Published in final edited form as:

Biometrics. 2015 December ; 71(4): 941–949. doi:10.1111/biom.12338.

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

Jinkyung Ha and

Int Med-Geriatric Medicine, University of Michigan, Ann Arbor, Michigan 48109, U.S.A

Alexander Tsodikov

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109, U.S.A

Jinkyung Ha: jinha@med.umich.edu; Alexander Tsodikov: tsodikov@umich.edu

Summary

Misclassified causes of failures are a common phenomenon in competing risks survival data such as cancer mortality. We propose new estimating equations for a semiparametric proportional hazards (PH) model with misattributed causes of failures. Unlike other methods, the estimator does not require any parametric assumptions on baseline cause-specific hazard rates. It is shown that the estimators for regression coefficients are consistent and asymptotically normal. Simulation results support the theoretical analysis in finite samples. The methods are applied to analyze prostate cancer survival.

Keywords

Competing risks; Misattribution of cause of failure; Estimating equations

1. Introduction

Misclassified causes of failures are common in competing risks data due to various reasons such as inaccuracies of cancer death certificates (Percy et al., 1981). 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).

The analysis of competing risks data where the cause of failure is misclassified according to the missing-at-random (MAR) mechanism (Little and Rubin, 1987) 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

8. Supplementary Material

Web Appendix A, referenced in Sections 3.1, 3.4, 4, and 5 is available with this paper at the Biometrics website on Wiley Online Library.

failure of interest is proportional to the one for the other cause (proportional hazards (PH) assumption). Lu and Tsiatis (2005) derived a semiparametric efficient score function that incorporates covariates into the probability of missing cause of failure on covariates. Recently, several authors have investigated the approaches that 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 MAR 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 MAR assumption may be violated. In the prostate cancer example misclassification depends on the unobserved type of failure. The PH assumption is also often unrealistic. Motivated by the prostate cancer problem we develop a robust approach to the problem that relies neither on the PH assumption between the competing causes of failure, nor on the MAR mechanism.

As we lay the groundwork for the new robust approach we first consider nonparametric maximum likelihood estimation (NPMLE, Section 3.1) and the model where baseline hazards are restricted by the PH assumption (Section 3.2). We show that the standard NPMLE and profile likelihood approach (Section 3.1) fails to yield useful estimates without further assumptions. The problem is solved either under the PH assumption linking the baseline hazards (Section 3.2) or using the newly proposed robust methods (Section 3.3) based on weighted martingale estimating equations that work without any restrictions on the baseline hazards. For identifiability we supplement survival data with an independent external dataset informing the model on the misclassification rates through a review of medical records. In one version of the robust approach (Section 3.3.1) we used the form of the weights motivated by the PH-based solution. The performance of the proposed estimates is studied by asymptotic analysis (Section 4) and simulations (Section 5). Finally, we analyzed real prostate cancer data using the proposed methodology and conducted sensitivity analyses.

2. Assumptions and notation

In this article, 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). Without misclassification, the data for individual i consist of (T_i, Ω_i, Z_i) , where T_i is time of failure (regardless of type) or censoring, Ω_i is an indicator taking values 0, 1 or 2 corresponding to censoring, type 1 or 2 failure, respectively. let Z be a vector of covariates, $(Z_1, Z_2, \dots, Z_p)^T$. With the understanding that all expressions are conditional on the subject-specific Z we will drop Z, i , and regression coefficients from expressions in the sequel except when exposing them is essential. Let $T^j, j = 0, 1, 2$, be the potential competing failure times that correspond to the three causes of failure so that $T = \min_j T^j$.

Generally, estimation of the distributions of T^j is subject to the well known non-identifiability issue (Tsiatis, 1975; Prentice et al., 1978). Namely, any joint distribution $f_j(t)dt = \Pr\{T \in [t, t+dt), \Omega = j\}$ can be reproduced by a set of independent competing risks T^j , given covariates. We assume throughout the paper that, in the absence of misclassification, T^j are independent, given $Z, j = 0, 1, 2$. However, the risks will become dependent under misclassification. Under the independent competing risks assumption we have the equality of the so called crude (λ_j^c) and net (λ_j) cause specific hazards

$$\lambda_j^c(t) := \frac{\Pr\{T \in [t, t+dt), \Omega = j | T \geq t\}}{dt} = \lambda_j(t) := \frac{\Pr\{T^j \in [t, t+dt) | T \geq t\}}{dt}, \quad (1)$$

where $:=$ is equality by definition. Suppose that the covariate effects are multiplicative, resulting in two PH models

$$\begin{aligned} \lambda_1(t|Z) &= h_1(t) \exp(\beta^T Z) := h_1(t) \theta_1, \\ \lambda_2(t|Z) &= h_2(t) \exp(\gamma^T Z) := h_2(t) \theta_2, \end{aligned} \quad (2)$$

where h_1, h_2 are the baseline hazard functions, and β, γ are regression coefficients. Note that we let both models incorporate the same Z without loss of generality, as restricting some of the regression coefficients to zero can reproduce arbitrary overlapping or non-overlapping covariate structure in the two models. Also, Z can be time dependent. Capital letters Λ, H will be used to denote respective cumulative hazards. When misclassification is a possibility, instead of the true failure-type indicator Ω we observe a noisy indicator ω . Then, the observed data are $\{T_i, \omega_i, Z_i\}$, independent across i .

