Effects of a Misattributed Cause of Death on Cancer Mortality

by

Jinkyung Ha

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Professor Alexander Tsodikov, Chair
Professor Jack D. Kalbfleisch
Professor Jeremy M.G. Taylor
Associate Professor David Mendez

ACKNOWLEDGEMENTS

I would like to express my appreciation to my academic advisor Dr. Alexander Tsodikov. With his experience and knowledge, he steered me in the right direction at every stage of the dissertation. He has also exposed me to participate in various projects with the Cancer Intervention and Surveillance Modeling Network team. It was a valuable experience as I am starting out my career in this field. Without his great guidance and encouragement, I would never have accomplished this work.

I also would like to extend my gratitude to Professor Jack D. Kalbfleisch, Professor Jeremy M.G. Taylor and Associate Professor David Mendez for being on my committee and providing helpful suggestions.

Finally, I want to thank my family for their love and support all the way in my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

Prostate cancer is the second leading cause of cancer mortality among American men. Statistics from the Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute show that US prostate cancer mortality rates over the period of increased utilization of prostate-specific antigen (PSA) screening follow the unimodal shape of the incidence rates (Figure 1.1) . However, changes in the incidence due to screening are expected to lag those of mortality as patients with prostate cancer typically have good prognosis.

One of possible explanations for the observed trend could be a misattribution of the underlying cause of death in prostate cancer deaths. With the introduction of PSA screening in the late 1980s, increasingly many men get a diagnosis of prostate cancer. Although many of these men will die from causes other than prostate cancer, because prostate tumors are often slow growing, a proportion of these deaths is likely to be misattributed to prostate cancer just because the men were diagnosed with the disease. This is often referred to as over-attribution. Feuer et al. (1999) argued that this phenomenon would lead to a peak in mortality coinciding with the peak in incidence even if the misattributtion rate were a constant. A recent review of death certificates in New Mexico and a descriptive study by Hoffman et al. (2003) indicated

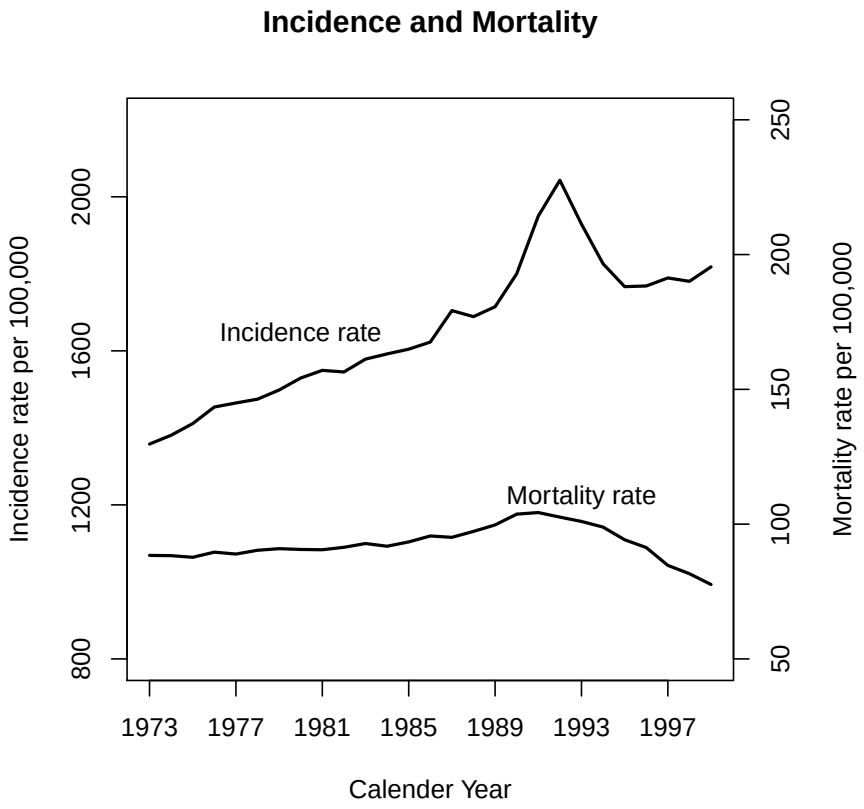

Figure 1.1: Prostate cancer incidence and mortality rates for men aged between 50 and 84 by calendar year.

that misattribution bias may explain up to a half of mortality increase before 1995.

Motivated by the suggestion we study a competing risks survival model where other causes of a death may be misattributed as a death of interest (prostate cancer) in Chapter II and III. Then, we incorporate the adjusted survival model into a mortality model, and assess the effects of over-attribution on cancer mortality in Chapter IV. To be specific,

Chapter II. We derive isotonic estimation of survival in univariate model under a misattribution of cause of death

Under misattribution the observed cause-specific hazard is distorted compared to

the true one. Nonparametric maximum likelihood estimation (NPMLE) yields consistent but non-monotonic estimates for cause-specific survival function, defined as *naïve* estimates. *Constrained* NPMLE is generally used to solve the problem of non-monotonicity. In Chapter II, we study some interesting topics observed through these estimators. Monotonicity of *naïve* estimates is not guaranteed even with large samples. *Constrained* estimator to which EM algorithm also converges is not consistent although it is monotonic. We consider other *isotonic* estimation approaches to achieve both consistency and monotonicity using the supremum (SUP) method and the Pooled-Adjacent-Violators (PAV) algorithm.

Chapter III. We develop an estimating equation for the semiparametric proportional hazards model assuming nonparametric cause-specific baseline hazards under a misattribution of cause of death

Analysis of competing risks data has received considerable attention when causes of failures are misclassified. However, two assumptions are generally required. The first is a missing-at-random (MAR) assumption: the probability of missing cause of failure does not depend on the failure type. Since we only assume that over-attribution exists in this article, the MAR assumption is violated. Another restriction is to make parametric assumptions on the cause-specific baseline hazards. However, as expected, it could yield a serious bias of estimator of interest due to misspecified model on nuisance parameter. In Chapter III, we develop an estimating equation approach with the Cox proportional hazards model which does not require any further assumption on cause-specific baseline hazards under over-attribution.

Chapter IV. We derive adjusted mortality rates by attribution bias using a statical model, a convolution of cause-specific survival and distributions for variables measured at diagnosis of tumor.

To assess the effects of misattribution on mortality rates, we use a mortality model which incorporates a cause-specific survival model under misattribution using the Kullback-Leibler's estimating equation introduced in Chapter III. PSA screening results in the lead time effect on survival. Hence, we present a cause-specific survival model which is a convolution of baseline survival in the absence of screening and the lead time distribution. With a variety of misattribution models, a sensitivity analysis is performed to assess the effect of attribution bias on recent trends in mortality rates using data from the Surveillance, Epidemiology, and End Results (SEER) Program.

CHAPTER II

Isotonic Estimation of Survival under a Misattribution of Cause of Death

2.1 Introduction

The rise and fall of US prostate cancer mortality follow the shape of incidence in the PSA-era. The cause for the trends in prostate cancer mortality rates is unclear. Several authors have indicated that incorrectly classified cause of death for prostate cancer survivors may have played a role in the observed recent peak and decline of prostate cancer mortality. A recent review of death certificates in New Mexico and a descriptive study by Hoffman et al. (2003) showed that about 5.6% misattribution bias is observed among the deaths in 1995.

Misattribution of the cause of death leads to a loss of monotonicity in naïve estimators of cause-specific survival. While there have been studies on inference under a masked cause of failure (e.g., Dinse (1982); Goetghebeur and Ryan (1995); Flehinger et al. (1998); Dewanji and Sengupta (2003); Craiu and Duchesne (2004)), isotonic methods dealing with the monotonicity constraint received little attention. The reason for this is the belief that the problem is of a small-sample origin, and that non-monotonicity disappears in large samples as all methods approach the naïve NPMLE asymptotically. We found this generally not to be true in our setting.

A nonparametric maximum likelihood estimation (NPMLE) approach was used

for the post-treatment survival under misattribution. Since unknown true cause of death is a missing data problem, we developed an EM algorithm in Section 2.3.2. Small-sample and asymptotic properties of the procedure are discussed in Section 2.3.3. Expecting EM to solve an unconstrained (naïve) NPMLE problem we found to our surprise that EM and constrained NPMLE are equivalent and both give biased solutions in the continuous setting. We studied other isotonic estimation techniques to remedy the problem in Section 2.4, and also explored their asymptotic properties. A simulation study is done to investigate the performance of the estimators in Section 2.5, and the methods are applied to real data from the SEER program in Section 2.6.

2.2 Assumption and notation

Suppose that we observe n independent individuals and that each individual can fail from one of two possible causes, which we term type 1 (prostate cancer) and 2 (other cause), respectively, or can be subject to an independent right censoring mechanism (type 0 failure). The observed data for individual i can be represented as (T_i, ω_i), where T is time of failure or censoring (whichever comes first), and $\omega = 0, 1, 2$ is the observed failure type. Define the corresponding true failure type as Ω which is not observed for some individuals. The true cause-specific hazard for type j failure is given by

$$\lambda_j(t) = \lim_{h \to 0} \frac{1}{h} \Pr[t \leq T < t + h, \Omega = j | T \geq t] \quad \text{for} \quad j = 0, 1, 2.$$

We assume the competing true causes of failure to be independent. Under this assumption, the crude hazards will be equal to their net counterparts. By the non-identifiability aspect of competing risks (Tsiatis (1975)) the dependence of risks cannot be recovered without additional assumptions. The same maximum likelihood

solution is achieved for the crude cause-specific hazards regardless of whether competing risks are assumed dependent or independent (Prentice et al. (1978)).

Typically, there are two types of misattribution in this setting. Assuming our interest focuses on a failure from prostate cancer (type 1 cause), define *over-attribution* as a death from other cause (type 2) attributed to prostate cancer (type 1). On the other hand, *under-attribution* is defined as a death from prostate cancer (type 1) attributed to other cause (type 2). Hoffman et al. (2003) found no *under-attribution* in recent prostate cancer data, and we consider *over-attribution* as the only type operating; that is,

$$\begin{aligned} &\Pr[\omega = 1 | T = t, \Omega = 2] = r(t) \\ &\Pr[\omega = 2 | T = t, \Omega = 1] = 0 \end{aligned} \tag{2.1}$$

where r is considered to be a known function of t. However, the results are easily extendable when *under-attribution* is also present.

Goetghebeur and Ryan (1995), Dewanji and Sengupta (2003) and Gao and Tsiatis (2005) proposed semi- and nonparametric inference procedures under a 'missing-at-random' mechanism of Little and Rubin (1987). This, however, is not generally the case unless the second probability in (2.1) is $1 - r(t)$; more precisely, the chance of observing failure from type j cause does not depend on its true failure types.

Under misattribution the observed cause-specific hazard is distorted compared to the true one

$$\lambda_j^{obs}(t) = \lim_{h \to 0} \frac{1}{h} \Pr[t \leq T_i < t + h, \omega = j | T_i \geq t] =$$

$$\begin{cases} \lambda_0(t), & \omega = 0, \quad \text{censoring} \\ \lambda_1(t) + r(t)\lambda_2(t), & \omega = 1, \quad \text{observed prostate cancer death} \\ \lambda_2(t)(1 - r(t)), & \omega = 2, \quad \text{observed other cause death.} \end{cases} \tag{2.2}$$

Note that

$$\lambda_1(t) + \lambda_2(t) = \lambda_1^{obs}(t) + \lambda_2^{obs}(t) = \lambda(t), \tag{2.3}$$

the marginal hazard of any type, as misattribution is a re-distribution of cases over the causes of failure.

2.3 Nonparametric maximum likelihood estimation

Let $N_j(t)$ be the process counting failures of type j, and $Y(t)$ be the at risk process. In the nonparametric context the model is parameterized by the jumps of the cumulative hazard function $d\Lambda_j(t) = \lambda_j(t)dt$. In terms of the hazard function, the loglikelihood can be written as

$$l = \int dN_1(t) \log[d\Lambda_1(t) + r(t)d\Lambda_2(t)] + dN_2(t) \log[\bar{r}(t)d\Lambda_2(t)] - Y(t)d\Lambda(t), \tag{2.4}$$

where $d\Lambda = d\Lambda_1 + d\Lambda_2$ is the jump of the marginal hazard of any type, $\bar{r} = 1 - r$, and the integral is taken over t and all the three terms. By taking derivatives of the loglikelihood with respect to $d\Lambda_j(t)$ for all distinct observed failure times t

$$\begin{aligned} \frac{\partial l}{\partial d\Lambda_1(t)} &= -Y(t) + \frac{dN_1(t)}{d\Lambda_1(t) + r d\Lambda_2(t)} \\ \frac{\partial l}{\partial d\Lambda_2(t)} &= -Y(t) + \frac{dN_2(t)}{d\Lambda_2(t)} + \frac{r dN_1(t)}{d\Lambda_1(t) + r d\Lambda_2(t)} \end{aligned} \tag{2.5}$$

and setting them to zero, we get the naïve nonparametric maximum likelihood estimates (naïve NPMLE) for type 1 and 2 failure hazard rates

$$\begin{aligned} d\tilde{\Lambda}_1(t) &= \frac{1}{Y(t)} \{dN_1(t) - \text{Odds}[r(t)] dN_2(t)\} \\ d\tilde{\Lambda}_2(t) &= \frac{dN_2(t)}{\bar{r}(t)Y(t)}, \end{aligned} \tag{2.6}$$

where $\text{Odds}[r] = r/(1-r)$. Here and in the sequel $1/Y$ is assumed to be 0 when $Y = 0$. Note that naïve NPMLE is a linear function of empirical estimates for the

observed hazard rates

$$
\begin{aligned}
d\tilde{\Lambda}_1(t) &= d\hat{\Lambda}_1^{obs}(t) - \text{Odds}[r(t)]d\hat{\Lambda}_2^{obs}(t) \\
d\tilde{\Lambda}_2(t) &= [\bar{r}(t)]^{-1}d\hat{\Lambda}_2^{obs}(t),
\end{aligned} \tag{2.7}
$$

where $\hat{\Lambda}_j^{obs}(t) = dN_j(t)/Y(t)$ are the Nelson-Aalen estimators for the hazard of the observed failures of type j. The naïve NPMLE is unique, and the corresponding information matrix is positive-definite.

Note that we define the naïve estimator as the one derived from the equation (2.6) even when some of dN_j's are zeros. When $dN_1(t_\alpha)$ is zero at a time t_α, the naïve estimate is negative $-\text{Odds}[r(t_\alpha)]d\hat{\Lambda}_2^{obs}(t_\alpha)$. It seems reasonable to have the corresponding estimate forced to zero to respect the monotonicity restriction on the cumulative hazard (cf. constrained estimator defined below). Unfortunately, allowing for negative $d\tilde{\Lambda}_1$ is essential for the consistency of the estimator $\int_0^t d\tilde{\Lambda}_1(s)$, which will be discussed in Section 2.3.3.

Note that from (2.7) the monotonicity is violated when

$$
d\hat{\Lambda}_1^{obs}(t) < d\hat{\Lambda}^{obs}(t)r(t) \quad \text{or} \quad \bar{r}(t)dN_1(t) < r(t)dN_2(t), \tag{2.8}
$$

where $d\hat{\Lambda}^{obs} = d\hat{\Lambda}_1^{obs} + d\hat{\Lambda}_2^{obs} = (dN_1 + dN_2)/Y = dN/Y$ is the Nelson-Aalen estimate of the marginal hazard of any type.

When violation of monotonicity at a time t_α is corrected by setting $d\Lambda_1(t_\alpha) = 0$, the model assumes all failures of type 1 at t_α are misattributed. Under this assumption, $Yrd\hat{\Lambda}^{obs} = Yrd\tilde{\Lambda}$ represents an estimate of the expected number of failures of type 1, by virtue of (2.3), while the left part of (2.8) is the observed counterpart of the same quantity. Hence, monotonicity is violated whenever the expected number of failures of type 1 is greater than the observed number of failures under the assumption of full misattribution. In case of untied data, according to

(2.8), violations of monotonicity occur whenever $dN_1(t) = 0$ and $dN_2(t) = 1$, i.e. whenever type 2 failure (other cause) is observed.

2.3.1 Constrained estimation

Define a constrained estimator $\hat{\Lambda}_j$, $j = 1, 2$ by maximizing the likelihood under the constraint of monotone Λ_js:

$$\max_{d\Lambda_j(t)\geq 0,\, j=1,2} l(\Lambda_1, \Lambda_2). \tag{2.9}$$

Note that the likelihood (2.4) is a sum of independently parameterized terms $l = \sum_\alpha l_\alpha(d\Lambda_1(t_\alpha), d\Lambda_2(t_\alpha))$ over distinct event times t_α. Therefore enforcement of the restriction proceeds separately for each such event time. The likelihood has the unique point of maximum defined by the score equations. When this point is outside of the admissible subset of parameters defined by $d\Lambda_1(t) \geq 0$, the constrained maximum must be on the border. Therefore, constrained solution corresponds to setting any monotonicity violator in the naïve NPMLE to zero, $d\hat{\Lambda}_1(t_\alpha) = 0$. When $d\Lambda_1(t_\alpha) = 0$, the optimal $d\hat{\Lambda}_2(t_\alpha) = d\hat{\Lambda}^{obs}(t_\alpha)$. As noted above in this case the model presumes all observed type 1 failures are misattributed true type 2 failures in which case their incidence is the same as the observed marginal one at t_α.

As a result, the non-negative counterpart of the naïve NPMLE (2.7) becomes

(2.10)

$$d\hat{\Lambda}_1(t) = d\tilde{\Lambda}_1(t) I\left[\bar{r}(t) dN_1(t) \geq r(t) dN_2(t)\right]$$

$$d\hat{\Lambda}_2(t) = d\tilde{\Lambda}_2(t) I\left[\bar{r}(t) dN_1(t) \geq r(t) dN_2(t)\right] + d\hat{\Lambda}^{obs}(t) I\left[\bar{r}(t) dN_1(t) < r(t) dN_2(t)\right],$$

where I is an indicator function $I(A) = 1$ if A is true and 0 otherwise. The role of indicator function in the constrained NPMLE for type 1 failure is to force the naïve estimates to zero whenever they are negative.

If data are untied, any non-empty event time point will be populated by only one failure, either type 1 or type 2, but not both types. It is clear that in this case (2.10) becomes

$$d\hat{\Lambda}_i(t) = d\hat{\Lambda}_i^{obs}(t), \quad i = 1, 2, \tag{2.11}$$

i.e. the constrained estimators for cause-specific (net) hazards for the true failure type coincide with the (crude) estimators specific to the observed failure types. Misclassification mechanism makes the latter a distorted version of the former and leads to bias. Note that both types of estimators yield the same (asymptotically unbiased) Nelson-Aalen estimator for the marginal hazard

$$d\hat{\Lambda}(t) = \sum_{i=1}^{2} d\hat{\Lambda}_i(t) = \sum_{i=1}^{2} d\tilde{\Lambda}_i(t) = \sum_{i=1}^{2} d\hat{\Lambda}_i^{obs}(t) = \frac{dN(t)}{Y(t)}. \tag{2.12}$$

Therefore, because of the "zero sum game" expressed by (2.12), $\hat{\Lambda}_i$, will be biased in opposite directions for $i = 1$ vs. 2. We will study the bias later in greater detail.

