric transformational model for the marginal distribution and a copula model for the joint distribution of time-to each-event. In our setting, since death due to prostate cancer can

only occur following metastasis but not vice versa, these events represent recurrent events that are partially unobserved, rather than semi-competing risks.

Multi-stage models have been a popular tool (Andersen et al 1991; Xu et al 2010; Kalbfleisch and Prentice 2011; Hougaard and Hougaard 2000) whose simplest non-trivial example is the illness-death model (Andersen and Keiding 2002). It models the transition intensities between any two stages of the disease process. However, the limitation of these transition intensities is that they only consider two stages at a time rather than a joint progression process. This is usually referred to as the “Markovian” assumption (Andersen et al 1991). There are recent developments to handle situations where the Markovian assumption is invalid including a nonparametric estimation of the transition intensities (Meira-Machado et al 2006), the use of Gamma frailty (Xu et al 2010), and a branch of shared-frailty effect (Nielsen et al 1992; Hougaard and Hougaard 2000; Liu et al 2004; Govindarajulu et al 2011).

To address the issue of latent stage, Hu and Tsodikov (2014b) proposed a semi-parametric regression model that is based on Markov modulated processes with non-parametric time transformation models in the setting of metastasis - recurrence paradigm where they model time-to-recurrence as a marked- end-point. We extended Hu and Tsodikov’s work to model the cancer process with more than two stages that can be latent or observed, terminal and non-terminal. Using non-parametric maximum likelihood (NPMLE) (Tsodikov 2003; Zeng and Lin 2006; Chen 2009), we provide an estimating procedure for the effect of treatment and other covariates on time-to-metastasis, time-to-cancer-diagnosis and time-to-death, while taking into account the dependence structure between them.

In our example for real data analysis, death due to other causes were treated as censors. This presents the issue of competing risks. From literature review, there are two common approaches to address this issue: Switch to cumulative incidence functions (Fine and Gray 1999), or use copulas to explain the dependence between potential competing risks. The former is not ideal because inference would depend on censoring, which is unpleasant to keep track because they would change with different population and cause distribution. On the other hand, because of the non-identifiability issue, defining a model based on copulas would depend on strong assumptions; thus we decided to do a sensitivity analysis instead to access how the estimate would be effected if the risks were indeed dependent.

In Section 2, we provide notation and model description (2.1), data structure (2.2), counting processes (2.3), joint likelihood (2.4), score functions and estimating procedure (2.5), and asymptotic properties (2.6). In Section 3, we use Monte Carlo simulations to verify the model’s performance and its robustness against dependence censoring. In Section 4, we present an application to prostate cancer using SEER data to demonstrate the model and its ability to handle issues in multi-stage time-to-event data.

2 Methods

2.1 Notation and Model

Let T_U be the unobservable time-to-metastasis-onset T_1^* and T_2^* be the potentially (due to censoring) observable times of diagnosis and death, respectively. For clarity, we denote $\mathbf{Z}_1$ as the sets of covariates that affect all stages, while $\mathbf{Z}_2$ refers to covariates that start after diagnosis (e.g. for SEER analysis, $\mathbf{Z}_1$ includes racial categories and $\mathbf{Z}_2$ includes treatments). We recognized the potential pitfall of including internal covariates (Kalbfleisch and Prentice 2011) in $\mathbf{Z}_2$. However, this is not the case for SEER because the treatment choices are not manifestation of any biological process within the patient, thus is plausible to be treated as external. Let $\mathbf{Z} = (\mathbf{Z}_1 \cup \mathbf{Z}_2)$ be the covariate vector.

Hu and Tsodikov (2014b) provide mechanistic justification for use of common baseline hazard across hazards to different stages as a surrogate of the same underlying disease pattern driving the time to diagnosis, metastasis and death. The assumption can be understood as a single nonparametrically specified time scale for all disease-related events.

We relax this assumption of common hazard by creating additional parameters δ_1 and δ_2 for the nonparametric baseline cumulative hazard function $H(t)$ (instantaneous hazard $h(t)$). Denote cumulative hazard for metastasis, diagnosis, and death respectively as $H(t)$, $H(t)^{\delta_1}$, and $H(t)^{\delta_2}$. Thus, the analogous instantaneous hazards are $h(t)$, $\delta_1 H(t)^{\delta_1-1} h(t)$, and $\delta_2 H(t)^{\delta_2-1} h(t)$. This parameterization can flexibly account for the various stretching and shrinking in baseline hazards for different cancer events, while saving us the computational burden of estimating multiple baseline hazards nonparametrically. Using $I(\cdot)$ as an indicator function and a series of Cox proportional hazards models, the hazards for different cancer events are parameterized as follows: (1) The time-to-metastasis-onset

$$\lambda_U(t|Z) = h(t) \left(\eta I(t < T_1) + \tilde{\eta} I(t \geq T_1) \right); \quad (2) \text{ the time-to-diagnosis}$$

$$\lambda_1(t|Z, T_U) = \delta_1 H(t)^{\delta_1-1} h(t) \theta \mu^{I(t \geq T_U)}; \quad (3) \text{ the time-to-death}$$

$$\lambda_2(t|Z, T_U, T_1) = \delta_2 H(t)^{\delta_2-1} h(t) \gamma I(t \geq \max\{T_U, T_1\}), \text{ where } \eta = e^{\beta_{\eta_1} Z_1}, \tilde{\eta} = e^{\beta_{\eta_1} Z_1 + \beta_{\eta_2} Z_2},$$

$$\theta = e^{\beta_{\theta 0} + \beta_{\theta 1} Z_1}, \mu = e^{\beta_{\mu 0} + \beta_{\mu 1} Z_1}, \text{ and } \gamma = e^{\beta_{\gamma 0} + \beta_{\gamma 1} Z_1}.$$

Let $\beta = [\beta_{\eta}, \beta_{\theta}, \beta_{\mu}, \beta_{\gamma}, \delta_1, \delta_2]$ be the coefficients vector. Let

$$S_x(t|\cdot) = e^{-\int_0^t \lambda_x(s|\cdot) ds} \text{ and } f_x(t|\cdot) = \lambda_x(t|\cdot) S_x(t|\cdot) \text{ be the corresponding conditional survival and density functions, respectively, for hazard functions } \lambda_x(t|\cdot).$$

2.2 Observed Data Structure

Let V^* be the censoring time, independent of (T_1^*, T_2^*) given $\mathbf{Z}$; and ξ be the maximum follow-up time of the study. For each subject $i = 1, 2, \dots, n$ in the study, we observe $\{T_{1i}, T_{2i}, \delta_{1i}^{(L)}, \delta_{1i}^{(D)}, \delta_{2i}^{(L)}, \delta_{2i}^{(D)}, Z_i\}$, where $T_{1i} = \min(T_1^*, V_i^*, \xi)$ is the time-to-first-observable-

event (diagnosis or censoring); $T_{2i} = \min(T_{2i}^*, V_i^*, \xi)$ is the time-to-second-observable-event (death or censoring); $\delta_{1i}^{(L)} = I(T_{1i} = T_{1i}^*, T_{Ui} > T_{1i})$ is an indicator of diagnosis without metastasis; $\delta_{1i}^{(D)} = I(T_{1i} = T_{1i}^*, T_{Ui} \leq T_{1i})$ is an indicator of diagnosis with metastasis; $\delta_{2i}^{(L)} = I(T_{2i} = T_{2i}^*)\delta_{1i}^{(L)}$ is an indicator of observing death after diagnosis without metastasis (i.e., metastasis occurs some time between diagnosis and death); and $\delta_{2i}^{(D)} = I(T_{2i} = T_{2i}^*)\delta_{1i}^{(D)}$ is an indicator of observing death after diagnosis with metastasis.