In general, failures can be misclassified to either type 1 or type 2 cause with different probabilities. Motivated by the prostate cancer example (Hoffman et al., 2003) we assume that only type 2 failures (other cause) can be misattributed to type 1 cause (prostate cancer), but not vice versa. This type of misclassification is called *over-attribution* (Hoffman et al., 2003) as it inflates the observed incidence of type 1 failure. Extensions to the general case of two-way misclassification are straightforward. In the over-attribution-only model, we have

$$\begin{aligned} \Pr[\omega_i = 1 | \Omega_i = 2, T_i = t, Z_i] &:= r(Z), \\ \Pr[\omega_i = 2 | \Omega_i = 1, T_i = t, Z_i] &= 0. \end{aligned}$$

Because under over-attribution the probability of missing cause of failure depends on the failure type, the model does not follow the MAR mechanism. Define the observed hazard rates $\lambda_j^{obs}(t|Z) := \frac{\Pr\{T \in [t, t+dt), \omega = j | T \geq t, Z\}}{dt}$.

Then, given covariates,

$$\lambda_j^{obs}(t|Z) = \begin{cases} \lambda_0(t|Z), & j=0, \text{ censoring} \\ \lambda_1(t|Z) + r(Z)\lambda_2(t|Z), & j=1, \text{ over-attributed, inflated hazard} \\ \bar{r}(Z)\lambda_2(t|Z), & j=2, \text{ deflated hazard,} \end{cases} \quad (3)$$

where $\bar{r} \equiv 1 - r$ is the probability of correct attribution. Clearly, in the absence of covariates, the misclassification mechanism r is not identifiable jointly with the nonparametrically specified hazards. Indeed, in this case for any r in a certain range, and a set of observed hazard rates λ_j^{obs} , one can find a set of latent rates λ_j matching the observed ones, using the following relationship (from (3)),

$$\begin{aligned} \lambda_1(t) &= \lambda_1^{obs}(t) - \text{Odds}[r]\lambda_2^{obs}(t) \\ \lambda_2(t) &= \bar{r}^{-1}\lambda_2^{obs}(t), \end{aligned} \quad (4)$$

where $\text{Odds}[r] = r/\bar{r}$. External information on r is needed to resolve the issue such as studies where r is estimated by re-classifying patients' causes of death going back to their medical records (Percy et al., 1981; Hoffman et al., 2003; Fall et al., 2008). In this paper we use data from Hoffman et al. (2003) on the number of correctly classified causes out of total diagnoses examined to provide the information on r .

3. Estimation

3.1 NPMLE

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 . A common counting process argument yields the full loglikelihood proportional to

$$\begin{aligned} \ell_n &= \sum_{i=1}^n \sum_{j=1}^2 \{I(\omega_i=j) \log d\Lambda_{ij}^{obs}(T_i) - \Lambda_{ij}^{obs}(T_i)\} \\ &= \sum_{i=1}^n \sum_{j=1}^2 \int \{dN_{ij}(t) \log d\Lambda_{ij}^{obs}(t) - Y_i(t) d\Lambda_{ij}^{obs}(t)\}, \end{aligned} \quad (5)$$

where T_i is the event time for subject i regardless of the type, and Λ_{ij}^{obs} is model-based cumulative hazard under misclassification, given by (3) in its subject-specific form. In the rest of the paper, i is assumed to be $1, \dots, n$ unless noted otherwise. Then, the score function with respect to a (finite or infinite-dimensional) generic model parameter κ becomes

$$\frac{\partial \ell_n}{\partial \kappa} = \sum_i \sum_{j=1}^2 \int \left[\frac{\partial}{\partial \kappa} \log d\Lambda_{ij}^{obs}(t) \right] dM_{ij}(t)$$

for $\kappa = dH_1(\cdot), dH_2(\cdot), \beta$ and γ .

A formal nonparametric maximum likelihood (NPMLE) argument involves treating baseline cumulative hazards as step-functions with steps at the times of failure, and differentiating

over a set of jumps $dH_1(t)$, $dH_2(t)$ at time-points t where failures are observed. In order to streamline the asymptotic argument for NPMLE we define a functional derivative $\frac{\partial}{\partial dH(s)}$ that extends the idea of differentiating a step function over the jump at time s to smooth functions $H(t)$ in the Supplementary Materials. Unfortunately, as we will see below, the NPMLE argument does not produce useful estimators here. Now we have the score equations with respect to the functional part of the model as

$$\begin{aligned} \frac{\partial \ell_n}{\partial dH_j(s)} &= \sum_i \sum_{j=1}^2 \int \left\{ \left[\frac{\partial}{\partial dH_j(s)} \log d\Lambda_{ij}^{obs}(t) \right] dN_{ij}(t) - Y_i(t) \left[\frac{\partial}{\partial dH_j(s)} d\Lambda_{ij}^{obs}(t) \right] \right\} \\ &= \sum_i \sum_{j=1}^2 \int \left[\frac{\partial}{\partial dH_j(s)} \log d\Lambda_{ij}^{obs}(t) \right] dM_{ij}(t), \end{aligned} \quad (6)$$

where $dM_{ij}(t) = dN_{ij}(t) - Y_i(t)d\Lambda_{ij}^{obs}(t)$ is a martingale increment under the true model. Specifically for the model (2), (3), using Supplementary Materials, we have

$$\begin{aligned} \frac{\partial}{\partial dH_1(s)} \log d\Lambda_1^{obs}(t) &= \frac{\theta_1 dI(t-s)}{\theta_1 dH_1(s) + r\theta_2 dH_2(s)}, \quad \frac{\partial}{\partial dH_1(s)} \log d\Lambda_2^{obs}(t) = 0, \\ \frac{\partial}{\partial dH_2(s)} \log d\Lambda_1^{obs}(t) &= \frac{r\theta_2 dI(t-s)}{\theta_1 dH_1(s) + r\theta_2 dH_2(s)}, \quad \frac{\partial}{\partial dH_2(s)} \log d\Lambda_2^{obs}(t) = \frac{dI(t-s)}{dH_2(s)}. \end{aligned} \quad (7)$$