2.3.2 EM algorithm

Several papers have developed nonparametric estimates using EM algorithm when there are some missing failure types under a non-missing at random mechanism; see, for example, Dinse (1982) and Craiu and Duchesne (2004). Pretending Ω_i is observed, define the processes counting the true failures $N_j^0(t)$, where j is the type of failure. Then the complete data loglikelihood is

$$l = \int dN_1^0(t) \log[d\Lambda_1(t)] + dN_2^0(t) \log[d\Lambda_2(t)] - Y(t)[d\Lambda_1(t) + d\Lambda_2(t)]. \tag{2.13}$$

The estimates are improved by maximizing conditional expectation of the complete loglikelihood given observed data.

Suppose a failure of type $\omega = 1$ or 2 is observed ($\omega = 1$ corresponds to the failure type of interest while $\omega = 2$ corresponds to other cause of failure). Given this

information and the model assumptions, the distribution of the unknown true cause of failure Ω takes the form

$$P_{\Omega|\omega}(t) = \Pr(\Omega|\omega, T = t) = \frac{r(t)^{I(\Omega=2)}\lambda_\Omega(t)}{\lambda_1(t) + r(t)\lambda_2(t)} \times I(\omega = 1) + I(\Omega = \omega = 2). \tag{2.14}$$

Consequently, imputation of the unobserved $dN_k^0(t)$, $k = 1, 2$ is given by

$$\mathrm{E}\left\{ dN_k^0 \middle| Y, dN_1, dN_2 \right\} = \frac{r^{I(k=2)} d\Lambda_k}{d\Lambda_1 + r d\Lambda_2} dN_1 + I(k = 2) dN_2, \tag{2.15}$$

where the dependence on t is suppressed for brevity. Assuming $d\Lambda_k^m$ in the E-step (2.15) is indexed by the current iteration number, m, and maximizing the complete data likelihood (2.13) with dN_k^0 replaced by (2.15) (M-Step), we obtain the EM iteration sequence

$$d\Lambda_k^{m+1} = \frac{r^{I(k=2)} d\Lambda_k^m}{d\Lambda_1^m + r d\Lambda_2^m} d\hat{\Lambda}_1^{obs} + I(k = 2) d\hat{\Lambda}_2^{obs}. \tag{2.16}$$

A common perception is that EM solves an unrestricted MLE problem. If this were true, EM algorithm would converge to the naïve estimator (2.6). In our case, however, counter to this intuition, the EM solves the constrained problem (2.9) resulting in the estimator (2.10) that is biased. The key to this unexpected property is that the iterations (2.16) have two fixed points. The problem at hand is simple enough so we can find the fixed points explicitly. To do so assume $d\Lambda_k^{m+1}$ and $d\Lambda_k^m$ have a common limit $d\Lambda_k$, $k = 1, 2$, as $m \to \infty$. Substituting the common limit into (2.16), written for a particular point t, and solving the resultant equations for $d\Lambda_k(t)$ results in two distinct solutions. Let us write the first equation explicitly

$$d\Lambda_1(t) = \frac{d\Lambda_1(t)}{d\Lambda_1(t) + r d\Lambda_2(t)} d\hat{\Lambda}_1^{obs}. \tag{2.17}$$

First, assuming $d\Lambda_1(t) > 0$ cancels from (2.17), this equation enforces that the observed hazard is equal to the predicted marginal one $d\hat{\Lambda}_1^{obs} = d\Lambda_1(t) + r d\Lambda_2(t)$.

This, jointly with the second equation

$$d\Lambda_2(t) = \frac{r d\Lambda_2(t)}{d\Lambda_1(t) + r d\Lambda_2(t)} d\hat{\Lambda}_1^{obs} + d\hat{\Lambda}_2^{obs} \tag{2.18}$$

gives the naive estimator (2.6). Second, note that $d\Lambda_1(t) = 0$ also satisfies (2.17)! Solving the second equation (2.18) under $d\Lambda_1(t) = 0$ results in $d\Lambda_2(t) = d\Lambda^{obs}(t)$ indicating that the whole marginal incidence is explained by type 2 failures. Under a violation of monotonicity when (2.8) is satisfied, this coincides with the constrained estimator.

Technically, when monotonicity is not violated, EM converges to the naïve estimator represented by the first fixed point. When a violation is present, the first fixed point is outside of the restricted parameter space. Trying to reach it, the EM algorithm hits the border $d\Lambda_1 = 0$. At this point it is stuck in the restricted subspace because (2.17) is satisfied, and converges to the second fixed point. The proof that (2.10) is actually the point of convergence is given in Appendix .1.

2.3.3 Asymptotic properties

Let $\mathbf{A}(t) = (A_1(t), A_2(t))$ denote the compensator processes for $\mathbf{N}(t) = (N_1(t), N_2(t))$

$$A_1(t) = \int_0^t Y(s)\{d\Lambda_1(s) + r(s)d\Lambda_2(s)\}$$
$$A_2(t) = \int_0^t Y(s)(1 - r(s))d\Lambda_2(s).$$

We have $\mathbf{N}(t) = \mathbf{A}(t) + \mathbf{M}(t)$ where $\mathbf{M}(t) = (M_1(t), M_2(t))$ is a vector of 2 local square integrable martingales (Andersen and Gill (1982)). Naïve NPMLE of $\mathbf{\Lambda}(t)$ is given by $\int_0^t I(Y(s) > 0) d\tilde{\mathbf{\Lambda}}(s)$. By Rebolledo's theorem (Andersen et al. (1993)), $\sqrt{n}(\tilde{\mathbf{\Lambda}}(t) - \mathbf{\Lambda}(t))$ converges weakly to a zero-mean Gaussian vector process whose

covariance function can be consistently estimated by

$$n \begin{pmatrix} \int_0^t \frac{1}{Y(s)^2}\{dN_1(s) + \text{Odds}^2[r(s)]dN_2(s)\} & -\int_0^t \text{Odds}^2[r(s)]\frac{dN_2(s)}{r(s)Y(s)^2} \\ -\int_0^t \text{Odds}^2[r(s)]\frac{dN_2(s)}{r(s)Y(s)^2} & \int_0^t \text{Odds}^2[r(s)]\frac{dN_2(s)}{[r(s)Y(s)]^2} \end{pmatrix}. \tag{2.19}$$

Negative covariance terms off the diagonal are a consequence of the components of $\tilde{\Lambda}$ taking simultaneous jumps in opposite directions with the occurrence of type 2 failure. Consistency of the naïve estimates follows an application of Lenglart's inequality. Furthermore, they attain the Cramer-Rao lower bound. However, we cannot directly apply the results to constrained estimates because of the monotonicity restriction. It turns out that estimates show different behavior in the continuous and discrete case. The indicator function in constrained estimates (2.10) can be represented by summation over subject-i-specific processes $N_{ik}(t)$, $k = 1, 2$

$$I\left(\sum_{i=1}^{n} dN_{i1}(t) - \text{Odds}[r(t)]dN_{i2}(t) > 0\right). \tag{2.20}$$

Under the continuous time setting where no two events can occur at the same time, (2.20) behaves like a linear function in the sense that I and Σ can be interchanged. This results in the asymptotic bias of the estimator. To put it more accurately, as we show in Appendix .2, $\text{E}[\hat{\mathbf{\Lambda}}(t) - \mathbf{\Lambda}(t)]$ converges in probability to

$$\left(\int_0^t r(s)d\Lambda_2(s), \int_0^t -r(s)d\Lambda_2(s)\right),$$

as $n \to \infty$, which can be consistently estimated by

$$\left(\int_0^t \text{Odds}[r(s)]\frac{dN_2(s)}{Y(s)}, \int_0^t -\text{Odds}[r(s)]\frac{dN_2(s)}{Y(s)}\right).$$

Namely, the same amount of bias is added to both type 1 and 2 estimator but in an opposite direction.

The vector process $\sqrt{n}\left(\hat{\Lambda}_1(t) - (\int_0^t d\Lambda_1(s) + r(s)d\Lambda_2(s)), \hat{\Lambda}_2(t) - \int_0^t (1 - r(s))d\Lambda_2(s)\right)$ is a martingale that converges weakly to a zero-mean Gaussian vector process with

a consistent estimate of covariance as follows:

$$n \begin{pmatrix} \int_0^t \frac{dN_1(s)}{Y^2(s)} & 0 \\ 0 & \int_0^t \frac{dN_2(s)}{Y^2(s)} \end{pmatrix}. \tag{2.21}$$

Discrete time setting, on the contrary, yields asymptotically unbiased estimators. Discrete random variable, T, in survival analysis often arises due to rounding off measurements or grouping of failure times into intervals. In case of a continuous model coarsened to a discrete one, average probabilities corresponding to grouping intervals represent the target of estimation, and bias is understood relative to this target. In discrete time setting, the summation in (2.20) can no longer be interchanged with the indicator function since empirical probabilities of events occurring at the same time are not negligible asymptotically. Let $\Delta(t_k)$ denote the observation time interval, $t_{k+1} - t_k$, between two adjacent event times that may in general possess some multiplicity (ties). The asymptotic properties of Nelson-Aalen or Kaplan-Meier estimator for discrete lifetimes are based on $\Delta(t_k)$ and $\Pr\{T \in [t_k, t_{k+1})\}$ becoming negligibly small uniformly in probability as the sample size increases. However, the monotonicity restriction indicator (2.20) converges to 1 in probability only if the number of failures in $[t_k, t_{k+1})$ is allowed to accumulate. Consider the simple case: T grouped into intervals of unit length and no censoring. Let $\Delta N_j(t)$ denote the number of failures of type j at the interval $[t, t+1)$. Suppose the random vector $\big(\Delta N_1(t), \Delta N_2(t)\big)$ conditional on $Y(t)$ has the following distribution:

$$\begin{aligned} &Pr(\Delta N_1(t) = x_1, \Delta N_2(t) = x_2 | Y(t) = n) \\ &= \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} [\lambda_1(t) + r(t)\lambda_2(t)]^{x_1} [(1 - r(t))\lambda_2(t)]^{x_2} [1 - \lambda_1(t) - \lambda_2(t)]^{n - x_1 - x_2} \end{aligned}$$

where $\lambda_j(t) = Pr(t \le T < t+1, \Omega = j | T \ge t)$ for $j = 1, 2$. Maximizing $\prod_{t \ge 0} Pr(\Delta N_1(t), \Delta N_2(t) | Y(t))$ under non-negative restriction on λ, we obtain an

estimator for $\lambda_1(t)$ as

$$\frac{1}{Y(t)}\{\Delta N_1(t) - \text{Odds}[r(t)]\Delta N_2(t)\}I[\bar{r}(t)\Delta N_1(t) \geq r(t)\Delta N_2(t)]$$

that is almost identical with the constrained NPMLE in the continuous setting except that the monotonicity constraint I converges to 1 in probability as sample size increases (see discussion below). This makes the above estimator asymptotically equivalent to the naïve (unbiased) one.

2.4 Isotonic estimation

In addition to the constrained NPMLE, there are other approaches to achieve monotonicity. One of them is to replace a naïve estimate at time t with the maximum value of estimates up to t: that is,

$$\hat{\Lambda}_1^S(t) = \sup_{0\leq s\leq t} \tilde{\Lambda}_1(s).$$

This is motivated by Lin and Ying (1994). The sup-estimator $\hat{\Lambda}_1^S$ (SUP) is consistent. Indeed, $\hat{\Lambda}_1^S(t) = \tilde{\Lambda}_1(\tau_t)$ for some $\tau_t \leq t$. Then, since Λ_1 is increasing and $\tau_t \leq t$, we have

$$\tilde{\Lambda}_1(t) - \Lambda_1(t) \leq \hat{\Lambda}_1^S(t) - \Lambda_1(t) \leq \sup_{0\leq x\leq t} (\tilde{\Lambda}_1(x) - \Lambda_1(x)),$$

and consistency of the sup-estimator follows from the uniform consistency of the naïve one.

Lin and Ying (1994) argue that $\hat{\Lambda}_1^S$ is asymptotically equivalent to $\tilde{\Lambda}_1$ in the sense that $\sqrt{n}(\hat{\Lambda}_1^S - \Lambda_1)$ converges to the same limiting distribution as $\sqrt{n}(\tilde{\Lambda}_1 - \Lambda_1)$. Finding the distribution of a supremum of a stochastic process over a finite interval is a challenge. Such properties are only known for a very restricted set of processes such as the Brownian motion possibly with a linear drift and stationary Gaussian processes with very specific correlation structures (Adler (1990)). Our $\sqrt{n}(\hat{\Lambda}_1^S - \Lambda_1)$ process is

non-stationary and only becomes Gaussian in the limit. Lin and Ying (1994) argued heuristically that the sup-estimator is asymptotically equivalent to the naïve one, and their argument is applicable to our case. According to the argument $\hat{\Lambda}_1^S - \tilde{\Lambda}_1$ goes to zero in probability faster than $1/\sqrt{n}$ because the natural $1/\sqrt{n}$ rate of convergence of both estimators to the common limit Λ_1 is accelerated by the convergence of τ_t to t. We have found in simulations that the variance of the sup-estimator is very close to the naïve one (2.19), and we have adopted the working hypothesis of equivalence implied by the heuristic argument of Lin and Ying (1994).

Another isotonic estimator, PAV, $\hat{\mathbf{\Lambda}}_1^P$, can be constructed using the so-called pool-adjacent-violators algorithm; see Ayer et al. (1955) and Barlow et al. (1972) (pp. 13-5). $\hat{\mathbf{\Lambda}}^P$'s are derived as a set of $\mathbf{\Lambda}$'s minimizing a weighted sum of squared deviations of the isotonic estimate from the naïve estimate

$$\hat{\mathbf{\Lambda}}^P = \arg \min_{\{\mathbf{\Lambda} \in \Psi\}} \sum_{k=1}^{K} [\mathbf{\Lambda}(t_k) - \tilde{\mathbf{\Lambda}}(t_k)]^{\otimes 2} \mathbf{W}(t_k), \tag{2.22}$$

where Ψ is the class of non-decreasing functions, and $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^T$, and $\mathbf{W}$ is a vector of non-negative weights. This is a promising approach since the naïve estimator is consistent. Using Doob-Meyer decomposition we have the corresponding linear regression model with correlated errors

$$\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}(t)$ is an asymptotically Gaussian martingale (noise) with covariance obtained from a large sample limit of (2.19). The naïve estimator $\tilde{\mathbf{\Lambda}}$ given by (2.6) is a saturated solution to the above regression problem. Therefore, absent restrictions, the solution to (2.22) is the naïve estimator regardless of $\mathbf{W}$ as long as weights are non-negative. In the above approach the naïve estimator serves as the response variable.

Alternatively, the problem can be re-formulated as a regression with the observed hazard

$$d\hat{\mathbf{\Lambda}}^{obs} = (d\hat{\Lambda}_1^{obs}, d\hat{\Lambda}_2^{obs})' = \left(\frac{dN_1}{Y}, \frac{dN_2}{Y}\right)'$$

as a response, where the prime means a transposed vector. We have the model

$$\begin{pmatrix} \frac{dN_1(t)}{Y(t)} \\ \frac{dN_2(t)}{Y(t)} \end{pmatrix} = \begin{pmatrix} 1 & r(t) \\ 0 & 1-r(t) \end{pmatrix} \begin{pmatrix} d\Lambda_1(t) \\ d\Lambda_2(t) \end{pmatrix} + d\boldsymbol{\epsilon}(t), \tag{2.23}$$

where $\boldsymbol{\epsilon}$ is asymptotically Gaussian with uncorrelated components and the covariance matrix given by $(2.21)/n$. Since the naïve estimate $\tilde{\Lambda}_2$ does not violate the restrictions, we will apply the corrections to $\tilde{\Lambda}_1$ only, setting $\hat{\Lambda}_2^P(t) = \tilde{\Lambda}_2(t)$. The contribution to the sum of squares containing Λ_1 has the form

$$\left(\int d\Lambda_1 + r d\Lambda_2 - \frac{dN_1}{Y}\right)^2 W,$$

where W is some weight. On substitution of the naïve estimator for Λ_2, the above expression turns into

$$\left(\int d\Lambda_1 - d\tilde{\Lambda}_1\right)^2 W,$$

and the problem is equivalent to (2.22). We found hardly any difference between the optimal weights derived based on the error covariance matrices and constant weights, so constant weights are used for simplicity.

The consistency of $\hat{\Lambda}_1^P$ follows from Theorem 1.6 (Barlow et al. (1972)). The details are presented in Appendix .3.

2.5 Simulation

Simulated data were generated from a population with constant cause-specific hazards, $\lambda_1 = 0.05$ and $\lambda_2 = 0.15$, and independent censoring which is uniformly distributed on $(0, 15)$. In the continuous time setting, the observation time T is

obtained by a minimum of these three event times. Five choices of misattribution were examined: $(\psi_0, \psi_1) = (0.1, 0), (0.3, 0), (0.5, 0), (0.5, -0.5)$ and $(0.5, -0.1)$ in $logit\ r(t) = (logit\ \psi_0) + \psi_1 t$. First three choices represent constant misattribution 0.1, 0.3, and 0.5. In the last two models, misattribution decreases as the survival time is prolonged, which yield 0.24 and 0.43 misattribution on average, respectively. We carried out 2000 simulations with three different sample sizes, 100, 200 and 500. The means of the naïve, constrained (EM) and isotonic estimators using the supremum (SUP) function and the Pool-Adjacent-Violator (PAV) algorithm over 2000 simulations are calculated, and values at time 4.32, $Pr(T \leq 4.32) = 0.7$ in the continuous setting, are reported in Table 2.1. Constant versus variable weights yield very similar results, so only constant weight PAV estimates are shown. Asymptotic variances are given as the first value in parentheses. They are estimated based on the asymptotic distribution (2.19) and (2.21), for the naïve and the constrained estimator, respectively. Naïve variance estimators were used with the isotonic methods for comparison and as an approximation. The empirical standard deviations based on 2000 simulations are also given as the second value in parentheses.

The naïve nonparametric maximum likelihood estimates are in excellent agreement with the true value. On the other hand, the constrained estimates are seriously biased, although they have uniformly less variability than the other estimators. Interestingly, the bias keeps increasing with sample size, and converges to the theoretically predicted $\int_0^t r(s) d\Lambda_2(s)$. Both of the SUP and PAV estimator seem to behave similar to the naïve estimates. However, the rate of decrease in bias with sample size for the SUP estimator becomes slower than the one for the PAV estimator as misattribution increases.

Using the same simulated data, we again compared the properties of the estimators

in the discrete time case. In this setting, T is only observed as an integer value through a ceiling function; $\lceil T \rceil = \min\{n \in \mathbf{Z} | T \leq n\}$ where $\mathbf{Z}$ is the set of integers. The corresponding simulation-based estimates at time 7, $Pr(\lceil T \rceil \leq 7) = 0.82$, are reported in Table 2.2.