2.3 Counting Process Notation

In our motivating example, SEER registry only include subject-level data after cancer diagnosis. Thus, we use a separate SEER population count data to track the risk-set size at each time point where our cancer cases file occur. This essentially provides a life-table type of for cancer diagnosis event, following by subject-level data for death-due-to-cancer.

From the population data, let $Y_{z_1}(t)$ denote the population counts for subgroup that takes value z_1 for covariate set Z_1 at time t . From the cases file data, denote $N_{z_1}(t) = \sum_{i: Z_{1i} = z_1} I(T_{1i}^* < t)$ the counting process for number of diagnoses from subgroup z_1 . After diagnosis, we have subject-level processes: Denote $N_{2i}(t) = I(T_{2i}^* < t)N_{1i}(t)$ and $Y_{2i}(t) = I(T_{1i} < t < T_{2i})$ the counting process and at-risk process for death event in subject i , respectively.

Note that both $N_{2i}(t)$ and $Y_{2i}(t)$ can be further broken down into two types (death after metastasis-then-diagnosis or diagnosis-then-metastasis).

2.4 Likelihood Construction

We integrate out time-to-metastasis-onset, T_U , by the appropriate time interval since the exact time is not observable. Denote $f_x f_y f_z(t_x, t_y, t_z)$ the joint probability of the sequence of event x , y , and z occurring at time t_x , t_y , t_z respectively. For ease of notation, we shorten $H(t_x)$ to H_x henceforth. The contribution of each subject to the likelihood falls into one of the following five scenarios (see Section 1 of Supplementary Materials for derivation):

1. Subject is diagnosed without metastasis at t_1 , then metastasizes, and dies at t_2 :

$$\begin{aligned} \mathcal{L}_1^* &= \int_{t_1}^{t_2} f_U f_1 f_2(t_u, t_1, t_2) dt_u = \\ &\delta_1 \delta_2 H_1^{\delta_1 - 1} H_2^{\delta_2 - 1} h_1 h_2 \tilde{\eta} \theta \gamma e^{-H_1(\eta - \tilde{\eta}) - H_1^{\delta_1} \theta - H_2^{\delta_2} \gamma} \int_{t_1}^{t_2} h_u e^{-H_u \tilde{\eta} + H_u^{\delta_2} \gamma} du. \end{aligned}$$

This joint probability can be written as $\mathcal{L}_1^* = \mathcal{L}_1 \cdot \tilde{\mathcal{L}}_1$, where $\mathcal{L}_1$ is the probability of observing up to time-of-diagnosis:

$$\mathcal{L}_1 = S_U f_1(t_1, t_1) = \delta_1 H_1^{\delta_1 - 1} h_1 \theta e^{-H_1 \eta - H_1^{\delta_1} \theta}.$$

2. Subject is diagnosed without metastasis at t_1 , may or may not have metastasis and is censored at t_2 :

$$\begin{aligned} \mathcal{L}_2^* &= \int_{t_1}^{t_2} f_U f_1 S_2(t_u, t_1, t_2) dt_u + \int_{t_2}^{\infty} f_U f_1 S_2(t_u, t_1, t_2) dt_u \\ &= \delta_1 H_1^{\delta_1 - 1} h_1 \theta e^{-H_1(\eta - \tilde{\eta}) - H_1^{\delta_1} \theta} \left[\int_{t_1}^{t_2} \tilde{\eta} h_u e^{-H_u \tilde{\eta} + H_u^{\delta_2} \gamma - H_2^{\delta_2} \gamma} du + e^{-H_2 \tilde{\eta}} \right]. \end{aligned}$$

Alternatively, this can be specified as $\mathcal{L}_2^* = \mathcal{L}_1 \cdot \widetilde{\mathcal{L}}_2$.

3. Subject is diagnosed with metastasis at t_1 and dies at t_2 :

$$\begin{aligned} \mathcal{L}_3^* &= \int_0^{t_1} f_U f_1 f_2(t_u, t_1, t_2) dt_u \\ &= \delta_1 \delta_2 H_1^{\delta_1 - 1} H_2^{\delta_2 - 1} h_1 h_2 \eta \theta \mu \gamma e^{-H_1^{\delta_1} \theta \mu + H_1^{\delta_2} \gamma - H_2^{\delta_2} \gamma} \int_0^{t_1} h_u e^{-H_u \eta - H_u^{\delta_1} \theta \bar{\mu}} du. \end{aligned}$$

Alternatively, this can be specified as $\mathcal{L}_3^* = \mathcal{L}_3 \cdot \widetilde{\mathcal{L}}_3$, where

$$\mathcal{L}_3 = \int_0^{t_1} f_U f_1(t_u, t_1) dt_u = \delta_1 H_1^{\delta_1 - 1} h_1 \eta \theta \mu e^{-H_1^{\delta_1} \theta \mu} \int_0^{t_1} h_u e^{-H_u \eta - H_u^{\delta_1} \theta \bar{\mu}} du.$$

4. Subject is diagnosed with metastasis at t_1 and is censored at t_2 :

$$\begin{aligned} \mathcal{L}_4^* &= \int_0^{t_1} f_U f_1 S_2(t_u, t_1, t_2) dt_u \\ &= \delta_1 H_1^{\delta_1 - 1} h_1 \eta \theta \mu e^{-H_1^{\delta_1} \theta \mu + H_1^{\delta_2} \gamma - H_2^{\delta_2} \gamma} \int_0^{t_1} h_u e^{-H_u \eta - H_u^{\delta_1} \theta \bar{\mu}} du. \end{aligned}$$

Alternatively, this can be specified as $\mathcal{L}_4^* = \mathcal{L}_3 \cdot \widetilde{\mathcal{L}}_4$.

5. Subject is censored at t_1 before any event is observed:

$$\begin{aligned}\mathcal{L}_5 &= \int_0^{t_1} f_U S_1 S_2(t_u, t_1, t_1) dt_u + S_U S_1 S_2(t_1, t_1, t_1) \\ &= \eta e^{-H_1^1 \theta \mu} \int_0^{t_1} \frac{-H_u \eta - H_u^1 \theta \bar{\mu}}{h_u e} du + e^{-H_1 \eta - H_1^1 \theta}.\end{aligned}$$