The presence of $dI(t-s)$ in the numerators of (7), on substitution into (6), will erase the integration and leave the term under the integral evaluated at $t=s$, leading to the score equations

$$\begin{aligned} \sum_i \frac{\theta_{i1}}{d\Lambda_{i1}^{obs}(t)} dN_{i1}(t) &= \sum_i Y_i(t)\theta_{i1} \Rightarrow \sum_i \rho_i(t) dM_{i1}(t) = 0 \\ \sum_i \left\{ \frac{r_i\theta_{i2}}{d\Lambda_{i1}^{obs}(t)} dN_{i1}(t) + \frac{dN_{i2}(t)}{dH_2(t)} \right\} &= \sum_i Y_i(t)\theta_{i2} \Rightarrow \sum_i (1-\rho_i(t)) dM_{i1}(t) + dM_{i2}(t) = 0, \end{aligned}$$

where θ_{i1} , θ_{i2} and r_i are subject-specific quantities dependent on Z_i , and the weight

$$\rho_i(t) = \frac{\theta_{i1} dH_1(t)}{\theta_{i1} dH_1(t) + r_i \theta_{i2} dH_2(t)} \quad (8)$$

represents the conditional probability of no misclassification (that the true failure is of type 1, given that observed cause is of the same type 1). The score equations for κ are

$$\begin{aligned} \sum_i \rho_i(t) dM_{i1}(t) &= 0, \quad \sum_i (1-\rho_i(t)) dM_{i1}(t) + dM_{i2}(t) = 0, \\ \sum_i \int Z_i \rho_i(t) dM_{i1}(t) &= 0, \quad \sum_i \int Z_i \{ (1-\rho_i(t)) dM_{i1}(t) + dM_{i2}(t) \} = 0. \end{aligned} \quad (9)$$

However, there is a problem with the equations (9). In the continuous model data are untied, and N_1 , N_2 processes never jump at the same time, $N_j(t) = \sum_i N_{ij}(t)$. When $dN_1(t) = 0$, $dN_2(t) > 0$, from the first equation in (9) it has to be $d\Lambda_{i1}^{obs}(t) = 0$ uniformly over i , which cannot happen with H_1 , H_2 being independent of i . If the violation of the first equation is ignored,

the equation only being activated when $dN_1 > 0$, an (restricted) estimator will emerge that sets $dH_j = 0$ whenever $dN_j = 0, j = 1, 2$, and coinciding with the estimator obtained as if there were no misclassification ($r = 0$). In other words, such estimator will be biased when $r > 0$. This means the equations (9) do not have a reasonable solution in the presence of covariates. We note that in the absence of covariates, $d\Lambda_{i1}^{obs}(t)$ does not depend on i , and the contradiction outlined above is resolved by allowing $d\Lambda_1^{obs}(t) = 0$ that implies negative baseline cumulative hazard jumps. The resultant estimator for the cumulative hazards in the absence of covariates does exist and is consistent (Ha and Tsodikov (2012)). The restricted estimator in the univariate context is also biased and is the same as the one derived from the equality of observed and latent hazards that is only valid in the absence of misclassification.

3.2 Restricted Hazard Ratio

The results of Section 3.1 mean that no reasonable estimators exist in the NPMLE context when the hazard ratio $\phi(t) := \frac{dH_2(t)}{dH_1(t)}$ is unrestricted. Invariance of the MLEs ensures that the NPMLE problems are inherited by the H_1, ϕ parameterization. One simple way to avoid this problem is to put a parametric assumption on $\phi(t)$. The most commonly used assumption for $\phi(t)$ is that the ratio between two cause-specific baseline hazards is constant $h_2(t) = h_1(t) \exp(\phi)$. It is often referred to as the proportionality assumption (Goetghebeur and Ryan, 1995; Lu and Tsiatis, 2005). Under this PH assumption, the martingales become

$$\begin{aligned} dM_{i1}^*(t) &= dN_{i1}(t) - Y_i(t)(\theta_{i1} + r_i \theta_{i2} e^\phi) dH_1(t) \\ dM_{i2}^*(t) &= dN_{i2}(t) - Y_i(t) \bar{r}_i \theta_{i2} e^\phi dH_1(t). \end{aligned}$$

Here we use $*$ to mark quantities parameterized through ϕ . Unlike $dM_{ij}(t) = dN_{ij}(t) - Y_i(t) d\Lambda_{ij}^{obs}$, $dM_{ij}^*(t)$ is a martingale only when ϕ is correct. In the partial and profile likelihood approaches under PH assumption, we first derive the estimating equations for κ , given ϕ . Estimation of ϕ will be discussed later in the paper in Section 3.4.

3.2.1 Profile Likelihood—Under the PH assumption, with H_1 as the only functional parameter remaining in the model, we have the likelihood

$$\ell_n^F = \sum_i \sum_{j=1}^2 \int \{dN_{ij} \log d\Lambda_{ij}^{obs*}(t) - Y_i(t) d\Lambda_{ij}^{obs*}(t)\},$$

where

$$d\Lambda_{ij}^{obs*}(t) = (\theta_{i1} + r_i \theta_{i2} e^\phi) dH_1(t) \text{ and } d\Lambda_{i2}^{obs*}(t) = \bar{r}_i \theta_{i2} e^\phi dH_1(t).$$

The corresponding score equation for $\eta = dH_1(\cdot)$, β and γ is

$$\frac{\partial \ell_n^F}{\partial \eta} = \sum_i \sum_{j=1}^2 \int \left[\frac{\partial}{\partial \eta} \log d\Lambda_{ij}^{obs*}(t) \right] dM_{ij}^*(t)$$

resulting in

$$\begin{aligned} \sum_i dM_{i1}^*(t) + dM_{i2}^*(t) &= 0, & U_n^F(\beta) &:= \sum_i \int Z_i \rho_i dM_{i1}^*(t) = 0, \\ U_n^F(\gamma) &:= \sum_i \int Z_i \{(1 - \rho_i) dM_{i1}^*(t) + dM_{i2}^*(t)\} = 0, \end{aligned} \quad (10)$$

where

$$\rho_i = \frac{\theta_{i1}}{\theta_{i1} + r_i \theta_{i2} e^\phi}. \quad (11)$$