Overall, the constrained estimator has uniformly the smallest variance, at the cost of the highest small sample bias representing an extreme point on the bias-variance trade-off. The SUP estimator is more small-sample efficient than the PAV estimator in both continuous and discrete case. Unlike the continuous case, consistency of the constrained estimator is guaranteed in the discrete case. However, the rate of decrease in bias is slower than the other two isotonic estimators. In fact, it has the highest small-sample bias. Among the estimators that guarantee monotonicity, the largest bias occurs at the smallest sample size of 100 and largest attribution bias of 0.5, where mean difference between the SUP and PAV estimates is also the largest. In this case, none of the estimators seems to be appropriate. It is reasonable to take an average of the SUP and PAV estimators because their biases go in opposite directions (Table 2.3), and they have the same asymptotic properties as the naive estimator in the sense that $\sqrt{n}(\hat{\Lambda}_1^S + \hat{\Lambda}_1^P)/2 - \tilde{\Lambda}_1)$ converges to zero in probability by the heuristic argument of Lin and Ying (1994). By the same argument the asymptotic variance of the averaged estimator is the same as the naïve one.

Although the estimates were presented at a single time point in each setting, they all have similar results at any time point. Since type 2 failure is not of primary interest, and the properties of the estimators are similar to those for type 1 failure, the corresponding results are not shown.

2.6 Real Data Example

We illustrate our results using data from the Surveillance, Epidemiology, and End Results (SEER) cancer registry program. Staying close to the misattribution study of Hoffman et al. (2003), we looked at 1,094 New Mexico residents diagnosed with prostate cancer in 1993. There are two causes of failure: death due to prostate cancer (type 1) and death due to causes other than prostate cancer (type 2). We tried two values for misattribution, 0.05 and 0.1, within the range deemed plausible based on Hoffman et al. (2003).

Figure 2.1 displays the naïve and constrained nonparametric maximum likelihood estimates for the cumulative type 1 failure hazard. In addition, the Nelson-Aalen estimates for the observed type 1 failure are also presented (dotted lines). We have omitted the isotonic estimates since they are very close to the naïve estimates. Note that monotonicity for the naïve estimates still breaks down even though the sample size is large and misattribution is low, as expected. Predictably, it gets worse as misattribution increases. In this example, the failure time is recorded by month, ranging from 0 to 132, making it close to the continuous model setting. Less discrepancy between the naïve and constrained estimates would be expected if failure time is grouped into larger bins such as year, or if more data is available. It is not surprising that the naïve estimates are nondecreasing when using the data on 9 SEER registries, 25,088 men, not just New Mexico. The corresponding estimates and 95% point-wise confidence limits are presented in Figure 2.2.

2.7 Discussion

In this project, we have considered the problem of fitting a hazard model to competing risks data when there is a misattribution of failure type. The nature of the

misattribution mechanism makes the true failure status missing, but not at random, with the observed distorted status and time of failure providing some information.

We first presented the naïve nonparametric maximum likelihood estimator. We found that it violates the monotonicity constraint although it is consistent. Contrary to common belief, in this particular problem setting, violation of monotonicity never goes away with increasing sample size, even though the size of negative jumps of the cumulative hazard goes to zero. Following a common recipe we then derived the constrained nonparametric maximum likelihood estimator which achieved monotonicity and showed smaller variability, and surprisingly a substantial bias in the continuous setting. This led us to consider isotonic approaches using the supremum function and the pool-adjacent-violators algorithm, both providing consistent and monotonic estimates.

Another surprising finding is that the EM algorithm perceived as a local maximization method actually solves the constrained problem and converges to the biased constrained estimator. We traced the phenomenon to the specific structure of the EM iterative equations that have two fixed points, one of them associated with the unrestricted (naïve) solution, while the other fixed point corresponds to the restricted non-negative one. When the naïve solution is not admissible (non-monotonic), the iteration path of the EM is eventually locked in the restricted subspace associated with the second fixed point.

It has been generally taken for granted that constrained NPMLE and the EM algorithm have useful properties such as consistency. Consistency is perceived to be a consequence of the estimator satisfying the monotonicity constraint for large samples. However, in the present setting, we found that the constrained estimator is no longer consistent, but only when the failure time follows a continuous model.

These methods do behave well in the discrete setting when information is allowed to accumulate at distinct time points. This is counterintuitive to the properties of standard estimators such as Nelson-Aalen or Kaplan-Meier which serve both discrete and continuous cases well.

If one is prepared to overlook a violation of monotonicity of the cumulative hazard estimates, the naive NPMLE is preferred to any other estimates. However, interpreting such estimates involves reading a monotonic curve visually into the estimate without formalizing the procedure. With this project we are offering a tool to do this rigorously. We have found that the isotonic estimators obtained by using the supremum method and the pool-adjacent-violator algorithm behave similar to the naïve estimator yet provide monotonic cumulative hazards. From a purely function approximation standpoint, the issue is not much of a problem in large samples, since the naïve estimator is consistent.

We defined 'the discrete setting' as the case where the number of failures can accumulate over a discrete mesh of the observed failure times as the sample size increases. In this setting, the constrained estimator is consistent and attains the Cramer-Rao lower bound. However, we are still faced with non-monotonicity problem in small samples. An example showed a discrepancy between the naive and constrained estimates even with a sample size over a thousand (Figure 2.1), resulting in a substantial small-sample bias (Table 2.2). Thus, isotonic estimators are still useful in the discrete setting.

A simulation study showed that both SUP and PAV estimators, in most cases, work well even with a sample size as small as 100. The SUP estimator was found to be uniformly more efficient. Nevertheless, we prefer the PAV for several reasons. The PAV technique takes the naive estimates as a foundation and provides the amount

of smoothing that is just right to fix them. Thus, the bias of this estimator is less affected by small samples and high misclassification rates, and its small-sample variance is readily available through the model-based estimator (2.19). However, the approach will eventually break down under small-sample high misclassification challenge (see Table 2.2, 0.5 misattribution, sample size of 100, discrete setting). Our practical recommendation is to consider the naïve and PAV estimators, and report the PAV results if they are close to the naïve. If the PAV deviates from the naïve indicating a small sample bias, the SUP is expected to be biased in the opposite direction. Therefore in this situation we recommend reporting an average of SUP and PAV.

We argued that the naive estimator for $d\Lambda_1(t_\alpha)$ is negative if $dN_1(t_\alpha)$ is zero. It is generally true when misattribution is assumed to be known or estimated with a parametric model, constant or a function of t. Interestingly, this problem would not necessarily arise if $r(t_\alpha)$ is non-parametrically defined. Namely, if $r(t_\alpha)$ is estimated only with data at t_α, it becomes zero when $dN_1(t_\alpha)$ is zero. Thus, it makes the naive estimate for $d\Lambda_1(t_\alpha)$ equal to zero.

Throughout this work, we have assumed that the misattribution may depend on the time of failure and the true failure type. It is, however, possible that it is affected by other factors. Dependence on such covariates can be incorporated, for example, by assuming a logistic model for r, and a proportional hazard model for the true failure types. This refinement will be addressed in the next section.

Table 2.1: Simulation means for various estimators of the cumulative type 1 failure hazards at time 4.32 in the continuous setting with a sample size $n = 100$, 200 and 500 under misattribution, $logit\ r(t) = (logit\ \psi_0) + \psi_1 t$. The average of the standard error estimates and sample standard deviations are also given in parentheses. The true value is 0.216.

(ψ_0, ψ_1)	Estimator	$n = 100$	$n = 200$	$n = 500$
(0.1, 0)	Naïve	.216 (.074, .075)	.215 (.052, .053)	.216 (.033, .033)
	Constrained (EM)	.281 (.073, .073)	.280 (.051, .053)	.281 (.033, .032)
	SUP	.223 (.074, .073)	.219 (.052, .052)	.218 (.033, .033)
	PAV	.216 (.074, .073)	.215 (.052, .053)	.216 (.033, .033)
(0.3, 0)	Naïve	.214 (.097, .099)	.217 (.069, .070)	.216 (.044, .043)
	Constrained (EM)	.410 (.088, .089)	.411 (.063, .064)	.411 (.040, .040)
	SUP	.233 (.097, .091)	.227 (.069, .067)	.221 (.044, .042)
	PAV	.212 (.097, .095)	.216 (.069, .068)	.216 (.044, .042)
(0.5, 0)	Naïve	.214 (.129, .127)	.217 (.091, .093)	.216 (.058, .057)
	Constrained (EM)	.540 (.102, .101)	.540 (.072, .073)	.541 (.046, .045)
	SUP	.249 (.129, .110)	.237 (.091, .084)	.226 (.058, .054)
	PAV	.209 (.129, .119)	.215 (.091, .089)	.216 (.058, .056)
(0.5, -0.5)	Naïve	.214 (.091, .092)	.217 (.064, .067)	.216 (.041, .040)
	Constrained (EM)	.390 (.083, .083)	.391 (.059, .061)	.392 (.037, .037)
	SUP	.224 (.091, .088)	.221 (.064, .066)	.217 (.041, .040)
	PAV	.215 (.091, .090)	.217 (.064, .066)	.216 (.041, .040)
(0.5, -0.1)	Naïve	.214 (.118, .116)	.219 (.083, .084)	.217 (.053, .052)
	Constrained (EM)	.505 (.097, .097)	.506 (.069, .070)	.507 (.044, .043)
	SUP	.241 (.118, .104)	.233 (.083, .079)	.224 (.053, .050)
	PAV	.212 (.118, .111)	.218 (.083, .081)	.217 (.053, .051)

SUP=supremum, $\hat{\Lambda}_1^S(t)$, PAV=Pool-Adjacent-Violators algorithm, $\hat{\Lambda}_1^P(t)$

Table 2.2: Simulation means for various estimators of the cumulative type 1 failure hazards at time 7 in the discrete setting with a sample size $n = 100$, 200 and 500 under misattribution, $logit\ r(t) = (logit\ \psi_0) + \psi_1 t$. The average of the standard error estimates and sample standard deviations are also given in parentheses. The true value is 0.305.

(ψ_0, ψ_1)	Estimator	$n = 100$	$n = 200$	$n = 500$
(0.1, 0)	Naïve	.301 (.103, .100)	.304 (.074, .076)	.304 (.047, .045)
	Constrained (EM)	.314 (.102, .094)	.308 (.073, .074)	.304 (.046, .045)
	SUP	.308 (.103, .096)	.306 (.074, .075)	.304 (.047, .045)
	PAV	.294 (.103, .098)	.302 (.074, .075)	.303 (.047, .045)
(0.3, 0)	Naïve	.303 (.137, .139)	.310 (.097, .098)	.305 (.061, .062)
	Constrained (EM)	.338 (.124, .120)	.322 (.089, .090)	.307 (.056, .059)
	SUP	.320 (.137, .126)	.316 (.097, .093)	.306 (.061, .060)
	PAV	.288 (.137, .134)	.302 (.097, .098)	.303 (.061, .061)
(0.5, 0)	Naïve	.310 (.183, .183)	.307 (.128, .128)	.306 (.081, .078)
	Constrained (EM)	.381 (.144, .144)	.339 (.101, .108)	.313 (.064, .073)
	SUP	.345 (.183, .155)	.323 (.128, .114)	.310 (.081, .075)
	PAV	.284 (.183, .177)	.289 (.128, .127)	.299 (.081, .080)
(0.5, -0.5)	Naïve	.302 (.105, .108)	.305 (.075, .074)	.302 (.047, .046)
	Constrained (EM)	.312 (.102, .102)	.308 (.073, .072)	.302 (.046, .046)
	SUP	.304 (.105, .106)	.306 (.075, .073)	.302 (.047, .046)
	PAV	.301 (.105, .107)	.305 (.075, .073)	.302 (.047, .046)
(0.5, -0.1)	Naïve	.300 (.152, .154)	.308 (.108, .107)	.304 (.068, .068)
	Constrained (EM)	.347 (.131, .130)	.327 (.094, .096)	.308 (.059, .065)
	SUP	.320 (.152, .139)	.315 (.108, .102)	.306 (.068, .066)
	PAV	.290 (.152, .147)	.301 (.094, .104)	.303 (.068, .068)

SUP=supremum, $\hat{\Lambda}_1^S(t)$, PAV=Pool-Adjacent-Violators algorithm, $\hat{\Lambda}_1^P(t)$

Table 2.3: Simulation means for the average between SUP and PAV estimators of the cumulative type 1 failure hazards at time 7 in the discrete setting with a sample size $n = 100, 200$ and 500 under misattribution, $logit\ r(t) = (logit\ \psi_0) + \psi_1 t$. The average of the standard error estimates and sample standard deviations are also given in parentheses. The true value is 0.305.

(ψ_0, ψ_1)	Estimator	$n = 100$	$n = 200$	$n = 500$
(0.5, 0)	SUP	.345 (.183, .155)	.323 (.128, .114)	.310 (.081, .075)
	PAV	.284 (.183, .177)	.289 (.128, .127)	.299 (.081, .080)
	(SUP+PAV)/2	.307 (.183, .160)	.310 (.128, .120)	.302 (.081, .078)

SUP=supremum, $\hat{\Lambda}_1^S(t)$, PAV=Pool-Adjacent-Violators algorithm, $\hat{\Lambda}_1^P(t)$

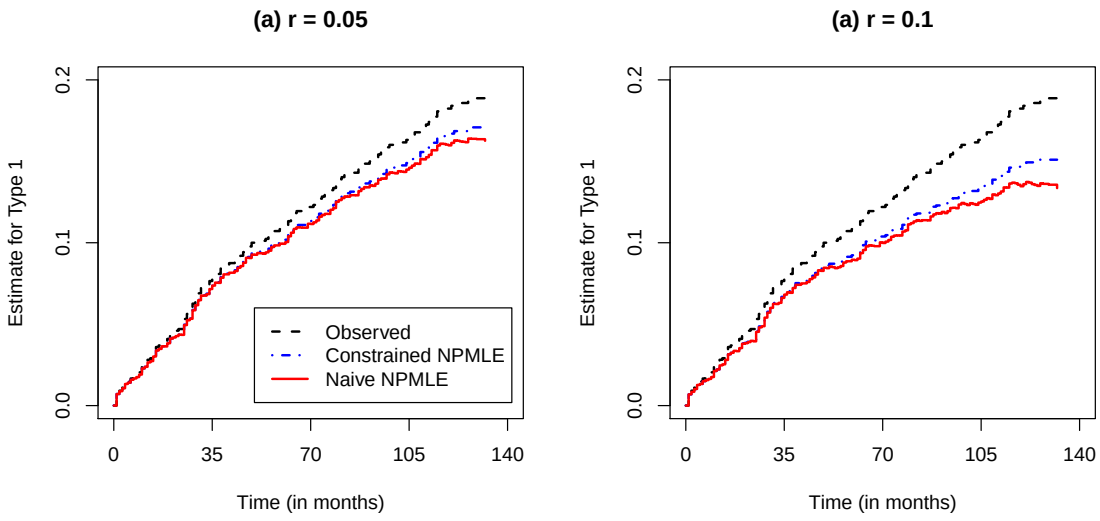

Figure 2.1: Naïve and constrained NPMLE of cumulative type 1 failure hazards when (a) $r = 0.05$ and (b) $r = 0.1$ with a sample size of 1,094 in the discrete setting.

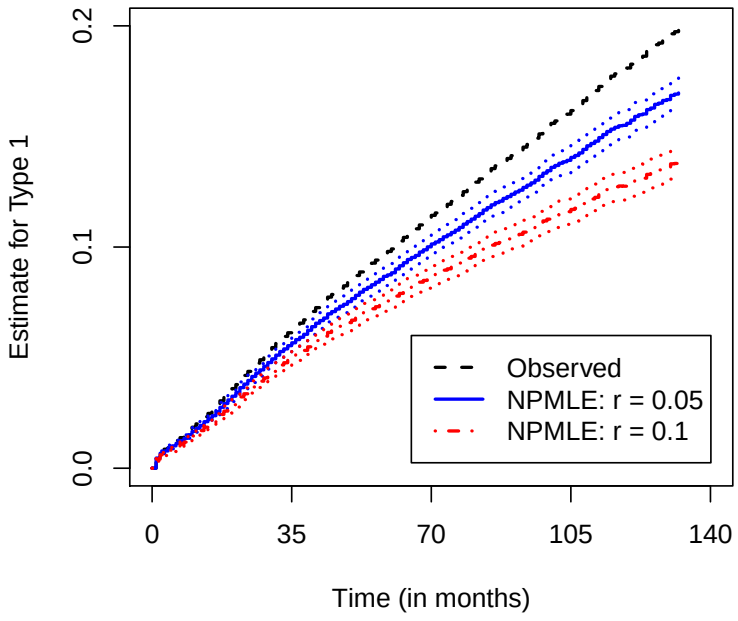

Figure 2.2: Naïve NPMLE of cumulative type 1 failure hazards when $r = 0.05$ and 0.1 with a sample size of 25,088 in the discrete setting. The dotted lines represent 95% point-wise confidence limits.

CHAPTER III

Semiparametric Estimation in the Proportional Hazard Model Accounting for a Misclassified Cause of Failure

3.1 Introduction

Competing risks data arise in many clinical studies where time to failure is of primary interest. The Cox (1972; 1975) proportional hazards model has been widely used; the effect of covariates is estimated using partial likelihood. Misclassified causes of failures are very common in competing risks data due to various reasons such as inaccuracies of cancer death certificates (Percy et al. (1981)). For example, many authors have pointed out that with the introduction of prostate-specific antigen (PSA) screening in the late 1980s, a proportion of deaths may be mistakenly classified as prostate cancer just because the men were diagnosed with the disease (Feuer et al. (1999); Hoffman et al. (2003)).

Under the missing-at-random assumption (Little and Rubin (1987)), in the sense that the probability of missing cause of failure does not depend on the failure type, analysis of competing risks data has received considerable attention. Lu and Tsiatis (2001) used multiple imputation procedures to impute missing causes of failures using the probability that a missing cause is the cause of interest deduced from the complete cases. Goetghebeur and Ryan (1995) proposed an approach that utilizes two types of partial likelihoods assuming that the baseline hazard function for the failure of

interest is proportional to the one for the other cause (proportionality assumption). Lu and Tsiatis (2005) derived a semiparametric efficient score function which also allows the dependence of the probability of missing cause of failure on covariates. Recently, several authors have investigated the approaches which do not require any parametric assumptions between the two baseline hazard functions. Gao and Tsiatis (2005) derived the augmented inverse probability weighted complete-case-estimator with linear transformation models under the missing-at-random mechanism. Lu and Liang (2008) applied their approach to the semiparametric additive hazards model. In addition, Chen et al. (2009) adopted the sieve approach by approximating the two baseline hazard rates using piecewise constant functions.

However, the missing-at-random (MAR) assumption may not be true in practice such as our prostate cancer data mentioned above. A proportionality assumption is often unrealistic. Therefore, the primary goal of the present study is to develop an estimating equation approach for estimation of regression parameters in Cox proportional hazards model that does not require a proportionality assumption under a non missing-at-random (NMAR) mechanism. First, we show that the standard partial likelihood is no longer identical with the profile likelihood and introduce a semiparametric efficient score function which requires a parametric assumption between the two baseline hazard functions. In the next section, we define a function motivated by the Kullback-Leibler divergence. By making a few adjustments to the corresponding score functions, we propose Kullback-Leibler estimating equation which does not require any assumption for baseline hazards. Martingale theory is applied to establish consistency and asymptotic properties of the resulting estimators. A simulation is provided to study the performance of the estimators.