Thus, the contribution of subject i to the log-likelihood is:

$$\begin{aligned}\ell_i &= \delta_{1i}^{(L)} \delta_{2i}^{(L)} \log(\mathcal{L}_{1i}^*) + \delta_{1i}^{(L)} (1 - \delta_{2i}^{(L)}) \log(\mathcal{L}_{2i}^*) + \delta_{1i}^{(D)} \delta_{2i}^{(D)} \log(\mathcal{L}_{3i}^*) + \quad (1) \\ &\delta_{1i}^{(D)} (1 - \delta_{2i}^{(D)}) \log(\mathcal{L}_{4i}^*) + (1 - \delta_{1i}^{(L)} - \delta_{1i}^{(D)}) \log(\mathcal{L}_{5i}) \\ &= \left\{ \delta_{1i}^{(L)} \log(\mathcal{L}_{1i}) + \delta_{1i}^{(D)} \log(\mathcal{L}_{3i}) + (1 - \delta_{1i}^{(L)} - \delta_{1i}^{(D)}) \log(\mathcal{L}_{5i}) \right\} + \\ &\left\{ \delta_{1i}^{(L)} \left[\delta_{2i}^{(L)} \log(\widetilde{\mathcal{L}}_{1i}) + (1 - \delta_{2i}^{(L)}) \log(\widetilde{\mathcal{L}}_{2i}) \right] \right\} + \\ &\left\{ \delta_{1i}^{(D)} \left[\delta_{2i}^{(D)} \log(\widetilde{\mathcal{L}}_{3i}) + (1 - \delta_{2i}^{(D)}) \log(\widetilde{\mathcal{L}}_{4i}) \right] \right\}.\end{aligned}$$

$$\begin{aligned}&= \int_0^\xi [\log \Theta_{1i}(H_t, \beta) + \log dH_t] dN_{1i}(t) - Y_{1i}(t) \Theta_{1i}(H_t, \beta) dH_t + \quad (2) \\ &\int_0^\xi [\log \Theta_{2i}(H_t, \beta, t_1) + \log dH_t] dN_{2i}(t) - Y_{2i}(t) \Theta_{2i}(H_t, \beta, t_1) dH_t.\end{aligned}$$

If we denote $\ell_{1i}, \ell_{2i}^{(L)}, \ell_{2i}^{(D)}$ as the quantities in each line of equation (1), it is easy to see that ℓ_{1i} is the contribution from diagnosis at t_1 , while $\ell_{2i}^{(L)}$ and $\ell_{2i}^{(D)}$ are the contribution from the subsequent time segment between diagnosis and death.

This is re-written using counting processes in equation (2), where $\Theta_{1i}(H_t, \beta) dH_t$ is the hazard of subject i being diagnosed at time t and $\Theta_{2i}(H_t, \beta, t_1) dH_t$ is the hazard of subject i dying of cancer, given diagnosis at time t_1 with status (Met- or Met+). These quantities can be derived easily as follows:

$$\theta_1(H_t, \beta) = \begin{cases} \frac{\mathcal{L}_1}{\mathcal{L}_5^{h_1}} = \frac{\delta_1 H_t^{\delta_1 - 1} \theta}{1 + \eta \int_0^t h_u e^{\eta(H_t - H_u) + \theta \bar{\mu} \left(H_t^{\delta_1} - H_u^{\delta_1} \right) du}}, & \text{if Met-} \\ \frac{\mathcal{L}_3}{\mathcal{L}_5^{h_1}} = \frac{\delta_1 H_t^{\delta_1 - 1} \eta \theta \mu \int_0^t h_u e^{\eta(H_t - H_u) + \theta \bar{\mu} \left(H_t^{\delta_1} - H_u^{\delta_1} \right) du}}{1 + \eta \int_0^t h_u e^{\eta(H_t - H_u) + \theta \bar{\mu} \left(H_t^{\delta_1} - H_u^{\delta_1} \right) du}}, & \text{if Met +} \end{cases}$$

$$\theta_2(H_t, \beta, t_1) = \begin{cases} \frac{\mathcal{L}_1^*}{\mathcal{L}_2^{*h_2}} = \frac{\delta_2 H_t^{\delta_2 - 1} \tilde{\eta} \gamma \int_{t_1}^t h_u e^{\tilde{\eta}(H_t - H_u) - \gamma \left(H_t^{\delta_2} - H_u^{\delta_2} \right) du}}{1 + \tilde{\eta} \int_{t_1}^t h_u e^{\tilde{\eta}(H_t - H_u) - \gamma \left(H_t^{\delta_2} - H_u^{\delta_2} \right) du}}, & \text{if Met-} \\ \frac{\mathcal{L}_3^*}{\mathcal{L}_4^{*h_2}} = \delta_2 H_t^{\delta_2 - 1} \gamma, & \text{if Met +} \end{cases}$$

(see Section 2 of Supplementary Materials for details of derivation).

For ease of notation in the rest of this paper, we shorten

$\theta_{z_1}[H_t, z_1, \beta]$ and $\theta_2[H_t, \beta, t_1]$ to $\theta_{z_1}(t)$ and $\theta_2(t)$ respectively.

As discussed in section 2.3 and motivated by the structure of SEER database, we do not have subject-level information before cancer diagnosis. Thus, we utilized an additional SEER dataset of population count to construct life-table type of data for event of diagnosis. Let Z_1 denotes set of covariates available at population level before diagnosis. We stratified all subjects by Z_1 and set $\theta_{z_1}(t) = \theta_{1i}(t)$ where $Z_{1i} = z_1$. Once diagnosis occurs, SEER follows individual subject; thus we have notation i for subject-level data. The joint log-likelihood in equation (2) is rewritten as follows:

$$\ell = \int_0^\xi \sum_{z_1 \in Z_1} \left[\log \theta_{z_1}(t) + \log dH_t \right] dN_{z_1}(t) - Y_{z_1}(t) \theta_{z_1}(t) dH_t + \quad (3)$$

$$\sum_{i=1}^n \left[\log \theta_{2i}(t) + \log dH_t \right] dN_{2i}(t) - Y_{2i}(t) \theta_{2i}(t) dH_t.$$

2.5 Nonparametric maximum likelihood estimator (NPMLE)

2.5.1 Score Functions—Denote partial derivatives

$\theta_{\star, H}(t) = \frac{\partial \theta_{\star}(t)}{\partial dH(x)}$ and $\theta_{\star, \beta}(t) = \frac{\partial \theta_{\star}(t)}{\partial \beta}$ (see Section 3 of Supplementary Materials for close form expressions.)

Then the score function with respect to $H(x)$ is

$$U_{H(x)} = \frac{\partial \ell}{\partial dH(x)} = \sum_{z_1 \in Z_1} \left\{ \left[\frac{dN_{z_1}(x)}{dH(x)} - Y_{z_1}(x) \Theta_{z_1}(x) \right] + \int_x^\xi \frac{\Theta_{z_1, H}(t)}{\Theta_{z_1}(t)} dM_{z_1}(t) \right\} \quad (4)$$

$$+ \sum_{i=1}^n \left\{ \left[\frac{dN_{2i}(x)}{dH(x)} - Y_{2i}(x) \Theta_{2i}(x) \right] + \int_x^\xi \frac{\Theta_{2i, H}(t)}{\Theta_{2i}(t)} dM_{2i}(t) \right\}$$

2.5.2 Estimation Procedure—We set equations (4) to zero and solve for NPMLE of the jump in the baseline hazard $d\hat{H}(x)$ and the covariates effects $\hat{\beta}$. In particular,

$$d\hat{H}(x) = \frac{\sum_{z_1 \in Z_1} dN_{z_1}(x) + \sum_{i=1}^n dN_{2i}(x)}{\sum_{z_1 \in Z_1} Y_{z_1}(x) \omega_{z_1}[\bar{H}(\xi), x, \beta] \Theta_{z_1}(x) + \sum_{i=1}^n Y_{2i}(x) \omega_{2i}[\bar{H}(\xi), x, \beta, t_{1i}] \Theta_{2i}(x)}, \quad (5)$$