From the first equation, we can derive the Breslow-type estimator

$$d\hat{H}_1(t) = \frac{dN_1(t) + dN_2(t)}{\sum_i Y_i(t) (\theta_{i1} + \theta_{i2} e^\phi)}. \quad (12)$$

Misclassification at any t is a zero-sum game redistributing the causes within a given total. It is not surprising then that the above Breslow-type estimator could be derived from the equality between the hazards of failure regardless of the type expressed using observed or true hazards $d\Lambda_{i1}^{obs}(t) + d\Lambda_{i2}^{obs}(t) = d\Lambda_{i1}(t) + d\Lambda_{i2}(t)$. The MLE of $(\beta, \gamma)^T$ is obtained by substitution of $d\hat{H}_1(t)$ considered as a function of $(\beta, \gamma)^T$, into U_n^F .

3.2.2 Partial Likelihood—Applying the arguments of Cox (1975), consider the following partial log likelihood

$$\begin{aligned} \ell_n^P &= \sum_i \sum_{j=1}^2 \int dN_{ij}(t) \log \frac{d\Lambda_{ij}^{obs}(t)}{\sum_k Y_k(t) d\Lambda_{kj}^{obs}(t)} \\ &= \sum_i \int dN_{i1}(t) \log \frac{\theta_{i1} + r_i \theta_{i2} e^\phi}{\sum_j Y_j(t) (\theta_{j1} + r_j \theta_{j2} e^\phi)} + dN_{i2}(t) \log \frac{\theta_{i2}}{\sum_j Y_j(t) \theta_{j2}}. \end{aligned}$$

Unlike the profile likelihood, it does not include the infinite dimensional parameters. The estimator of the combined parameter vector (β, γ) can be directly obtained by maximizing ℓ_n^P .

Motivated by the form of the score equations (9) we can identify a set of augmented martingale equations that would allow joint estimation of H_1, H_2, β and γ such that, on exclusion of H_1, H_2 from the system, the score equations for the partial likelihood would emerge:

$$\begin{aligned} \sum_i dM_{i1}^*(t) &= 0, \quad \sum_i dM_{i2}(t) = 0, \quad U_n^P(\beta) := \sum_i \int Z_i \rho_i dM_{i1}^*(t) = 0, \\ U_n^P(\gamma) &:= \sum_i \int Z_i \{ (1 - \rho_i) dM_{i1}^*(t) + dM_{i2}(t) \} = 0, \end{aligned} \quad (13)$$

where the first two martingale equations are for $dH_1(t)$ and $dH_2(t)$, respectively. From the first two equations, we can have the Breslow-type estimators

$$d\hat{H}_1^P(t) = \frac{dN_1(t)}{\sum_i Y_i(t)(\theta_{i1} + r_i \theta_{i2} e^\phi)} \quad \text{and} \quad d\hat{H}_2^P(t) = \frac{dN_2(t)}{\sum_i \bar{r}_i Y_i(t) \theta_{i2}}. \quad (14)$$

By replacing $dH_1(t)$ and $dH_2(t)$ in U_n^P with the Breslow-type estimators, we get the partial likelihood score equation $\partial \ell_n^P / \partial \beta = 0$ and $\partial \ell_n^P / \partial \gamma = 0$. The system of equations (13) represents linear martingale transforms under the true model that ensures their unbiasedness and consistency. Because the jumps of the hazards are not synchronized, however, $d\hat{H}_2^P(t)$ is not equal to $e^\phi d\hat{H}_1^P(t)$. That is the estimated model is not PH. However, the true model and the large-sample limit of the estimates do follow the PH assumption. When the PH assumption is not satisfied, the estimators based on the partial likelihood are generally biased. The partial likelihood ℓ_n^P , however, does not allow to estimate ϕ (see Section 3.4).

3.3 Weighted Martingale Estimating Equations

Since estimating equations (10) and (13) are based on ϕ , consistency of estimators holds only when ϕ is correctly specified. To derive robust estimates with respect to ϕ , we temper with the weights embedded into the martingale transform

$$\frac{\partial \ell_n}{\partial \kappa} = \sum_i \sum_{j=1}^2 \int \underbrace{\left[\frac{\partial}{\partial \kappa} \log d\Lambda_{ij}^{obs}(t) \right]}_{\text{weight}} dM_{ij}(t).$$

The proposed weighted estimating equations for $dH_1(t)$, $dH_2(t)$, β and γ are equations (9) except that the weights ρ are now specified as independent of the model's infinite dimensional parameters H_1, H_2 . In this scenario, from the first two equations (9), we have closed form estimators for dH_1 and dH_2 as follows:

$$\begin{aligned} d\hat{H}_1^W(t) &= \frac{1}{S(t)} \left\{ \sum_i \rho_i(t) dN_{i1}(t) \sum_i Y_i(t) \theta_{i2} - dN(t) \sum_i r_i \rho_i(t) Y_i(t) \theta_{i2} \right\}, \\ d\hat{H}_2^W(t) &= \frac{1}{S(t)} \left\{ dN(t) \sum_i \rho_i(t) Y_i(t) \theta_{i1} - \sum_i \rho_i(t) dN_{i1}(t) \sum_i Y_i(t) \theta_{i1} \right\}, \\ S(t) &= \sum_i \rho_i(t) Y_i(t) \theta_{i1} \sum_i Y_i(t) \theta_{i2} - \sum_i Y_i(t) \theta_{i1} \sum_i r_i \rho_i(t) Y_i(t) \theta_{i2}. \end{aligned} \quad (15)$$

Now, the specific robust estimators will depend on the choice of weights ρ . Two options are offered in the following Sections 3.3.1 and 3.3.2.