3.2 Assumption and notation

In this project, we consider a sample of n independent individuals, each of whom can fail from one of two possible causes, which we term type 1 (prostate cancer) and 2 (other causes), respectively, or can be subject to independent right censoring mechanism (type 0). Typically, the data for individual i consist of $(T_i, \Omega_i, Z_i, X_i)$, where T_i is time of failure or censoring, Ω_i is an indicator taking values 0, 1 or 2 which correspond to censoring, type 1 or 2 failure, respectively, Z and X are p and q-variate vector of covariates related to cause-specific hazards, possibly time dependent. To avoid nonidentifiability problems, we assume independence of competing risks in the rest of study.

Suppose the effect of (Z, X) on the cause-specific hazard rates is multiplicative, namely,

$$\lambda_1(t|Z, X) = \lambda_1(t) \exp(\beta^T Z(t)) := \lambda_1(t)\theta_1(t; \beta),$$
$$\lambda_2(t|Z, X) = \lambda_2(t) \exp(\gamma^T X(t)) := \lambda_2(t)\theta_2(t; \gamma),$$

where $:=$ is equality by definition, and X may or may not overlap with Z. When there are some misclassified failures, we observe the indicator ω_i for subject i instead of true failure-type indicator Ω_i. Then, the observed data are summarized as $\{T_i, Z_i, X_i, A_i, \omega_i\}$, independent across i. Here, A are auxiliary covariates that are not used to model the hazards but may be used to describe the missingness mechanism.

In general, failures can be misclassified to either type 1 or type 2 cause with different probabilities. For simplicity in the present development we assume that only type 1 failures can be misattributed to type 2 cause (*over-misattribution*). This assumption is motivated by prostate cancer data (Hoffman et al. (2003)). Over-

misattribution can depend on T and $W = (Z, X, A)$; that is,

$$Pr[\omega_i = 1|\Omega_i = 2, T_i = t, W_i] = r(T_i = t, W_i) := r_i(t),$$
$$Pr[\omega_i = 2|\Omega_i = 1, T_i = t, W_i] = 0.$$

Unfortunately, this misattribution probability cannot be estimated with the data structure of this project. External studies are necessary to establish the underlying true causes of failures for a random sample of the data (Percy et al. (1981); Hoffman et al. (2003); Fall et al. (2008)). To avoid the nonidentifiability problem, we assume r is a known function through this study.

3.3 Parameter Estimation

Let $(N_{i1}(t), N_{i2}(t))$ be the counting process for failures where $N_{ij}(t)$ represents the number of observed type j failures for individual i up to time t. To simplify the notation, we write $\lambda_{i1}(t;\beta)$ and $\lambda_{i2}(t;\gamma)$ instead of $\lambda_1(t)\theta_{i1}(t;\beta)$ and $\lambda_2(t)\theta_{i2}(t;\gamma)$, respectively.

It is straightforward to show that the full loglikelihood is proportional to

$$\sum_{i=1}^{n} \int dN_{i1}(t) \log\{\lambda_{i1}(t;\beta) + r_i(t)\lambda_{i2}(t;\gamma)\} + dN_{i2}(t) \log\{\bar{r}_i(t)\lambda_{i2}(t;\gamma)\}$$
$$- \sum_{j=1}^{n} Y_j(t)\{\lambda_{j1}(t;\beta) + \lambda_{j2}(t;\gamma)\}dt,$$

where $\bar{a} = 1-a$ for any a. The corresponding score functions for $\eta = (\lambda_1(t), \lambda_2(t), \beta, \gamma)$

are given by

$$
\begin{aligned}
&\sum_i dN_{i1}(t)\rho_i(t;\eta) - Y_i(t)\lambda_{i1}(t;\beta)dt = 0,\\
&\sum_i dN_{i1}(t)\bar{\rho}_i(t;\eta) + dN_{i2}(t) - Y_i(t)\lambda_{i2}(t;\gamma)dt = 0,\\
&\sum_i \int Z_i(t)\{dN_{i1}(t)\rho_i(t;\eta) - Y_i(t)\lambda_{i1}(t;\beta)dt\} = 0,\\
&\sum_i \int X_i(t)\{dN_{i1}(t)\bar{\rho}_i(t;\eta) + dN_{i2}(t) - Y_i(t)\lambda_{i2}(t;\gamma)dt\} = 0,
\end{aligned}
\tag{3.1}
$$

where

$$
\rho_i(t;\eta) = \frac{\lambda_{i1}(t;\beta)}{\lambda_{i1}(t;\beta) + r_i(t)\lambda_{i2}(t;\gamma)}. \tag{3.2}
$$

Unfortunately, we cannot derive a profile likelihood by eliminating functional nuisance parameters ($\lambda_1(t)$, $\lambda_2(t)$) without a restriction. Restricting the hazard ratio to $\phi(t)$, we have the profile likelihood, L_n for ($\phi(t) = \lambda_2(t)/\lambda_1(t)$, β, γ) assuming that $\hat{\lambda}_1(t;\phi(t),\beta,\gamma)dt = dN(t)/\sum_j Y_j(t)\{\theta_{j1}(t;\beta) + \theta_{j2}(t;\gamma)\phi(t)\}$ is the NPMLE of $\lambda_1(t)dt$ for a given value of ($\phi(t)$, β, γ),

$$
\begin{aligned}
\log L_n = \sum_{i=1}^{n} \int & dN_{i1}(t)\log\{\theta_{i1}(t;\beta) + r_i(t)\theta_{i2}(t;\gamma)\phi(t)\} + dN_{i2}(t)\log\{\bar{r}_i(t)\theta_{i2}(t;\gamma)\phi(t)\}\\
&- dN_i(t)\log\{\sum_j Y_j(t)\{\theta_{j1}(t;\beta) + \theta_{j2}(t;\gamma)\phi(t)\}\},
\end{aligned}
$$

where $dN_i(t) = dN_{i1}(t) + dN_{i2}(t)$ and $\sum_i dN_i(t) = dN(t)$. Note that this is not the standard partial likelihood built with the conditional probabilities of an event of specified type given the type of event. In fact, it is equivalent to the more informative partial likelihood (Kalbfleisch and Prentice (1980), pp.170-1) that it utilizes the conditional probabilities of an event of specified type, given that one event occurs. The score function of $\phi(t)$ is given by the functional derivative

$$
\frac{\partial \log L_n}{\partial \phi(t)} = \sum_i dN_{i1}(t)\bar{\rho}_i(t;\phi(t),\beta,\gamma) + dN_{i2}(t) - Y_i(t)\theta_{i2}(t)\phi(t)\hat{\lambda}_1(t;\phi(t),\beta,\gamma)dt = 0,
$$

where $\rho_i(t;\phi(t),\beta,\gamma) = \theta_{i1}(t;\beta)/[\theta_{i1}(t;\beta) + r_i(t)\theta_{i2}(t;\gamma)\phi(t)]$. Unfortunately, these equations do not lead to suitable estimators. In the classical survival setting, it is generally assumed that there is at most one jump for each subject at time t. Asymptotically unbiased nonparametric maximum likelihood estimator (NPMLE) for $\phi(t)$ cannot be derived in the continuous setting. It leads to a breakdown of the EM algorithm as described in the case of no covariates (Chapter II).

One simple way to avoid this problem is to put a parametric assumption on $\phi(t)$. Let α denote the unknown parameters for time t: $\phi(t;\alpha)$. The most commonly used assumption for $\phi(t;\alpha)$ is that the ratio between two cause-specific baseline hazards is constant.

$$\lambda_2(t) = \lambda_1(t)\exp(\alpha).$$

It is often referred to as the proportionality assumption (Goetghebeur and Ryan (1995); Lu and Tsiatis (2005)). Then, we have the following score functions for $\xi = (\alpha, \beta, \gamma)$, $U_n(\xi)$

(3.3)

$$\frac{\partial \log L_n}{\partial \alpha} = \sum_i \int dN_{i1}(t)\bar{\rho}_i(t;\xi) + dN_{i2}(t) - Y_i(t)\theta_{i2}(t;\gamma)\exp(\alpha)\hat{\lambda}_1(t;\xi)dt = 0,$$

$$\frac{\partial \log L_n}{\partial \beta} = \sum_i \int Z_i(t)\{dN_{i1}(t)\rho_i(t;\xi) - Y_i(t)\theta_{i1}(t)\hat{\lambda}_1(t;\xi)dt\} = 0,$$

$$\frac{\partial \log L_n}{\partial \gamma} = \sum_i \int X_i(t)\{dN_{i1}(t)\bar{\rho}_i(t;\xi) + dN_{i2}(t) - Y_i(t)\theta_{j2}(t)\exp(\alpha)\hat{\lambda}_1(t;\xi)dt\} = 0$$

where

$$\rho_i(t;\xi) = \frac{\theta_{i1}(t;\beta)}{\theta_{i1}(t;\beta) + r_i(t)\theta_{i2}(t;\gamma)\exp(\alpha)},$$

$$\hat{\lambda}_1(t;\xi)dt = \frac{dN(t)}{\sum_j Y_j(t)\{\theta_{j1}(t;\beta) + \theta_{j2}(t;\gamma)\exp(\alpha)\}}.$$

The profile likelihood estimator $\hat{\xi}_n$ is semiparametric efficient within the class of regular and asymptotically linear (RAL) estimators for ξ (Appendix .4). Note that

$dN_1\rho$ and $dN_1\bar{\rho}+dN_2$ can be considered as an estimate of the number of true failures from type 1 and 2 cause, $d\hat{N}_1^0$ and $d\hat{N}_2^0$, respectively. The estimators $\hat{\xi}_n$, $d\hat{N}_1^0$ and $d\hat{N}_2^0$ are consistent and asymptotically unbiased, given correct specification of $\phi(t)$.

3.4 Kullback-Leibler Estimator

It is convenient to introduce the following notation

$$S(t;m_j)=\sum_i Y_i(t)\theta_{ij}(t;m_j), \qquad S_A(t;m_j)=\sum_i Y_i(t)A_i(t)\theta_{ij}(t;m_j),$$

for $j=1,2$, $m_1=\beta$, $m_2=\gamma$ and $A\in\{Z,X,r,\bar{r},\bar{r}X\}$.

Consider the martingale processes

$$\begin{aligned} M_{i1}(t;\eta^0) &= \int_0^t dN_{i1}(s) - Y_i(s)\{\lambda_{i1}^0(s;\beta^0)+r_i(s)\lambda_{i2}^0(s;\gamma^0)\}ds, \\ M_{i2}(t;\eta^0) &= \int_0^t dN_{i2}(s) - Y_i(s)\bar{r}_i(s)\lambda_{i2}^0(s;\gamma^0)ds, \end{aligned} \tag{3.4}$$

where $Y_i(s)$ indicates whether or not individual i is at risk at time s, $Y_i(s)=I(T_i\geq s)$, and a^0 denotes the true value of a.

The true functions $\lambda_{i1}^0(t;\beta^0)+r_i(t)\lambda_{i2}^0(t;\gamma^0)$ and $\bar{r}_i(t)\lambda_{i2}^0(t;\gamma^0)$ represent the true hazards of observed type 1 and type 2 failure for subject i, respectively. Their empirical estimates can be defined as $dN_{i1}(t)/Y_i(t)$ and $dN_{i2}(t)/Y_i(t)$ where $Y_i(t)$ is the empirical estimate for the survival function $F_i^0(t)=P(T_i\geq t)$.

Introduce a function $\upsilon(x)=\log x-x+1$ which has a unique maximum of 0 when $x=1$. Using this function, we define a mean risk functional

$$\begin{aligned} KL_\infty = -\int E_W \Bigg[\Bigg\{ (\lambda_1^0(t|Z)+r(t,W)\lambda_2^0(t|X))\upsilon\left(\frac{\lambda_1(t|Z)+r(t,W)\lambda_2(t|X)}{\lambda_1^0(t|Z)+r(t,W)\lambda_2^0(t|X)}\right) \\ + \bar{r}(t,W)\lambda_2^0(t|X)\upsilon\left(\frac{\lambda_2(t|X)}{\lambda_2^0(t|X)}\right)\Bigg\} F^0(t)\Bigg] dt, \end{aligned}$$

where W is the combined covariate vector $W=(Z,X,A)$, and $\lambda_1^0(t|Z)$=$\lambda_1^0(t;\beta^0)$ and $\lambda_2^0(t|X)$=$\lambda_2^0(t;\gamma^0)$ are the true hazards. KL_∞ represents an expectation of $\upsilon(x)$ with

x being a hazard ratio of a model versus the true model at an observed data point averaged over the data generating mechanism. The true model quantities indexed by 0 will be replaced by their empirical counterparts to derive the estimates. It is proportional to

(3.5)

$$-\int E_W[\{(\lambda_1^0(t|Z)+r(t,W)\lambda_2^0(t|X))\log(\lambda_1(t|Z)+r(t,W)\lambda_2(t|X)) + \bar{r}(t,W)\lambda_2^0(t|X)\log(\bar{r}(t,W)\lambda_2(t|X)) - (\lambda_1(t|Z)+\lambda_2(t|X))\}F^0(t)]dt.$$

KL_∞ is similar to the *Kullback-Leibler divergence*. It is often considered a measure of "distance" between the *true* probability distribution and some *model* probability distribution, although it does not have the properties of a mathematical distance. In our case the divergence is between $\{\lambda_1^0(t|Z)+r(t,W)\lambda_2^0(t|X)\}F^0(t)$ and $\{\lambda_1(t|Z)+r(t,W)\lambda_2(t|X)\}F^0(t)$, and between $\bar{r}(t,W)\lambda_2^0(t|X)F^0(t)$ and $\bar{r}(t,W)\lambda_2(t|X)F^0(t)$. Clearly, KL_∞ is minimized when the *model* and *true* functions are same ($v(x)=0, x=1$): $\lambda_1^0(t|Z)=\lambda_1(t|Z)$ and $\lambda_2^0(t|X)=\lambda_2(t|X)$ which is a desirable result. Note that this approach yields the maximum likelihood estimator when the true quantities are replaced by their empirical counterparts. The empirical counterpart of KL_∞ in (3.5) turns out to be the negated full loglikelihood divided by sample size n.

The first functional derivative of KL_∞ with respect to $\lambda_1(t)$ is given by

(3.6)
$$E_W[\{-(\lambda_1^0(t|Z)+r(t,W)\lambda_2^0(t|X))\frac{\lambda_1(t|Z)}{\lambda_1(t|Z)+r(t,W)\lambda_2(t|X)} + \lambda_1(t|Z)\}F^0(t)]$$

(3.7)
$$=E_W[\{-(\lambda_1^0(t|Z)+r(t,W)\lambda_2^0(t|X))\left(1-\frac{r(t,W)\lambda_2(t|X)}{\lambda_1(t|Z)+r(t,W)\lambda_2(t|X)}\right) + \lambda_1(t|Z)\}F^0(t)].$$

It is straightforward to show that an empirical version of (3.6) is identical with the full likelihood score function of $\lambda_1(t)$ in (3.1). However, asymptotic bias of nuisance

parameter $\phi(t)$ is a major problem in deriving consistent estimators from a profile likelihood. Thus, we eliminate $\phi(t)$ by modifying the form of KL_∞ and replace the model-based denominator $\lambda_1(t|Z) + r(t,W)\lambda_2(t|X)$ with the corresponding true hazard function $\lambda_1^0(t|Z) + r(t,W)\lambda_2^0(t|X)$ in the above equation. This does not disrupt the estimating mechanism in the sense that KL_∞ is still minimized when $\lambda_1^0(t|Z) = \lambda_1(t|Z)$ and $\lambda_2^0(t|X) = \lambda_2(t|X)$. However, the empirical version of KL_∞ will now be different and will no longer have MLE form. Note that this adjustment makes the equation (3.6) identically zero. The modified equation (3.7) is used as the basis for the corresponding empirical counterpart. Applying similar modifications to the equations for $\lambda_2(t)$, β and γ, we have the following *Kullback-Leibler* estimating equations for η, $U_n^{KL} = \sum_i U_i^{KL}(.)$:

$$
\begin{aligned}
&\sum_i dM_{i1}(t;\eta) = 0,\\
&\sum_i dM_{i2}(t;\eta) = 0,\\
&\sum_i \int Z_i(t) dM_{i1}(t;\eta) = 0,\\
&\sum_i \int X_i(t) dM_{i2}(t;\eta) = 0.
\end{aligned}
\tag{3.8}
$$

We denote the *Kullback-Leibler* estimator (KLE) representing the solutions of (3.8) by $\hat{\lambda}_1^{KL}$, $\hat{\lambda}_2^{KL}$, $\hat{\beta}_n^{KL}$ and $\hat{\gamma}_n^{KL}$. It can be also derived by replacing model-based ρ_i in (3.1) with the empirical $(dN_{i1} - Y_i r_i \lambda_{i2} dt)/dN_{i1}$.

From the first two equations in (3.8), we have Breslow-type estimators for fixed (β, γ).

$$
\begin{aligned}
\hat{\lambda}_1^{KL}(t;\beta,\gamma)dt &= \frac{1}{S(t;\beta)}\left\{dN_1(t) - \frac{S_r(t;\gamma)}{S_{\bar{r}}(t;\gamma)} dN_2(t)\right\},\\
\hat{\lambda}_2^{KL}(t;\gamma)dt &= \frac{dN_2(t)}{S_{\bar{r}}(t;\gamma)}.
\end{aligned}
$$

Note that $\hat{\lambda}_1^{KL}$ can be negative; thus, the corresponding estimates of cumulative hazards $\int \hat{\lambda}_1^{KL} dt$ can decrease at some time intervals. Interestingly, the negative estimates are necessary to preserve consistency of $\int \hat{\lambda}_1^{KL} dt$. Isotonic approaches such as the supremum function and pooled-adjacent-violators algorithm can be used to achieve monotonicity (Chapter II).