Where $\omega_{\star}[\bar{H}(\xi), x, \beta] = 1 - \frac{\int_x^\xi \frac{\Theta_{\star, H}(t)}{\Theta_{\star}(t)} dM_{\star}(t)}{\Theta_{\star}(x)}$, constructed using the following martingales

$$dM_{z_1}(t) = dN_{z_1}(t) - Y_{z_1}(t) \Theta_{z_1}(H_t, z_1, \beta) dH_t, \text{ for } z_1 \in Z_1$$

$$dM_{2i}(t) = dN_{2i}(t) - Y_{2i}(t) \Theta_{2i}(H_t, \beta, t_{1i}) dH_t, \text{ for } i = 1, 2, \dots, n.$$

Maximization of the log-likelihood with respect to both H and β simultaneously is done via an iterative reweighting algorithm (Chen 2009) as follows:

1. For a candidate $\hat{\beta}^*$, start with weight $\omega_{\star}(0) = 1$ for all m, k, i and $d\hat{H}(t)_{(0)}$ the Nelson-Aalen estimator of baseline hazard.
2. Repeat until convergence of $\hat{H}(t)$: At iteration s , fix weights $w_{(s)}$ and calculate the jump in baseline hazard at time t , $d\hat{H}(t)_{(s+1)}$, using equation (5). Repeatedly solving this equation across t where events/censors happen gives the full path of $\hat{H}_{(s+1)}$. Use $\hat{H}_{(s+1)}$ to update $w_{(s+1)}$.
3. Calculate the log-likelihood given in equation (3) using $\hat{\beta}^*$ and the converged $\hat{H}(t)$.
4. The final $\hat{\beta}$ maximizes the log-likelihood.

2.6 Asymptotic properties

The asymptotic properties of $\hat{\beta}$ are derived using Bagdonavicius and Nikulin (1999). The details and notation are shown in Section 4 of Supplementary Materials.

Under suitable regularity conditions, we have as $n \rightarrow \infty$,

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \Sigma_*^{-1} \Sigma^* (\Sigma_*^{-1})'), \quad (6)$$

where Σ^* and Σ_* can be consistently estimated by

$$\begin{aligned} \hat{\Sigma}^* = & n^{-1} \int_0^\xi \left\{ \sum_{z_1 \in Z_1} \left[\frac{\dot{\theta}_{z_1}(x; \hat{\beta}, \hat{H})}{\theta_{z_1}(x; \hat{\beta}, \hat{H})} - Q(x; \hat{\beta}, \hat{H}) \right] \right\}^{\otimes 2} Y_{z_1}(x) \theta_{z_1}(x; \hat{\beta}, \hat{H}) d\hat{H}(x) \\ & + \sum_{i=1}^n \left\{ \frac{\dot{\theta}_{2i}(x; \hat{\beta}, \hat{H})}{\theta_{2i}(x; \hat{\beta}, \hat{H})} - Q(x; \hat{\beta}, \hat{H}) \right\}^{\otimes 2} Y_{2i}(x) \theta_{2i}(x; \hat{\beta}, \hat{H}) d\hat{H}(x), \end{aligned}$$

$$\hat{\Sigma}_* = n^{-1} \int_0^\xi \frac{\dot{S}_2(x; \hat{\beta}, \hat{H}) S_0(x; \hat{\beta}, \hat{H}) - \dot{S}_1^{\otimes 2}(x; \hat{\beta}, \hat{H})}{S_0(x; \hat{\beta}, \hat{H})} d\hat{H}(x).$$

Proof of the consistency and weak convergence of NPMLE $\hat{\Omega} = (\hat{\beta}, \hat{H}_t)$ is done in Hu and Tsodikov (2014a). Assuming regularity conditions hold, the main result is as follows:

1. With probability one, $\hat{\beta}$ converges to β^0 , $\hat{H}_t$ converges to $\hat{H}_t^0$ uniformly in the interval $[0, \xi]$, where H_t^0, β^0 are the true values of H_t, β .
2. $n^{1/2}\{\hat{\beta} - \beta^0, \hat{H}_t - H_t^0\}$ converges weakly to a zero-mean Gaussian process. In addition, $n\mathcal{E}^T(\mathcal{J}_n)^{-1}\mathcal{E}$ converges in probability to the asymptotic variance-covariance function of the linear functional $n^{1/2}\{\int_0^\xi a^T(\hat{\beta} - \beta^0) + b(t)^T(\hat{H}_t - H_t^0)\}$, where a is real vector, $b(t)$ is a function with bounded total variation in $[0; \xi]$, $\mathcal{J}_n$ is the negative Hessian matrix of the observed log-likelihood function with respect to $\hat{\Omega}$.

3 Simulation Studies

3.1 Different baseline hazards and sample sizes

We performed Monte Carlo simulations using R software to assess the proposed methodology. For each scenario, 1000 repetitions with sample sizes 200 and 500 were performed. Each scenario included covariates $Z_1 \sim U(0,5)$ and an increasing or decreasing baseline hazard following a Weibull distribution $h_t = 0.02t$ (increasing) or $h_t = 0.05(t/10)^{-1/2}$

(decreasing). Time-to-metastasis-onset, time-to-diagnosis, and time-to-death were generated under Cox proportional hazards using parameters $\beta_n = 0.9$, $\beta_\theta = [0.5, 0.8]$, $\beta_\mu = [0.5, 0.6]$, and $\beta_\gamma = [0.1, 0.7]$. We set $\delta_1 = \delta_2 = 1$ (ie. proportional hazards for diagnosis, metastasis, and death) because these parameters were found to be 1 from our motivating SEER analysis. Independent censoring time followed a uniform distribution $U(0, 20)$ for the increasing hazard and $U(0, 10)$ for the decreasing hazard scenario, which yielded approximately 15% censoring before the first event (diagnosis) and 30% censoring before the second event (death). The convergence criterion was set at 10^{-5} for change in baseline hazard estimation and improvement in the full likelihood.

The simulation results are summarized in Table 1. The iterative reweighting algorithm (Section 2.5.2) works well for both scenarios of increasing and decreasing baseline hazards. The estimates for covariate effects on time-to-latent-metastasis, time-to-diagnosis, and time-to-death are almost unbiased at small sample size ($n=200$) and further reduce bias for larger sample size ($n=500$). The empirical standard errors (ESE) are close to the mean of estimated theoretical errors (ASE), which validate the performance of the variance estimators. Coefficient estimators improve with increasing sample size as the estimated variances decrease and 95% confidence intervals tighten. The 95% coverage probabilities for all coefficient estimators remain as expected for both sample sizes.