3.3.1 Weighted Martingale method—In this method (called ‘Weighted Martingale’) we use the weights (11) derived under the PH assumption. Note that estimates of cumulative hazards (15) are consistent even with misspecified ϕ because the martingale structure in (9) does not depend on ϕ unlike in the equations (10) and (13). The PH assumption and ϕ was only used to suggest the form of the weights that would be nearly optimal in models that do not show dramatic violations of the PH assumption. This makes the resultant estimators consistent under an unrestricted (non-PH) model that was ill served by the NPMLE approach, i.e. the Weighted Martingale method represents a fix for the problem of Section 3.1. Also, in the absence of covariates, the estimators coincide with the univariate NPMLEs derived in Ha and Tsodikov (2012). Still required for the weights, ϕ is estimated jointly with the other model parameters, even if the model is deemed non-proportional (in which case it represents a kind of average hazard ratio). This estimation is accomplished in Section 3.4.

3.3.2 Unweighted Martingale method—The weighted robust method can be uncoupled from ϕ when the weights are not motivated by the PH assumption. A natural special case is when all weights are set to 1 (or constant $\rho_i = \rho$). This leads to simpler equations

$$\begin{aligned} U_n^C(\beta) &= \sum_i \int Z_i \left\{ dN_{i1}(t) - Y_i(t)(\theta_{i1} d\hat{H}_1^C(t) + r_i \theta_{i2} d\hat{H}_2^C(t)) \right\} = 0, \\ U_n^C(\gamma) &= \sum_i \int Z_i \left\{ dN_{i2}(t) - Y_i(t) \bar{r}_i \theta_{i2} d\hat{H}_2^C(t) \right\} = 0, \end{aligned} \quad (16)$$

where

$$d\hat{H}_1^C(t) = \frac{1}{\sum_i Y_i(t) \theta_{i1}} \left\{ dN_1(t) - \frac{\sum_i r_i Y_i(t) \theta_{i2}}{\sum_i \bar{r}_i Y_i(t) \theta_{i2}} dN_2(t) \right\}, \quad d\hat{H}_2^C(t) = \frac{dN_2(t)}{\sum_i \bar{r}_i Y_i(t) \theta_{i2}}. \quad (17)$$

In this form the martingale-based equations resemble a common construction used in transformation models (Lin and Ying, 1994; Chen et al., 2002).

3.4 Estimation of ϕ

The score equations for ϕ can be derived from the profile and partial likelihood.

$$\frac{\partial \ell_n^F}{\partial \phi} = \sum_i \int dN_{i1}(t) (1 - \rho_i) + dN_{i2}(t) - dN_i(t) \frac{\sum_j Y_j(t) \theta_{j2} e^\phi}{\sum_j Y_j(t) (\theta_{j1} + \theta_{j2} e^\phi)} = 0 \quad (18)$$

$$\frac{\partial \ell_n^P}{\partial \phi} = \sum_i \int dN_{i1}(t) \left\{ (1 - \rho_i) - \frac{\sum_j Y_j(t) r_j \theta_{j2} e^\phi}{\sum_j Y_j(t) (\theta_{j1} + r_j \theta_{j2} e^\phi)} \right\} = 0. \quad (19)$$

However, the second equation (19) based on the partial likelihood does not lead to a suitable estimator for ϕ . For example, the equation is satisfied when $\beta = \gamma$ ($\theta_1 = \theta_2$), and the partial-likelihood based weights are used. The problem is that type 2 failures make no contribution in the score equation while ϕ depends on both baseline hazards. However, we can use the first equation (18) for ϕ jointly with profile (10), partial (13) and (11)-weighted equations

(9). The Unweighted martingale method does not require estimation of φ . With a fixed r , the profile likelihood approach results in the most efficient semiparametric estimator within the class of regular and asymptotically linear (RAL) estimators for (φ, β, γ) (see Supplementary Materials).

3.5 Misattribution

Misattribution represents a missing data mechanism. Our key assumption is that misattribution depends on the true cause of failure and covariates Z but does not depend on the failure times, given the cause of failure and Z . Specifically, no misclassification of the prostate cancer death is allowed (no under-attribution), while other cause of death can be misclassified as a prostate cancer death (over-attribution). With survival data the true cause of failure is unobserved, and the probabilities of the two types of misclassification are not equal; therefore, this mechanism is not missing at random (NMAR).

Misattribution is not identifiable from survival data $\{T_i, \omega_i, Z_i\}$ when φ is unknown. For example, when $\beta = \gamma$ and r_i is constant, $\logit r = \psi_0$, $\partial \ell_n^F / \partial \psi_0$ and $\partial \ell_n^F / \partial \phi$ are identical as $\int dN_1 r e^{\varphi} = \int dN_2 (1 + r e^{\varphi})$. This shows that either φ or r has to be informed by external data. The analysis of medical records and registered cause of death for deceased subjects (Hoffman et al., 2003) represents a binary supervised classification problem with the true class (failure type) being known from medical records and the registered cause of death representing the classification decision. In this context, we can formulate a binary regression model for the subsample of patients who have truly died from other causes ($\Omega = 2$). The two outcomes in this binary model are “misclassified as prostate cancer” $\{\omega = 1 \text{ \& } \Omega = 2\}$ with probability r , or “not misclassified” $\{\omega = 2 \text{ \& } \Omega = 2\}$ with probability $r = 1 - r$. Suppose such a subsample (called external data in the sequel) is available $\{\omega_i, \Omega_i = 2, Z_i\}$, $i = n + 1, \dots, n + n_r$ (data from (Hoffman et al., 2003) will be used in the example). By the misattribution assumptions stated above, r only depends on observed data in the subsample. Then, using the logistic link, the estimate of over-attribution probability r is derived by maximizing the log likelihood representing the binary classification decisions as follows:

$$\ell_{n_r}^R = \sum_{i=n+1}^{n+n_r} I(\omega_i=1, \Omega_i=2) \log r_i + I(\omega_i=2, \Omega_i=2) \log \bar{r}_i$$

where $\log r_i / \bar{r}_i = \psi_0 + \psi_1 Z_1 + \dots + \psi_p Z_p$. The corresponding additional estimation (score) equations providing information on r are

$$\begin{aligned} U_{n_r}^R(\psi_0) &= \sum_{i=n+1}^{n+n_r} I(\omega_i=1, \Omega_i=2) \bar{r}_i - I(\omega_i=2, \Omega_i=2) r_i = 0 \\ U_{n_r}^R(\psi_x) &= \sum_{i=n+1}^{n+n_r} I(\omega_i=1, \Omega_i=2) \bar{r}_i Z_{xi} - I(\omega_i=2, \Omega_i=2) r_i Z_{xi} = 0, \quad x=1, \dots, p. \end{aligned} \quad (20)$$

4. Asymptotic properties

We assume r is estimated from external data as discussed in Section 3.5. The external data set has no survival information and therefore does not contribute information on the

parameters of the survival model $\chi = (\varphi, \beta, \gamma)^T$, so that the corresponding component of the information matrix $I_{\psi\chi} = E[-\dot{U}_{nr}^R(\chi)_\chi |_{\chi=\chi^0}] = 0$. Here and in the sequel, $\dot{f}(x)_x$ denotes the first derivative of f with respect to x .

Since $\text{Cov}(dN_{ij}, dN_{kl}) = 0$ unless $i = k$ and $j = l$, we have that $((\psi^{\hat{A}} - \psi^0), n^{1/2}(\chi^{\hat{A}} - \chi^0))^T$ is asymptotically normal with mean zero and covariance maxtrix $I^{A-1}V^A I^{A-1T}$ (the sandwich estimator) where

$$I^A = \begin{pmatrix} I_{\psi\psi}^R & 0 \\ I_{\chi\psi}^A & I_{\chi\chi}^A \end{pmatrix} \quad \text{and} \quad V^A = \begin{pmatrix} I_{\psi\psi}^R & 0 \\ 0 & V_{\chi\chi}^A \end{pmatrix} \quad (21)$$

with A taking values F (for profile likelihood), P (for partial likelihood) and W (for weighted estimating equation). The components of the information and covariance matrices entering the sandwich estimator are derived in Supplementary Materials.

Consistency and asymptotic normality of the estimators follow from the arguments of Andersen and Gill (1982) and Lu and Ying (2004).

5. Simulation

The univariate covariates Z were chosen to have a normal distribution with mean 1 and standard deviation 1. Given Z , the failure time T_1 of main interest follows an exponential distribution with hazard function $\lambda_1(t|Z) = 0.5 \exp(Z)$.

Two different survival models are considered: PH and non-PH. In the PH model the log hazard ratio $\varphi = 0.7$ is constant, and the other failure time T_2 has hazard function $\lambda_2(t|Z) = 0.5 \exp(0.7)$. In the non-PH model the log hazard ratio is time-varying $\varphi = 2t$ and the other failure time T_2 has hazard function $\lambda_2(t|Z) = 0.5 \exp(2t)$. Censoring is assumed independent and uniformly distributed on $(0, 2)$ in the PH scenario and $(0, 1.8)$ in the non-PH scenario. This yields, on average, 46 ~ 47% and 32% failures of type 1 and 2, respectively.

Misattribution is assumed to follow a logistic model with the log odds ratio of misattribution of $-1.7 + 0.5Z$.

First, we fit the logistic model to external data by solving the score equations (20). Then, four different methods/models are applied to fit the survival data: (1) Profile (Section 3.2.1), (2) Partial (Section 3.2.2), (3) Weighted Martingale (Section 3.3.1), and (4) Unweighted Martingale (Section 3.3.2), as specified in the left half of Table 1. The first 3 methods/models in this list require estimation of the hazard ratio φ (Section 3.4). The first two of them assume a PH model (constant φ), while the third one is a non-PH model (unrestricted baseline hazards) that uses a (generally misspecified) PH model to estimate and specify the weights. The last Unweighted method/model is based on a non-PH model that uses constant weights.

We carried out 1, 000 simulation replicates in each of the experiments.

Nonproportionality of the true model (time-varying φ column of Table 1) leads to a misspecified model behind the Profile and the Partial methods. This results in biased estimates of log hazard ratio (0.802, 0.652 instead of the true $2t$, for Profile and Partial methods, respectively, Table 1). Both Weighted and Unweighted martingale approaches do not restrict the relationship between the baseline hazards for the two types of failure and the corresponding estimates are consistent under a non-PH as well as PH true model and show a small finite-sample bias (Table 1).

Under the PH model, the profile likelihood estimator χ^F is the most efficient (constant φ column of Table 1). However, it is most affected by misspecified φ . Although partial likelihood estimator is somewhat more robust in this case, consistency still requires correct parametric assumption on φ .

The Weighted estimator, using PH-based weights, exhibited smaller bias and variance than the Unweighted estimator in scenarios with time-dependent hazard ratio $\varphi(t)$, under the chosen true model.

The small-sample bias of the misattribution estimators $\hat{\psi}$ does not affect the other estimators much. There is reasonable precision even with sample sizes n_r of the external data set as small as 50 (See Supplementary Materials) for the effects of different sample size). This lends credibility to the strategy of conducting a small study to generate external data with known misattribution status as in our real data example below.