Estimating equations for β and γ are summarized as

$$\sum_i \int \left\{ dN_{i1}(t) - Y_i(t) r_i(t) \theta_{i2}(t,\gamma) \frac{dN_2(t)}{S_{\bar{r}}(t,\gamma)} \right\} \left\{ Z_i(t) - \frac{S_Z(t;\beta)}{S(t;\beta)} \right\} = 0,$$
$$\sum_i \int dN_{i2}(t) \left\{ X_i(t) - \frac{S_{\bar{r}X}(t;\gamma)}{S_{\bar{r}}(t;\gamma)} \right\} = 0.$$

KLE is computationally efficient because concave functions can be deduced from the above equations. First, it is easy to see that $\hat{\gamma}_n^{KL}$ maximizes the following function

$$\sum_i \int dN_{i2}(t) \log \frac{\theta_{i2}(t;\gamma)}{S_{\bar{r}}(t;\gamma)}.$$

For a fixed $\hat{\gamma}_n^{KL}$, $\hat{\beta}_n^{KL}$ is obtained by maximizing

$$\sum_i \int \left\{ dN_{i1}(t) - Y_i(t) r_i(t) \theta_{2i}(t;\hat{\gamma}_n^{KL}) \frac{dN_2(t)}{S_{\bar{r}}(t;\hat{\gamma}_n^{KL})} \right\} \log \frac{\theta_{i1}(t;\beta)}{S(t;\beta)}.$$

3.5 Asymptotic Properties

3.5.1 Profile likelihood estimator

If the model $\phi(t;\alpha) = \exp(\alpha)$ is correctly specified, the martingale process (3.4) can be written as

$$M_{i1}(t;\xi^0) = \int dN_{i1}(t) - Y_i(t)[\theta_{i1}(t;\beta^0) + r_i(t)\theta_{i2}(t;\gamma^0)\exp(\alpha^0)]\lambda_1^0(t)dt,$$
$$M_{i2}(t;\xi^0) = \int dN_{i2}(t) - Y_i(t)\bar{r}_i(t)\theta_{i2}(t;\gamma^0)\exp(\alpha^0)\lambda_1^0(t)dt.$$

Then, U_n is a martingale transform at the true value.

$$U_n(\xi^0) =$$

$$\sum_i \int dM_{i1}(t;\xi^0)\frac{\dot{h}_{i1}(t;\xi)_\xi}{h_{i1}(t;\xi)} + dM_{i2}(t;\xi^0)\frac{\dot{h}_{i2}(t;\xi)_\xi}{h_{i2}(t;\xi)} - dM_i(t;\xi^0)\frac{\sum_j Y_j(t)\dot{h}_{j\cdot}(t;\xi)_\xi}{\sum_j Y_j(t)h_{j\cdot}(t;\xi)}\Bigg|_{\xi=\xi^0},$$

where $dM_i = dM_{i1} + dM_{i2}$, $h_{i1}(t;\xi) = \theta_{i1}(t,\beta) + r_i(t)\theta_{i2}(t,\gamma)\exp(\alpha)$ and $h_{i2}(t;\xi) = \bar{r}_i(t)\theta_{i2}(t,\gamma)\exp(\alpha)$. Here and in the sequel, $\dot{f}(t;x)_x$ denotes the first derivative of f with respect to x: $\partial f(t;x)/\partial x$. Consistency and asymptotic normality of the estimators $\hat{\xi}_n$ are easily proved by using the arguments of Andersen and Gill (1982). The asymptotic covariance matrix of $n^{1/2}(\hat{\xi}_n - \xi^0)$ is the inverse of $I(\xi^0) = E[-n^{-1}\dot{U}_n(\xi)_\xi|_{\xi=\xi^0}]$.

The assumption of constant baseline hazard ratio is crucial in the profile likelihood approach. If the assumption is violated, it is clear that the asymptotic properties discussed above are not valid anymore. Specifically, $\hat{\xi}_n$ is asymptotically consistent for a value ξ^* that maximizes the function $E[\log L_n(\xi)]$, where the expectation is taken under the true underlying distribution (van der Vaart (1998), pp.55-6). Namely, $\xi^* = (\alpha^*, \beta^*, \gamma^*)$ is the unique solution to the equations

$$\int E\left[Y(t)\{\theta_1(t;\beta^0)\lambda_1^0(t) + r(t)\theta_2(t;\gamma^0)\lambda_2^0(t)\}\left\{\frac{\dot{h}_1(t;\xi)_\xi}{h_1(t;\xi)} - \frac{E[Y(t)\dot{h}_\cdot(t;\xi)_\xi]}{E[Y(t)h_\cdot(t;\xi)]}\right\}\right]$$
$$+E\left[Y(t)\bar{r}(t)\theta_2(t;\gamma^0)\lambda_2^0(t)\left\{\frac{\dot{h}_2(t;\xi)_\xi}{h_2(t;\xi)} - \frac{E[Y(t)\dot{h}_\cdot(t;\xi)_\xi]}{E[Y(t)h_\cdot(t;\xi)]}\right\}\right]dt = 0.$$

$n^{1/2}(\hat{\xi}_n - \xi^*)$ is asymptotically normal with mean zero and with a covariance matrix, $\Sigma^* = I^{*-1}V^*I^{*-1T}$, the so-called "sandwich" estimator. $I^*(\xi^*)$ can be consistently estimated by $-n^{-1}\partial^2 \log L_n(\xi)/\partial\xi^2|_{\xi=\xi^*}$. The main difference is the matrix V^* because $U_n(\xi^*)$ is no longer a martingale integral due to a misspecified model. However, $n^{-1/2}U_n(\xi^*)$ is asymptotically equivalent to $n^{-1/2}\sum_i u_i(\xi^*)$ by an argument similar to theorem 2.1 by Lin and Wei (1989). By the multivariate Central Limit Theorem,

it follows an asymptotically zero mean normal distribution with covariance matrix $E[u(\xi^*)^{\otimes 2}]$ where $A^{\otimes 2} = AA^T$. It is consistently estimated by $n^{-1}\sum_i \hat{u}_i(\xi^*)^{\otimes 2}$, where

$$\hat{u}_i(\xi) = U_i(\xi) - \int Y_i(t)\left\{\dot{h}_i(t;\xi)_\xi - \frac{\sum_j Y_j(t)\dot{h}_j(t;\xi)_\xi}{\sum_j Y_j(t)h_j(t;\xi)}h_i(t;\xi)\right\}\frac{dN(t)}{\sum_j Y_j(t)h_j(t;\xi)}$$

and $U_i(\xi)$ is the contribution from the ith observation to the score function $U_n(\xi)$.

3.5.2 Kullback-Leibler estimator

The KL estimating equation U_n^{KL} is a martingale at the true value:

$$U_n^{KL}(\beta^0) = \sum_i \int \left\{dM_{i1}(t) - Y_i(t)r_i(t)\theta_{i2}(t;\gamma^0)\frac{dM_2(t)}{S_{\bar{r}}(t;\gamma^0)}\right\}\left\{Z_i(t) - \frac{S_Z(t;\beta^0)}{S(t;\beta^0)}\right\} = 0,$$

$$U_n^{KL}(\gamma^0) = \sum_i \int dM_{i2}(t)\left\{X_i(t) - \frac{S_{\bar{r}X}(t;\gamma^0)}{S_{\bar{r}}(t;\gamma^0)}\right\} = 0.$$

The consistency of estimators can be proved in almost identical fashion to that of Lu and Ying (2004). It then follows that $n^{-1/2}U_n^{KL}$ converges weakly to a $(p+q)$-variate normal with mean zero and with a covariance matrix V^{KL}. However, V^{KL} is no longer equivalent to I^{KL} which is the limit of $-n^{-1}\dot{U}_n^{KL}(\xi)_\xi|_{\xi=\xi^0}$. This is a typical characteristic of $Z-$*estimators* (for zero) obtained by solving estimating equations. Under suitable classical conditions (van der Vaart (1998), pp.67-70), $n^{1/2}(\hat{\beta}_n^{KL} - \beta^0)$ converges weakly to a p-variate normal with mean zero and with a covariance matrix $\Sigma^{KL}|_{\beta\beta} = [I^{KL-1}V^{KL}I^{KL-1T}]|_{\beta\beta}$, the "sandwich" estimator;

$$I^{KL} = \begin{pmatrix} I_{\beta\beta}^{KL} & I_{\beta\gamma}^{KL} \\ 0 & I_{\gamma\gamma}^{KL} \end{pmatrix}, \qquad V^{KL} = \begin{pmatrix} V_{\beta\beta}^{KL} & 0 \\ 0 & I_{\gamma\gamma}^{KL} \end{pmatrix},$$

where $A|_{\beta\beta}$ denotes the upper left-hand quadrant of the matrix A corresponding to β. The formulations of I^{KL} and V^{KL} are given in the Appendix .5.

3.6 Simulation

The univariate covariate Z was chosen to take values 1 or 0 with equal probabilities 0.5 and $X = 0$. Given Z, the cause-of-interest failure time T_1 follows an exponential

distribution with hazard function $\lambda_1(t|Z) = \lambda_1 \exp(\beta Z)$. The other failure time T_2 follows an exponential distribution with hazard function $\lambda_2(t|X) = \lambda_2$ in scenario 1 and a Gompertz distribution with hazard function $\lambda_2(t|X) = \alpha_1 \exp(\alpha_2 t)$ in scenario 2. Thus, a proportionality assumption is not correct in scenario 2. Also, assume that there is no censoring. Then, for the observed failure time $T = \min(T_1, T_2)$, misattribution is assumed to have a logistic regression model: $logit\{r(T, W)\} = \psi_0 + \psi_1 T + \psi_2 Z + \psi_3 A$ where the auxiliary covariate A follows a standard normal distribution. We chose $\lambda_1 = 1, \lambda_2 = 0.7, \alpha_1 = \alpha_2 = 0.5, \psi_0 = 0.1, \psi_1 = \psi_2 = -0.5$ and $\psi_3 = 0.1$. For each $\beta \in \{-2, 0, 2\}$, these yield, on average, $(63\%, 41\%, 25\%)$ and $(63\%, 40\%, 24\%)$ failures from the cause of type 2 in scenario 1 and 2, respectively. We carried out $1,000$ simulations with a sample size of 300. The simulation results are summarized in Table 3.1: the sample standard deviations and the average of the standard error estimates are given in parenthesis.

The simulation results are consistent with the theoretical results. Both profile and KL estimators work well when $\phi(t)$ is correctly specified and the profile likelihood estimator $\hat{\beta}_n$ is more efficient as expected. KL estimator is almost as efficient as the profile one under the null hypothesis $H_0 : \beta^0 = 0$.

In case of misspecified $\phi(t)$, $\hat{\beta}_n$ is biased while $\hat{\beta}_n^{KL}$ is unbiased asymptotically. Interestingly, the KL estimator is more efficient than the profile one in this simulation.

3.7 Discussion

A general estimating equation procedure has enjoyed considerable attention in semiparametric analysis of survival data. In cases where the partial likelihood approach cannot be used to eliminate a nuisance parameter, estimating equations based on a martingale structure have been considered as an alternative (Lin and Ying

(1994); Chen et al. (2002)). Our Kullback-Leibler estimating procedure is similar in spirit to their approaches. However, we provided a motivation for derivation of the equations using Kullback-Leibler divergence.

We found that the standard profile likelihood is not applicable with misattributed causes of failures. As is well known, Breslow estimator of baseline hazard rate, $d\hat{\Lambda}$, in the Cox model is not consistent even when there is no attribution bias (Burr (1994)). The main problem is that its asymptotically unbiased nonparametric estimator of $\phi(t)$ cannot be derived in the continuous survival setting targeting step-function estimators for hazards. This fact has led many authors to a parametric $\phi(t)$ or use smoothing. A key contribution of this article is that it provides a formal estimating procedure for all cumulative hazards in the step-function setting without any assumption restricting the ratio.

The score function derived from the profile likelihood is semiparametric efficient, whereas consistency of the resulting estimators relies entirely on a parametric assumption for $\phi(t)$. A simulation study showed that the profile estimator can be seriously biased and lose efficiency when a proportionality assumption on $\phi(t)$ is violated.

Although Goetghebeur and Ryan (1995) proposed partial likelihood approach which is quite robust against misspecification, inconsistency of the estimator with a misspecified $\phi(t)$ is still an issue. Of course, the model assumption for $\phi(t)$ can be relaxed, if necessary; a piecewise constant model (Chen et al. (2009)) or smoothing is a possibility. However, our approach is still appealing when restrictions are undesirable for risk of bias. Also, our method can be used to suggest a pertinent form of $\phi(t)$ as a model-building step before this form is enforced in a parametric fashion.

Derivation of the estimating equations from the KL function can be done in a

variety of ways. Further research is needed to understand which ways lead to better estimators and whether the concept of efficiency can be formulated in a setting where MLE does not make sense. A study of efficiency among RAL estimators assuming no parametric assumptions on the cause-specific baseline hazards is worth pursuing. We compared our approach to other martingale-based estimating equations with some examples, and Kullback-Leibler estimators turn out to be most efficient (the results are omitted from this article).

Table 3.1: Simulation results for the covariate effect estimated using the profile likelihood and Kullback-Leibler estimator based on 1000 simulations with $n = 300$. The sample standard deviations and the average of the standard error estimates are given in parenthesis.

$\phi(t)$	β	$E[r]$	$E[\bar{\rho}]$	Profile	KLE
Constant	-2	0.341	0.364	-2.042 (.377, .386)	-2.063 (.458, .468)
	0	0.395	0.218	0.001 (.174, .173)	0.001 (.175, .174)
	2	0.443	0.129	2.006 (.180, .182)	2.012 (.208, .207)
Time-dependent	-2	0.336	0.364	-2.246 (.504, .531)	-2.064 (.452, .475)
	0	0.383	0.206	0.015 (.182, .180)	0.012 (.171, .170)
	2	0.435	0.118	2.108 (.189, .189)	2.016 (.197, .196)

$\phi(t) = \lambda_2(t)/\lambda_1(t)$: a proportionality assumption is violated when $\phi(t)$ is time-dependent.

$E[r]$: mean of estimates of the probability that failures due to type 2 cause are misclassified as type 1 cause.

$E[\bar{\rho}]$: mean of estimates of the probability that observed type 1 failures are misclassified as type 2 cause.

CHAPTER IV

Adjusted Prostate Cancer Mortality Rates under Misattributed Cause of Death

4.1 Introduction

Prostate cancer is the second leading cause of cancer mortality among American men. Since the prostate-specific antigen (PSA) screening test was introduced in the late 1980s, increasingly many men get a diagnosis of prostate cancer. Despite the PSA test's successful dissemination, there has been considerable debate and uncertainty as to its benefits.

Generally, mortality response to a diagnostic intervention lags incidence as it takes the duration of post-treatment survival for the mortality effect to exhibit itself. However, Statistics from the Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute show that US prostate cancer mortality trends follow the incidence rates (Figure 1.1) that peak for a few years around 1991 when PSA testing was at its highest intensity.

The most credible explanation for the trend in mortality rates seems to be the attribution bias in the cause of death. Namely, some of deaths of other causes are mistakenly attributed to prostate cancer just because this disease had been previously diagnosed. Feuer et al. (1999) argued that this phenomenon would lead to a peak in mortality coinciding with the peak in incidence even if the attribution bias were

constant. Hoffman et al. (2003) reviewed the New Mexico Bureau of Vital Statistics' (BVS) assignment of causes of deaths. They have concluded that misattribution could account for 53% of the observed increase in mortality rates. However, little has been done to model the effect quantitatively in its dynamics. Motivated by these studies, we explore the possibility that incorrect attribution of cause of death may have made a substantial contribution to the recent mortality trends, using a statistical model.

The primary goal of the present article is to estimate the adjusted mortality rates in the presence of misattribution of cause of death. We first present a mortality model derived from the prostate cancer specific survival function. Using Kullback-Leibler's estimating equation in Chapter III, misattribution can be explicitly introduced into a semiparametric survival model. New survival estimates adjusted for misattribution are used to derive adjusted mortality rates. Finally, we perform the sensitivity analysis to study variations in the mortality trends over the plausible range of misattribution.

4.2 Mortality Model

Conditional on birth year x, cancer mortality rate by age (at death) a_M is the hazard function, λ_M, given by,

$$\begin{aligned}\lambda_M(a_M|x)da_M &= \lambda_M(a_M, t_M)da_M \\ &= -\frac{dG_M(a_M|x)}{G_M(a_M|x)},\end{aligned}$$

where $t_M = x + a_M$ is the calendar time (year) of death, and $G_M(a_M|x)$ is the probability of surviving from prostate cancer for a man at age a_M born in year x.

This can be represented by the following convolution

$$G_M(a_M|x) = \int_0^{a_M} \sum_{\phi} f_I(a_I, \phi|x) \sum_{\tau} f_\tau(\tau|a_I, \phi, x) G(a_M - a_I|a_I, \phi, \tau, x) da_I + G_I(a_M|x), \tag{4.1}$$

where for a x-year birth cohort,

- $f_I(a_I, \phi|x)$ is the joint probability of age (a_I) and disease characteristics (ϕ) at incidence. In this article, ϕ is represented by four possible combinations of stage and grade classification: local/regional (LR) and distant (D) stage, and well or moderately (WM) differentiated and poorly differentiated or undifferentiated (PU) grade of the cancer: i.e. $\phi = D/PU, D/WM, LR/PU$ or LR/WM.

- $f_\tau(\tau|a_I, \phi, x)$ is the treatment model providing probability of administering treatment τ, given disease presentation at diagnosis. Using SEER data, we classify treatments into three major categories: Watchful Waiting (WW), Radiation Therapy (RT), and Radical Prostatectomy (RP): i.e. $\tau = WW, RT$ or RP. Most likely RT includes a hormonal regimen.

- $G(a_M - a_I|a_I, \phi, \tau, x)$ is the prostate cancer-specific survival function from the point of diagnosis. It describes the probability for a man to survive at least by a_M, given a diagnosis of prostate cancer at the age a_I with ϕ-stage/grade tumor in year $t_I = x + a_I$ and the treatment τ.

- $G_I(a_M|x)$ is the prostate cancer incidence survival function. It represents the probability for a man to have no diagnosis of prostate cancer by the age a_M.

The first term in (4.1) models survival of a man with prostate cancer, while the second term represents the possibility that prostate cancer is not diagnosed by the age a_M.

4.2.1 Survival Model

We consider two competing causes of failures, type 1 cause of interest (prostate cancer) and type 2 cause (other causes), both of which can be censored. The type 1 is our main interest. Let Ω denote the indicator of event: 0 if censored, 1 if type 1 cause, and 2 if type 2 cause. T is the time from diagnosis to death or censoring, and Z is a p-variate vector of covariates. Then, the observed data post-diagnosis are summarized as (T, Ω, Z). To avoid nonidentifiability problems, risks are assumed to be conditionally independent given Z in the rest of the article. Under the proportional hazards assumption, the cause-specific hazards for type 1 and type 2 are given by

$$\begin{aligned} \lambda_j(t|Z) &= \lim_{h\to 0} h^{-1} Pr[t \le T < t+h, \Omega = j | T \ge t, Z] \\ &= \begin{cases} \lambda_1(t)\exp(\beta^T Z) := \lambda_1(t)\theta_1 & \text{if } j = 1, \\ \lambda_2(t)\exp(\gamma^T Z) := \lambda_2(t)\theta_2 & \text{if } j = 2, \end{cases} \end{aligned} \tag{4.2}$$

where ':=' is equality by definition. PSA screening results in the so-called lead time effect on survival. Lead time measures an advance in the diagnosis of prostate cancer due to screening generally accompanied by a favorable stage shift. The bias would be operating even if early detection and treatment were of no benefit. Tsodikov et al. (2006) estimated lead time distribution for prostate cancer. We make the lead time adjustment to the baseline survival G_b (for clinical diagnosis in the absence of screening) as follows

$$\begin{aligned} G(t|a_I, \phi, \tau, x) &= G(t|a_I, \phi, \tau, t_I) \\ &= G_{LT}(t|a_I, \phi, t_I) + \int_0^t f_{LT}(s|a_I, \phi, t_I) G_b(t-s|a_I+s, \phi, \tau, t_I) ds, \end{aligned}$$

where lead time survival function G_{LT} and probability density function f_{LT} depends on age at diagnosis, tumor characteristics and year of diagnosis. Figure 4.1 presents

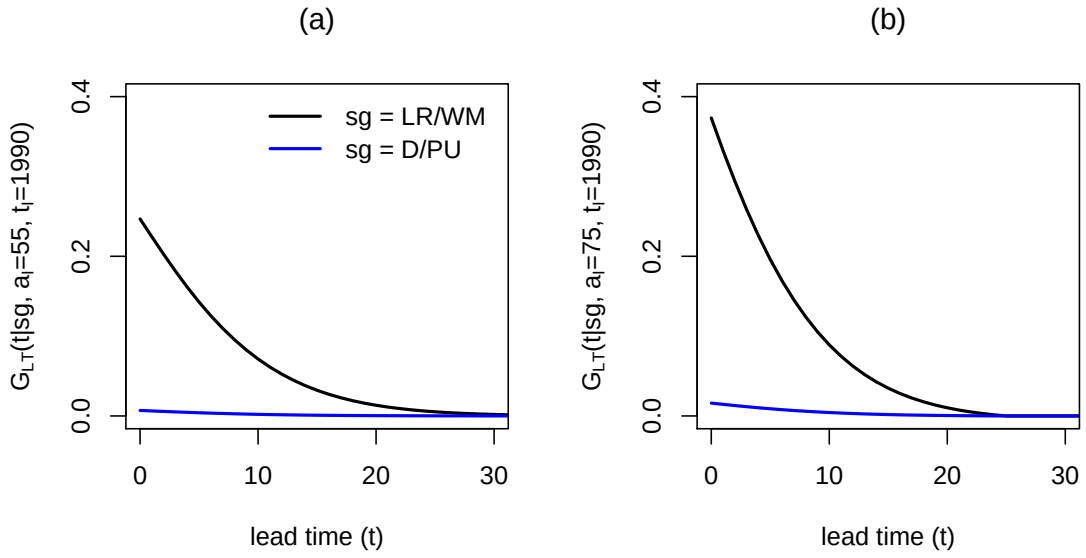

Figure 4.1: Lead time survival probability in year 1990 of diagnosis, t_I, for patients with local/regional stage and well/moderately differentiated grade (LR/WM) and distant stage and poorly/undifferentiated (D/PU) grade of tumor, ϕ, diagnosed at the age of a_I, (a) 55 and (b) 75.

estimates of lead time survival probability for a man diagnosed LR/WM or D/PU tumor at age 55 or 75 in 1990 (Tsodikov et al. (2006)). As expected, lead time is likely to be larger for older patients.