3.2 Informative censoring

We performed simulation to assess the performance of the estimating procedures and prediction function in the presence of time-dependent covariates and dependent censoring. The simulated variables are similar to SEER example as follows: 3 race indicators $Z_{race} \sim Unif(1, 3)$, and 4 treatment indicators $Z_{trt} \sim Unif(1, 4)$. Risk of metastasis is

$$\lambda_U(t_u | Z) = h(t_u) e^{\beta_\eta Z_{race} + \beta_\eta Z_{trt} I_{t_1 < t_u}} \text{ where } h(t) = 0.02t. \text{ Risk of diagnosis is}$$

$$\lambda_1(t_1 | Z, t_u) = h(t_1) e^{\beta_\theta Z_{race} + \beta_\mu Z_{race} I_{t_1 > t_u}}. \text{ Risk of death is}$$

$$\lambda_2(t_2 | Z, t_u, t_1) = h(t_2) e^{\beta_\gamma Z_{race} + \beta_\gamma Z_{trt} I_{t_2 > \max(t_u, t_1)}}.$$

For the independent censoring scenario, censoring time $V^* \sim Unif(0, 30)$ results in 15 % censored before diagnosis and 30 % censored before death. For the dependent censoring scenarios, we use the following copula for bivariate survival data:

$$S(t, v) = G\left\{G^{-1}[S_2(t)] + G^{-1}[S_V(v)]\right\} \text{ where } G(s) = \left(\frac{\phi}{\phi + s}\right)^\rho, S_2 \text{ and } S_V \text{ are survival functions for death-due-to-prostate-cancer and death-due-to-other-causes (ie censors) respectively. Thus, smaller } \rho \text{ corresponds to higher dependence.}$$

Table 2 shows the simulation results from 1000 runs with sample size 500. The performance of coefficient and variance estimators remains robust with dependent censoring. The effects of time-dependent covariates are accurately captures with finite sample size.

4 Application to SEER prostate cancer data between 1988–2003

4.1 Data

We applied our proposed model to the motivating setting of prostate cancer cases file and population-at-risk counts available from SEER registry. The proposed model in its current specification (Figure 1) is tailored to this disease since virtually all deaths from prostate cancer occur after the cancer has metastasized. SEER, controlled by the National Cancer Institute, registers yearly incidence and population data on various cancers in the U.S. starting from time of diagnosis. We use the following two datasets:

SEER incidence data: This dataset provided information on race, gender, age, and stage (metastasis status) at diagnosis, and subsequent survival time of men who were diagnosed with prostate cancer between 1973–2003. Out of all recorded cases in this time period ($n=450,386$), we concentrated on men between the ages of 50–85 ($n=428,643$). Because we used age 50 as the starting time of study, the time-to-event can also be interpreted as age-at-event minus 50. Subjects with unknown race, age, or treatment were removed under the assumption of missing at random ($n=329,442$ remained).

SEER population data: A large part of the population at risk was missing since the incidence dataset only followed subjects who were diagnosed with cancer. Therefore, we use a second SEER data that tabulate population counts stratified by race, gender, age, and year. We subset the population from which the cancer cases originated (i.e. U.S. male population age 50–85 between 1988–2003) to provide risk-set count before diagnosis ($Y_z(t)$ in eq.3). The risk set for death-due-to-cancer after diagnosis ($Y_2(t)$ in eq.3) is the counting process of diagnosed cases in cases data. Although it is possible to focus only on the case file starting from diagnosis, it would mean losing a lot of information regarding the latent metastasis process which occurs before diagnosis.

Besides clinical variables, another factor that could potentially confound the effect of treatment on survival is the use of prostate-specific antigen (PSA) screening which began around 1988. The benefits of PSA screening are still debatable since it can result in overdiagnosis (i.e. detection of disease that would not become symptomatic during a patient's lifetime in the absence of screening). Overdiagnosis can positively affect the overall survival from diagnosis under screening compared to no screening prior to 1988 (largely an artifact). To avoid confounding, we focused on cases with diagnosis after 1988 ($n=299,922$). Amongst these men, 94.8% were diagnosed without metastasis and 5.2% were diagnosed with metastasis. Out of 299,922 men, 25,020 (7.7%) were recorded to die from prostate cancer by 12/31/2003. Deaths from other causes and those alive at the end of study were censored. Even though this data has significantly more censoring than simulation studies, we expect unbiased estimates due to the large sample size in SEER.

The stratification by races and treatments is as follow: 83.0% of men were white, 11.8% of men were black, and 5.2% of men were of other race. After diagnosis of prostate cancer, 30.9% of the men received no treatment, 35.2% received radiotherapy, 32.0% had surgery, and 1.9% had both radiotherapy and surgery. Note that unlike race, treatments are time-dependent (zero prior to diagnosis and thus only affect subsequent time intervals between

diagnosis and death). These specifications can easily be accommodated by taking race as a covariate Z_1 in the hazard of initial diagnosis $\lambda_1(t|Z_1)$, and time-dependent treatment as an additional covariate Z_2 in the hazard of metastasis $\lambda_U(t|Z_1, Z_2)$ (if metastasis occurs after cancer diagnosis) and of death $\lambda_2(t|Z_1, Z_2)$.

4.2 SEER Analysis Results

The proposed model fitting is summarized in Table 3, 4, and 5. The coefficients denote instantaneous relative risks with respect to a baseline group which is white men who did not receive treatment. A positive coefficient signifies higher risk and thus a shorter time to event. The risk of metastasis (Table 3) changes at point of diagnosis (due to introduced interventions). Similarly, the risk of being diagnosed of cancer (Table 3) increases in all races once metastasis occurred.

Even though statistically significant, parameters δ_1 and δ_2 were both estimated to be close to 1. This means that, beside proportionality, there is not much evidence of drastic changes in baseline hazards between metastasis, diagnosis, and death.

Race was significantly associated with time to each of the three stages of prostate cancer progression. White men were found to have lower risk of metastasis overall compared to same-age black men ($\hat{\beta}_{\eta_{black}} = 0.19, P < 0.001$) and same-age men from other races ($\hat{\beta}_{\eta_{other}} = 1.05, P < 0.001$). White men were also found to have lower risk of being diagnosed with prostate cancer prior to metastasis compared to black men and men from other races ($\hat{\beta}_{\theta_{black}} = 0.40, P < 0.001$; $\hat{\beta}_{\theta_{other}} = 0.61, P < 0.001$). The occurrence of latent metastasis had an effect on time-to-diagnosis: After occurrence of metastasis, the risk of being diagnosed with prostate cancer increases across all races: most among black men ($\hat{\beta}_{\mu_{Intercept} + black} = 7.37, P < 0.001$), less among white men ($\hat{\beta}_{\mu_{Intercept}} = 5.94, P < 0.001$), and least among men of other races ($\hat{\beta}_{\mu_{Intercept} + other} = 5.10, P < 0.001$). Instantaneous risk of death after diagnosis and after the onset of metastasis (both events need to happen in order for this risk to be non-zero) is higher among black men compared to white men ($\hat{\beta}_{\gamma_{black}} = 0.07, P = 0.01$) and lowest among men from other races ($\hat{\beta}_{\eta_{other}} = -0.86, P < 0.001$).

Treatment administered after being diagnosed with prostate cancer affected overall survival through two pathways: (1) indirectly via time-to-metastasis (if first diagnosed without metastasis), which is a required condition specified in our model before death can occur, and (2) directly on time to death. When first diagnosed with local cancer, only radiotherapy was found to delay time to metastasis ($\hat{\beta}_{\eta_{radiotherapy}} = -0.57, P < .001$). Surgery (radical prostatectomy) by itself or in combination with radiotherapy did not decrease risk of metastasis, but significantly increased ($\hat{\beta}_{\eta_{surgery}} = 1.39, P < .001$; $\hat{\beta}_{\eta_{combo}} = 3.04, P < .001$).