Weighted estimators have uniformly smaller variability than the partial likelihood estimator that, unlike the Profile estimator, is not fully efficient in the misattribution models.

More details of the simulation study are presented in the Supplementary Materials. These include showing bias with the correctly specified model getting smaller with sample size, showing that redundant regression terms (misattribution in the true model does not depend on covariates, while the fitted model does include them) leads to loss in efficiency, and showing that misspecified model for misattribution leads to substantial bias in all estimates.

6. Example

Here we apply the proposed method to analyze the effects of covariates on prostate cancer specific survival for men aged 50 to 84. Prostate cancer incidence data is available from the Surveillance, Epidemiology, and End Results (SEER) 9 cancer registry. This data set includes over 300,000 cases diagnosed from 1973 through 2000 from 9 registries: Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco - Oakland, Seattle - Puget Sound and Utah.

There are two competing causes of failure in survival data in addition to censoring: type 1 cause of interest (prostate cancer) and type 2 cause (other causes of death). Both types of failure are subject to censoring. The survival time T is the time from diagnosis to death or censoring, and covariates of interest are age ($Z_1 = \text{age} - 50$), stage of tumor ($Z_2 = 1$ if distant stage) and treatment ($Z_3 = 1$ if treated) at diagnosis. Treatments include radiation therapy and radical prostatectomy. Radiation is usually followed by a hormonal regimen.

Over-attribution of the cause of death to prostate cancer (or any cancer in general) is a well recognized problem in cancer registry data (Feuer et al., 1999; Hoffman et al., 2003). In cancer registries misattribution occurs when the person filling out the death certificate tends to attribute death to cancer if the deceased person had a cancer diagnosis. The true cause of death is not available in cancer registries, but can be recovered in a thorough review of the medical records for the deceased person by an experienced oncologist. We utilized an external dataset from (Hoffman et al., 2003) to provide information on misattribution. Hoffman and colleagues retrospectively reviewed medical records to classify the cause of death for men with prostate cancer who died in New Mexico in 1995. Using a constant misattribution model the MLE of Section 3.5 yields the log-odds of over-attribution $\psi_0 = -2.819$. This estimation was done jointly with the survival model parameters.

Ideally, the external dataset should be representative of the survival registry data. So it made sense to select the year of prostate cancer diagnosis (DxY) before the year of death of 1995 in the study by (Hoffman et al., 2003) in such a way that these dead patients could have plausibly been diagnosed in that year DxY. An additional complication is that survival data are confounded by PSA screening leading to early detection of cancer and longer post-diagnosis survival times as a result. PSA screening started in 1988, and its utilization was increasing until the 90-ies. These considerations led us to select cases diagnosed in year 1990 from the SEER registries for the analysis.

Naive (ignoring misattribution), unweighted and weighted estimates for prostate cancer deaths are presented in Table 2. It turns out that the effect of age and stage on prostate cancer specific survival is under-estimated due to the misclassification, while the treatment effect is over-estimated. In view of the risk that external data may not be representative of the population of the cancer registry, we conducted a sensitivity analysis with respect to r (Table 2). We found that the attenuation of the covariate effects and the over-optimistic assessment of the effect of treatment exacerbate with the increase in the rate of over-attribution. For example with over-attribution probability as high as 0.2, treatment effect is no longer significant and its point estimate is virtually zero.

This highlights the importance of accounting for misclassification of the cause of death in cancer registry survival studies, and that nested medical record review studies are needed by the cancer registries to provide the unbiased assessment of treatment and covariate effects in such studies.

7. Discussion

Our proposed weighted martingale estimating equations procedures are similar in spirit to early approaches to transformation models designed to avoid the complexity of NPMLE (Lin and Ying, 1994; Chen et al., 2002). In our case though the NPMLE approach fails without parametric assumptions, and the proposed methods solve the problem without putting additional constraints on the baseline hazards specific to failure types.

The standard profile and partial 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). In the general model

with unrestricted hazards, the main problem is that no useful NPMLE estimators exist in the continuous survival setting targeting step-function estimators for hazards. This fact has led many authors to impose parametric restrictions on the hazard ratio φ or use smoothing. A key contribution of this article is that it provides a formal semiparametric estimating procedure for all cumulative hazards in the step-function setting without any assumptions restricting the ratio.

Under the PH assumption, the score function derived from the profile likelihood is semi-parametrically efficient. A simulation study showed that the profile estimator can be seriously biased and lose efficiency when the proportionality assumption on φ is violated. The model assumption for φ 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 and assumptions that are difficult to verify are undesirable for risk of bias. Also, our method can be used to suggest a pertinent form of $\varphi(t)$ as a model-building step before this form is enforced in a parametric fashion.

Derivation of the estimating equations 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.

The methodology of the paper can be extended to modeling multiple causes of failure subject to an non-MAR missing data mechanism. The extension is straightforward albeit cumbersome. The current model could be considered as a model for dependent observed competing risks, where the dependence is explained by misattribution. Shared frailty has been a popular instrument to model dependent competing risks. The model of dependence can be refined assuming that latent competing risks, given misattribution status, are also dependent.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This paper is supported by the grant 5U01CA157224 (CISNET) from the National CancerInstitute.

References

- Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. *Annals of Statistics*. 1982; 10:1100–1120.
- Burr D. On inconsistency of Breslow's estimator as an estimator of the hazard rate in the Cox model. *Biometrics*. 1994; 50:1142–1145. [PubMed: 7786994]
- Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. *Biometrika*. 2002; 89:659–668.
- Chen P, He R, Shen J, Sun J. Regression analysis of right-censored failure time data with missing censoring indicators. *Acta Mathematicae Applicatae Sinica*. 2009; 25:415–426.
- Cox D. Partial likelihood. *Biometrika*. 1975; 62:269–76.