4.3 Misattribution

Here we assume that some of type 2 failures are mistakenly attributed to type 1 cause, and thus the observed type 1 hazard is over-estimated. Under this scenario, we cannot observe the true indicator Ω. Assuming that such an attribution bias can depend on survival time T since diagnosis of prostate cancer, the probability of misattribution can be written as

$$r(t) = Pr[\omega_i = 1 | \Omega_i = 2, T_i = t],$$

where ω is observed indicator of the cause of death. This is often referred to as *over-misattribution* since the number of failures from cause-of-interest (type 1) is inflated. To be specific, the observed hazards are expressed in terms of true cause-specific hazards as follows:

$$\begin{aligned} &\lambda_1(t|Z_i) + r(t)\lambda_2(t|Z_i) && \text{if } \omega_i = 1 \quad \text{; observed prostate cancer death} \\ &(1 - r(t))\lambda_2(t|Z_i) && \text{if } \omega_i = 2 \quad \text{; observed other death.} \end{aligned}$$

To derive adjusted estimates, we use an estimating equation motivated by a Kullback-Leibler information argument (Chapter III) which leads to consistent estimates without any further assumption constraining the two cause-specific baseline hazards functions. First, we define the counting process $N_{ij}(t)$ as the process counting failures of type j cause for a subject i by time t and $Y_i(t)$ as the at-risk indicator process $I(T_i \geq t)$. The corresponding martingale processes are given by

$$\begin{aligned} \sum_i \int_0^t dM_{i1}(s;\eta^0) &= \sum_i \int_0^t dN_{i1}(s) - Y_i(s)\{\theta_{i1}^0 \lambda_1^0(s) + r(s)\theta_{i2}^0 \lambda_2^0(s)\}ds, \\ \sum_i \int_0^t dM_{i2}(s;\eta^0) &= \sum_i \int_0^t dN_{i2}(s) - Y_i(s)(1 - r(s))\theta_{i2}^0 \lambda_2^0(s)ds, \end{aligned}$$

where η^0 is the true value of $\eta = (\lambda_1(\cdot), \lambda_2(\cdot), \beta, \gamma)$.

Kullback-Leibler function (KL_∞) introduced in Chapter III describes the distance between a model-based and true hazards. The attractive feature of KL_∞ is that full-likelihood function converges to the negated KL_∞ as sample size increases; thus, estimators can be derived from an empirical version of KL_∞. Using this approach,

we have proposed estimating equations for η as follows:

$$\begin{aligned}
&\sum_i dM_{i1}(t;\eta) = 0,\\
&\sum_i dM_{i2}(t;\eta) = 0,\\
&\sum_i \int Z_i dM_{i1}(t;\eta) = 0,\\
&\sum_i \int Z_i dM_{i2}(t;\eta) = 0.
\end{aligned} \tag{4.3}$$

From the above equations, Kullback-Leibler estimating equations (KLE) for β and γ are given by

$$\begin{aligned}
&\sum_i \int dN_{i1}(t)\{Z_i - \bar{Z}(t;\beta)\} - dN_{i2}(t)\frac{r(t)}{1-r(t)}\{\bar{Z}(t;\gamma) - \bar{Z}(t;\beta)\} = 0,\\
&\sum_i \int dN_{i2}(t)\{Z_i - \bar{Z}(t;\gamma)\} = 0,
\end{aligned}$$

where

$$\bar{Z}(t;\beta) = \frac{\sum_j Y_j(t) Z_j \theta_{j1}}{\sum_j Y_j(t)\theta_{j1}}, \qquad \bar{Z}(t;\gamma) = \frac{\sum_j Y_j(t) Z_j \theta_{j2}}{\sum_j Y_j(t)\theta_{j2}}.$$

Note that the regression effect for type 2 cause, γ, is not affected by misattribution as long as it does not depend on covariates. Given regression coefficients $\hat{\beta}$ and $\hat{\gamma}$, we have the following Breslow-type estimator for cause-specific baseline hazards from the first two equations in (4.3)

$$\begin{aligned}
\hat{\lambda}_1(t;\hat{\beta},\hat{\gamma})dt &= \frac{1}{\sum_j Y_j(t)\hat{\theta}_{j1}}\left\{dN_1(t) - \frac{r(t)}{1-r(t)}dN_2(t)\right\},\\
\hat{\lambda}_2(t;\hat{\gamma})dt &= \frac{1}{1-r(t)}\frac{dN_2(t)}{\sum_j Y_j(t)\hat{\theta}_{j2}},
\end{aligned}$$

where $\hat{\theta}_{i1} = \exp(\hat{\beta}^T Z_i)$ and $\hat{\theta}_{i2} = \exp(\hat{\gamma}^T Z_i)$. Then, the estimator for the survival function for type 1 cause (prostate cancer) is

$$\begin{aligned}
\hat{G}_1(t|Z) &= \exp\left\{-\int_0^t \hat{\lambda}_1(s;\hat{\beta},\hat{\gamma})\exp(\hat{\beta}^T Z)ds\right\}\\
&= \exp\{-\hat{\Lambda}_1(t;\hat{\beta},\hat{\gamma})\exp(\hat{\beta}^T Z)\}.
\end{aligned}$$

Note that the above estimates can be increasing even though they are consistent. This is because the Breslow-type estimator for type 1 hazards admits negative jumps. This problem also occurs in the case of no covariates. I proved that constrained estimator obtained via EM algorithm is not consistent although monotonicity is guaranteed in Chapter II. Therefore, an estimator $\hat{\mathbf{\Lambda}}_1^P$ based on the pooled-adjacent violated (PAV) algorithm is advocated to achieve both consistency and monotonicity. For K distinct observation time points $t_1, t_2, \cdots, t_K$, $\hat{\mathbf{\Lambda}}_1^P$'s are derived as a set of $\mathbf{\Lambda}_1 = (\Lambda_1(t_1), \Lambda_1(t_2), \cdots, \Lambda_1(t_K))$ minimizing a sum of squared deviations of the isotonic estimates from the unrestricted ones $(\hat{\Lambda}_1)$

$$\hat{\mathbf{\Lambda}}_1^P = \min_{\mathbf{\Lambda}_1 \in \Psi} \sum_{k=1}^{K} [\Lambda_1(t_k) - \hat{\Lambda}_1(t_k)]^2,$$

where Ψ is the class of non-decreasing functions, $\Lambda_1(t_1) \leq \Lambda_1(t_2) \leq \cdot \leq \Lambda_1(t_K)$. The consistency and asymptotic normality of PAV estimator are proved in Chapter II.

4.4 Analysis of SEER data

Here we apply the proposed methods to estimate the mortality rates of prostate cancer for men aged between 50 and 84. Prostate cancer incidence data is available from the Surveillance, Epidemiology, and End Results (SEER) 9 database. In this data set, more than 300,000 cases diagnosed from 1973 through 1999 are available for 9 registries: Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, and Utah. Age distribution in the U.S. population in year 2000 is used to age-adjust observed or predicted rates presented in this article.

Empirical estimates were used for disease presentation at diagnosis $(f_I(a_I, \phi|x))$ and treatment distribution $(f_\tau(\tau|a_I, \phi, x))$.

- Let us denote by $R(a_I, t_I)$ the number of men at risk of prostate cancer and

by $PC(a_I, \phi, t_I)$ the number of men newly diagnosed with the stage/grade ϕ tumor at age a_I in year $t_I = x + a_I$. Then, the disease presentation at diagnosis distribution $f_I(a_I, \phi|x)$ is estimated as

$$\hat{f}_I(a_I, \phi|x) = \frac{PC(a_I, \phi, t_I)}{R(a_I, t_I)},$$

- Similarly, the treatment probability is estimated as

$$\hat{f}_\tau(\tau|a_I, \phi, x) = \frac{PC(a_I, \phi, \tau, t_I)}{PC(a_I, \phi, t_I)},$$

where $PC(a_I, \phi, \tau, t_I)$ is the number of men diagnosed with ϕ-stage/grade of prostate cancer at age a_I in year t_I and received τ treatment.

- Also, $PC(a_I, t_I)$ denotes the number of men newly diagnosed with prostate cancer at age a_I in year $t_I = x + a_I$. Using the Kaplan-Meier method, the incidence survival probability is estimated as

$$\hat{G}_I(a_M|x) = \prod_{a_I \le a_M} \left(1 - \frac{PC(a_I, t_I)}{R(a_I, t_I)}\right).$$

- We adopt Cox proportional hazard model for survival model with covariates $Z = \{a_I, \phi, \tau\}$. If $t_I < 1988$ in the pre-PSA era, the effect of t_I is modeled non-parametrically. On the other hand, last observation carried forward approach is used to specify baseline survival in the PSA era. If $t_I \ge 1988$ in the PSA era, the baseline disease-specific survival is frozen in calendar time at $t_I = 1987$, just prior to the advent of PSA screening.

$$G_{[t_I]}(t|Z = \{a_I, \phi, \tau\}) =$$
$$\begin{cases} \exp\left\{-\int_0^t \lambda_{1[t_I]}(s)\exp(\beta_{[t_I]}^T Z)ds\right\} & \text{for} \quad t_I < 1988, \\ G_{LT}(t|a_I, \phi, t_I) & \\ \quad + \int_0^t f_{LT}(s|a_I, \phi, t_I)\exp\left\{-\int_0^t \lambda_{1[1987]}(s)\exp(\beta_{[1987]}^T Z)ds\right\} & \text{for} \quad t_I \ge 1988. \end{cases}$$

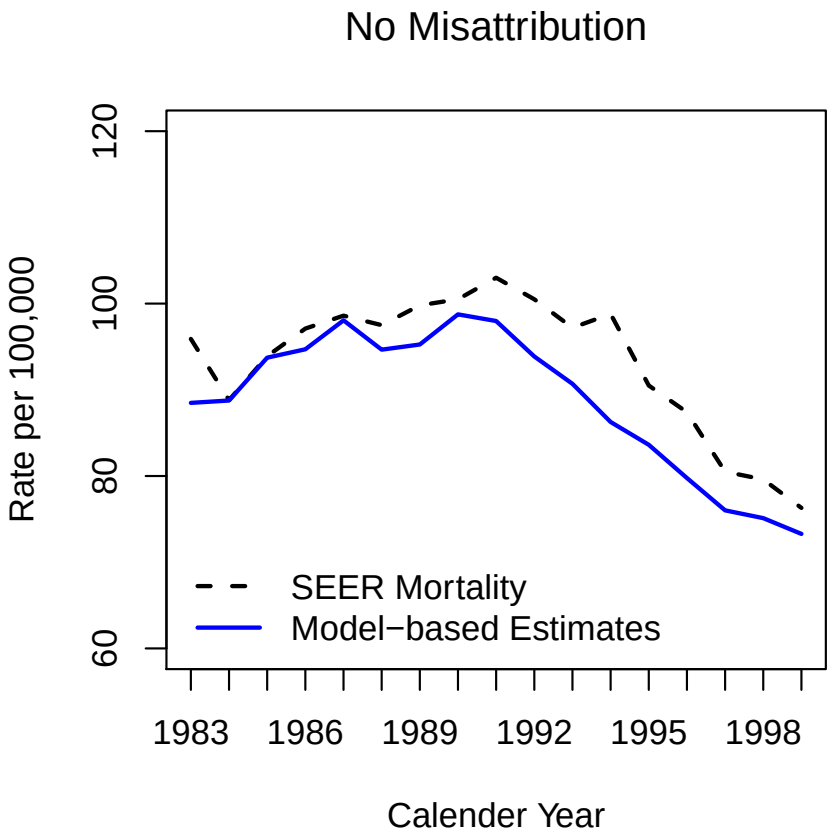

Figure 4.2: Age-adjusted observed mortality rates of SEER 9 registries (dotted line) and model-based estimates assuming no misattribution (solid line).

Because the data is available from SEER beginning in 1973, the predictions before 1973 use $t_I = 1973$; i.e. the earliest observed year effect is carried backwards. For a small fraction of missing values in stage/grade and treatment variables, we assume a 'missing-at-random' mechanism. Figure 4.2 shows the observed SEER 9 and predicted prostate cancer mortality without misattribution using our model (4.1) for men between 50 and 84 from 1983 to 1999.

4.4.1 Analysis with misattribution

Applying estimating approaches discussed in the previous section, we have the following adjusted survival probability

$$\begin{aligned}
&\hat{G}(a_M - a_I | t_I, Z = (a_I, \phi, \tau)) \\
&\quad = \exp\left\{ - \int_0^{a_M - a_I} \hat{\lambda}_{1[t_I]}(s; \hat{\beta}_{[t_I]}, \hat{\gamma}_{[t_I]}) \exp(\hat{\beta}_{[t_I]}^T Z) ds \right\} \\
&\quad = \exp\left\{ - \int_0^{a_M - a_I} \hat{\lambda}_{1[t_I]}(s|Z) \right\}.
\end{aligned}$$

Also, we denote the unadjusted estimators derived under no misattribution $r = 0$ by $\dot{\lambda}$ and $\dot{G}$. Then, observed hazards can be estimated as

$$\begin{array}{ll}
\hat{\lambda}_{1[t_I]}(t|Z) + r_{[t_I]}(t)\hat{\lambda}_{2[t_I]}(t|Z) & \text{for } t_I < 1988, \\
\hat{\lambda}_{1[1987]}(t|Z) + r_{[t_I]}(t)\hat{\lambda}_{2[1987]}(t|Z) & \text{for } t_I \geq 1988 \text{ and } r_{[1987]}(t) > 0, \\
\dot{\lambda}_{1[1987]}(t|Z) + r_{[t_I]}(t)\dot{\lambda}_{2[1987]}(t|Z) & \text{for } t_I \geq 1988 \text{ and } r_{[1987]}(t) = 0.
\end{array} \tag{4.4}$$

Note that the proportional hazard (PH) assumption for observed failures is valid only when $\beta_{[t_I]} = \gamma_{[t_I]}$ or $r_{[t_I]} = 0$.

The model for misattribution is informed by external death certificate review studies such as Hoffman et al. (2003) that specify the misattribution rate r. However, since such studies are limited to small populations, they may not generalize to the U.S. population. Furthermore, the exact mechanism or degree to which attribution bias occurs is unknown. Motivated by these limitations, we assess a variation in mortality rates over the plausible range of misattribution models motivated by the external studies.

Many papers have discussed plausible explanations for attribution bias. Hoffman et al. (2003) found that the number of prostate cancer deaths was significantly inflated from 125 to 146 (16.8%) in 1995 year of death. Furthermore, a time-dependent

model is generally accepted; misattribution is higher for newly diagnosed cases and fades with time (Feuer et al. (1999); Fall et al. (2008)).

Based on these arguments, we chose 12 hypothetical models which depend on year of diagnosis and survival time (in years) as follows.

$$r_{[t_I]}(t) = \{logit^{-1}(\psi_0 + \psi_1 t)\} I(\psi_3 \leq t_I \leq \psi_4),$$

$$(4.5) \qquad (\psi_0, \psi_1, \psi_2, \psi_3) = \begin{cases} (logit0.02, 0, 0, 1999), & (logit0.05, 0, 0, 1999) \\ (logit0.02, 0, 1986, 1995), & (logit0.05, 0, 1986, 1995), \\ (logit0.02, 0, 1988, 1995), & (logit0.05, 0, 1988, 1995), \\ (-3.5, -0.3, 0, 1999), & (-2.5, -0.5, 0, 1999), \\ (-3.5, -0.3, 1986, 1995), & (-2.5, -0.5, 1986, 1995), \\ (-3.5, -0.3, 1988, 1995), & (-2.5, -0.5, 1988, 1995), \end{cases}$$

where $logit^{-1}$ is an inverse $logit$ function and $I(A)$ is an indicator function which takes 1 if A is true. These models represent a constant misattribution or a decreasing one over survival time. We also model a smooth secular trend in mortality.

Figure 4.3 and 4.4 show adjusted estimates for observed and true mortality rates by various misattribution models defined above. Comparing with our model-based estimates for observed mortality rates (blue lines), adjusted true rates are all under-predicted due to the assumption of inflated prostate cancer deaths although they differ somewhat by the magnitude of bias. Constant and time-dependent models show similar patterns for the same period (t_I).

The peak of mortality rates becomes less pronounced under misattribution in the PSA-era, $1986 \leq t_I \leq 1995$ or $1988 \leq t_I \leq 1995$. Especially dramatic changes are observed in time-decreasing models in the PSA-era; $\{logit^{-1}(-2.5 - 0.5t)\} I(1986 \leq t_I \leq 1995)$ and $\{logit^{-1}(-2.5 - 0.5t)\} I(1988 \leq t_I \leq 1995)$. However, when misattribution occurs steadily in all years, the adjusted observed and true mortality

rates are quite parallel to each other; $r_{[t_I]}(t) = 0.02,\ 0.05,\ logit^{-1}(-3.5 - 0.3t)$ and $logit^{-1}(-2.5 - 0.5t)$. Thus, the shape of mortality is not altered if misattribution is time-homogeneous.