Similar effects of surgical removal of the primary tumor increasing time to metastasis have been reported in cancer literature (O'Reilly et al 1994; Smolle et al 1997; Biki et al 2008;

Neeman and Ben-Eliyahu 2013) and could be overlooked in other models that do not account for the latent metastasis process. This may also be due to a con-founder in the physicians' decision to perform surgeries only on more severe prostate cancer diagnoses or support the theory that removing the primary tumor along with its foreign antigens weakens the signal for the immune system's attack, making other body parts more susceptible to metastasis. On the other hand, surgery was found to significantly increase overall survival time ($\hat{\beta}_{\gamma surgery} = -3.05, P < .001$). A combination of both surgery and radio-therapy was also significantly beneficial ($\hat{\beta}_{\gamma combo} -2.41, P < .001$). Radiotherapy by itself slightly prolonged the survival rate compared to no treatment, but the effect was small compared to the other two treatment options ($\hat{\beta}_{\gamma radiotherapy} -0.07, P = 0.005$).

Due to the potential confounder of physicians subjectivity and other unaccounted covariates, the interpretation of treatment difference may suffer if interpreted as treatment effect in the context of randomized clinical trial. However, if we are willing to assume that treatment is given based on observation of the covariates included in the model and thus a function of these covariates only, and ignore the extent to which treatment depends on unobserved characteristics that are not controlled in the model, then we can believe they are rough estimate of treatment effect. We pointed out that beyond interpreting the result with caution, there is not much we can do using observational registry data.

If we adopt the view that there are unaccounted covariates that influence treatment choice, then we can use treatment as a surrogate for these covariates. This means we do not interpret treatment coefficient as purely treatment effect in clinical trial, but as indication that perhaps more aggressive treatment is given to patients that are worse off, and thus were adjusted for because we want to adjust for everything significant, via treatment as covariate that has both treatment effect in clinical trial and effect of other covariates for which that it was used as surrogate.

Beside race and the use of PSA screening, there are other potential con- founders that could bias treatment effects which were not captured in our model. One is the tumor grade (i.e. level of differentiation of tumor cells). There is a biological concensus that tumors containing well-differentiated cells are less aggressive due to the close resemblance of these cells to normal cells. Another factor that could not easily be captured is the doctor's decision in treatment choice which may reflect disease severity (not captured by stage), experience, and patient decision. For the purpose of treatment comparisons, statistical models assume similar representation of patients in each treatment group; however, this may not be the case in reality. Physicians usually recommend more aggressive treatments when patients are diagnosed with more advanced disease status and are young and healthy enough to endure it. Treatment regimens may also change when a patient's condition declines, improves, or experiences side effects. Currently SEER data does not include more clinical information to adjust for this subjectivity or changing treatment regimen.

4.3 Predicted Survival Rates

From our fitted model, we can derive closed-form formulas to predict the time to each disease stage. This carries clinical significance because these survival rates could be used to

inform treatment decisions at each time point for future patients diagnosed with prostate cancer. In this section, we derive the formula for survival probabilities at time t for diagnosis and death in the SEER example. These predicted survival rates were compared against the empirical survival rates to validate our proposed model. The survival probability for latent metastasis is not shown since we do not have empirical time-to- metastasis-onset in SEER data to compare against.

The predicted survival rate for incidence of cancer diagnosis ($\mathcal{L}_5$ from section 2.4) is given by:

$$\hat{S}_1(t_1|Z) = \hat{\eta} e^{-\hat{H}_1^{\hat{\delta}_1 \hat{\mu}}} \int_0^{t_1} \hat{h}_u e^{-\hat{H}_u \hat{\eta} - \hat{H}_u^{\hat{\delta}_1 \hat{\mu}} \hat{\theta}} du + e^{-\hat{H}_1 \hat{\eta} - \hat{H}_1^{\hat{\delta}_1 \hat{\mu}} \hat{\theta}}. \quad (7)$$

Figure 2 plots the Kaplan-Meier (KM) curve and predicted survival rate for prostate-cancer diagnosis incidences in each racial group (eq. 7). Note that all the survival probabilities are very close to 1 because we used the full US male population ages 50–85 as the risk-set, which is much larger than the size of incidence dataset. This figure validates the estimating procedure since the predicted and empirical survival rates match closely for White and Black men. There is a slight deviation between the predicted survival rates and empirical KM curve amongst men from other races, which could be due to the relatively much smaller sample size in this group compare to the group of White or Black men.

The conditional survival rate given time-of-diagnosis can be derived and used to compare the how different treatments affect time-to-death among those that were diagnosed of cancer at the same time. This conditional survival rate is derived from $\mathcal{L}_2^* / \mathcal{L}_1$ (section 2.4), which gives:

$$\hat{S}_2(t_2|t_1, T_u > t_1, Z) = \hat{\eta} \int_{t_1}^{t_2} \hat{h}_u e^{-\hat{\eta}(\hat{H}_u - \hat{H}_1) - \hat{\gamma}(\hat{H}_2^{\hat{\delta}_2} - \hat{H}_u^{\hat{\delta}_2})} du + e^{-\hat{\eta}(\hat{H}_2 - \hat{H}_1)}. \quad (8)$$

Figure 3 shows the empirical KM curves for death-due-to-prostate-cancer incidence among men diagnosed without metastasis at 60 years old, along with predicted survival rates (eq. 8). Both were stratified by race and treatment. Some of the treatments are indistinguishable and thus difficult to be compared against KM curves, especially for the stratified groups that have small sizes. The validity of the prediction function was further confirmed via simulations that resemble SEER setting but with equally distributed sample size between stratified groups (results not shown).

5 Discussion

We presented an NPMLE estimator and proposed a model that simultaneously estimated time to different events, where events were both latent and observable. Assuming that the

underlying biological cause that drives these events is the same across stages, our proposed model used the same baseline hazard for all stages to construct the joint likelihood of different disease timelines. In practical terms, the assumption of a common baseline hazard allows for gains in efficiency compared to cause-specific models, and closed-form likelihood expressions. This assumption can be relaxed by using a parametric transformation for the baseline hazards of different events. In the likelihood, the latent disease event is integrated within the appropriate time interval anchored by observed events. The non-parametric baseline hazard and covariate effects are iteratively estimated using a weighted Breslow-type estimator derived from the score function. We verified the performance of this estimating algorithm through Monte Carlo simulations at two sample sizes and types of baseline hazards. Application of the proposed model to SEER prostate cancer cases among U.S. men ages 50–85 revealed different ways through which treatments ultimately effected overall survival.

Our model differs from the existing models in that it can flexibly describe the relationship between outcomes in multiple stages beyond the fact that one outcome can prevent another from being observed. In SEER analysis, we specified that cancer death occurs after metastasis. This dependence can be even more nuanced by specifying, for example, that the quickness by which metastasis occurs affects how quickly death subsequently occurs. In cancer studies, where it is intuitively and biologically plausible that the “competing” stages are dependent, our model offers a convenient way to summarize these effects compared to existing methods such as cumulative incidence functions in a competing risk model or transitional intensity in an illness-death model. Moreover, the alterable components in the proposed model’s stage-specific hazards and incorporation of time-dependent covariates allow for easy adaptability to reflect covariates’ effects on different parts of the disease process.

The existing methods also fail to explicitly model the time-to-a-latent- event. In certain diseases, these latent outcomes may carry significant quality-of-life importance. Therefore, knowing how covariates effect these unseen events is sometimes as important as studying the observable outcomes. Our current model is equipped to infer this and incorporate scenarios where the event is partially missing in a subset of the population.