- Fall K, Stromberg F, Rosell J, Andren O, Varenhorst E. Reliability of death certificates in prostate cancer patients. *Scandinavian Journal of Urology and Nephrology*. 2008; 42:352–357. [PubMed: 18609293]
- Feuer EJ, Merrill RM, Hankey BF. 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*. 1999; 91:1025–1032. [PubMed: 10379965]
- Gao G, Tsiatis AA. Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*. 2005; 92:875–891.
- Goetghebuer E, Ryan L. Analysis of competing risks survival data when some failure types are missing. *Biometrika*. 1995; 82:821–833.
- Ha J, Tsodikov A. Isotonic estimation of survival under a misattribution of cause of death. *Lifetime Data Analysis*. 2012; 18:58–79. [PubMed: 22094534]
- Hoffman RM, Stone SN, Hunt WC, Key CR, Gilliland FD. Effects of misattribution in assigning cause of death on prostate cancer mortality rates. *Annals of Epidemiology*. 2003; 13:450–454. [PubMed: 12875804]
- Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika*. 1994; 81:61–71.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. New York: Wiley; 1987.
- Lu K, Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*. 2001; 57:1191–1197. [PubMed: 11764260]
- Lu K, Tsiatis AA. Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Analysis*. 2005; 11:29–40. [PubMed: 15747588]
- Lu W, Liang Y. Analysis of competing risks data with missing cause of failure under additive hazards model. *Statistica Sinica*. 2008; 18:219–234.
- Lu W, Ying Z. On semiparametric transformation cure models. *Biometrika*. 2004; 91:331–343.
- Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health*. 1981; 71:242–250. [PubMed: 7468855]
- Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 34:541–554. [PubMed: 373811]
- Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*. 1975; 72:20–22.

Table 1

Simulation results for various estimators of model parameters (regression coefficients ψ , β , γ and log hazard ratio φ) using profile, partial, weighted martingale equations jointly with the logistic estimation equation for ψ . True β and γ are 1 and 0, respectively. Both the true and the fitted models use covariate (Z)-dependent misattribution. The log-odds ratio of misattribution in the true model is $-1.7 + 0.5Z$. The sample sizes of the survival and external data sets are $n = 500$ and $n_r = 150$, respectively. True log hazards ratio $\varphi = 0.7$ in the time-independent case (PH) and $\varphi = 2T$ in the time-varying case (non-PH). The sample standard deviations (SSD) and average of the standard error estimates (SEE) are presented in parenthesis.

Assumed Model/Method, Estimating Equations	Martingale Weights	Baseline Hazard(s)	Finite-dim Parameters	True model used in simulations	
				constant φ	time-varying φ
Logistic (Misattribution), External data, (20)	-	-	ψ_0	-1.771 (.371,.367)	-1.776 (.371,.358)
			ψ_1	0.533 (.300,.300)	0.531 (.333,.315)
	PH-based (11), (18)	$\hat{H}_1(t)$, (12)	φ^F	0.693 (.229,.229)	0.802 (.233,.228)
			β^F	0.999 (.097,.093)	1.089 (.097,.094)
			γ^F	0.006 (.114,.114)	-0.068 (.112,.115)
Partial, (13) PH, Const φ	PH-based (11), (18)	$\hat{H}_1(t)e^{\varphi} \rightarrow \hat{H}_2(t)$ (14)	φ^P	0.695 (.238,.236)	0.652 (.227,.233)
			β^P	1.000 (.102,.097)	1.011 (.093,.095)
			γ^P	0.007 (.116,.116)	-0.001 (.124,.120)
Weighted Martingale, (9)	PH-based (11), (18)	$\hat{H}_1(t)$, $\hat{H}_2(t)$ (15) with (11)	φ^W	0.692 (.236,.235)	0.636 (.220,.229)
			β^W	0.999 (.101,.097)	1.001 (.091,.093)
			γ^W	0.007 (.115,.116)	0.002 (.123,.119)
Unweighted Martingale, (16)	Const	$\hat{H}_1(t)$, $\hat{H}_2(t)$ (17)	β^C	1.003 (.107,.102)	1.005 (.094,.097)
			γ^C	0.004 (.119,.120)	-0.002 (.126,.123)

Table 2

Real data analysis and sensitivity analysis with respect to the probability of misattribution. Naïve, Weighted and Unweighted estimates obtained from a survival sample ($n = 11,485$) cancer patients from SEER registry, and an external data set ($n_r = 373$) from Hoffman et al. (2003) study. Naïve model assumes no misattribution. Last two columns (fixed misattribution) represent the results of a sensitivity analysis.

Method:	Naïve	Weighted	Unweighted	Sensitivity Analysis (Weighted)	
Misattribution: r	–	$r \hat{=} 0.056$	$r \hat{=} 0.056$	Fixed $r = 0.1$	Fixed $r = 0.2$
Hazard ratio: $\phi^{\wedge}$	–	0.127 (.119)	–	0.266 (.124)	1.060 (.181)
Prostate Cancer Death					
Age - 50	0.025 (.003)	0.015 (.004)	0.018 (.004)	0.006 (.004)	–0.011 (.004)
Distant Stage	2.181 (.050)	2.417 (.078)	2.386 (.078)	2.674 (.068)	3.584 (.132)
Treatment	–0.368 (.051)	–0.293 (.062)	–0.323 (.062)	–0.203 (.071)	0.018 (.093)
Other Death					
Age - 50	0.076 (.002)	0.077 (.002)	0.076 (.002)	0.078 (.002)	0.075 (.002)
Distant Stage	0.194 (.052)	0.185 (.052)	0.194 (.052)	0.184 (.052)	0.226 (.051)
Treatment	–0.516 (.033)	–0.521 (.033)	–0.516 (.033)	–0.525 (.032)	–0.517 (.031)