It should be noted that adjusted estimates for observed mortality rates are quite different across misattribution models. In fact, the marginal effect of observed mortality rates are not affected by misattribution since it is based on observed number of prostate cancer deaths. However, proportional hazard assumption is violated under misattribution (4.4). Thus, naive approach (ignoring misattribution) leads to biased estimates of observed rates (Figure 4.2) by misspecified model due to attribution bias. Among misattribution models defined in (4.5), $0.02I(1986 \leq t_I \leq 1995)$, $0.02I(1988 \leq t_I \leq 1995)$, $\{logit^{-1}(-2.5 - 0.5t)\}I(1988 \leq t_I \leq 1995)$ and $\{logit^{-1}(-3.5 - 0.3t)\}I(1986 \leq t_I \leq 1995)$ yield adjusted estimates close to observed SEER 9 mortality rates. Furthermore, these misattribution assumptions eliminate the 'hump' in mortality rates during 1988-1994, and yield 8.11%, 6.87%, 9.73% and 5.32% inflated mortality rates in 1995, respectively.

We can calibrate the misattribution model to observed rates. This results in less misattribution in the beginning and higher misattribution in the middle of the PSA-era. For example, using the models in (4.5), we define

$$r_{[t_I]}(t) =$$

$$\begin{cases} 0.02I(1986 \leq t_I \leq 1990) + 0.05I(1991 \leq t_I \leq 1994) \\ \{logit^{-1}(-3.5 - 0.3t)\}I(1986 \leq t_I \leq 1990) + \{logit^{-1}(-2.5 - 0.5t)\}I(1991 \leq t_I \leq 1995) \end{cases}.$$

Adjusted estimates for observed and true rates based on the above model are presented in Figure 4.5. This scenario explains 16.05% and 10.69% age-adjusted increase in mortality rates in 1995, respectively.

4.5 Discussion

Our modeling approach for mortality predictions is based on a cause-specific survival model using a convolution of baseline survival in the absence of screening and the lead time. Mortality rates are adjusted for attribution bias which is explicitly introduced into the survival model. A key contribution of this project is that it provides a formal method for adjusting mortality rates for misattributed causes of deaths. Unfortunately, misattribution cannot be estimated with SEER data only. External studies such as Hoffman et al. (2003) and Fall et al. (2008) are needed, that use verified information on the true underlying causes of death, to quantify the magnitude of misattribution. To address generalizability of such external data, we performed a sensitive analysis over plausible range of misattribution models in this study.

We derived adjusted mortality rates with 14 different misattribution models that are constant or decreasing in survival time. To see if the peak in mortality is induced by misattributed deaths in the PSA-era, we assumed that misattribution varies by year of diagnosis. Our sensitivity analysis showed that constant misattribution cannot explain the mortality trend. Remarkable changes were observed under misattribution changing with calendar time. Under this scenario, attribution bias can explain the increase and subsequent fall in prostate cancer mortality rates in the late 1980s and early 1990s.

In this article, we assess constant and time-decreasing misattribution which can differ by year of diagnosis. However, there are other possible explanations for the mechanism. It may depend on age at diagnosis since prostate cancer is rarer among younger patients and thus more likely to be considered as a cause of death. Con-

versely, deaths among the elder may not be investigated as thoroughly as those among the younger (Fall et al. (2008)). Moreover, over-reported prostate cancer deaths are probably higher for men with slowly progressing tumor or under conservative management (Fall et al. (2008); Newschaffer et al. (2000)). Men with localized/regional cancer are less likely to die of the disease. If men do not get aggressive treatments such as radiation therapy or radical prostatectomy, physicians may be more likely to attribute the death to prostate cancer because they perceived ineffectiveness of the non-curative approach. Moreover, increased media attention given to prostate cancer may contribute to increased over-attribution. Any misattribution model can be explicitly incorporated in the survival model. We introduced a general misattribution model that can incorporate covariates into the misattribution mechanism and model the survival time in Chapter III.

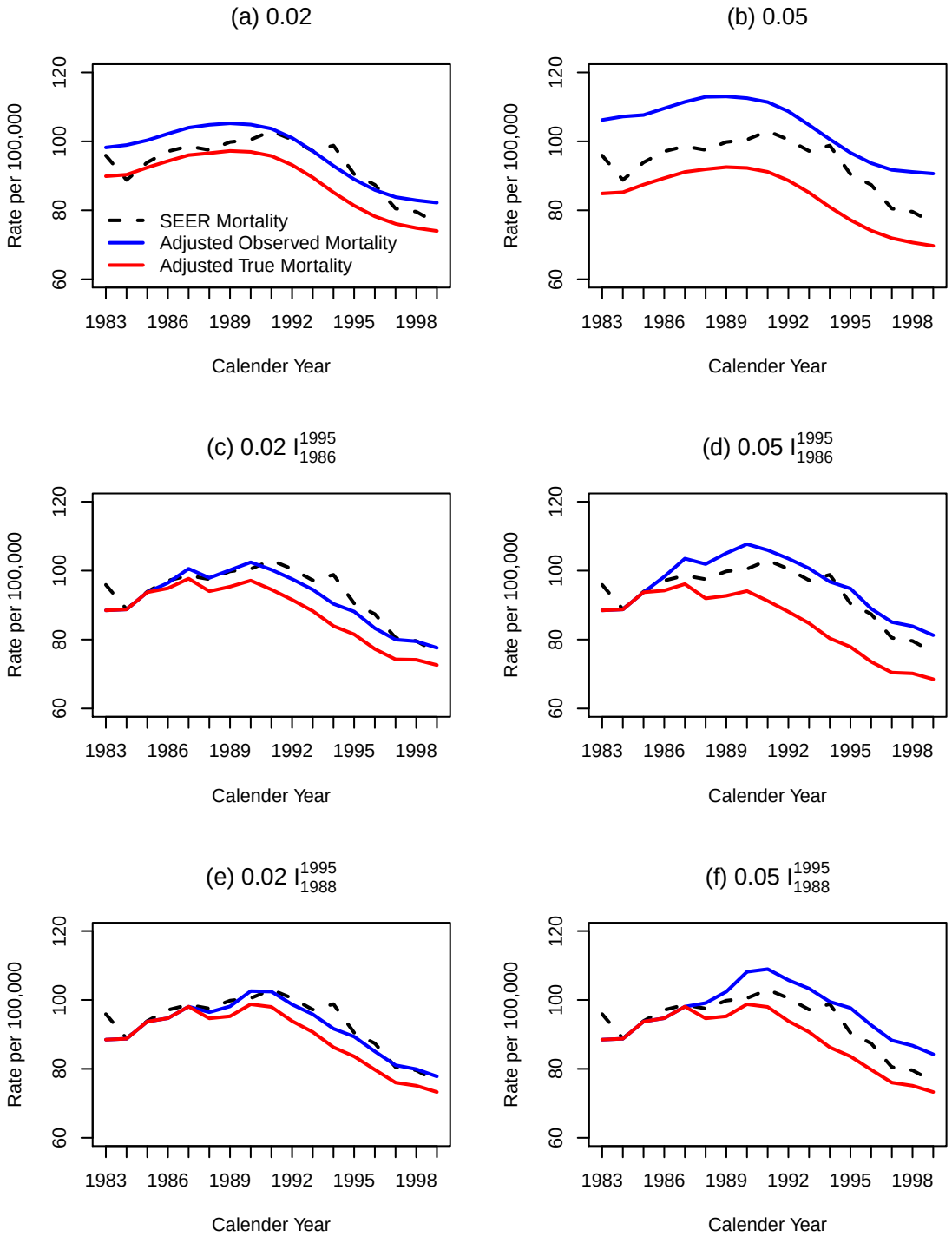

Figure 4.3: Adjusted estimates for observed (blue) and true (red) mortality rates by misattribution: (a) 0.02, (b) 0.05, (c) $0.02I(1986 \leq t_I \leq 1995)$, (d) $0.05I(1986 \leq t_I \leq 1995)$, (e) $0.02I(1988 \leq t_I \leq 1995)$ and (f) $0.05I(1988 \leq t_I \leq 1995)$ where t_I is year of diagnosis.

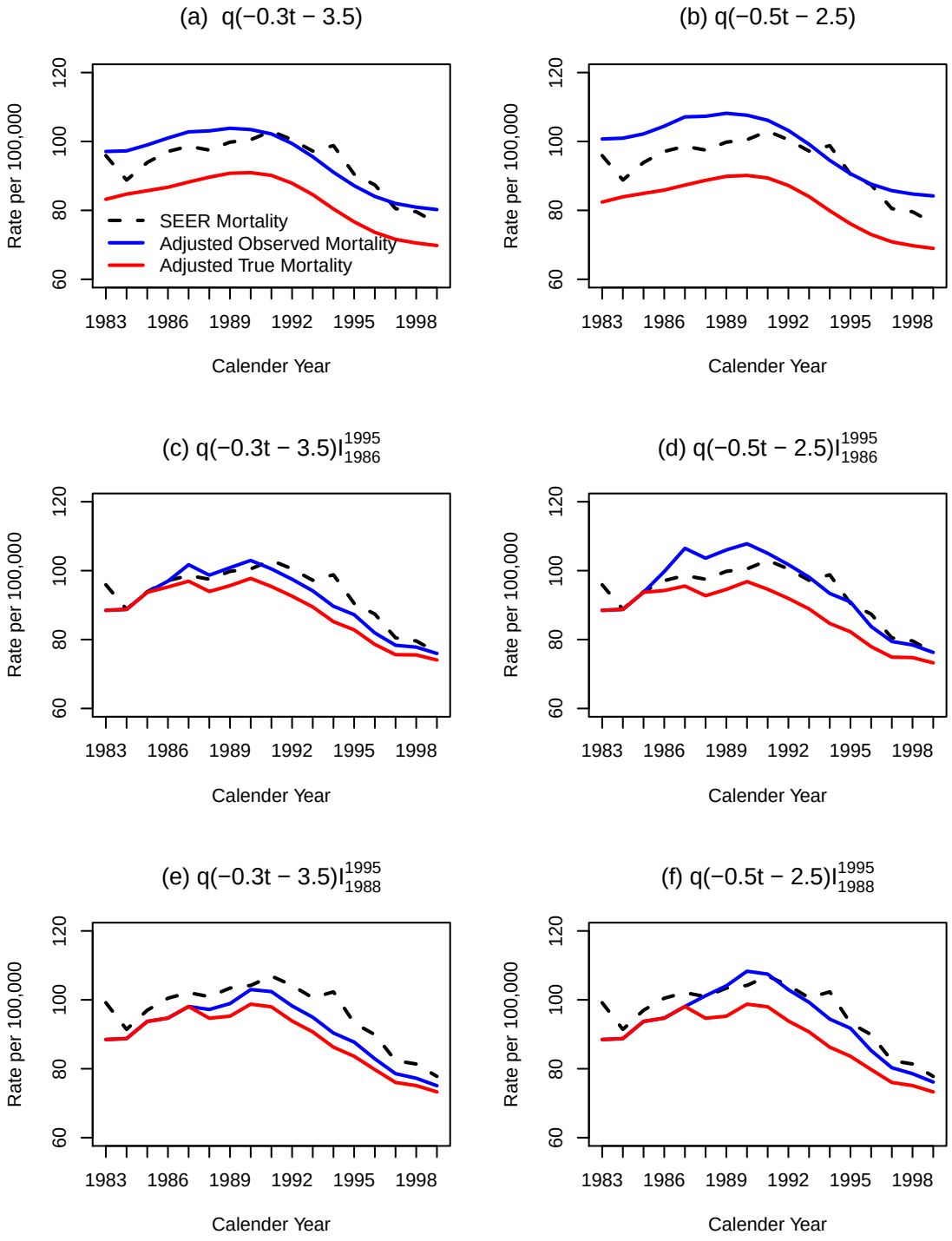

Figure 4.4: Adjusted estimates for observed (blue) and true (red) mortality rates by misattribution: (a) $q(-0.3t - 3.5)$, (b) $q(-0.5t - 2.5)$, (c) $q(-0.3t - 3.5)I(1986 \leq t_I \leq 1995)$, (d) $q(-0.5t - 2.5)I(1986 \leq t_I \leq 1995)$, (e) $q(-0.3t - 3.5)I(1988 \leq t_I \leq 1995)$ and (f) $q(-0.5t - 2.5)I(1988 \leq t_I \leq 1995)$ where t_I is year of diagnosis, t is survival time (in year) and q is an inverse logit function ($logit\ q(x) = x$)

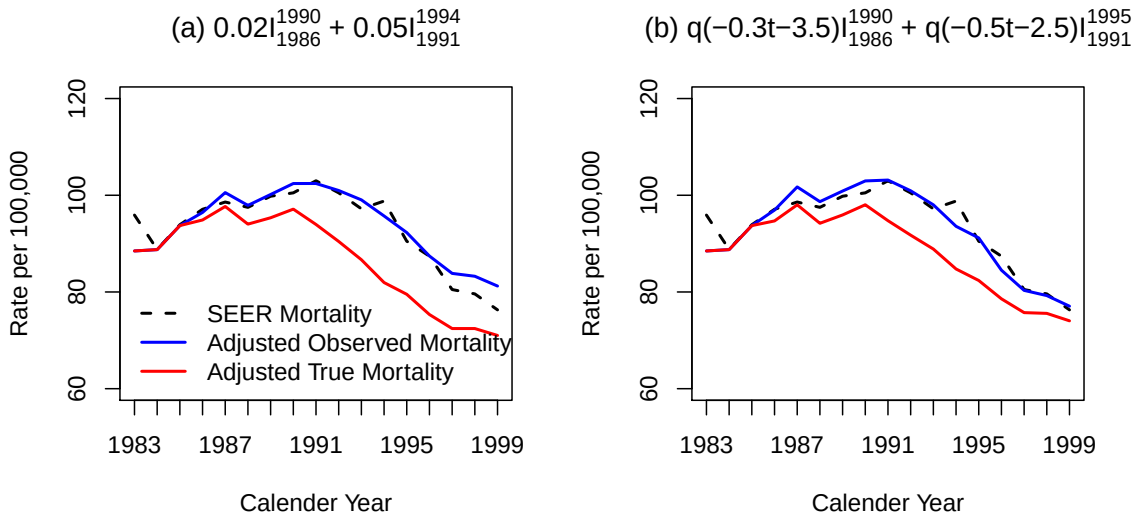

Figure 4.5: Adjusted estimates for observed (blue) and true (red) mortality rates by misattribution: (a) $0.02I(1986 \leq t_I \leq 1990) + 0.05I(1991 \leq t_I \leq 1994)$ and (b) $q(-0.3t - 3.5)I(1986 \leq t_I \leq 1990) + q(-0.5t - 2.5)I(1991 \leq t_I \leq 1995)$ where t_I is year of diagnosis, t is survival time (in year) and q is an inverse logit function ($logit\ q(x) = x$)

CHAPTER V

Discussion

Our study is motivated by the recent trend in US prostate cancer mortality rates which follows the incidence rates. Many studies have brought forward this phenomenon to understand the role of PSA screening in the observed rise and fall of US prostate cancer mortality rates. One possibility is that the peak in mortality and its decline phase observed currently is an artefact of the cause of death misattribution. The overdiagnosis and the sharp peak in cancer incidence resulting from screening would lead to similar behavior of mortality in the presence of misattribution of some of those excess incident cases at death as (falsely) related to prostate cancer. However due to the complexity of the phenomenon, it did not appear possible to provide an answer without modeling. To assess this hypothesis, we first developed the methodology to address the problem of estimating mortality rates where other cause of death can be misattributed to prostate cancer.

We first studied a competing risks survival (prostate cancer v.s other causes) under misattribution using univariate and multivariate non- and semi-parametric setting, including the Cox proportional hazards (PH) model. We showed that available standard approaches depend on the ratio between the two hazards $\phi(t) = \lambda_2(t)/\lambda_1(t)$. The profile likelihood for the Cox PH model is a function of the ratio $\phi(t)$. Under

a univariate model, the conditional expectation of true failure indicator in E step is updated with the conditional probability (2.14) which depends on $\phi(t)$.

The main theoretical challenge emerging in the univariate nonparametric setting is that with the continuous survival times, the asymptotically unbiased estimator for $\phi(t)$ is not feasible without restrictions on the $\phi(t)$. It is easy to see that a na"ive nonparametric estimate of $\phi(t)$ is either 0 or ∞ assuming all failures are correctly specified.

Another challenge that due to over-attribution expected to be operating with prostate cancer data, the missing-at-random (MAR) assumption that was extensively utilized in the previous literature is wrong in our case.

Many conventional analytic tools break down in this situation. If a parametric model for $\phi(t)$ is assumed, leaving the hazards otherwise unspecified, both the EM algorithm and partial likelihood approach will work. However, without additional restrictions, the Nonparametric Maximum Likelihood Estimator for the survival function satisfying monotonicity constraints exhibits bias, and so does the EM algorithms, a primary tool to deal with estimation in the presence of missing data.

We developed a number of estimation approaches to yield monotonic and consistent cause-specific survival estimates avoiding any assumptions restricting the ratio of hazards between the causes of death.

Having tackled the theoretical development and implementation of the newly developed methodology in statistical software we turned our attention to the real life problem that motivated this research in the first place.

Our modeling approach for mortality predictions is based on a cause-specific survival model using a convolution of baseline survival in the absence of screening and the lead time. Mortality rates are adjusted for attribution bias which is explicitly in-

troduced into the survival model. The survival estimates adjusted for misattribution are used to correct the survival component of the mortality model trying to explain the inflated morality rates in the 90-ies. We explicitly modeled a smooth effect of year of diagnosis in survival model to allow for a slow secular trend in mortality rates.

Throughout the dissertation, we assumed that misattribution is a known function. Knowledge of misattribution probabilities is essential to obtain valid estimates of mortality rates in the nonparametric context. Unfortunately, these probabilities are not identifiable with the SEER data, even if they are constant. Although there are special cases in which they are estimable with the observed data (e.g., Dinse (1986); Craiu and Duchesne (2004)), missing-at-random assumption is in general necessary to avoid nonidentifiability problems. Therefore we relied on published studies reviewing cause of death decisions on a the random subsample of patients and re-classifying t he cause of death by an expert panel. An example of such death certificate review studies is (Hoffman et al. (2003)). However, concerns of small sample size and potential representativeness of the certificate study population lead us to sensitivity analysis with respect to the missing data mechanism. We derived adjusted mortality rates with 14 different misattribution models that are constant or decreasing in survival time. To see if the peak in mortality is induced by misattributed deaths in the PSA-era, we assumed that misattribution varies by year of diagnosis. Our sensitivity analysis showed that constant misattribution cannot explain the mortality trend. Remarkable changes were observed under misattribution changing with calendar time. Under this scenario, attribution bias can explain the increase and subsequent fall in prostate cancer mortality rates in the late 1980s and early 1990s.

Several assumptions can explain the observed dynamics of mortality. We assessed constant and time-decreasing misattribution that can differ by year of diagnosis. Other possible explanations for the mechanism include dependence on age at diagnosis since prostate cancer is rare in younger patients and thus may be more likely considered as a cause of death. Also, deaths among the elderly may not be investigated as thoroughly as those in the younger patient population (Fall et al. (2008).