Currently, the earliest event of interest in our model is metastasis (or diagnosis if it happens prior). We recognize that there is a disease-free period before tumor is initiated. However, modeling this process means making assumption about covariates that is never observed. Our model addresses the challenge of latent variable time-to-metastasis because it was observed indirectly via metastasis status at cancer diagnosis time. If we introduce a point of tumor initiation, where risk of diagnosis and metastasis are zero, we would need even stronger assumption than the current model to describe it.

While we denote these stages as unique, the demarcation of these stages is not always clear. Such is the case with metastasis in our SEER example. Treatments taken from the time-of-diagnosis, while not affecting time-to-metastasis in patients whose cancer had already metastasized prior to diagnosis, should have an effect in the rest. This ability to select different covariates for each stage, and at the same time allow latter stages to be dependent

upon earlier stages, may reveal the effect of treatment at different stages and help guide clinical decision making.

The optional pathways in the joint likelihood further makes our proposed model adaptable to describe different disease processes. In some cancers, death occurs after metastasis. In others, death may occur at any point after diagnosis. To model the latter, an additional term to the likelihood can reflect the possibility of a patient dying without intermediate stages, while the estimation procedure remains the same. This can also be generalized to model different segment of the diseases, such as the time between recurrence and death. Furthermore, while our current model assumes proportional hazards between stages via the common baseline hazard, this assumption can be relaxed by using parametric transformations to link different stages' baseline hazards as described in Hu and Tsodikov (2014a).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by National Cancer Institute grant U01CA199338 (CISNET).

References

- Andersen PK, Keiding N (2002) Multi-state models for event history analysis. *Statistical Methods in Medical Research* 11(2):91–115 [PubMed: 12040698]
- Andersen PK, Hansen LS, Keiding N (1991) Non-and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous markov process. *Scandinavian Journal of Statistics* pp 153–167
- Bagdonavicius VB, Nikulin MS (1999) Generalized proportional hazards model based on modified partial likelihood. *Lifetime Data Analysis* 5(4):329–350 [PubMed: 10650741]
- Biki B, Mascha E, Moriarty DC, Fitzpatrick JM, Sessler DI, Buggy DJ (2008) Anesthetic technique for radical prostatectomy surgery affects cancer recurrence: A retrospective analysis. *The Journal of the American Society of Anesthesiologists* 109(2):180–187
- Chen YH (2009) Weighted breslow-type and maximum likelihood estimation in semipara- metric transformation models. *Biometrika* 96(3):591–600
- Chen YH (2012) Maximum likelihood analysis of semicompeting risks data with semipara- metric regression models. *Lifetime data analysis* 18(1):36–57 [PubMed: 21850528]
- Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65(1):141–151
- Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94(446):496–509
- Fine JP, Jiang H, Chappell R (2001) On semi-competing risks data. *Biometrika* 88(4):907–919
- Govindarajulu US, Lin H, Lunetta KL, D'Agostino R (2011) Frailty models: applications to biomedical and genetic studies. *Statistics in medicine* 30(22):2754–2764 [PubMed: 21786277]
- Hougaard P, Hougaard P (2000) *Analysis of multivariate survival data*, vol 564 Springer New York
- Hu C, Tsodikov A (2014a) Joint modeling approach for semicompeting risks data with missing nonterminal event status. *Lifetime data analysis* 20(4):563–583 [PubMed: 24430204]
- Hu C, Tsodikov A (2014b) Semiparametric regression analysis for time-to-event marked endpoints in cancer studies. *Biostatistics* 15(3):513–525 [PubMed: 24379192]

- Kalbfleisch JD, Prentice RL (2011) The statistical analysis of failure time data, vol 360 John Wiley & Sons
- Liu L, Wolfe RA, Huang X (2004) Shared frailty models for recurrent events and a terminal event. *Biometrics* 60(3):747–756 [PubMed: 15339298]
- Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C (2006) Nonparametric estimation of transition probabilities in a non-markov illness–death model. *Lifetime Data Analysis* 12(3):325–344 [PubMed: 16917736]
- Neeman E, Ben-Eliyahu S (2013) Surgery and stress promote cancer metastasis: new out-looks on perioperative mediating mechanisms and immune involvement. *Brain, behavior, and immunity* 30:S32–S40
- Nelsen RB (1999) An Introduction to Copulas. Springer Science & Business Media
- Nielsen GG, Gill RD, Andersen PK, Sørensen TI (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* pp 25–43
- Oakes D (1989) Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84(406):487–493
- O'Reilly MS, Holmgren L, Shing Y, Chen C, Rosenthal RA, Moses M, Lane WS, Cao Y, Sage EH, Folkman J (1994) Angiostatin: a novel angiogenesis inhibitor that mediates the suppression of metastases by a lewis lung carcinoma. *cell* 79(2):315–328 [PubMed: 7525077]
- Peng L, Fine JP (2007) Regression modeling of semicompeting risks data. *Biometrics* 63(1):96–108 [PubMed: 17447934]
- Smolle J, Soyer HP, Smolle-Jittner F, Rieger E, Kerl H (1997) Does surgical removal of primary melanoma trigger growth of occult metastases? an analytical epidemiological approach. *Dermatologic surgery* 23(11):1043–1046 [PubMed: 9391562]
- Tsodikov A (2003) Semiparametric models: a generalized self-consistency approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(3):759–774 [PubMed: 21127741]
- Xu J, Kalbfleisch JD, Tai B (2010) Statistical analysis of illness-death processes and semi- competing risks data. *Biometrics* 66(3):716–725 [PubMed: 19912171]
- Zeng D, Lin D (2006) Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* 93(3):627–640

Fig. 1.

The stages in cancer progression where we assume death from cancer occurs due to metastasis. Start time can be birth or the age where the disease risk originates. The exact time of metastasis onset is unknown and can occur prior to or following diagnosis of disease.

Fig. 2.

Survival estimates by race for time-to-diagnosis. Our method closely matches Kaplan-Meier (KM) estimates for each race group where follow-up time starts after age 50.

Fig. 3. Predicted survival rates and empirical Kaplan-Meier curves for death-from-prostate-cancer among men who were diagnosed at 60. Plot A is for white men and plot B is for black men.

Table 1

Simulation 1

N	Statistics	Metastasis	Diagnosis				Death	
		$\beta_{\eta 1}$	$\beta_{\theta 1}$	$\beta_{\theta 1}$	$\beta_{\mu 0}$	$\beta_{\mu 1}$	$\beta_{\gamma 0}$	$\beta_{\gamma 1}$
	True	0.9	0.5	0.8	0.5	0.6	0.1	0.7
Increasing baseline hazard $h_t = 0.02t$								
200	Bias	0.013	0.023	0.008	0.061	0.006	-0.045	0.020
	ESE	0.094	0.312	0.088	0.740	0.380	0.433	0.096
	ASE	0.093	0.301	0.087	0.742	0.332	0.437	0.095
	CP(%)	94.2	94.8	95.7	95.5	94.4	95.7	95.1
500	Bias	0.006	0.003	0.006	0.027	-0.004	-0.021	0.007
	ESE	0.059	0.190	0.058	0.471	0.223	0.276	0.058
	ASE	0.058	0.188	0.054	0.449	0.200	0.271	0.058
	CP(%)	94.1	94.6	92.7	94.8	93.5	95.4	95.7
Decreasing baseline hazard $h_t = 0.05(t/10)^{-1/2}$								
200	Bias	0.003	-0.021	0.010	0.004	0.032	-0.064	0.018
	ESE	0.094	0.295	0.083	0.741	0.371	0.454	0.094
	ASE	0.094	0.291	0.085	0.790	0.346	0.448	0.092
	CP(%)	94.7	94.6	95.8	96.1	94.0	94.7	94.8
500	Bias	0.004	0.002	0.005	-0.010	0.004	-0.010	0.005
	ESE	0.058	0.188	0.055	0.497	0.219	0.296	0.057
	ASE	0.059	0.183	0.053	0.473	0.206	0.281	0.057
	CP(%)	95.6	94.8	94.3	94.7	94.6	93.3	94.3