Over-reported prostate cancer deaths are probably higher for men with slowly progressing tumors or under conservative management (Fall et al. (2008); Newschaffer et al. (2000)). Men with localized/regional cancer are less likely to die of the disease. If men do not get curative aggressive treatment such as radiation therapy or radical prostatectomy, physicians may be more likely to attribute the death to prostate cancer because of perceived ineffectiveness of the non-curative approach. Increased media attention given to prostate cancer may contribute to increased over-attribution. Any misattribution model can be explicitly incorporated in the survival model. We introduced a general misattribution model that can incorporate covariates into the misattribution mechanism and model the survival time. We only consider over-misattribution. Several authors have raised the possibility of both over- and under-misattribution in cancer death certificates (Percy et al. (1981); Fall et al. (2008); Hoffman et al. (2003)), particularly with data before PSA. Our modeling approach can be extended to cover this case. It would be interesting to study under what condition observed mortality rates would stay intact despite misattribution as a result of over- and under- misattribution balancing each other out.

The fundamental character of cancer is that it is a latent chronic disease with substantial unobserved heterogeneity. The progress towards reduced cancer mortality and cancer burden has centered on a combined use of more efficacious treatments

as determined by randomized clinical trials (RCT) and the early detection of cancer as evaluated in screening trials. This strategy has lead to an increase in the cost of treatment resulting from a surge in the incidence of early stage low grade cancers that are treated on a just in case basis without sufficient evidence for efficiency of a variety of treatment options. As modern biomarker research translated into ever more sensitive cancer screening tests presents unprecedented early detection opportunities, we are seeing cancers that were never accessible before. Confronted with changes in the heterogeneity of the incident disease much of which is now indolent cancers, we are using scientific tools that were perfected when most cancers were relevant and treatment success was reached under the strategy the earlier and the more aggressive, the better.

Because an early detected cancer can be very different from the symptomatic one, treatments may show substantial variability in the outcomes and side effects. The majority of early stage cancers may be over-diagnosed and over-treated resulting in a substantial burden to the health system, inefficient use of resources, and potentially avoidable harm to a large fraction of patients due to the side effects of over-treatment. Dependent on the disease progression mechanism, early treatment may also be harmful overall. For example, early hormonal therapy may provoke a transformation of indolent prostate cancer cells into androgen independent ones in response to the androgen deprivation stress.

Until a precise biological reason for this variability can be found, and individualized treatments developed to target them, patients, physicians and policy makers must make decisions under uncertainty based on available data.

The controversy of prostate cancer resulting from the latent heterogeneity and high prevalence of the disease needs to be addressed by development of efficient

management of the disease. The arsenal of actions ranges from aggressive treatment of all early stage cancers and struggling to identify which treatments work in a dynamically heterogeneous pool of incident diseases, active surveillance and deferred treatment strategies informed by patient's monitoring, to canceling early detection efforts altogether. The question of what is an optimal investment of resources is an enormous challenge. This challenge cannot be addressed without an integrative data analysis and modeling approach. Traditionally, treatment efficacy is measured in Randomized Clinical Trials (RCT), while the effectiveness of cancer early detection modalities is evaluated in screening trials such as PLCO and ERSPC trials in prostate cancer, currently showing conflicting results. Both approaches taken in isolation are likely to be underpowered, inefficient, and perhaps biased.

This dissertation work has provided a step toward better understanding of the mortality phenomenon by supplying rigorous methods that peal off yet another element of complexity in the study of chronic diseases and their management strategies. The novel more precise statistical methodology will contribute to the development of optimal cancer management strategies informed by integrative statistical modeling of disease onset, progression, diagnosis, treatment and survival, and assessment of the interaction between treatment and diagnostic interventions and the heterogeneity of the disease.

While the problem of prostate cancer will not be solved any time soon, we have provided a step forward to the solution in a structured and rigorous way.

APPENDICES

.1 Convergence of EM algorithm

The likelihood is maximized only when $d\hat{\Lambda}(t) = d\hat{\Lambda}^{obs}(t)$. Therefore, it is sufficient to show the convergence of estimator for $d\Lambda_1$.

Case 1. $\bar{r}(t)dN_1(t) \geq r(t)dN_2(t)$

Without loss of generality, we assume that $d\Lambda^{(0)} \geq 0$. We prove first that $d\Lambda_1(t)^m < d\Lambda_1(t)^{m+1} \leq d\tilde{\Lambda}_1(t)$ if $0 < d\Lambda_1(t)^m < d\tilde{\Lambda}_1(t)$ for any $d\Lambda_2(t)^m$. First inequality can be easily proved by using the following equation: $d\Lambda_1(t)^m + d\Lambda_2(t)^m = \frac{dN_1(t)+dN_2(t)}{Y(t)}$. The proof of second inequality is based on the observation that $d\Lambda_1^{m+1}[.]$ is a strictly increasing function of $d\Lambda_1(t)^m$. Thus, $d\Lambda_1^{m+1}[d\tilde{\Lambda}_1(t)] \geq d\Lambda_1^{m+1}[x]$ for $x \leq d\tilde{\Lambda}_1(t)$ where the equality holds only if $x = d\tilde{\Lambda}_1(t)$. Similarly $d\Lambda_1(t)^m > d\Lambda_1(t)^{m+1} \geq d\tilde{\Lambda}_1(t)$ if $d\Lambda_1(t)^m > d\tilde{\Lambda}_1(t)$ for any $d\Lambda_2(t)^m$. However, $d\Lambda_1(t)^{m+1}$ turns out to be zero whenever $d\Lambda_1(t)^m$ is.

Case 2. $dN_1(t) - Odds[r(t)]dN_2(t) < 0$

Similar arguments for case 1 can be used to prove that $0 \leq d\Lambda_1(t)^{m+1} < d\Lambda_1(t)^m$ if $d\Lambda_1(t)^m > 0$ where the equality holds only if $d\Lambda_1(t)^{m+1} = 0$ (*). Similarly, if $d\tilde{\Lambda}_1(t) < d\Lambda_1(t)^m < 0$, then $d\Lambda_1(t)^m < d\Lambda_1(t)^{m+1} \leq 0$. However, if $d\Lambda_1(t)^m < d\tilde{\Lambda}_1(t)$, then $d\Lambda_1(t)^m > d\Lambda_1(t)^{m+1}$. When $d\Lambda_1(t)^{m+1} < -\text{Odds}[r(t)]d\hat{\Lambda}^{obs}(t)$, $d\Lambda_1(t)^{m+1}$ becomes greater than zero, so it goes back to the first case (*).

.2 Bias for constrained NPMLE in a continuous time case

We first fix a continuous time interval $\mathcal{T} = [0, \tau]$ for a given terminal time $0 < \tau < \infty$. Let $\{\mathcal{F}_t; t \in \mathcal{T}\}$ be a filtration of the probability space. We have

$$
\begin{aligned}
&\mathrm{E}[\hat{\Lambda}_1(t)] \\
&= \int_0^t \mathrm{E}\left[\frac{I(Y(s)>0)}{Y(s)} \times \right. \\
&\qquad \left. \mathrm{E}\left[\left\{dN_1(s) - \frac{r(s)}{1-r(s)} dN_2(s)\right\} I\left(\sum_{i=1}^n dN_{i1}(s) - \frac{r(s)}{1-r(s)} dN_{i2}(s) > 0\right) \middle| \mathcal{F}_{s-}\right]\right] \\
&= \sum_{i=1}^n \int_0^t \mathrm{E}\left[\frac{I(Y(s)>0)}{Y(s)} \times \right. \\
&\qquad \left. \mathrm{E}\left[\left\{dN_{i1}(s) - \frac{r(s)}{1-r(s)} dN_{i2}(s)\right\} I\left(dN_{i1}(s) - \frac{r(s)}{1-r(s)} dN_{i2}(s) > 0\right) \middle| \mathcal{F}_{s-}\right]\right] \\
&= \sum_{i=1}^n \int_0^t \mathrm{E}\left[\frac{I(Y(s)>0)}{Y(s)} \mathrm{E}\left[dN_{i1}(s) \middle| \mathcal{F}_{s-}\right]\right]
\end{aligned}
$$

Therefore, $\mathrm{E}[\hat{\Lambda}_1(t) - \Lambda_1(t)]$ turns out to be

$$\sum_{i=1}^n \int_0^t -Pr(Y(s)=0) d\Lambda_1(s) + Pr(Y(s)>0) r(s) d\Lambda_2(s).$$

Similarly, $\mathrm{E}[\hat{\Lambda}_2(t) - \Lambda_2(t)]$ is reduced to

$$
\begin{aligned}
&\sum_{i=1}^n \int_0^t \mathrm{E}\left[\frac{I(Y(s)>0)}{Y(s)} \times \right. \\
&\qquad \left. \mathrm{E}\left[\{dN_{i1}(s) + dN_{i2}(s)\} I\left(dN_{i1}(s) - \frac{r(s)}{1-r(s)} dN_{i2}(s) < 0\right) \middle| \mathcal{F}_{s-}\right]\right] - d\Lambda_2(s) \\
&= \sum_{i=1}^n \int_0^t \mathrm{E}\left[\frac{I(Y(s)>0)}{Y(s)} \mathrm{E}\left[dN_{i2}(s) \middle| \mathcal{F}_{s-}\right]\right] - d\Lambda_2(s) \\
&= \sum_{i=1}^n \int_0^t -Pr(Y(s)=0) d\Lambda_2(s) - Pr(Y(s)>0) r(s) d\Lambda_2(s).
\end{aligned}
$$

.3 Asymptotic distribution of estimator using the pool-adjacent-violators algorithm

For any $a > 0$, $\hat{\Lambda}_1^S(t) - \frac{a}{\sqrt{n}}$ and $\hat{\Lambda}_1^S(t) + \frac{a}{\sqrt{n}}$ are isotonic functions for $t \in \mathcal{T}$, and $\hat{\Lambda}_1^P$ is an isotonic regression of $\tilde{\Lambda}_1$. By Theorem 1.6 (Barlow et al. 1972),

$$\begin{aligned} if \quad & \hat{\Lambda}_1^S(t) - \frac{a}{\sqrt{n}} \leq \tilde{\Lambda}_1(t) \leq \hat{\Lambda}_1^S(t) + \frac{a}{\sqrt{n}}, \\ then \quad & \hat{\Lambda}_1^S(t) - \frac{a}{\sqrt{n}} \leq \hat{\Lambda}_1^P(t) \leq \hat{\Lambda}_1^S(t) + \frac{a}{\sqrt{n}}. \end{aligned}$$

Since $\sqrt{n}(\hat{\Lambda}_1^S(t) - \tilde{\Lambda}_1(t))$ converges to zero in probability, $\sqrt{n}(\hat{\Lambda}_1^S(t) - \hat{\Lambda}_1^P(t))$ also converges to zero in probability. Therefore, $\sqrt{n}(\hat{\Lambda}_1^P(t) - \Lambda_1(t))$ weakly converges to the same distribution as $\sqrt{n}(\tilde{\Lambda}_1(t) - \Lambda_1(t))$ by Slutsky's theorem. Consistency of the estimator $\hat{\Lambda}_1^P$ can be easily proved using similar arguments.

.4 Semiparametric Efficiency

Notation is similar to those used in Lu & Tsiatis (2005) and Tsiatis (2006). The model is characterized by the $1 + p + q$ parameter of interest ξ and the infinite dimensional nuisance parameters $\{\lambda_1(t), \lambda_{0|W}(t|w), p_W(w)\}$, where $\lambda_{0|W}$ denotes a hazard function for censoring and $p_W(w)$ denotes the marginal density of W. The density of a single data item is given by

$$\begin{aligned} & p(T = t, \omega, W = (z, x, a)) = \\ & p_W(w)\{(\theta_1(t;\beta) + r(t,w)\theta_2(t;\gamma)\exp(\alpha))\lambda_1(t)\}^{I(\omega=1)}\{\bar{r}(t,w)\theta_2(t;\gamma)\exp(\alpha)\lambda_1(t)\}^{I(\omega=2)} \\ & \times \exp\left\{-\int_0^t (\theta_1(t;\beta) + \theta_2(t;\gamma)\exp(\alpha))\lambda_1(s)\right\} \lambda_{0|W}(t|w)^{I(\omega=0)} \exp\{-\Lambda_{0|W}(t|w)\} \end{aligned}$$

The nuisance tangent space can be written as a direct sum of three orthogonal linear spaces, namely

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s},$$

where Λ_{1s} is associated $\lambda_1(t)$, Λ_{2s} associated with $\lambda_{0|W}(t|w)$, Λ_{3s} associated with $p_W(w)$. Specifically, the space Λ_{1s} is

$$\Lambda_{1s} = \{\int a(t)dM(t) \qquad \text{for all } (1+p+q)\text{-dimensional functions } a(t)\}.$$

where $dM = dM_1 + dM_2$. The score vector for β evaluated at the true model is given by

$$\begin{aligned} U_\beta &= \int Z(t)\frac{\theta_1(t;\beta)}{\theta_1(t;\beta) + r(t,W)\theta_2(t;\gamma)\exp(\alpha)}dM_1(t) \\ &= \int Z(t)\rho(t;\xi)dM_1(t). \end{aligned}$$

This is orthogonal to Λ_{2s} and Λ_{3s}. Therefore, the efficient score, derived as the residual after projecting U_β onto Λ (or in this case Λ_{1s}), is given by

$$U_{eff} = \int \{Z(t)\rho(t;\xi) - a^*(t)\}dM_1(t) - a^*(t)dM_2(t)$$

where

$$a^*(t) = \frac{E[Y(t)Z(t)\theta_1(t;\beta)]}{E[Y(t)\{\theta_1(t;\beta) + \theta_2(t;\gamma)\exp(\alpha)\}]}.$$

.5 Covariance Terms of KL Estimator

$$I^{KL}_{\beta\beta} = \int E\left[Y(t)\theta_1(t;\beta)\left(Z(t) - \frac{s_Z(t;\beta)}{s(t;\beta)}\right)^{\otimes 2}\right]\lambda_1(t)dt\Big|_{\eta=\eta^0},$$

$$I^{KL}_{\beta\gamma} = \int -E\left[Y(t)r(t)\theta_2(t;\gamma)\left(Z(t) - \frac{s_Z(t;\beta)}{s(t;\beta)}\right)\left(X(t) - \frac{s_{X\bar{r}}(t;\gamma)}{s_{\bar{r}}(t;\gamma)}\right)^T\right]\lambda_2(t)dt\Big|_{\eta=\eta^0},$$

$$I^{KL}_{\gamma\gamma} = \int E\left[Y(t)(1-r(t))\theta(t;\gamma)\left(X(t) - \frac{s_{X\bar{r}}(t;\gamma)}{s_{\bar{r}}(t;\gamma)}\right)^{\otimes 2}\right]\lambda_2(t)dt\Big|_{\eta=\eta^0},$$

$$\begin{aligned} V^{KL}_{\beta\beta} = I^{KL}_{\beta\beta} &+ \int E\left[Y(t)r(t)\theta(t;\gamma)\left(Z(t) - \frac{s_Z(t;\beta)}{s(t;\beta)}\right)^{\otimes 2}\right]\lambda_2(t)dt \\ &+ \int E\left[Y(t)(1-r(t))\theta(t;\gamma)\left(\frac{s_{Zr}(t;\gamma)}{s_{\bar{r}}(t;\gamma)} - \frac{s_Z(t;\beta)}{s(t;\beta)}\frac{s_r(t;\gamma)}{s_{\bar{r}}(t;\gamma)}\right)^{\otimes 2}\right]\lambda_2(t)dt\Big|_{\eta=\eta^0}, \end{aligned}$$

where

$$a = E[n^{-1}A] \qquad \text{for } A \in \{S(t;\beta), S_Z(t;\beta), S_r(t;\gamma), S_{\bar{r}}(t;\gamma), S_{X\bar{r}}(t;\gamma)\}.$$

The estimates are obtained by substituting $\hat{\eta}_n^{KL} = (\hat{\lambda}_1^{KL}(t), \hat{\lambda}_2^{KL}(t), \hat{\beta}_n^{KL}, \hat{\gamma}_n^{KL})$ for η and replacing the expectation by its empirical counterpart.

Bibliography

Adler, R. J. (1990). *An Introduction to Continuity, Extrema and Related Topics for General Gaussian Processes*. IMS, Hayward, Ca.

Andersen, P. K., Borgan, O., D., G. R., and N., K. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10,** 1100–1120.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathmatical Statistics* **26,** 641–647.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. New York: Wiley.

Burr, D. (1994). On inconsistency of Breslow's estimator as an estimator of the hazard rate in the Cox model. *Biometrics* **50,** 1142–1145.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89,** 659–668.

Chen, P., He, R., Shen, J., and Sun, J. (2009). Regression analysis of right-censored

failure time data with missing censoring indicators. *Acta Mathematicae Applicatae Sinica* **25,** 415–426.

Craiu, R. V. and Duchesne, T. (2004). Inference based on the em algorithm for the competing risks model with masked causes of failure. *Biometrika* **91,** 543–558.

Dewanji, A. and Sengupta, D. (2003). Estimation of competing risks with general missing pattern in failure types. *Biometrics* **59,** 1063–1070.

Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure data. *Biometrics* **38,** 417–431.

Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *J American Statistical Association* **81,** 328–336.

Fall, K., Stromberg, F., Rosell, J., Andren, O., and Varenhorst, E. (2008). Reliability of death certificates in prostate cancer patients. *Scandinavian Journal of Urology and Nephrology* **42,** 352–357.

Feuer, E. J., Merrill, R. M., and Hankey, B. F. (1999). Cancer surveillance series: interpreting trends in prostate cancer-part II: Cause of death misclassification and the recent rise and fall in prostate cancer mortality. *J Natl Cancer Inst* **91,** 1025–1032.

Flehinger, B. J., Reiser, B., and Yashchin, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika* **85,** 151–164.

Gao, G. and Tsiatis, A. A. (2005). Semiarametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika* **92,** 875–891.

Goetghebeur, E. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82,** 821–833.

Hoffman, R. M., Noell, S. S., Hunt, W. C., Key, C. R., and Gilliland, F. D. (2003). Effects of misattribution in assigning cause of death on prostate cancer mortality rates. *Annals of Epidemiology* **13,** 450–454.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wildey.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Jornal of the American Statistical Association* **84,** 1074–1078.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81,** 61–71.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* **57,** 1191–1197.

Lu, K. and Tsiatis, A. A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Analysis* **11,** 29–40.

Lu, W. and Liang, Y. (2008). Analysis of competing risks data with missing cause of failure under additive hazards model. *Statistica Sinica* **18,** 219–234.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika* **91,** 331–343.

Newschaffer, C. J., Otani, K., McDonald, M. K., and T., P. L. (2000). Causes of death in elderly prostate cancer patients and in a comparison nonprostate cancer cohort. *Journal of the National Cancer Institute* **92,** 613–621.

Percy, C., Stanek, E. r., and Gloeckler, L. (1981). Accuracy of cancer death ceritificates and its effect on cancer mortality statistics. *American Journal of Public Health* **71,** 242–250.

Prentice, R. L., Kalbfleisch, J. D., Peterson, J. A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34,** 541–554.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences* **72,** 20–22.

Tsodikov, A., Szabo, A., and Wegelin, J. (2006). A population model of prostate cancer incidence. *Statistics in Medicine* **25,** 2846–2866.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.