ESE: empirical standard errors based on 1,000 estimates

ASE: average of estimated standard errors

CP: 95% Coverage Probability

Table 2

Simulation with dependence censoring and time-varying covariates (n=500)

Dependence (censor %)	Statistics	Metastasis						Diagnosis						Death					
		B_{η_1}	β_{η_2}	β_{η_3}	β_{η_4}	β_{η_5}	β_{η_6}	β_{η_0}	β_{η_1}	β_{η_2}	β_{η_3}	β_{η_4}	β_{η_5}	β_{η_6}	β_{η_7}	β_{η_8}	β_{η_9}	$\beta_{\eta_{10}}$	
Independent (30%)	True	0.2	0.5	0.4	-0.3	-1	-0.7	1	0.6	0.1	0.8	0.3	0.4	0.9	0.5	0.6	-0.6	-2	-1.4
	Bias	-0.01	0.00	0.00	0.00	-0.02	-0.03	0.01	0	0.01	0.04	-0.01	-0.06	-0.02	0.04	0.04	-0.02	-0.06	-0.05
	ESE	0.23	0.23	0.31	0.41	0.53	0.4	0.21	0.15	0.18	0.54	0.72	0.72	0.43	0.27	0.27	0.37	0.4	0.36
	ASE	0.22	0.23	0.29	0.4	0.45	0.4	0.21	0.15	0.18	0.48	0.65	0.69	0.4	0.24	0.25	0.37	0.35	0.34
	CP(%)	93.9	94.7	93.5	95.9	92.3	95.5	95.1	94.7	95.5	93.1	95.5	95.7	91.9	92.7	93.7	93.3	90.5	92.9
$\rho = 5$ (30%)	Bias	-0.01	0.01	0.03	-0.03	-0.08	-0.07	-0.01	0.01	0.00	0.06	-0.01	-0.04	-0.08	0.02	0.04	-0.01	-0.05	-0.02
	ESE	0.22	0.22	0.32	0.41	0.49	0.43	0.20	0.15	0.19	0.48	0.74	0.69	0.38	0.22	0.23	0.36	0.34	0.33
	ASE	0.20	0.22	0.27	0.36	0.41	0.36	0.20	0.15	0.18	0.46	0.64	0.67	0.35	0.21	0.22	0.33	0.31	0.30
	CP(%)	92.70	95.10	95.30	96.30	93.70	93.90	93.90	94.90	94.10	94.30	97.40	95.50	91.10	95.30	93.70	92.10	90.70	92.90
	$\rho = 2$ (30%)	Bias	-0.03	-0.01	0.02	-0.02	0.01	-0.01	-0.01	0.00	0.00	0.02	0.01	0.03	-0.12	0.06	0.07	0.01	-0.13
ESE		0.22	0.22	0.27	0.37	0.50	0.39	0.20	0.15	0.18	0.49	0.70	0.70	0.36	0.22	0.22	0.35	0.33	0.31
ASE		0.20	0.22	0.26	0.36	0.42	0.36	0.20	0.15	0.18	0.46	0.64	0.68	0.35	0.20	0.21	0.33	0.31	0.30
CP(%)		92.10	95.10	96.00	96.00	92.60	96.40	95.30	94.90	95.30	93.40	95.10	97.00	91.30	90.90	92.80	91.90	90.20	93.40
Bias		-0.01	-0.01	0.01	-0.01	-0.05	-0.09	-0.01	0.00	0.01	0.01	0.05	0.04	-0.16	0.08	0.09	0.00	-0.18	-0.08
$\rho = 1$ (30%)	ESE	0.20	0.23	0.29	0.44	0.51	0.43	0.20	0.15	0.18	0.44	0.66	0.75	0.38	0.21	0.24	0.36	0.34	0.33
	ASE	0.20	0.22	0.27	0.37	0.44	0.36	0.20	0.15	0.18	0.45	0.63	0.68	0.35	0.20	0.21	0.33	0.32	0.30
	CP(%)	95.20	94.00	95.40	94.60	90.70	94.00	94.40	95.60	95.60	93.80	95.20	96.70	90.50	92.30	90.00	92.10	89.60	90.20
	$\rho = 0.5$ (30%)	Bias	0.01	0.02	-0.02	-0.03	-0.18	-0.14	0.02	0.00	0.00	0.05	0.05	0.00	-0.24	0.15	0.15	0.00	-0.23
ESE		0.22	0.24	0.31	0.49	0.57	0.45	0.21	0.15	0.18	0.53	0.76	0.76	0.38	0.23	0.24	0.35	0.34	0.32
ASE		0.21	0.22	0.27	0.37	0.44	0.36	0.20	0.15	0.18	0.47	0.67	0.69	0.36	0.21	0.22	0.33	0.32	0.30
CP(%)		92.7	91.7	94	94.2	86.1	90.1	92.3	95.2	94.2	94.2	95	95	86.3	86.5	87.5	92.7	86.5	90.9

Table 3

SEER: Time-to-metastasis-onset

Before diagnosis (β_{η})

Covariate	Estimate	SE	P-value
Z_{Black}	0.19	0.03	<.001
Z_{Other}	1.05	0.08	<.001

After diagnosis ($\beta_{\tilde{\eta}}$)

Covariate	Estimate	SE	P-value
<i>Intercept</i>	6.78	0.01	<.001
$Z_{Surgery}$	1.39	0.27	<.001
$Z_{Radiotherapy}$	-0.57	0.03	<.001
Z_{Combo}	3.04	0.26	<.001

Table 4

SEER: Time-to-diagnosis

Change to baseline hazard

Parameter	Estimate	SE	P-value
δ_1	1.04	0.01	<.001

Before metastasis (β_η)

Covariate	Estimate	SE	P-value
<i>Intercept</i>	3.20	0.04	<.001
Z_{Black}	0.40	0.01	<.001
Z_{Other}	0.61	0.01	<.001

Additional change after metastasis (β_μ)

Covariate	Estimate	SE	P-value
<i>Intercept</i>	5.94	0.12	<.001
Z_{Black}	1.43	0.06	<.001
Z_{Other}	-0.84	0.22	<.001

Table 5

SEER: Time-to-death

Change to baseline hazard

Parameter	Estimate	SE	P-value
δ_2	1.02	0.01	<.001

After diagnosis and metastasis (β_{γ})

Covariate	Estimate	SE	P-value
<i>Intercept</i>	8.50	0.00	<.001
<i>Z_{Black}</i>	0.07	0.03	0.010
<i>Z_{Other}</i>	−0.86	0.05	<.001
<i>Z_{Surgery}</i>	−3.05	0.16	<.001
<i>Z_{Radiotherapy}</i>	−0.07	0.03	0.005
<i>Z_{Combo}</i>	−2.41	0.07	<.001